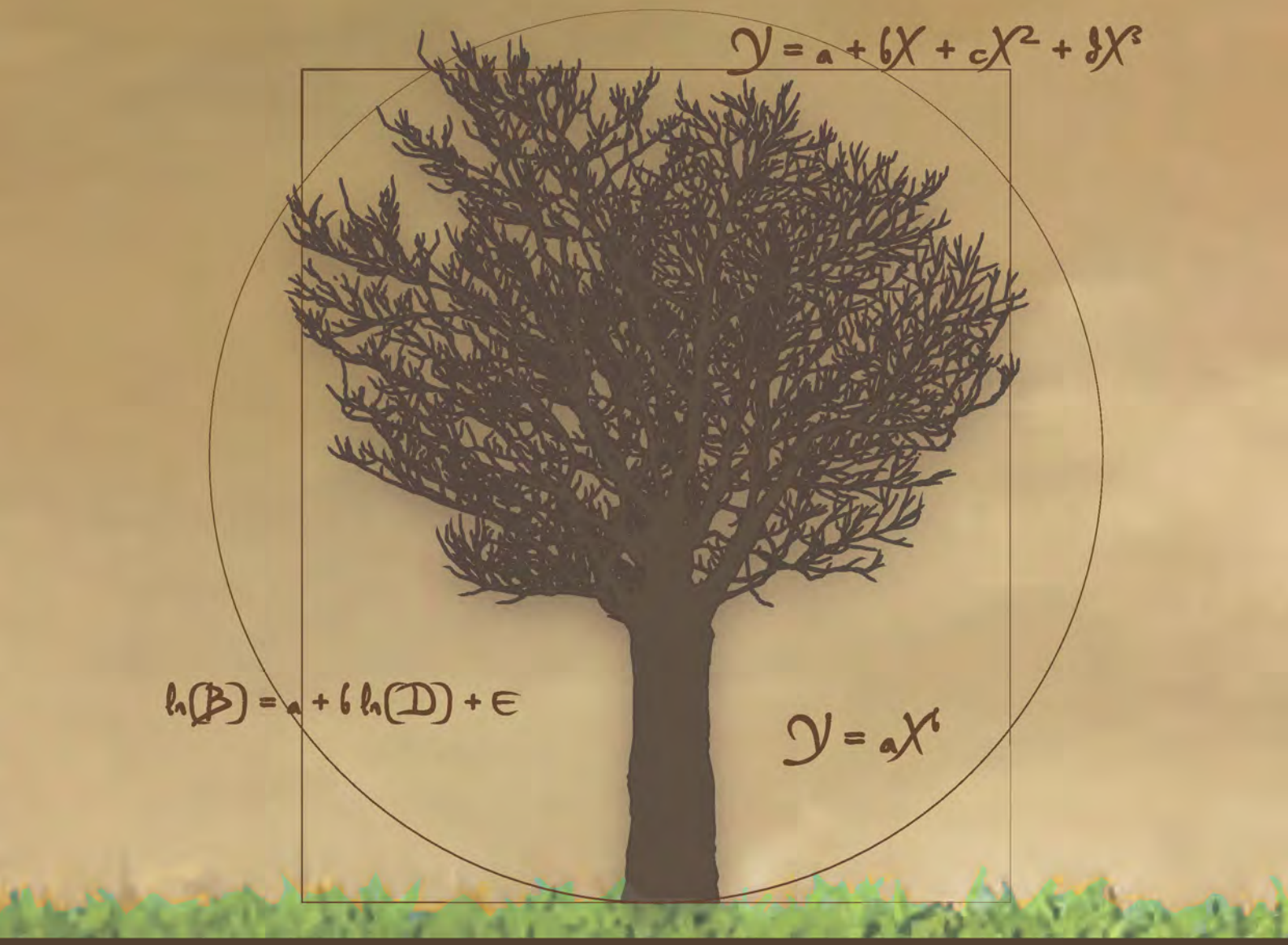
Résumé du manuel de construction d'équations allométriques pour l'estimation du volume et la biomasse des arbres

De la mesure de terrain à la prédiction



cirad

FAO

Résumé du manuel de construction d'équations allométriques pour l'estimation du volume et la biomasse des arbres: de la mesure de terrain à la prédiction

Gael Sola,
Département des forêts,
Organisation des Nations Unies pour l'alimentation et l'agriculture

Nicolas Picard,
Département Environnements et Sociétés,
Centre de Coopération Internationale en Recherche Agronomique pour le Développement

Laurent Saint-André,
UMR Eco&Sols,
Centre de Coopération Internationale en Recherche Agronomique pour le Développement
& UR1138 BEF, Institut National de la Recherche Agronomique

Matieu Henry,
Département des forêts,
Organisation des Nations Unies pour l'alimentation et l'agriculture

Organisation des Nations Unies pour l'alimentation et l'agriculture (FAO)
Viale delle Terme di Caracalla
00153 Rome, Italie

Centre de Coopération Internationale en Recherche Agronomique pour le Développement (CIRAD)
Campus international de Baillarguet
34 398 Montpellier Cedex, France

Citation recommandée: Sola G., Picard N., Saint-André L., Henry M. 2012. Résumé du manuel de construction d'équations allométriques pour l'estimation du volume et la biomasse des arbres: de la mesure de terrain à la prédiction. Organisation des Nations Unies pour l'alimentation et l'agriculture, et Centre de Coopération Internationale en Recherche Agronomique pour le Développement, Rome, Montpellier, 20 pp.

Préambule

Le manuel de construction d'équations allométriques pour l'estimation du volume et de la biomasse des arbres [1] est destiné aux étudiants, chercheurs et ingénieurs qui souhaitent acquérir les connaissances et la méthodologie nécessaires à l'établissement d'équations allométriques pour le cubage, la biomasse ou la minéralomasse des arbres. Il nécessite de faibles pré-requis grâce aux nombreux exemples et aux fiches techniques qui jalonnent l'ouvrage et permettent d'acquérir les connaissances par la pratique. Cette synthèse permet de se familiariser avec le cadre général de l'ouvrage en indiquant les principales étapes.

Un manuel pour améliorer les estimations du volume et de la biomasse des forêts

Les forêts offrent de nombreux services, en particulier la production de bois d'œuvre, le bois énergie et plus récemment le stockage de carbone. L'évaluation des ressources forestières et la quantification de ces services nécessitent une estimation rigoureuse du volume et de la biomasse des arbres. À l'échelle de la parcelle, de la forêt ou d'un ensemble de forêts, obtenir cette estimation de façon précise reste un défi à la fois de par les moyens techniques, humains et financiers, mais aussi par les outils scientifiques nécessaires.

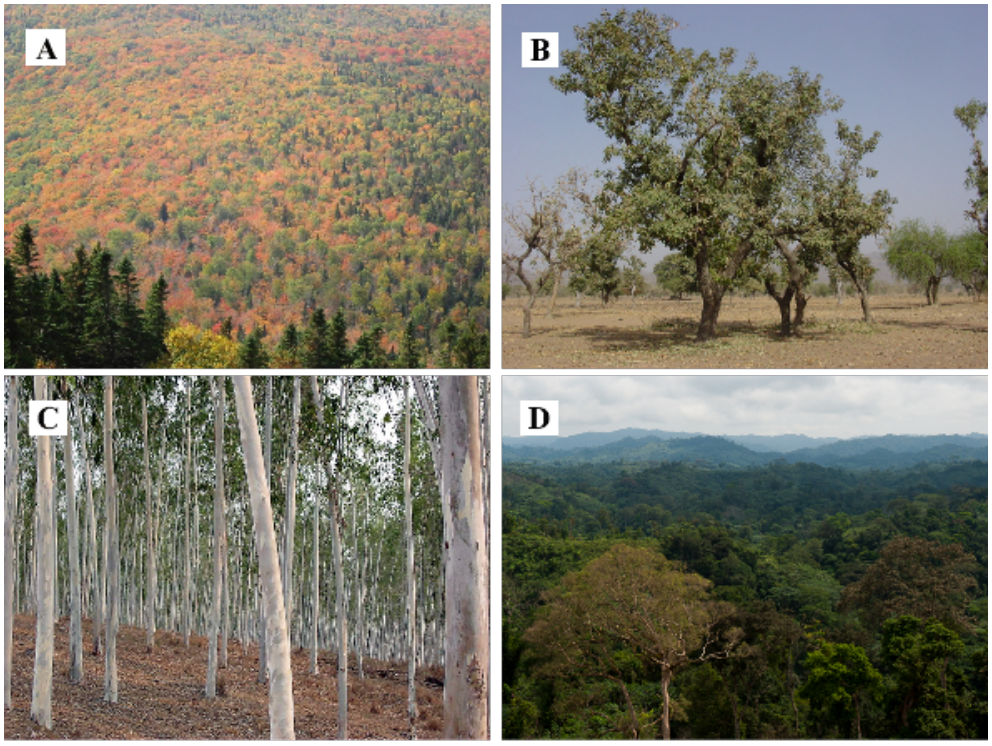


Figure 1. A: forêt tempérée continentale (photo: Bruno Locatelli); B: forêt tropicale sèche (photo: Régis Peltier); C: plantation d'eucalyptus (photo: Bruno Locatelli); D: forêt tropicale humide (photo: Gael Sola).

1. Picard N., Saint-André L., Henry M., 2012. Manuel de construction d'équations allométriques pour l'estimation du volume et la biomasse des arbres: de la mesure de terrain à la prédiction. Organisation des Nations Unies pour l'alimentation et l'agriculture, et Centre de Coopération Internationale en Recherche Agronomique pour le Développement, Rome, Montpellier, 218 pp.

La demande et la prise en compte nouvelle de ces nombreux services ont été suivi d'une mise à jour des outils. Ils découlent de l'évolution de la recherche dans le domaine des mathématiques, de la biologie et des sciences forestières pour apporter des réponses précises à la question de la variabilité de la biomasse suivant le type de forêt (figure 1), l'espèce et le compartiment de l'arbre.

Les stocks de carbone sont estimés à partir de la biomasse des arbres, c'est-à-dire de leur masse sèche de matière organique. Obtenir cette biomasse nécessite donc dans l'absolu de peser l'ensemble des éléments constitutifs d'un arbre (figure 2). Ces mesures deviennent difficiles voir impossibles à l'échelle d'une forêt pour deux raisons: (1) il s'agit de mesures destructives, à grande échelle c'est donc non souhaitable voire interdit et (2) le coût en temps et en main d'œuvre devient irréaliste. Pour les racines, le travail de mesure est encore plus fastidieux et bien souvent des facteurs multiplicatifs sont appliqués directement à la biomasse aérienne pour éviter les mesures.

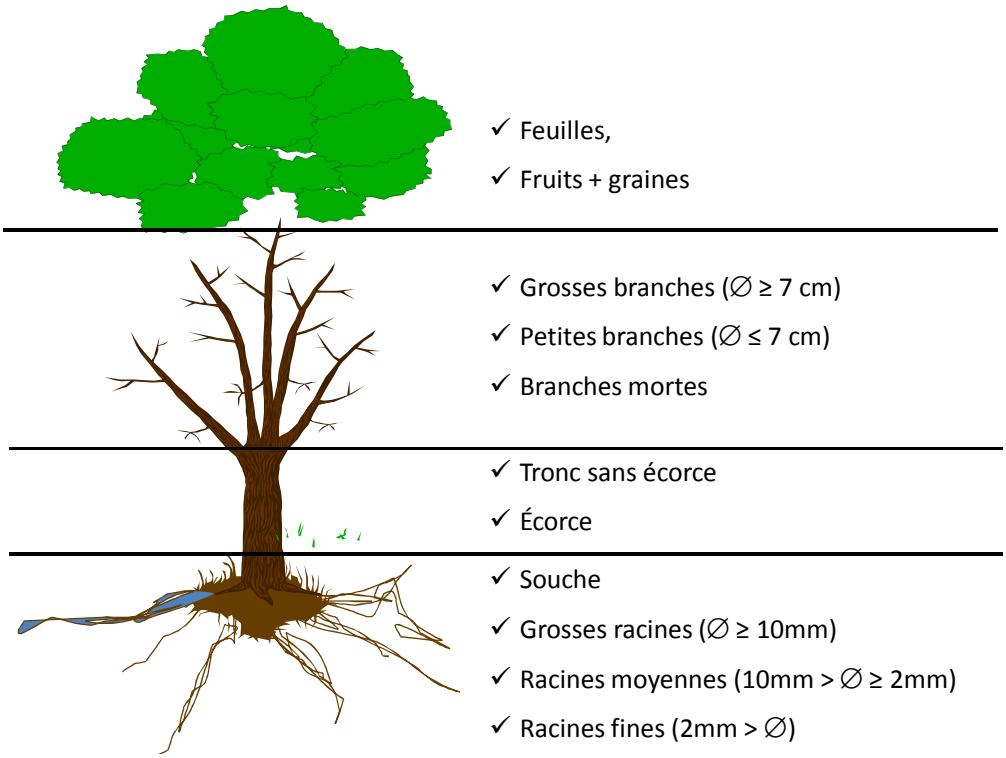


Figure 2. Les différents compartiments d'un arbre.

Une autre méthode plus accessible pour obtenir une estimation de la biomasse d'un arbre consiste à utiliser la relation entre le diamètre des arbres et leur biomasse. C'est à ce niveau qu'interviennent les équations allométriques. L'allométrie désigne la relation statistique qui existe entre deux caractéristiques de taille des individus d'une même population. Il est donc possible de définir une relation statistique entre certaines grandeurs faciles à mesurer à grande échelle (diamètre, hauteur, densité) et une grandeur difficile à mesurer telle que la biomasse ou le volume. Effectuer des mesures coûteuses et destructrices pourra ainsi être limité à un échantillon d'arbres et aura pour but d'ajuster des paramètres pour l'ensemble des arbres d'une zone donnée (voir l'encadré 1).

Encadré 1:
Des paramètres peuvent être ajustés pour obtenir une relation puissance entre la biomasse et le diamètre des arbres:

$$\underbrace{Biomasse}_{\substack{\text{variable à}\\ \text{expliquer}}} = b \times \underbrace{\text{Diamètre}}_{\substack{\text{variable}\\ \text{explicative}}}{}^{a}$$

L'ajustement consiste à déterminer statistiquement les paramètres a et b qui donnent la meilleure relation entre la biomasse et le diamètre pour une zone donnée.

Pour améliorer la connaissance sur les équations allométriques et faciliter leur développement, les auteurs du manuel proposent une méthodologie structurée en 7 étapes (figure 3), de la selection des variables explicatives à la détermination de l'équation allométrique qui y répondra le mieux, en passant par les mesures de terrain. Grâce aux outils statistiques qui y sont présentés, l'erreur commise sur la prédiction peut être calculée.

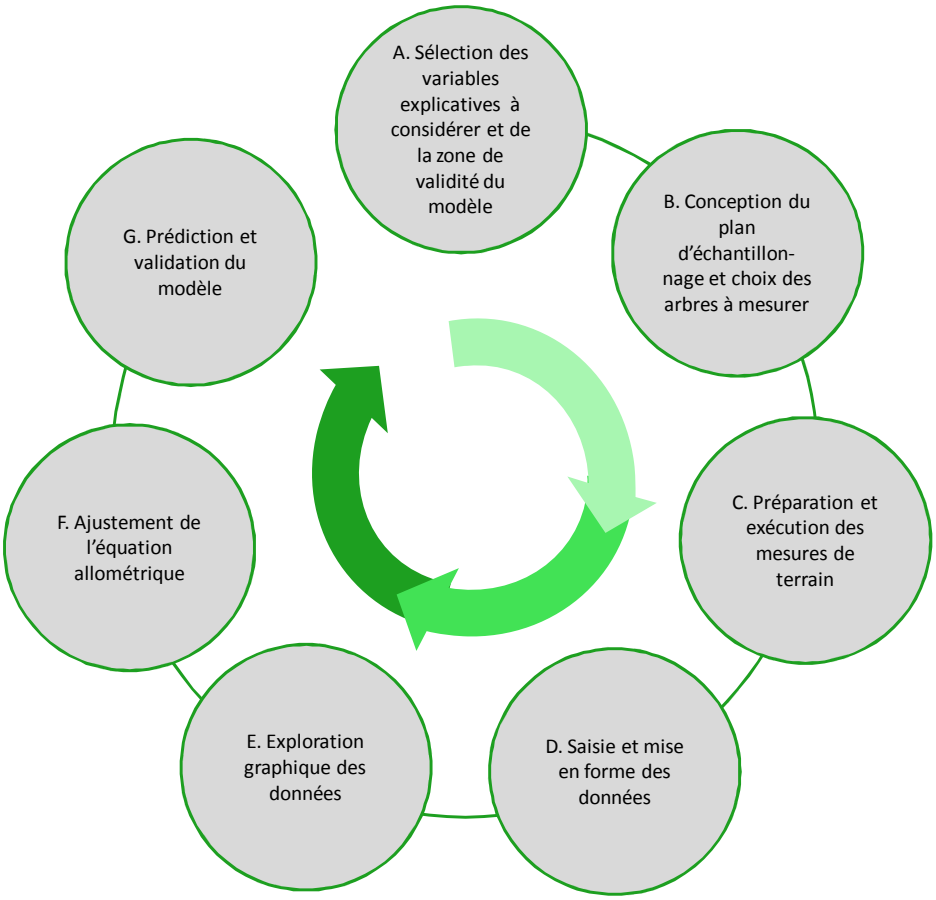


Figure 3. Les différentes étapes de la méthode présentée dans le manuel.

Selection des variables explicatives de la biomasse et de la zone de validité de l'équation

La pertinence des équations allométriques repose sur le fait qu'il existe une relation de proportionnalité entre les accroissements relatifs des dimensions d'un individu. Ainsi la biomasse d'un arbre est reliée à son diamètre (voir l'encadré 2). Mais trouver une relation

statistique entre une biomasse et des facteurs explicatifs n'a que peu de sens si cette relation n'a pas de réalité biologique et son utilisation peut aboutir à des erreurs. Les variables explicatives sont donc à chercher parmi celles qui ont une influence sur la croissance des arbres.

Encadré 2:
La relation de proportionnalité entre biomasse (B) et diamètre (D) s'exprime ainsi:

$$\frac{dB}{B} = a \times \frac{dD}{D}$$

Elle s'intègre en:

$$B = bD^a$$

où a et b sont les paramètres du modèle.

La croissance des arbres résulte de deux processus: la croissance en longueur, résultant de l'activité des bourgeons, et la croissance en épaisseur, provenant de l'activité du cambium. Cette croissance dépend du patrimoine génétique des arbres, de leur stade de développement (vieillissement des tissus) et de leur environnement (sol, atmosphère, concurrence des autres arbres, influence de l'homme sur cette concurrence, voire sur l'arbre lui-même). Il est possible de traduire ces facteurs en variables mathématiques, comme par exemple:

- la fertilité d'une station à travers l'âge de son peuplement et de sa hauteur dominante H_0 (loi de Eichhorn élargie, peuplement équiennes et monospécifiques),
- la densité du peuplement (facteur d'espacement de Hart-Becking, Reinecke Density Index RDI, coefficient d'élancement H/D, coefficient de robustesse $D^{1/2}/H$) pour tenir compte de la concurrence entre les arbres,
- le statut social des arbres (H/H_0 ou D/D_0, D_0 représentant le diamètre dominant du peuplement).

L'introduction de la fertilité permet notamment d'élargir la zone de validité d'un tarif et l'ensemble de ces facteurs peut être utilisé pour déterminer l'équation la plus précise. Une fois les facteurs choisis et la zone d'application du tarif déterminée (parcelle, massif forestier, ensemble de forêt, aire de répartition d'une espèce, etc.), l'étape suivante consiste à définir l'échantillonnage le mieux adapté.

Échantillonnage et stratification des arbres pour améliorer le compromis coût-précision

Deux sources d'erreur sont considérées dans le processus de modélisation de la biomasse ou du volume. La première erreur est associée au fait que seule une partie des arbres est mesurée, il s'agit de l'***erreur d'échantillonnage***. La seconde est associée au fait que le modèle final ne donnera toujours qu'une approximation du volume ou de la biomasse, il s'agit de l'***erreur de prédiction***. Les erreurs de mesure et de saisie ne peuvent pas être déterminées ou encadrées statistiquement et le manuel fournit de nombreux conseils pour les éviter au maximum. Concernant l'erreur d'échantillonnage, elle peut être réduite lors de la conception du plan d'échantillonnage, c'est-à-dire du choix des arbres qui feront partie de l'échantillon. L'erreur de prédiction est intimement liée à la construction du modèle.

Échantillonnage pour déterminer la biomasse d'un arbre

Taille de l'échantillon

Plus le peuplement est variable, plus le nombre d'arbres à échantillonner doit être élevé pour obtenir une même précision. Ainsi, un peuplement équienne monospécifique (e.g. plantation industrielle) nécessitera un échantillonnage plus réduit qu'une forêt tropicale humide pour une même erreur. Dans le premier cas les arbres sont plus proches les uns des autres au niveau de leur stock de carbone: ils ont le même âge, sont de la même espèce voire du même clone et ont quasiment tous le même statut social (voir la figure 1 photo C). En forêt tropicale humide (voir la figure 1 photo D), la diversité floristique est grande, les statuts sociaux peuvent être multiples, etc. La taille de l'échantillon doit donc être raisonnée en fonction de la diversité des arbres au sein du peuplement.

Choix des arbres

Le choix des arbres au sein de l'échantillon a également une influence sur l'erreur d'échantillonnage. La théorie montre que dans le cas d'une régression linéaire (1) la précision augmente avec la taille de l'échantillon et (2) il est préférable d'échantillonner les arbres sur toute la gamme de valeurs de la variable d'entrée (voir encadré 3). Par exemple, lorsque la variable d'entrée sélectionnée est le diamètre, il est préférable de sélectionner un nombre d'arbres constant par classe de surface terrière, pour mieux représenter les arbres de diamètre élevé.

Encadré 3:
Lors de l'échantillonnage, l'amplitude de l'intervalle de confiance sur la prédiction est une fonction inverse de l'écart-type de la variable d'entrée sélectionnée. La demi-amplitude au seuil α de l'intervalle de confiance de cette régression est définie par le calcul suivant:

$$t_{n-2}\frac{\hat{\sigma}}{S_X\sqrt{n}} \text{ avec } S_X^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 \text{ et } \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

où t_{n-2} est le quantile $1-\alpha/2$ d'une loi de Student à $n-2$ degrés de liberté, $\hat{\sigma}$ est l'écart-type empirique des résidus du modèle, n la taille de l'échantillon et S_X l'écart-type empirique de la variable d'entrée X au sein de l'échantillon.

On retrouve dans cette formule le fait que plus la taille de l'échantillon est grande ($\sqrt{n}$ augmente) et plus l'intervalle de confiance sur la prédiction est étroit (*i.e.* la précision est grande). Il en est de même avec l'écart-type S_X, ce qui signifie qu'il est préférable d'échantillonner les arbres sur toute la gamme de la variable d'entrée sélectionnée.

Stratification de la zone à inventorier

Dans certains cas, il est préférable de classifier les types de forêt pour faciliter l'échantillonnage et diminuer les coûts relatifs aux mesures. Toutes les variables explicatives peuvent être utilisées: âge du peuplement (principalement pour les plantations), fertilité, station, traitement sylvicole, variété ou espèce, altitude, profondeur de la nappe phréatique, etc. Suivant ces critères, une division de la zone à inventorier en différentes strates sera effectuée (sol riche, sol pauvre par exemple). Cette stratification est utile lorsque la variabilité de la biomasse est plus grande sur certaines strates que sur d'autres. En augmentant l'effort d'échantillonnage sur ces strates, il en découlera une meilleure précision de l'estimation pour une même taille d'échantillon (voir encadré 4).

Encadré 4:
En théorie, l'effort d'échantillonnage doit être proportionnel à l'écart-type au sein de chaque strate. En pratique, il dépend des moyens techniques, financiers et humains et des contraintes liées au terrain.

Échantillonnage pour l'estimation de la biomasse d'un peuplement

Pour l'estimation de la biomasse à l'échelle du peuplement, la mesure de chaque arbre devient irréalisable. Il est préférable de mesurer tous les arbres dans un espace donné appelé placette et de répéter cette unité pour atteindre la taille de l'échantillon souhaitée. Comme pour les inventaires forestiers, l'erreur sera ici déterminée en fonction du nombre de placettes (voir encadré 5).

Encadré 5:
Il existe une relation entre le nombre de placettes, l'erreur sur l'échantillonnage associée et le coefficient de variation du peuplement étudié. De manière simplifiée, cette formule permet de déterminer le nombre de placettes à sélectionner:

$$n \approx (\frac{2CV_B}{E})^2 + 1 \text{ avec } CV_B = \frac{S_B}{\bar{B}}$$

avec n le nombre de placettes, CV_B le coefficient de variation de la biomasse ($\bar{B}$ représente la biomasse moyenne d'une placette et S_B son écart-type) et E l'erreur liée à l'échantillonnage.

Le coefficient de variation d'une parcelle de surface A est donc l'élément essentiel pour construire le plan d'échantillonnage. La relation entre les deux est une fonction puissance:

$$CV_B = kA^{-c}$$

Dans la relation entre CV_B et A, c représente l'agrégation de la biomasse dans le peuplement. Si $c < 0,5$ le peuplement est agrégé, sinon sa répartition spatiale est régulière.

Dans ces calculs, l'agrégation du peuplement intervient dans le choix de la taille des placettes. Si le peuplement à une biomasse agrégée, pour un même effort d'échantillonnage l'erreur sera réduite si un grand nombre de placettes de petite taille est choisi, alors que dans le cas d'une répartition spatiale régulière de la biomasse, il vaudra mieux choisir peu de placettes mais avec une surface plus grande (figure 4).

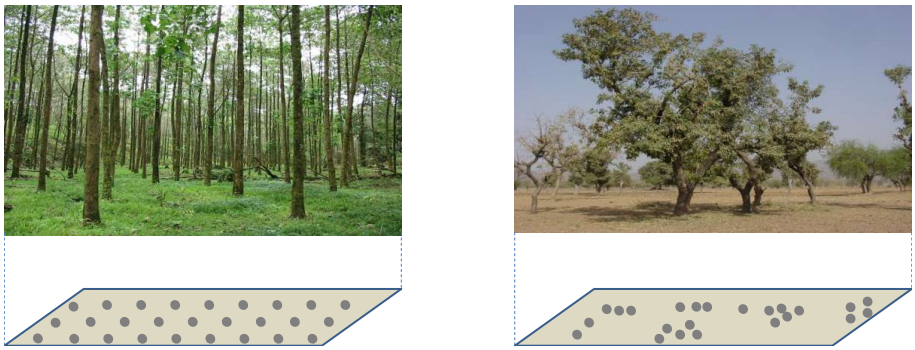


Figure 4. A gauche, le schéma représente un peuplement à structure régulière et à droite un peuplement à structure agrégée.

Il est également important de rappeler qu'en dehors de ces outils d'amélioration de la précision sur l'échantillonnage, la réalisation d'une équation allométrique dépend de nombreux autres facteurs tels que les facteurs financiers, humains, techniques mais aussi environnementaux. Chaque facteur imposera des limites et tout l'art de parvenir à la meilleure équation réside dans la recherche du compromis optimal.

Conseils pour la récolte des données sur le terrain et au laboratoire

Les erreurs de terrain sont coûteuses et ne peuvent pas être corrigées. Trois principes clés peuvent aider à les réduire au maximum:

→ il est préférable de peser tous les compartiments des arbres sur le terrain,

→ à chaque prélèvement d'un échantillon, il est préférable de peser systématiquement la masse totale de l'échantillon puis celle de l'aliquote pour suivre la perte d'humidité du matériel végétal,

→ les campagnes de biomasses étant coûteuses en moyens et en temps, d'autres mesures peuvent être faites pour éviter de revenir sur le terrain (profil de tige, échantillonnage pour la minéralomasse par exemple).

Les mesures de terrain sont préférentiellement destructives afin de pouvoir peser tous les compartiments sur place. Néanmoins, il n'est pas toujours possible (arbres trop lourds, interdiction de coupe) ou souhaitable (cas des forêts sèches) d'abattre tous les arbres. Dans un premier temps, des indications seront données pour des mesures directes sur le terrain. Quelques astuces pour pallier l'impossibilité de tout mesurer sur place seront ensuite présentées.

Encadré 6: quelques conseils avant de partir.

→ Il n'est pas rare que les contraintes de terrain (accès, etc.) remettent en cause l'échantillonnage initial.

→ Les différents compartiments de l'arbre n'ont pas tous la même densité ni le même taux d'humidité. Il est donc préférable de mesurer compartiment par compartiment.

→ Afin d'éviter un biais dans la sélection des échantillons, il est important que ce soit toujours la même personne qui échantillonne.

→ Pour éviter les pertes d'humidité, il est fortement recommandé d'utiliser une glacière si les échantillons ne peuvent être mesurés sur place.

→ La préparation en amont du matériel, des feuilles de saisie et des sacs contenant les futures aliquotes assurera un gain de temps sur le terrain.

Cas des mesures directes sur le terrain

Pour éviter les temps mort dans le travail des différentes équipes, une organisation en sept postes (figure 5) est proposée. Le nombre de personnes par poste ainsi que la chronologie d'intervention des différents acteurs est détaillée dans le manuel.

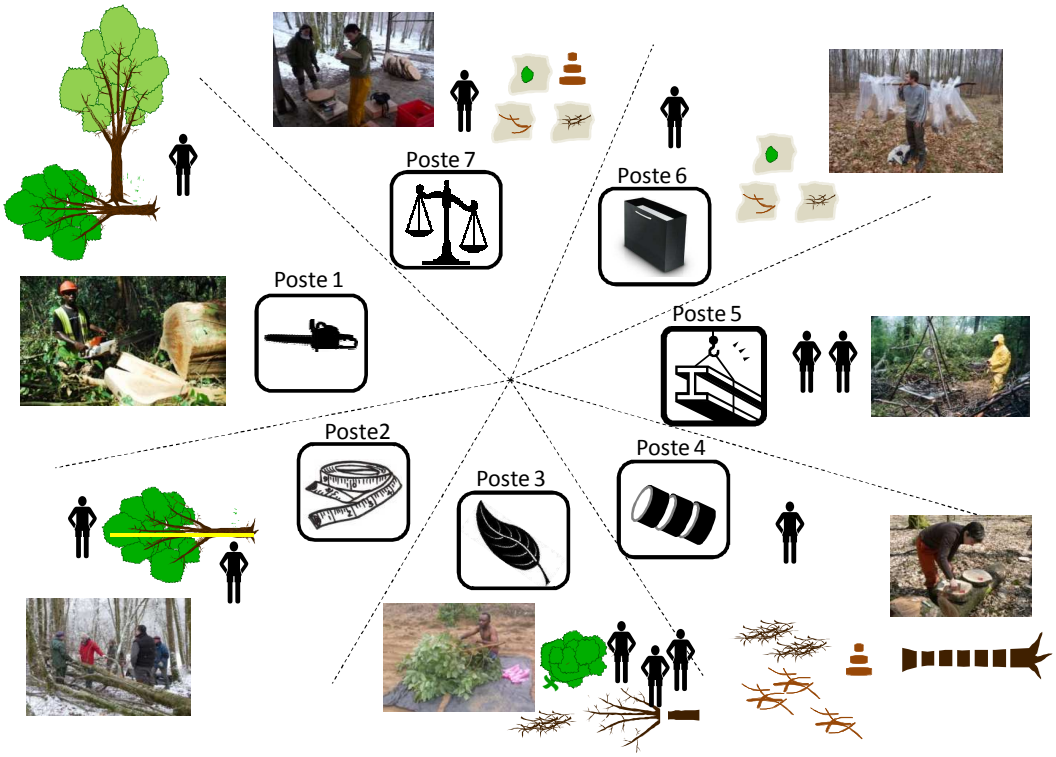


Figure 5. Poste 1: préparation du terrain et abattage des arbres (photo: M. Henry); poste 2: mesure sur arbres abattus (profils de tige, position des billons) (photo: M. Rivoire); poste 3: effeuillage et ébranchage (photo: R. D'Annunzio); poste 4: billonnage et étiquetage des rondelles (photo: C. Nys); poste 5: pesée des billons et des fagots (photo: B. Locatelli); poste 6: échantillonnage des branches (photo: M. Rivoire); poste 7: zone de pesée des échantillons (photo: M. Rivoire).

Cas des mesures semi-destructives

Plusieurs méthodes permettent d'éviter la découpe totale de l'arbre. Deux cas sont abordés ici, le cas des forêts sèches et les cas d'un arbre de trop grandes dimensions pour être pesé intégralement sur le terrain.

Dans le cas des ***forêts sèches***, les arbres sont émondés régulièrement mais leur abattage n'est souvent pas envisageable du fait de leur rareté. Les branches émondées sont utilisées pour calculer la biomasse des feuilles qui y sont présentent (la pratique veut que trois échantillons de feuilles de trois branches différentes soient requis pour former l'aliquote) et la biomasse ainsi que la masse volumique fraîche des branches. Sur les autres parties de l'arbre, le diamètre et la longueur sont mesurées (figure 6). Avec la densité des éléments émondés et le volume de tous les compartiments, la biomasse totale de l'arbre peut être estimée.

Dans le cas d'***arbres de dimensions trop grandes*** pour une pesée intégrale sur le terrain, les branches dont le diamètre est supérieur à 10 cm sont traitées différemment de celles dont le diamètre est inférieur. Ce seuil peut être modifié pour mieux correspondre aux réalités de terrain. La biomasse des éléments de diamètre supérieur à 10 cm est déduite de la mesure du volume (diamètre et longueur mesurés sur le terrain, voir figure 6) et du calcul de la masse volumique moyenne de l'arbre. Celle des éléments inférieurs à 10 cm est calculée à partir d'une régression entre leur diamètre basal et leur biomasse.

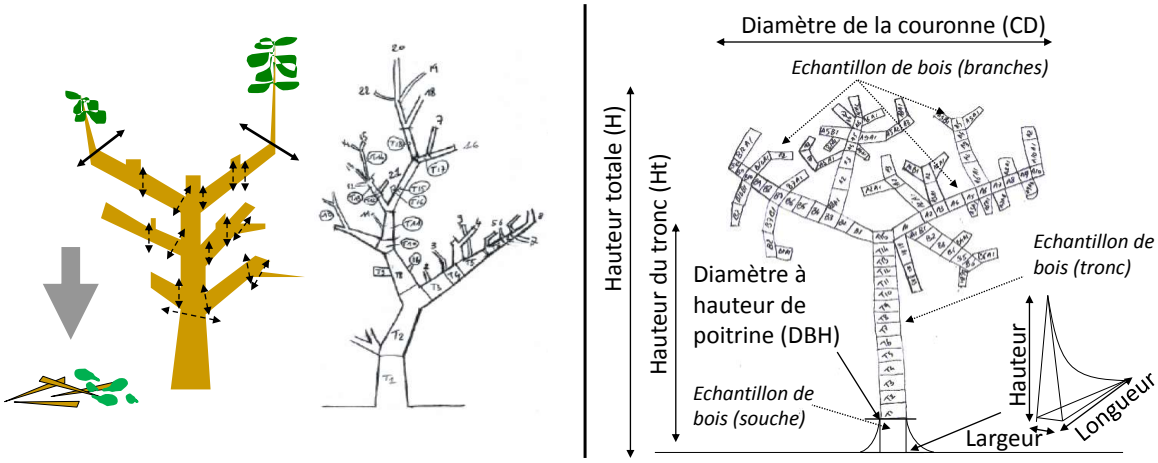
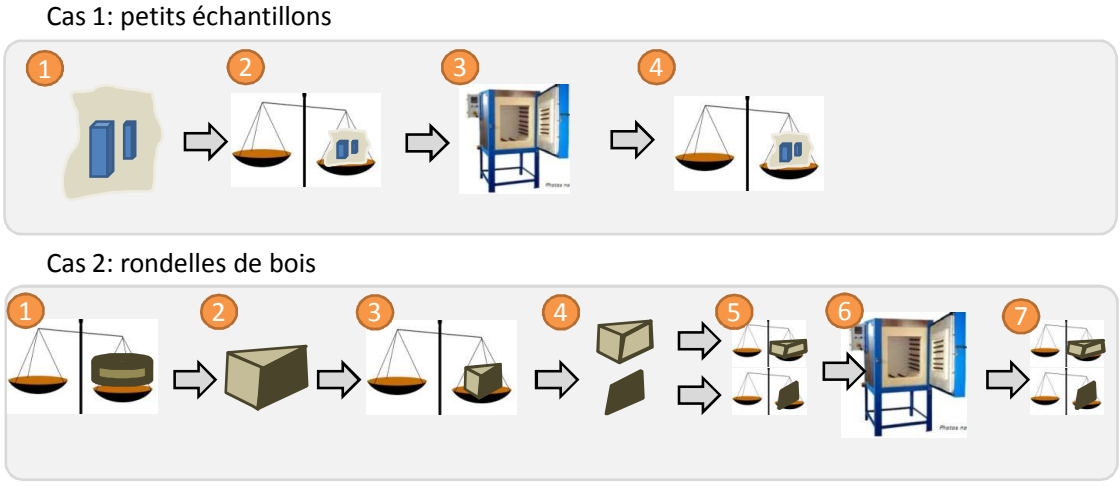


Figure 6. Partie gauche, séparation de la biomasse émondée et mesure des tronçons de l'arbre non émondé. Ces tronçons sont numérotés lors des mesures. Partie droite, numérotation des compartiments non coupés de l'arbre pour la mesure de leur longueur et de leur diamètre.

Mesures de laboratoire

Le travail de laboratoire consiste à mesurer le volume des échantillons, leur masse fraiche et leur masse une fois séchées à l'étuve (figure 7). Pour le passage à l'étuve, les feuilles et les fruits seront séchés à 70°C (65°C si des analyses chimiques sont prévues), pour les opérations de biomasse et pour le bois seulement, les échantillons seront séchés à 105°C.



Cas 1: petits échantillons

1. Echantillons positionnés dans un sac en papier
2. Pesée de l'échantillon humide et du sac en papier, ensemble et séparément
3. Séchage à l'étuve
4. Pesée

Cas 2: rondelles de bois

1. Pesée de la rondelle entière
2. Prélèvement d'un échantillon
3. Pesée d'un sous-échantillon
4. Séparation bois et écorce
5. Pesée séparée de la masse humide du bois et écorce
6. Séchage à l'étuve
7. Pesée séparée de la masse sèche du bois et écorce

Figure 7. Les différentes étapes du séchage et de la pesée des échantillons au laboratoire.

Saisie et mise en forme des données

Une fois la phase de terrain achevée, la préparation et la mise en forme des données est une étape relativement simple mais cruciale pour limiter les erreurs lors de la phase d'ajustement du modèle.

Les tables de données devront être emboitées si plusieurs niveaux de données sont considérés. La table contenant les données des arbres peut être emboitée dans une table pour les parcelles, elle-même emboitée dans une table pour le massif forestier.

Pour assurer une analyse correcte, le contrôle de la qualité et faciliter l'utilisation des données dans le futur, il est préférable de conserver les informations descriptives des données (méta-information). Celles-ci contiennent des informations telles que la nature des variables, la date de mesure, le nom de l'opérateur, etc. Lors de la saisie des données, les variables qualitatives doivent être bien différenciées des variables quantitatives. Il est préférable d'utiliser des codes comportant peu de caractères pour éviter les erreurs de saisie.

Figure 8. Opérateurs pendant la saisie des données (Photo: S. Giaccio).

Enfin, la double saisie par deux opérateurs différents (figure 8) est une méthode coûteuse mais qui assure une bonne fiabilité des données. Des procédures de contrôle de la qualité des données vont permettre d'identifier les valeurs irréalistes et les erreurs de saisie. Cette procédure fastidieuse peut être automatisée. L'analyse graphique représentant les variables deux à deux est également un bon moyen de détecter les erreurs.

Exploration graphique des données

Lorsque le jeu de données est prêt, la phase d'exploration graphique est une des étapes clés pour la modélisation de la biomasse. Les résultats d'un modèle sont composés de deux termes: ***la moyenne*** et l'***erreur*** (ou ***résidu***). L'exploration graphique permet d'identifier la forme de la moyenne et des résidus sans déterminer les paramètres du modèle. C'est lors de l'ajustement que les paramètres du modèle seront determinés. L'exploration graphique ne permet pas de trouver le meilleur modèle mais de retenir les trois ou quatre modèles qui apparaissent les plus pertinents. La méthode consiste à représenter les valeurs de la variable à expliquer en fonction de chaque variable explicative. L'analyse de la distribution des points permet d'identifier si la relation entre les variables suit une forme particulière, linéaire ou non et si la variance des résidus est constante ou non. Les graphiques suivants (figure 9) illustrent ces quatre types de relation.

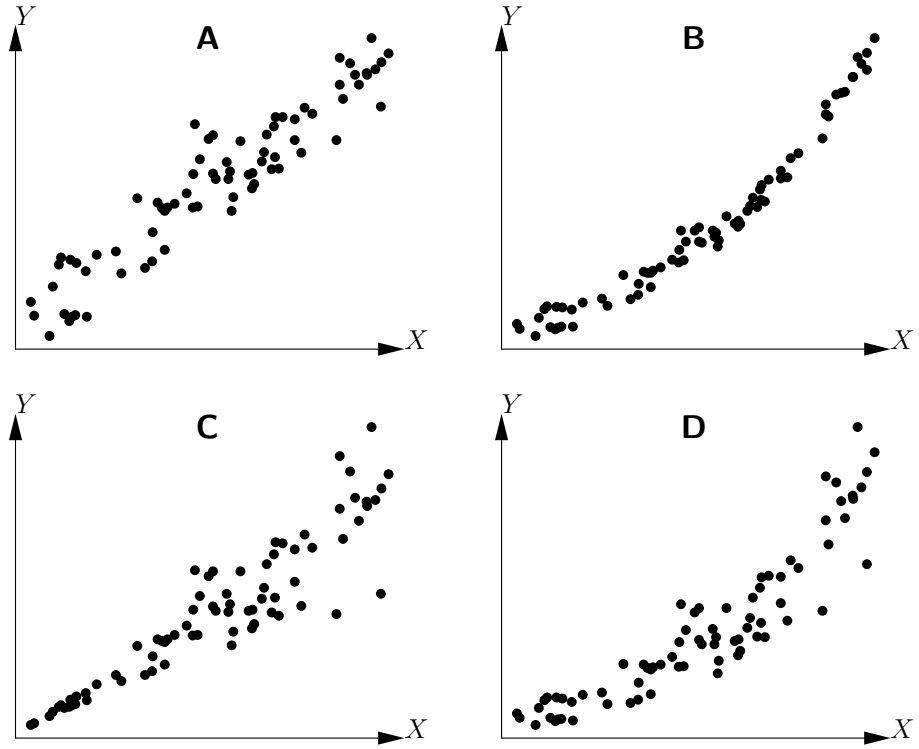


Figure 9. A: relation linéaire, variance des résidus constante; B: relation non linéaire, variance des résidus constante; C: relation linéaire, variance des résidus non constante, D: relation non linéaire, variance des résidus non constante.

Pour aller plus loin, il est également possible de représenter la variable à expliquer en fonction de:

→ variables couplées à partir des variables de base (comme par exemple D^2H),

→ variables transformées, la plus courante étant la transformation logarithmique, pour obtenir une relation linéaire (il est plus facile à l'œil nu de distinguer si une relation est linéaire ou non que d'identifier le type de relation à partir de courbes),

→ variables créées à partir de relations entre plusieurs variables interdépendantes ou non (sommes, multiplications, sommes + multiplications).

La phase d'exploration permet enfin de détecter des artefacts de modélisation, c'est-à-dire que certaines formes de nuages de points ne témoignent pas d'une relation linéaire entre les deux variables représentées alors que la modélisation statistique donne un coefficient de corrélation linéaire (R^2) élevé. Cette étape permet aussi de retrouver des erreurs de saisies flagrantes.

Ajustement de l'équation à partir des données

Après avoir identifié les variables explicatives susceptibles de générer un modèle pertinent, il va falloir utiliser des méthodes et tests statistiques pour déterminer les paramètres du modèle. L'ensemble de ces outils repose sur le comportement des résidus du modèle, une première partie leur est consacrée avant une présentation plus complète des différentes possibilités d'ajustement.

Définition et hypothèses sur les résidus

Les résultats du modèle sont des prédictions. Notre variable à expliquer (aussi appelée variable réponse du modèle) étant la biomasse, il va y avoir autant de biomasses mesurées sur le terrain que de biomasses prédites par le modèle (à partir des variables explicatives: diamètre, hauteur, densité, espèce, etc.). La différence entre chaque observation et sa prédiction est appelée résidu du modèle et il y a donc autant de résidus que d'observations (voir encadré 7).

Encadré 7:
À la i-ème observation de la variable réponse (B_i) correspond une prédiction par le modèle ($\hat{B}_i$) et un résidu (ε_i) défini par:

$$\varepsilon_i = B_i - \hat{B}_i$$

Les outils statistiques utilisés pour déterminer les paramètres du modèle sont basés sur le comportement de ces résidus et la vérification a posteriori de trois hypothèses:

1. ***Indépendance des résidus:*** cette hypothèse n'est généralement pas vérifiée par le calcul et doit être assurée lors de l'échantillonnage des arbres. Ils doivent être suffisamment éloignés les uns des autres pour que le choix d'un arbre n'influence pas celui de l'arbre suivant. En cas de doute, le test de Durbin-Watson permet de vérifier l'indépendance des résidus.

2. ***Distribution normale des résidus:*** cette hypothèse est vérifiée graphiquement par une représentation des quantiles empiriques en fonction des quantiles théoriques. Sur ce graphique, les points doivent s'aligner approximativement le long d'une droite (figure 10). Il existe également des tests statistiques pour vérifier la normalité des résidus (Kolmogorov-Smirnov, d'Agostino et al, etc.).

3. ***Variance constante des résidus:*** on parle également d'homoscédasticité et donc d'hétéroscédasticité lorsque cette hypothèse n'est pas vérifiée. L'homoscédasticité de la variance des résidus est contrôlée par le graphique des résidus en fonction des valeurs prédites. Ce graphique ne doit pas montrer de structure ou de tendance particulière du nuage de point (figure 10).

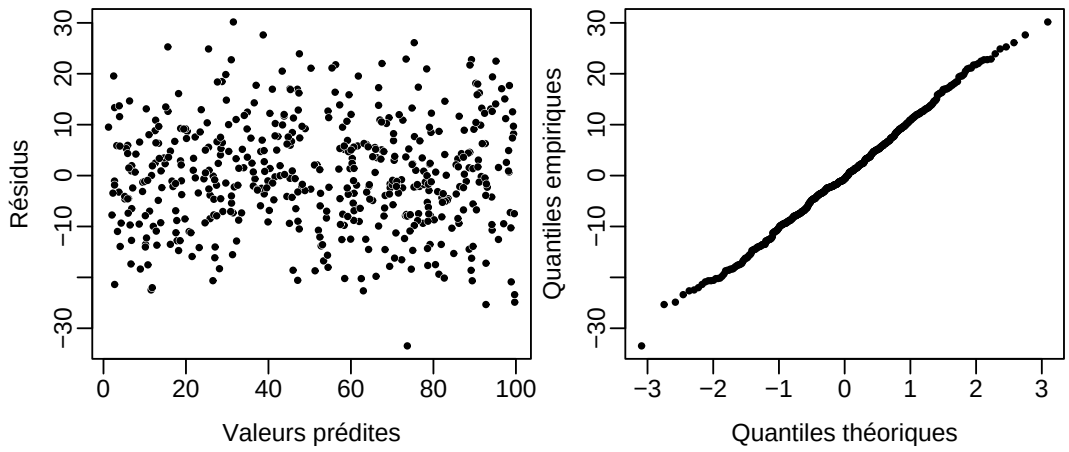


Figure 10. À gauche la variance des résidus est homogène, à droite la distribution des résidus est normale.

Les hypothèses deux et trois peuvent être vérifiées par simple visualisation graphique car les types de modèle présentés ci-après sont ***robustes***, c'est-à-dire qu'ils sont de qualité satisfaisante même si les hypothèses ne sont pas complètement vérifiées.

Les différents types d'ajustement

Les différents types d'ajustement présentés dans cette partie sont: (1) les régressions linéaires simples, multiples, pondérées ou avec modèle sur la variance et les régressions non linéaires avec exposant connu ou à estimer, puis (2) la sélection entre des modèles préexistants emboîtés, ayant la même variable réponse ou avec des variables réponses différentes. La méthode statistique adaptée à chaque type d'ajustement est présentée dans le tableau suivant (tableau 1).

Ajustement	**Méthode**	**Forme de l'équation**	**Loi des résidus**
Régression linéaire:			
Simple	Moindre carrés	$Y = a + bX + \varepsilon$	$\varepsilon \sim \mathcal{N}(0,\, \sigma)$
Multiple	Moindre carrés	$Y = a_0 + a_1X_1 + a_2X_2 + \ldots + a_pX_p + \varepsilon$	$\varepsilon \sim \mathcal{N}(0,\, \sigma)$
Pondérée	Moindre carrés pondérés	$Y = a_0 + a_1X_1 + a_2X_2 + \ldots + a_pX_p + \varepsilon$	$\varepsilon_i \sim \mathcal{N}(0,\, \sigma_i)$
Avec modèle sur la variance	Maximum de vraisemblance	$Y = a_0 + a_1X_1 + a_2X_2 + \ldots + a_pX_p + \varepsilon$	$\varepsilon \sim \mathcal{N}(0,\, kX_1^c)$
Régression non linéaire:			
Avec exposant c connu	Moindres carrés pondérés	$Y = f(X_1, X_2, \ldots, X_p; \theta) + \varepsilon$	$\varepsilon_i \sim \mathcal{N}(0,\, \mathrm{Var}(\varepsilon))$ $\mathrm{Var}(\varepsilon) = g(X_1, \ldots, X_p; \vartheta)$
Avec exposant c à estimer	Maximum de vraisemblance	$Y = f(X_1, X_2, \ldots, X_p; \theta) + \varepsilon$	$\varepsilon_i \sim \mathcal{N}(0,\, \mathrm{Var}(\varepsilon))$ $\mathrm{Var}(\varepsilon) = g(X_1, \ldots, X_p; \vartheta)$
Selection entre modèles préexistants:			
Emboités (linéaire)	Loi de Fisher	Teste si les paramètres ajoutés par le modèle complet au modèle emboité sont non différents de 0	
Emboités (non-linéaires)	Loi du χ^2	Teste si les paramètres ajoutés par le modèle complet au modèle emboité sont non différents de 0	
Variables réponses identiques	AIC	Le meilleur modèle sera celui qui minimise ce critère	
Variables réponses différentes	Indice de Furnival (F)	Le modèle avec la plus petite valeur de F sera considéré comme meilleur	

Tableau 1. Tableau présentant les différents tests statistiques utilisés pour l'ajustement des modèles. L'ensemble des symboles et des calculs sont expliqués dans le manuel.

Les méthodes de régression linéaire ou non

La méthode des ***moindres carrés*** consiste à calculer la somme des carrés des résidus (SCE), puis à trouver l'ensemble des paramètres qui minimise cette somme. Ce calcul est relativement facile mais ne convient qu'aux modèles les plus simples: régression linéaire simple ou multiple.

Dans le cas de la méthode des ***moindres carrés pondérés***, un poids positif est associé à chaque observation. Il est défini par une relation de proportionnalité avec l'inverse de la variance résiduelle. Cette relation peut être simplifiée pour les données biologiques telles que la biomasse ou le volume et il en résulte alors un exposant à identifier supplémentaire par rapport à la méthode précédente.

La méthode du ***maximum de vraisemblance*** revient à calculer les paramètres qui maximisent la vraisemblance des observations. La vraisemblance d'une observation est la densité de probabilité d'obtenir cette observation sous le modèle spécifié. La variable réponse est distribuée dans ce cas selon une loi normale dont l'espérance et l'écart-type dépendent de trois paramètres (θ,k,c). Cette méthode consiste donc à trouver le triplet de paramètres qui maximise cette fonction.

Dans le cas des modèles non-linéaires, que ce soit avec la méthode des moindres carrés pondérés ou celle du maximum de vraisemblance, une optimisation numérique est nécessaire pour vérifier que le minimum (respectivement le maximum) est bien atteint. Le manuel détaille plusieurs processus d'optimisation. Enfin, lors de la phase de test, plusieurs modèles linéaires et non-linéaires peuvent correspondre aux observations. Pour le choix du modèle final, le tableau 2 présente quelques avantages et inconvénients des modèles linéaires multiples et des modèles non linéaires.

	Avantages	**Inconvénients**
Régression linéaire:	Expression explicite des coefficients du modèle.	Contraintes sur la forme des résidus, peu de souplesse dans la forme du modèle.
Régression non linéaire:	Pas de restriction sur la forme du modèle pour la moyenne ou pour la variance.	Pas d'expression explicite des coefficients du modèle donc risque d'estimation erronée des paramètres.

Tableau 2. Tableau présentant quelques avantages et inconvénients des modèles linéaires multiples et non linéaires.

Sélection entre modèles préexistant

Dans le cas des ***modèles emboîtés***, la sélection revient à tester si les paramètres supplémentaires du modèle le plus complet sont significatifs ou non. L'hypothèse nulle est rejetée si la p-value du test (Fischer pour les modèles linéaires ou χ^2 pour les modèles non-linéaires) est inférieure au seuil de significativité (le modèle complet est donc le meilleur dans ce cas).

Dans le cas des ***modèles ayant la même variable réponse***, le modèle le plus pertinent est celui qui minimise le critère d'information d'Akaike (AIC). Ce critère est basé sur la vraisemblance du modèle utilisé.

Enfin, s'il s'agit de comparer des ***modèles avec des variables réponses différentes*** (l'une étant une variable transformée de l'autre), l'indice de Furnival peut être utilisé. Cet indice est défini pour un modèle dont les résidus ont une variance constante sans tenir compte de la forme de la transformation de variable. Le modèle avec la plus petite valeur de F sera considéré comme le meilleur.

Stratification et agrégation d'observations

Lorsque les données sont stratifiées, il existe deux méthodes d'ajustement de l'ensemble des données: (1) chaque strate est ajustée séparément et les méthodes décrites précédemment peuvent être utilisées ou (2) le jeu de données est analysé dans sa globalité. Dans le deuxième cas, l'analyse utilisée est appelée ***analyse de covariance***. Elle suppose que tous les résidus ont la même variance, dans et entre les différentes strates. Un test de Fisher est alors utilisé pour tester si les différents paramètres du modèle sont significativement différents de 0. Lorsque le nombre de strates est trop grand, d'autres solutions sont présentées dans le manuel. Ce cas peut se produire notamment pour la stratification par espèce d'arbre. Enfin le manuel présente également des solutions pour intégrer des modèles pour chaque compartiment de l'arbre en un modèle unique, comme le recours aux ***modèles multivariés***.

L'ensemble de ces outils permet donc de tester différentes modélisations à partir d'un jeu de données (figure 11) et de déterminer l'équation la plus précise, mais aussi de comparer plusieurs modèles déjà existant pour choisir celui qui correspondra le mieux à la zone d'étude.

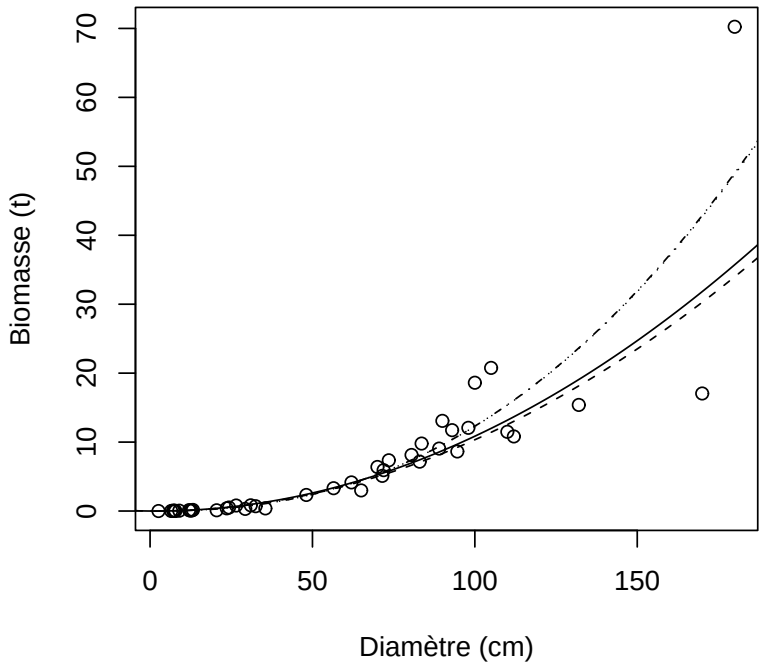


Figure 11. Prédiction de la biomasse par quatre différents tarifs (trait plein, tirets, pointillés, point-tirets) ajustés aux données de 42 arbres (points) mesurés au Ghana par Henry et al. (2010)[2].

Validation du modèle et encadrement des prédictions

Une fois le modèle définitif ajusté, il est possible de le ***valider*** si l'on dispose d'un jeu de données indépendant de celui utilisé pour l'ajustement. Il existe alors plusieurs critères permettant de comparer les prédictions du modèle aux observations: le biais, la somme des carrés des écarts résiduels, la variance résiduelle, l'erreur résiduelle ajustée, le R^2 ou encore

2. Henry, M., Besnard, A., Asante, W.A., Eshun, J., Adu-Bredu, S., Valentini, R., Bernoux, M. et Saint-André, L., 2010. Wood density, phytomass variations within and among trees, and allometric equations in a tropical rainforest of Africa. Forest Ecology and Management, 260(8): 1375-1388.

le critère d'information d'Akaike (AIC). Enfin, il ne faut pas oublier que les prédictions du modèle retenu ont une variabilité intrinsèque. Une fois le modèle achevé, il est donc conseillé de lui associer un indicateur d'incertitude (figure 12): ***l'intervalle de confiance à 95 %***. La méthode de calcul de l'intervalle de confiance diffère si la prédiction est basée sur un arbre moyen du peuplement ou sur un arbre pris au hasard et selon la forme du modèle, linéaire ou non.

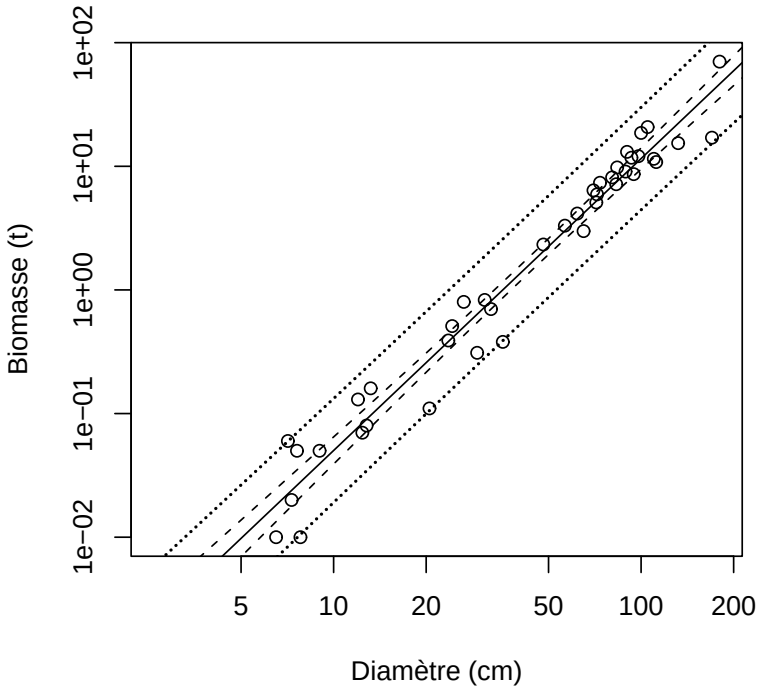
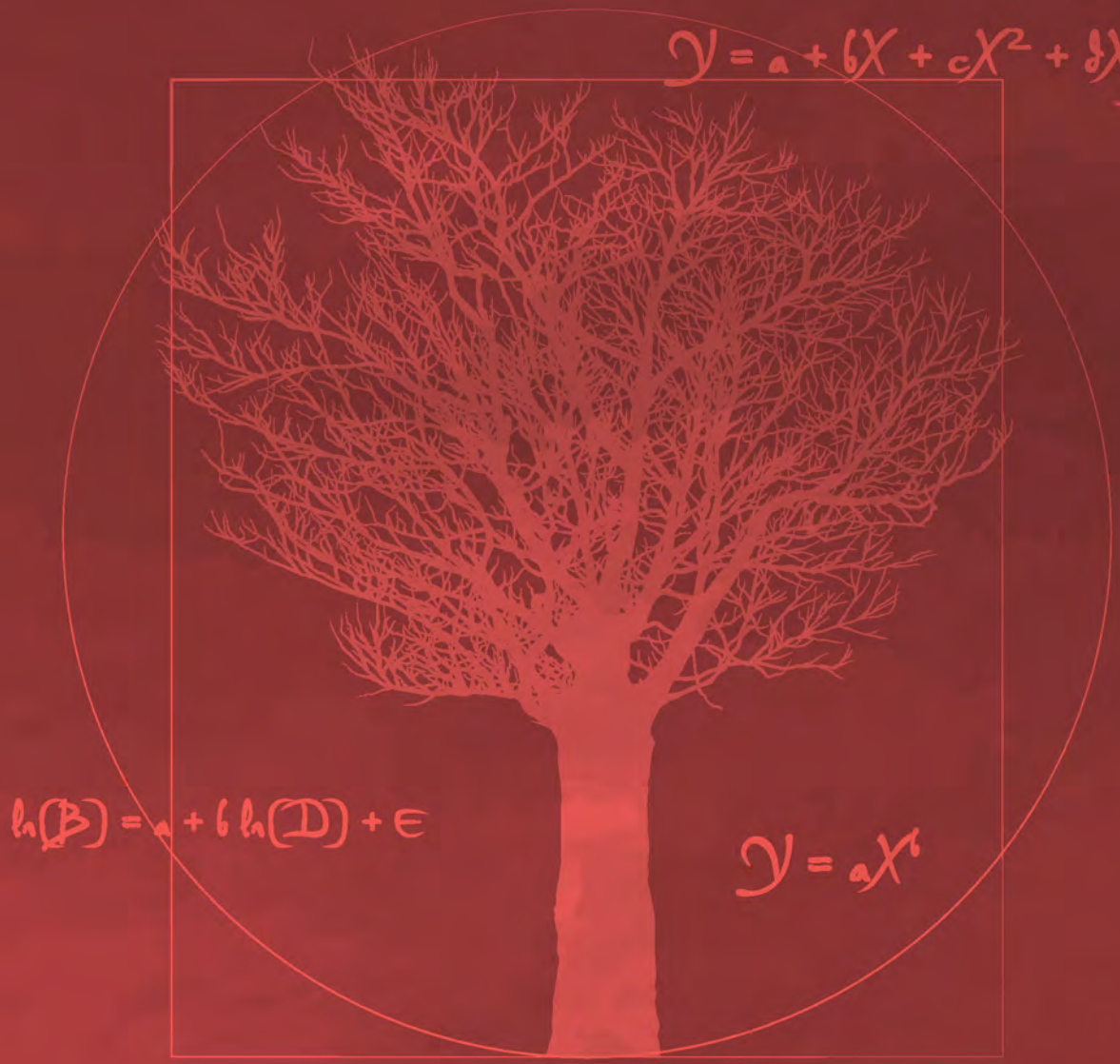


Figure 12. Données de 42 arbres (points) mesurés au Ghana, avec la prédiction (trait noir) et ses intervalles de confiance pour un arbre pris au hasard (pointillés) et pour l'arbre moyen (tirets)

Conclusion

Les outils statistiques et les différentes méthodes présentées dans le manuel de construction d'équations allométriques présentent l'originalité de fournir une démarche d'ensemble. L'objectif final de ce manuel est de fournir des outils pour augmenter la précision des estimations de stock de carbone forestier. Pour y parvenir, il est important de considérer la précision sur l'ensemble de la démarche qui aboutit aux équations: de l'échantillonnage aux statistiques en passant par le terrain. Tous les cas possibles ne sont pas étudiés car certaines situations représentent encore des défis techniques et scientifiques. De même, les statistiques sur le sujet sont en pleine évolution, bien que de nombreux progrès aient déjà été faits dans ce domaine. Lorsque les moyens de construire des équations allométriques robustes et précises seront largement répandus, de nouvelles perspectives pourront être explorées pour améliorer la comptabilisation carbone. En particulier, la création et l'entretien de bases de données regroupant l'ensemble des équations existantes permettra aux utilisateurs de trouver rapidement une équation précise adaptée à leur contexte.



E-ISBN 978-92-5-107347-6