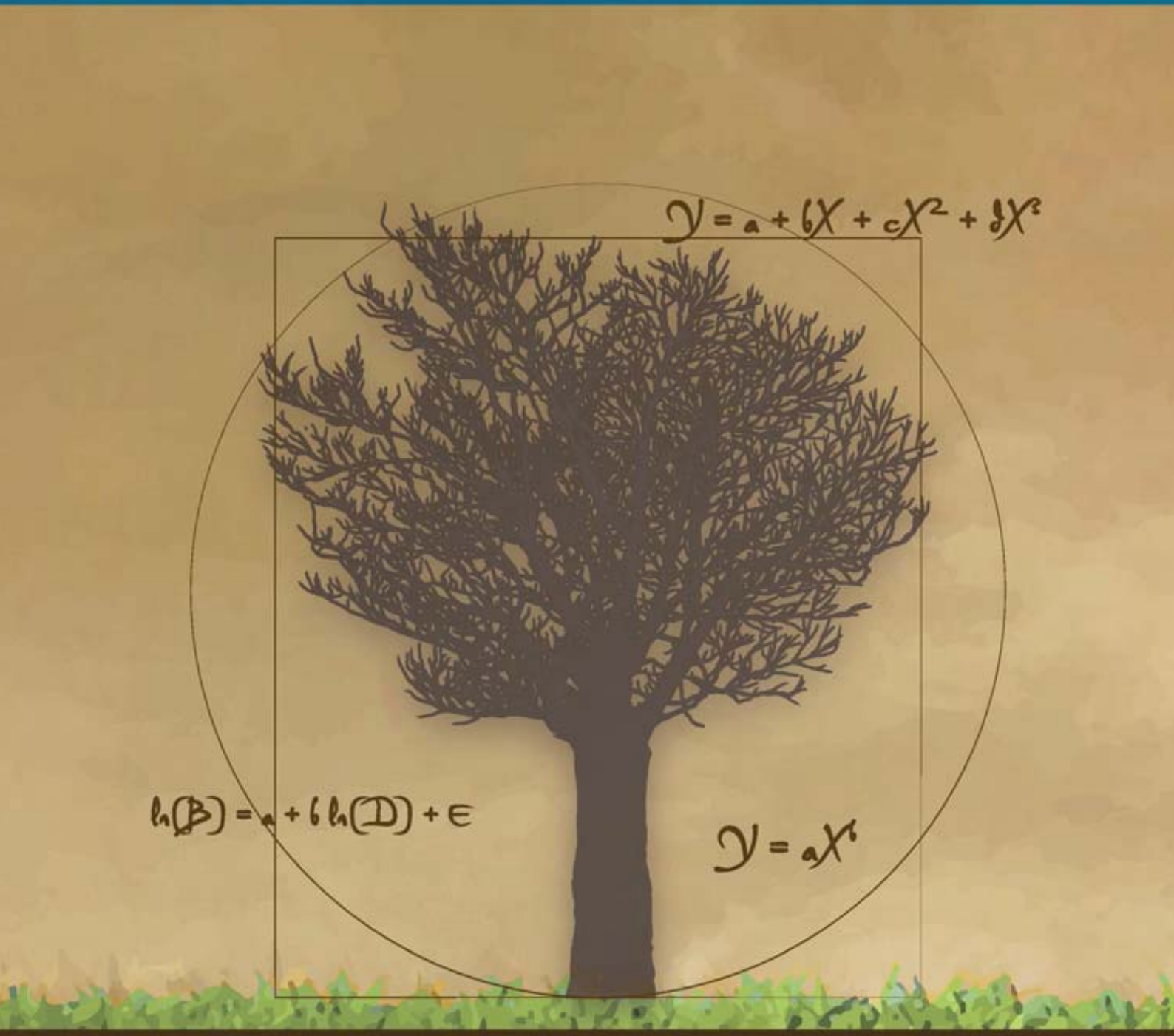
Manuel de construction d'équations allométriques pour l'estimation du volume et la biomasse des arbres

De la mesure de terrain à la prédiction



cirad

FAO

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

De la mesure de terrain à la prédiction

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

Août 2012

E-ISBN 978-92-5-207347-5

Organisation des Nations Unies pour l'alimentation et l'agriculture (FA0)
Viale delle Terme di Caracalla
00153 Rome, Italie

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

Crédit photographique: Stephen Adu-Bredu (photo 3.5), Rémi D'Annunzio (photo 3.4, figure 3.2), Astrid Genet (photos 3.13, 3.14), Matieu Henry (photos 3.8, 3.10), Christophe Jourdan (photos 3.11, 3.12, figure 3.8), Bruno Locatelli (photo 1.2), Claude Nys (photo 3.7, figure 3.2), Régis Peltier (photo 3.9), Jean-François Picard (photo 3.15, figure 3.2), Michaël Rivoire (photos 3.3, 3.5, 3.14, figure 3.2), Laurent Saint-André (photos 1.1, 3.3, 3.4, 3.6, 3.8, 3.11, figure 3.2).

Citation recommandée: 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, 220 pp.

Table des matières

Table des figures

Liste des photos

Liste des tableaux

Liste des encadrés du fil rouge

Préface

Dans le cadre de la Convention-cadre des Nations Unies sur les Changements Climatiques, les bénéfices potentiels pour les parties non visées à l'annexe I ayant diminué leurs émissions de gaz à effet de serre seront basés sur des résultats mesurés, reportés et vérifiés. La précision de ces résultats aura une influence majeure sur les compensations financières potentielles. Les mesures des stocks de carbone forestier prennent ainsi une importance accrue pour les pays qui projettent de contribuer à l'atténuation des changements climatiques *via* leurs activités forestières. Ces mesures font appel aujourd'hui à des techniques fonctionnant à différentes échelles, depuis les inventaires de terrain réalisés à une échelle locale jusqu'aux mesures de télédétection par satellite opérant à une échelle nationale ou sous-régionale, en passant par le laser ou le radar aéroporté. Les mesures indirectes des stocks de carbone forestier, telles que celles dérivées des indices satellitaires, Lidar ou radar, reposent sur des relations calibrées à partir de mesures de terrain. Il en va de même pour les inventaires. Il reste certain en fin de compte que toute mesure du carbone forestier nécessite que des arbres aient été pesés sur le terrain; cette étape constitue la pierre angulaire sur laquelle repose tout l'édifice de l'estimation des stocks de carbone forestier, quelles que soient les échelles considérées.

Aussi, les équations allométriques, qui permettent de prédire la biomasse d'un arbre à partir de caractéristiques dendrométriques plus faciles à mesurer (telles que son diamètre ou sa hauteur) sont des éléments clefs pour l'estimation de la contribution des écosystèmes forestiers au cycle du carbone. Ce manuel se propose de couvrir toutes les étapes de leur construction, depuis la mesure de la biomasse des arbres sur le terrain. Il devrait en particulier s'avérer très utile aux pays qui ne disposent pas encore des mesures et des modèles d'équations adaptés à leurs formations forestières.

Ce *Manuel de construction d'équations allométriques pour l'estimation du volume et la biomasse des arbres* est donc un guide pratique à l'attention des étudiants, techniciens et chercheurs qui travaillent sur l'évaluation des ressources forestières telles que le volume, la biomasse et les stocks de carbone, pour des objectifs commerciaux, bioénergétiques ou d'atténuation des changements climatiques. Les méthodes proposées dans ce manuel s'appliquent à la plupart des forêts et zones écologiques, avec un accent particulier sur les forêts tropicales qui, peut-être plus que les autres, nécessitent aujourd'hui un effort de la communauté internationale pour la mesure des stocks de carbone. Un fil rouge est proposé pour guider le lecteur: il s'agit d'un cas concret qui illustre les différentes questions liées à la réalisation des équations allométriques, l'échantillonnage, les mesures de terrain, la saisie des données, l'exploration graphique des données, l'ajustement des équations et leur utilisation pour la prédiction. Les données utilisées proviennent de trois sites très différents en termes de structure forestière et de moyens mis à disposition. Des conseils pratiques en découlent qui devraient permettre aux lecteurs d'affronter la majeure partie des problèmes rencontrés couramment. Il intéressera aussi les biométriciens forestiers dans la mesure où il contient,

non seulement des rappels exhaustifs sur la théorie mathématique de la régression et ses récents développements, mais aussi de nombreux conseils sur le choix et la manipulation des modèles de régression linéaire.

220 pages. Nombreuses illustrations. Bibliographie de 255 titres.

Francis Cailliez

Août 2012

Remerciements

Les auteurs souhaitent remercier l'Organisation des Nations Unies pour l'alimentation et l'agriculture pour avoir financé l'édition et la traduction de ce manuel.

Les auteurs souhaitent également remercier les personnes citées ci-dessous pour avoir contribué aux campagnes de terrains et aux données utilisées pour le fil rouge, pour avoir enrichi de leur propre expérience le contenu du manuel et pour avoir accepté de relire et traduire ce manuel: Dr. Stephen Adu-Bredu, Angela Amo-Bediako, Dr. Winston Asante, Dr. Aurélien Besnard, Fabrice Bonne, Noëlle Bouxiero, Emmanuel Cornu, Dr. Rémi D'Annunzio, Dr. Christine Deleuze, Serge Didier, Justice Eshun, Charline Freyburger, Dominique Gelhaye, Dr. Astrid Genet, Dickson Gilmour, Hugues Yvan Gomat, Dr. Christophe Jourdan, Dr. Jean-Paul Laclau, Dr. Arnaud Legout, Lawrence et Susy Lewis, Dr. Fleur Longuetaud, Dr. Raphaël Manlay, Jean-Claude Mazoumbou, Adeline Motz, Dr. Alfred Ngomanda, Dr. Yann Nouvellon, Dr. Claude Nys, Charles Owusu-Ansah, Thierry Paul, Régis Peltier, Dr. Jacques Ranger, Michaël Rivoire, Gaël Sola, Luca Birigazzi, Dr. Olivier Roupsard, Dr. Armel Thongo M'bou et Prof. Riccardo Valentini.

Les auteurs remercient ceux qui, en dépit des brefs délais impartis, ont apporté leurs commentaires et suggestions de correction, ainsi que leurs encouragements. Cet ouvrage a grandement bénéficié de leur précieuse contribution. Cependant, la responsabilité de son contenu incombe uniquement aux auteurs.

Les méthodes synthétiques présentées dans ce manuel ont été élaborées lors de campagnes de mesure financées par les projets suivants: ATP Carbone (CIRAD), MODEL-FOR (ONF), BIOMASSE OPE (ANDRA), SOERE F-ORE-T (GIP ECOFOR), EMERGE (ANR), WAFT (UE), ULCOS (UE), CarboAfrica (UE, contrat n° INCO-CT-2004-003729).

Préambule

Ce manuel est destiné aux étudiants, chercheurs ou ingénieurs qui souhaitent acquérir la méthodologie nécessaire à l'établissement des tarifs de cubages, biomasses ou minéralomasse. Ces modèles font l'objet d'un ouvrage unique car ils font tous appel au même principe: il s'agit d'estimer une donnée difficile à mesurer sur tous les arbres d'un peuplement (par exemple le volume) à partir de caractéristiques plus simples comme le diamètre de l'arbre à 1,30 m, sa hauteur ou son âge.

Basé sur un ensemble de publications de référence, ce manuel ne présente pas tous les cas possibles mais propose des techniques qui permettent de construire ces équations. Les références dans le texte sont précises (dans la mesure du possible: auteur, année, page) pour que le lecteur puisse facilement retrouver les informations. Un exemple concret (baptisé « fil rouge ») guide le lecteur pour acquérir ces connaissances par la pratique.

Les pré-requis pour ce manuel sont limités. Les logiciels utilisés dans le fil rouge sont Microsoft Excel pour la préparation des fichiers et R (R Development Core Team, 2005) pour l'ajustement des modèles. Les lignes de commande de R utilisées sont reproduites dans le fil rouge.

1

Les bases de l'estimation de la biomasse

À l'échelle d'une population, il existe une relation statistique entre les différentes mensurations d'un individu (Gould, 1966). Cette relation découle du développement ontogénique des individus, qui est, à la variabilité liée à l'histoire de vie près, le même pour tous. Ainsi, les proportions entre hauteur et diamètre, entre taille du houppier et diamètre, entre biomasse et diamètre, obéissent à une règle qui est la même pour tous les arbres vivants dans les mêmes conditions, du plus petit au plus gros (King, 1996; Archibald et Bond, 2003; Bohlman et O'Brien, 2006; Dietze *et al.*, 2008). Il s'agit du principe de base de l'allométrie, qui permet de prédire une mensuration d'un arbre (typiquement sa biomasse) en fonction d'une autre mensuration (par exemple son diamètre). Une équation allométrique est une formule qui formalise de manière quantitative cette relation. Dans le cas de la prédiction du volume, de la biomasse ou de la masse minérale nous parlerons dans ce manuel de tarifs de cubage, biomasse et minéralomasse respectivement. Il existe une définition plus restrictive de l'allométrie, qui consiste en une relation de la proportionalité entre les accroissements relatifs des mensurations (Huxley, 1924; Gayon, 2000). Si on note B la biomasse et D le diamètre, cette seconde définition signifie qu'il existe un coefficient a tel que:

$$\frac{\mathrm{d}B}{B} = a\frac{\mathrm{d}D}{D}$$

ce qui s'intègre en une relation puissance: $B = b \times D^a$. Avec cette définition restreinte, une équation allométrique est donc synonyme d'équation puissance (White et Gould, 1965). Le paramètre a donne le coefficient d'allométrie (proportionnalité entre les accroissements relatifs), tandis que le paramètre b indique une proportionnalité entre les grandeurs cumulées. Il est quelquefois nécessaire de rajouter une ordonnée à l'origine à cette relation qui devient $B = c + bD^a$, où c représente la biomasse de l'individu avant qu'il n'atteigne la hauteur à laquelle est mesurée le diamètre (par exemple 1,30 m si D a été pris à 1,30 m). La relation puissance renvoie à l'idée d'auto-similarité lors du développement des individus (Gould, 1971). Partant de ce principe, et en s'appuyant sur la « pipe theory » (Shinozaki *et al.*, 1964*a*,*b*), une théorie fractale de l'allométrie a été développée (West *et al.*, 1997, 1999; Enquist *et al.*, 1998, 1999). Sous certaines hypothèses de contraintes biomécaniques de stabilité des arbres et de résistance hydraulique dans les réseaux de cellules conductrices, cette théorie prédit une relation puissance avec un exposant égal à $a = 8/3 \approx 2{,}67$ entre la biomasse et le diamètre des arbres. Cette relation est intéressante car elle est fondée sur

des principes physiques et une représentation mathématique des réseaux de cellules dans les arbres. Elle est toutefois largement discutée, son caractère trop général étant quelquefois mis en défaut (Zianis et Mencuccini, 2004; Zianis *et al.*, 2005; Muller-Landau *et al.*, 2006), bien qu'il existe des possibilités d'avoir d'autres coefficients d'allométrie selon les hypothèses biomécaniques et hydrauliques retenues (Enquist, 2002).

Dans le cadre de ce manuel, nous adopterons la définition la plus large de l'allométrie, qui renvoie à une corrélation (linéaire ou non) entre les accroissements des mensurations des arbres. La relation puissance ne sera donc qu'une relation allométrique parmi d'autres. Quelle que soit la définition adoptée, l'allométrie renvoie au développement ontogénique des individus, c'est-à-dire à la croissance des arbres.

1.1 La « biologie »: loi de Eichhorn, site index...

La croissance des arbres est un phénomène biologique complexe (Pretzsch, 2009) qui résulte de l'activité des bourgeons (croissance primaire, ou accroissements en longueur des axes) et du cambium (croissance secondaire, ou accroissement en épaisseur des axes). Cette croissance des arbres est éminemment variable puisqu'elle dépend du patrimoine génétique de l'individu, de son environnement (sol, atmosphère), du stade de développement (vieillissement des tissus) et de l'action des hommes (modification de l'environnement ou de l'arbre lui-même comme les élagages ou les tailles).

Classiquement, pour les études de biomasse, les arbres sont divisés en compartiments homogènes: le bois de tronc, l'écorce, les branches vivantes, les branches mortes, les feuilles, les grosses et moyennes racines, et enfin les racines fines. La biomasse est un volume multiplié par une densité tandis que la minéralomasse est une biomasse multipliée par une concentration en éléments minéraux. Volume, densité et concentration évoluent non seulement en fonction des facteurs pré-cités (voir par exemple la revue de Chave *et al.*, 2009 sur la densité du bois) mais aussi au sein des arbres: entre compartiments, mais également en fonction de la position radiale (près de la moelle ou près de l'écorce), et de la position longitudinale (près du sol ou près du houppier), voir par exemple pour les concentrations en éléments minéraux: Andrews et Siccama (1995); Colin-Belgrand *et al.* (1996); Saint-André *et al.* (2002*b*); Augusto *et al.* (2008); ou pour la densité du bois: Guilley *et al.* (2004); Bergès *et al.* (2008); Henry *et al.* (2010); Knapic *et al.* (2011). Tout ceci a des implications sur les équations de biomasse et minéralomasse et ce chapitre a pour objet de rappeler quelques notions importantes en foresterie qui permettront ensuite de penser les modèles de ce manuel en termes « biologiques » (quels sont les facteurs de variation potentiels ?) et non pas en termes purement statistiques (quelle est la meilleure équation possible indépendamment de son degré de vraisemblance vis-à-vis des processus biologiques ?). La combinaison des deux objectifs est, *in fine*, le sens même de cet ouvrage.

1.1.1 Cas des peuplements équiennes et monospécifiques

Ce type de peuplement est caractérisé par une relative homogénéité de la population d'arbres: ils ont le même âge et sont majoritairement de la même espèce. La croissance de ces peuplements a été très largement étudiée (de Perthuis, 1788 *in* Batho et García, 2006) et les principes décrits ci-après ont une valeur relativement générique (Assmann, 1970; Dhôte, 1991; Skovsgaard et Vanclay, 2008; Pretzsch, 2009; García, 2011). Il est usuel de distinguer le peuplement dans son ensemble, puis l'arbre au sein du peuplement. Cette distinction permet de dissocier les différents facteurs qui interviennent dans la croissance des arbres: fertilité du site, pression globale au sein du peuplement, et statut social. La fertilité du site

au sens large comprend la capacité du sol à nourrir les arbres (en nutriments et en eau) ainsi que le climat général de la zone (éclairement, température et pluviométrie moyennes, récurrence habituelle des périodes de gel ou de sécheresse, etc.). La pression entre les arbres au sein du peuplement est appréciée par différents indices de densité du peuplement. Enfin, le statut social de chaque individu définit sa capacité à mobiliser les ressources dans son environnement proche.

Photo 1.1 – *Plantation d'eucalyptus au Congo. En haut, zone de Kissoko, illustration des mosaïques savane-plantation. En bas, zone de Kondi en cours d'exploitation illustrant les principaux débouchés pour le bois d'eucalyptus (rondins pour la pâte à papier et production de charbon de bois pour la ville de Pointe-Noire (photos: L. Saint-André).*

Croissance du peuplement

La notion de production en foresterie comprend le volume (ou la biomasse) sur pied ainsi que tout ce qui a été soustrait du peuplement au cours de sa vie (par mortalité ou par éclaircie). En général, cette notion de production, telle qu'elle figure dans les tables de production ou dans la plupart des modèles de croissance à base dendrométrique, ne comprend pas les chutes de litières (feuilles, branches, écorce) ni le turn-over des racines. Par contre, dans les modèles à base écophysiologique ou les études ayant pour objet les bilans de carbone et d'éléments minéraux au sein des peuplements, la production comprend également ce renouvellement des organes. Pour la suite de ce chapitre, nous considérons la production dans son acception réduite.

La production d'un peuplement monospécifique et équienne, pour une essence donnée, dans une région donnée, et dans une large gamme de sylviculture (dès lors que le couvert est fermé), est complètement déterminée par sa hauteur moyenne. Cette affirmation est connue sous le nom de loi de Eichhorn (1904), ou loi de Eichhorn élargie quand elle considère la hauteur dominante au lieu de la hauteur moyenne (Decourt, 1973). Elle signifie que la fertilité des différents sites d'une même région ne modifie que la vitesse de parcours de cette relation. Même si elle a remise en question (voir Assmann, 1970), il en résulte que la croissance en hauteur des arbres dominants (H_0) constitue le moteur principal de la plupart

des modèles de croissance à base dendrométrique (par exemple Dhôte, 1996; García, 2003; Saint-André *et al.*, 2008; Skovsgaard et Vanclay, 2008; Weiskittel *et al.*, 2009; García, 2011). Le principe est résumé par Alder (1980 *in* Pardé et Bouchon, 1988) au sein de la phrase suivante: « La relation hauteur / âge / indice de fertilité constitue l'élément fondamental de la prévision de l'accroissement des peuplements homogènes. On l'exprime ordinairement sous la forme d'un faisceau de courbes de fertilité ». Le fait que la croissance en hauteur dominante ne dépende que de la fertilité du site (au sens large, équivalent en anglais de « site index ») et de l'âge des peuplements est valable, en première approximation, dans la plupart des écosystèmes monospécifiques et équiennes tempérés ou tropicaux. Cela résulte de deux facteurs principaux: les arbres dominants, du fait de leur statut, sont moins sensibles à la compétition que les arbres dominés et, par ailleurs, la croissance en hauteur est également moins sensible à la sylviculture (sauf régime d'éclaircies particulier) que la croissance en diamètre des arbres. Il en résulte donc que la croissance en hauteur des arbres dominants traduit bien mieux la fertilité d'un site que la croissance moyenne en hauteur ou en diamètre. Pour arriver à la loi de Eichhorn, il est ensuite nécessaire de coupler l'accroissement en surface terrière (ou en volume) et l'accroissement en hauteur dominante. Cette relation est également stable pour une essence et une région données dans une large gamme de sylviculture (en première approximation dès lors que le couvert est suffisamment dense). Un exemple, pour le hêtre en France, est donné par Dhôte (1996).

Il existe cependant quelques exemples où la relation stricte $H_0 = f(\text{âge et fertilité})$ est mise en défaut: le pin Laricio dans le centre de la France (Meredieu *et al.*, 2003) et l'eucalyptus au Congo (Saint-André *et al.*, 2002*a*). Dans les deux cas, la croissance en hauteur dominante est également fonction de la densité du peuplement. L'hypothèse sous-jacente est liée à la faible fertilité des sols qui impliquerait une forte compétition pour l'accès aux ressources hydriques et minérales, même sur les arbres dominants. Depuis quelques années, il est également clairement mis en évidence, sur cette relation et celle qui relie l'accroissement en surface terrière à l'accroissement en hauteur dominante, un effet « date » du fait des changements globaux (voir par exemple Bontemps *et al.*, 2009, 2011; Charru *et al.*, 2010).

En résumé, même si elle est discutée et pas nécessairement aussi invariable qu'espéré, cette première « loi » est importante car elle permet ensuite, dans les tarifs de biomasse dits paramétrés, d'introduire la notion de fertilité *via* l'âge et la hauteur dominante des peuplements inventoriés (et accessoirement la densité) pour augmenter le caractère générique des équations construites.

Croissance des arbres au sein du peuplement

Lorsque la croissance du volume ou de la biomasse dans l'ensemble du peuplement est obtenue, il s'agit ensuite de la répartir entre les différents arbres. Les relations utilisées pour la croissance en diamètre individuelle sont souvent du type potentiel × réducteur, où le potentiel est donné par la croissance en surface terrière et / ou en hauteur dominante et les réducteurs sont fonction (*i*) d'un indice de densité et (*ii*) du statut social de l'arbre. Les indices de densité peuvent être tout simplement la densité du peuplement, mais les chercheurs ont élaboré d'autres indices comme le facteur d'espacement de Hart-Becking, basé sur la croissance des arbres hors peuplement (croissance libre), ou le RDI (« Reinecke density index »), basé sur la loi d'auto-éclaircie (croissance des arbres en peuplement hyper-dense) qui présentent tous deux l'avantage d'être moins dépendants de l'âge du peuplement que la densité elle-même (voir Shaw, 2006 ou plus généralement la revue bibliographique de Vanclay, 2009). Le statut social des arbres est en général exprimé par des ratios du type H/H_0

ou D/D_0 (où D est le diamètre de l'arbre, H sa hauteur, et D_0 le diamètre dominant du peuplement), mais d'autres relations peuvent également être utilisées. Par exemple, Dhôte (1990), Saint-André *et al.* (2002*a*) et plus récemment Cavaignac *et al.* (2012) utilisent un modèle linéaire segmenté pour traduire la croissance en épaisseur des arbres: en dessous d'un certain seuil de circonférence, les arbres sont surcimés et ne poussent plus; au delà, l'accroissement en surface terrière est une fonction linéaire de la circonférence des arbres. Cette relation traduit bien le fait que les arbres dominants poussent plus que les arbres dominés. Le seuil et la pente de la relation évoluent en fonction de l'âge des peuplements et de la sylviculture (éclaircies). La croissance en hauteur peut être estimée également *via* des relations du type potentiel × réducteur mais, en général, les modélisateurs utilisent des relations hauteur–circonférence (Soares et Tomé, 2002). Ces relations sont saturées (l'asymptote est égale à la hauteur dominante du peuplement) et curvilinéaires. Les paramètres de cette relation évoluent également en fonction de l'âge et de la sylviculture (Deleuze *et al.*, 1996).

En résumé pour ces deux autres relations qui donnent la dimension de chaque arbre au sein des peuplement, il faut retenir le point suivant: les indices de densité et de compétition (statut social) qui déterminent pour une large part la croissance individuelle des arbres au sein des peuplements sont des facteurs qu'il est également possible d'intégrer aux tarifs de biomasse. Les variables intéressantes à ce point de vue peuvent être: la densité du peuplement, le facteur d'espacement de Hart-Becking, le RDI, puis à une échelle individuelle: le coefficient d'élancement (H/D), la robustesse de l'arbre ($D^{1/2}/H$ — Vallet *et al.*, 2006; Gomat *et al.*, 2011), ou son statut social (H/H_0 ou D/D_0).

Répartition de la biomasse dans l'arbre

Enfin, à présent que la biomasse du peuplement a été répartie entre les arbres, il faut, pour chaque arbre, l'attribuer à chaque compartiment et la répartir le long des axes. Pour le tronc, une relation communément utilisée est la loi de Pressler (ou pour les écophysiologistes, son équivalent donné par le « pipe-model » de Shinozaki *et al.*, 1964*a*,*b*): (*i*) la surface du cerne augmente linéairement du haut de l'arbre jusqu'à la base fonctionnelle du houppier; (*ii*) elle reste ensuite constante de la base du houppier jusqu'au bas de l'arbre. Par conséquent, au fur et à mesure que l'arbre grandit, le tronc va devenir de plus en plus cylindrique puisque les largeurs des cernes sont plus fortes près du houppier qu'en bas. Cette loi de Pressler n'exprime toutefois qu'une répartition moyenne du bois le long de l'arbre (Saint-André *et al.*, 1999). En effet, pour les arbres dominants, la surface du cerne peut continuer d'augmenter sous la base du houppier et pour les dominés / surcimés, elle peut diminuer fortement. Dans les cas extrêmes, il est même possible que le cerne ne soit pas complet en bas des arbres, voire même qu'il soit manquant, comme par exemple chez le hêtre (Nicolini *et al.*, 2001). Par ailleurs, toute action sur le houppier (fortes ou faibles densités, éclaircies, élagages) va avoir des conséquences sur l'empilement des cernes et donc sur la forme du tronc (voir la revue de Larson, 1963, ou des exemples donnés par Valinger, 1992; Ikonen *et al.*, 2006). La densité du bois est également différente en haut et en bas de l'arbre (bois juvénile près du houppier, forte proportion de bois adulte en bas — Burdon *et al.*, 2004) mais va aussi varier selon les conditions de croissance des arbres (*via* des changements de proportion entre le bois d'été et de bois de printemps, ou des changements de structure et propriétés cellulaires, voir Guilley *et al.*, 2004; Bouriaud *et al.*, 2005; Bergès *et al.*, 2008 pour quelques publications récentes). Il en résulte, pour le tronc, qu'à dimensions égales (hauteur, diamètre, âge) des tiges, la biomasse sera ou non différente selon les conditions de croissance des arbres. Il est parfaitement possible qu'une augmentation du volume soit, par exemple, accompagnée par une baisse de densité (c'est un schéma classique sur les résineux)

et ne conduise pas à des différences majeures sur la biomasse des troncs. Pour les branches et les feuilles, la biomasse portée sera fortement fonction de l'architecture des arbres et par voie de conséquence de la densité du peuplement: à dimensions égales (hauteur, diamètre et âge) les arbres ayant poussé dans des peuplements ouverts auront plus de branches et de feuilles que les arbres ayant poussés dans des peuplements denses. Tout l'enjeu des recherches actuelles sur la biomasse consiste à identifier la part liée au développement intrinsèque de l'arbre (ontogénie) de la part liée aux facteurs environnementaux (Thornley, 1972; Bloom *et al.*, 1985; West *et al.*, 1999; McCarthy et Enquist, 2007; Savage *et al.*, 2008; Genet *et al.*, 2011; Gourlet-Fleury *et al.*, 2011). Pour les racines, leur biomasse dépend du biome, de la biomasse aérienne, du stade de développement et des conditions de croissance (cf. par exemple Jackson *et al.*, 1996; Cairns *et al.*, 1997; Tateno *et al.*, 2004; Mokany *et al.*, 2006).

De ces dernières notions, il faut retenir que les conditions de croissance vont non seulement influencer la quantité globale de biomasse produite mais aussi leur répartition au sein des arbres (proportion aérien / souterrain; empilement des cernes, etc.). Il sera donc absolument nécessaire de prendre en compte ces variations potentielles dans l'échantillonnage (en particulier pour le billonnage des troncs et le prélèvement des différentes aliquotes) mais aussi dans la construction des tarifs de façon à ce qu'ils restituent correctement les différents ratios de biomasse (aérien / souterrain; tronc / branches; feuilles / racines fines) en fonction des conditions de croissance.

1.1.2 Cas des peuplements inéquiennes et/ou plurispécifiques

Les notions décrites précédemment restent valables aussi pour les peuplements plurispécifiques et inéquiennes mais leur mise en équation devient difficile et la plupart du temps impossible sous la forme précédente (Peng, 2000). Par exemple, la notion de hauteur dominante est difficile à quantifier pour les peuplements irréguliers et / ou plurispécifiques (faut-il faire une hauteur dominante toutes essences confondues ? Essence par essence ?). De même, que signifie la surface terrière pour un peuplement fortement irrégulier comme il est possible d'en trouver en forêt tropicale humide ? Enfin, comment gérer le fait que l'âge des arbres est souvent inaccessible (Tomé *et al.*, 2006) ? Les modèles de croissance élaborés pour ces peuplements décomposent donc moins finement les différentes échelles (production de biomasse à l'échelle du peuplement, répartition entre arbres et allocation au sein des arbres) que ceux pour les peuplements réguliers. On peut distinguer trois types de modèles: (1) les modèles peuplement matriciels; (2) les modèles individus centrés qui, en général, dépendent des distances entre arbres; (3) les modèles de trouées (voir les différentes revues réalisées par Vanclay, 1994; Franc *et al.*, 2000; Porté et Bartelink, 2002). Les modèles de type matriciel regroupent les arbres par groupes fonctionnels (groupes ayant une stratégie de croissance commune) et par classes de dimension homogènes (en général le diamètre) et appliquent un système de matrices qui incluent le recrutement, la mortalité et le passage des individus d'un groupe à l'autre (voir par exemple Eyre et Zillgitt, 1950; Favrichon, 1998; Namaalwa *et al.*, 2005; Picard *et al.*, 2008). Pour les modèles individus centrés, la population des arbres est en général cartographiée et la croissance d'un arbre dépend de ses voisins (voir par exemple Gourlet-Fleury et Houllier, 2000 pour un modèle en forêt tropicale, ou Courbaud *et al.*, 2001 pour un modèle en forêt tempérée). Mais à l'instar des modèles développés pour les peuplements réguliers, il existe aussi des modèles individus centrés indépendants des distances (par exemple Calama *et al.*, 2008; Pukkala *et al.*, 2009; Vallet et Pérot, 2011; Dreyfus, 2012) et même des modèles intermédiaires (voir Picard et Franc, 2001; Verzelen *et al.*, 2006; Perot *et al.*, 2010). Enfin, dans les modèles de trouées, la forêt est représentée par un ensemble de cellules à différents stades du cycle sylvigénétique. La mortalité et le recrutement des arbres

y sont simulés stochastiquement tandis que la croissance des arbres suit des lois identiques à celles des modèles individus centrés et indépendants des distances (voir une revue dans Porté et Bartelink, 2002).

Le fait que ces peuplements soient plus compliqués à mettre en équation ne retire rien aux principes évoqués plus haut pour la construction des tarifs de volume, de biomasse ou de minéralomasse: (*i*) introduire la fertilité pour élargir la zone de validité des tarifs de biomasse; (*ii*) utiliser des indices de densité pour prendre en compte le degré de concurrence entre arbres; et (*iii*) tenir compte du statut social en plus des caractéristiques basiques des arbres (hauteur, diamètre).

En plus des contraintes liées à l'élaboration des tarifs développés pour les forêts monospécifiques, l'estimation de la biomasse en forêt plurispécifiques fait face à des contraintes additionnelles: la mise en place d'un échantillonnage adapté (quelles essences ? comment les regrouper en groupes dits fonctionnels ?) et l'accès au terrain (surtout en zone tropicale où ces peuplements sont souvent situés en zone de protection où l'abattage des arbres est très réglementé, voire interdit pour certaines essences).

Photo 1.2 – *Peuplements hétérogènes. À gauche, cas de peuplements plurispécifiques sur le Mont Saint-Anne au Québec; à droite, peuplements plurispécifiques et inéquiennes au Costa-Rica (photo: B. Locatelli).*

1.2 Choix de la méthode

1.2.1 Estimation de la biomasse d'un biome

Il n'existe pas « une » méthode pour estimer un stock de biomasse, mais plusieurs selon l'échelle considérée (Gibbs *et al.*, 2007). À l'échelle nationale et au-delà, des valeurs moyennes par biome sont le plus souvent utilisées (FAO, 2006): la quantité de biomasse est estimée en multipliant la superficie de chaque biome par la quantité de biomasse moyenne par unité de surface pour ce biome. Les quantités moyennes par biome sont elles-mêmes estimées à partir de mesures faites à une échelle plus restreinte. De l'échelle nationale à l'échelle du paysage, la télédétection peut permettre d'estimer la biomasse. Qu'il s'agisse de capteurs optiques satellitaires (Landsat, MODIS), d'images à haute résolution satellitaires (Ikonos, QuickBird) ou non (photographies aériennes), de capteurs radar ou micro-ondes

satellitaires (ERS, JERS, Envisat, PALSAR), ou de capteurs laser (Lidar), toutes ces méthodes supposent que des mesures de terrain sont disponibles pour caler des relations qui prédisent la biomasse en fonction des observations faites par les capteurs. Dans le cas des capteurs optiques satellitaires, des données de terrain sont nécessaires pour calibrer la relation entre la biomasse et les indices de végétation satellitaires (NDVI, NDFI, AVI, GVI, etc.) (Dong *et al.*, 2003; Saatchi *et al.*, 2007). Les images à haute résolution et les photographies aériennes fournissent des informations sur la taille des couronnes et la hauteur des arbres. Des données de terrain sont ensuite nécessaires pour relier ces informations à la biomasse (par exemple Bradley, 1988; Holmgren *et al.*, 1994; St.-Onge *et al.*, 2008; Gonzalez *et al.*, 2010). Il en va de même pour les informations sur la structure verticale de la forêt fournies par le Lidar, ou pour les informations sur la distribution verticale de l'eau contenue dans la végétation fournies par le radar ou les micro-ondes (par exemple Lefsky *et al.*, 2002; Patenaude *et al.*, 2004). Les méthodes de télédétection restent cependant limitées quant à la précision des mesures de biomasse (particulièrement les surfaces) et la différentiation des types de forêts en fonction des moyens techniques, financiers, des ressources humaines disponibles, de la couverture nuageuse et du risque de saturation des signaux utilisés pour certains types de végétation.

Ainsi, les méthodes d'estimation de la biomasse à l'échelle du paysage et au-delà reposent sur des mesures de terrain, à une échelle comprise entre le paysage et la parcelle. Dans cette gamme d'échelle, les estimations de la biomasse reposent sur des données d'inventaire forestier: inventaire d'un échantillon d'arbres si la surface est grande, ou inventaire en plein dans le cas contraire (en particulier dans les parcelles permanentes de quelques hectares). En deçà de cette échelle, ce sont des mesures individuelles de biomasse (pesée des arbres, pesée de la végétation du sous-bois) qui entrent en ligne de compte.

1.2.2 Estimation de la biomasse d'une forêt ou d'un ensemble de forêts

Les estimations de biomasse ou minéralomasse forestières basées sur les inventaires en forêt nécessitent de disposer

1. d'un inventaire exhaustif ou statistique des tiges présentes;
2. de modèles pour évaluer les stocks à partir des dimensions des individus mesurés;
3. d'une évaluation de la biomasse contenue dans la nécromasse (bois morts sur pied) et dans la végétation du sous-étage.

Nous avons concentré ce manuel sur le second aspect tout en sachant que la partie inventaire ou l'évaluation quantitative du sous-étage ne sont pas forcément faciles à réaliser en particulier dans les forêts très mélangées.

À partir des inventaires, deux grandes options peuvent être utilisées pour estimer les stocks de carbone ou d'éléments minéraux dans les arbres (MacDicken, 1997; Hairiah *et al.*, 2001; AGO, 2002; Ponce-Hernandez *et al.*, 2004; Monreal *et al.*, 2005; Pearson et Brown, 2005; Dietz et Kuyah, 2011): (1) l'utilisation de tarifs de biomasse / minéralomasse: cette solution est souvent adoptée car elle permet d'établir rapidement des bilans de carbone ou d'éléments minéraux au sein d'une parcelle à un instant donné. En général, tous les compartiments de l'écosystème sont considérés (aériens, souterrains, litières au sol, etc.). Les arbres sont spécifiquement abattus pour ces opérations. La définition des compartiments (découpes) peuvent varier selon l'application et le domaine d'intérêt (voir chapitre 3). (2) L'utilisation de modèles pour estimer successivement le volume des arbres, la densité du bois, et les teneurs en éléments minéraux. L'avantage de cette méthode est qu'elle dissocie les différentes composantes. Il est alors possible d'analyser l'influence de l'âge et des conditions de croissance indépendamment sur l'une ou l'autre des composantes. En général, seul le

tronc peut faire l'objet d'une modélisation détaillée (cerne, voir intra-cernes), la biomasse des autres compartiments étant estimée à partir de coefficients d'expansion volumique, de valeurs de densité moyenne du bois et de teneurs en éléments minéraux. Dans tous les cas, ces méthodes font largement appel à un grand type de modèle qui regroupe indifféremment les « tarifs de cubage, tarifs de biomasse, tarifs de minéralomasse, etc. » et qui fait l'objet de ce manuel.

Les tarifs de biomasse ou de minéralomasse s'apparentent beaucoup aux tarifs de cubages, modèles qui sont largement étudiés depuis près de deux siècles. Le premier tarif a été publié par Cotta, 1804 (*in* Bouchon, 1974) sur du hêtre (*Fagus sylvatica*). Le principe est de relier une grandeur difficile à mesurer (comme le volume de l'arbre, sa masse, ou sa teneur en éléments minéraux) à des grandeurs plus faciles à appréhender comme le diamètre à 1,30 m ou la hauteur de l'arbre. Si ces deux caractéristiques sont utilisées, on parle de tarif à deux entrées; si seulement le diamètre est utilisé, on parle alors de tarif à une entrée. En général, les corrélations sont bonnes et les fonctions les plus utilisées sont du type polynômial, logarithmique, ou puissance. Pour plus de détails, on peut se reporter aux revues proposées par Bouchon (1974); Hitchcock et McDonnell (1979); Pardé (1980); Cailliez (1980); Pardé et Bouchon (1988), et plus récemment par Parresol (1999, 2001).

Ces fonctions sont relativement simples mais présentent trois écueils majeurs. Premièrement, elles sont assez peu génériques: si on change d'espèce ou si on s'éloigne du domaine de calibration, les équations sont à utiliser avec précaution. Le chapitre sur l'échantillonnage donne quelques éclairages pour pallier ce problème. Le principe essentiel est de couvrir au maximum la variabilité des quantités étudiées.

Un deuxième écueil de ces fonctions réside dans la nature même des données qui sont traitées (volumes, masses, minéralomasses). En particulier, des problèmes d'hétéroscédasticité peuvent intervenir (c'est-à-dire variance non homogène des biomasses en fonction du régresseur). Cela influe peu sur la valeur des paramètres estimés: plus le nombre d'arbres échantillonnés est grand, plus la convergence vers les vrais paramètres est rapide (Kelly et Beltz, 1987). Cependant, tout ce qui concerne l'intervalle de confiance des estimations est affecté:

1. la variance des paramètres estimés n'est pas minimale;
2. cette variance est biaisée; et
3. la variance résiduelle est mal estimée (Cunia, 1964; Parresol, 1993; Gregoire et Dyer, 1989).

Ne pas corriger ces problèmes d'hétéroscédasticité a donc peu de conséquences sur l'estimation de la valeur moyenne de la biomasse ou du volume. En revanche, cela est absolument nécessaire pour obtenir des intervalles de confiance corrects autour des prédictions. Pour corriger ces problèmes d'hétéroscédasticité, deux méthodes sont souvent présentées: la première consiste à effectuer une pondération (par exemple par l'inverse du diamètre ou du diamètre au carré) mais tout réside dans la fonction de pondération et en particulier dans la puissance à appliquer; la seconde consiste à prendre le logarithme des termes de l'équation, mais dans ce cas, il est nécessaire de corriger les valeurs simulées pour retrouver une distribution normale des valeurs estimées (Duan, 1983; Taylor, 1986). De plus, il n'est pas rare que la transformation logarithmique ne conduise pas à un modèle linéaire (Návar *et al.*, 2002; Saint-André *et al.*, 2005).

Le troisième écueil est lié à l'additivité des équations. Les mesures de biomasse et ensuite les ajustements des fonctions sont souvent réalisés compartiment par compartiment. L'additivité des relations n'est pas immédiate et une propriété souhaitée du système d'équations est que la somme des prédictions des biomasses compartiment par compartiment soit égale

à la prédiction de la biomasse totale de l'arbre (voir Kozak, 1970; Reed et Green, 1985; Návar *et al.*, 2002). Trois solutions sont en général proposées (Parresol, 1999):

1. la biomasse totale est calculée en faisant la somme des biomasses compartiment par compartiment, et la variance de cette estimation utilise les variances calculées sur chaque compartiment et les covariances calculées deux à deux;
2. l'additivité est assurée en utilisant les mêmes régresseurs et les même poids pour toutes les fonctions, les paramètres de la fonction de biomasse totale étant la somme des paramètres obtenus pour chaque compartiment;
3. les modèles sont différents compartiment par compartiment mais sont ajustés conjointement et l'additivité est obtenue par des contraintes sur les paramètres.

Chaque méthode a ses inconvénients et ses avantages. Dans le cadre de ce manuel, nous ajusterons un modèle pour chaque compartiment et un modèle pour la biomasse totale en vérifiant que l'additivité est bien respectée. Pour illustrer ce manuel, un exemple concret (baptisé « fil rouge ») est utilisé tout au long du manuel. Il concerne un jeu de données obtenu dans le cadre d'une expérimentation menée au Ghana en forêt naturelle tropicale humide (Henry *et al.*, 2010).

1.2.3 Mesure de la biomasse d'un arbre

Le tarif de biomasse est l'outil qui assure le lien entre la mesure individuelle de la biomasse et l'estimation de la biomasse sur le terrain à partir de données d'inventaire. Peser des arbres pour en mesurer la biomasse fait donc partie intégrante de la démarche pour construire des équations allométriques et une bonne partie de ce manuel y sera consacré. Même si les principes généraux présentés dans le chapitre 3 (segmentation de l'arbre en compartiments ayant une densité de matière sèche homogène, mesure des ratios matière sèche sur volume frais pour des aliquotes et application d'une règle de trois...) devraient permettre de mesurer la biomasse de n'importe quel type de végétal arboré, il n'en reste pas moins que ce manuel n'abordera pas tous les cas particuliers. Les végétaux qui ne sont pas des arbres mais qui ont potentiellement la stature d'un arbre (bambous, rotins, palmiers, fougères arborescentes, Musaceae, *Pandanus sp.*, etc.) font partie de ces exceptions.

Les végétaux qui utilisent les arbres comme support pour leur croissance (épiphytes, plantes parasites, lianes, etc.) sont un autre cas particulier (Putz, 1983; Gerwing et Farias, 2000; Gerwing *et al.*, 2006; Gehring *et al.*, 2004; Schnitzer *et al.*, 2006, 2008). Leur biomasse devra être dissociée de celle de leur hôte.

Enfin les arbres creux, les arbres dont le tronc a une forme fortement différente d'un cylindre (comme *Swartzia polyphylla* DC.), les figuiers étrangleurs, etc., constituent autant d'exceptions pour lesquelles les tarifs de biomasse ne pourront pas être appliqués sans ajustement spécifique (Nogueira *et al.*, 2006).

2

Échantillonnage et stratification

L'échantillonnage consiste à prédire les caractéristiques d'un ensemble à partir d'une partie (l'échantillon) de cet ensemble. Typiquement, on veut estimer le volume de bois dans une forêt, mais il n'est pas possible de cuber tous les arbres un à un: on va donc cuber un échantillon d'arbres de la forêt, puis extrapoler l'estimation obtenue pour ces arbres à la forêt entière (CTFT, 1989, p.252). Comme le volume n'est mesuré que sur un échantillon d'arbres et non sur l'ensemble des arbres du peuplement, l'estimation du volume total que l'on obtient est entachée d'une *erreur d'échantillonnage*[1]. L'échantillonnage *stricto sensu* consiste:

1. à choisir au mieux les arbres qui feront partie de l'échantillon de mesure (on parle plutôt de *plan d'échantillonnage*),
2. à choisir la méthode de calcul (on parle plutôt d'*estimateur*) du volume total à partir des mesures,

de manière à minimiser l'erreur d'échantillonnage.

Dans la théorie d'échantillonnage classique, les volumes des N arbres du peuplement sont des données fixées: la seule source de variation des estimations est l'échantillonnage, de sorte qu'un échantillonnage exhaustif donnerait toujours la même estimation. Nous adopterons ici l'approche dite de super-population, qui a vu le jour dans les années 1970 (Cochran, 1977). Elle consiste à considérer que les volumes des N arbres qui composent le peuplement sont des variables aléatoires, de sorte que le peuplement observé n'est qu'une réalisation parmi d'autres tirée d'une super-population. Cette approche permet de s'affranchir de certaines approximations et de définir un plan d'échantillonnage optimal (ce qui n'est souvent pas possible dans l'approche classique), mais elle a l'inconvénient de conduire à de mauvaises solutions si le modèle de super-population adopté n'est pas conforme à la réalité.

Le choix d'une méthode d'échantillonnage découle de l'objectif fixé. Il faut donc en principe commencer par se demander à quoi vont servir les tarifs de cubage ou de biomasse que l'on se propose de construire. S'agit-il de prédire les caractéristiques d'un arbre particulier dont les variables d'entrée sont connues ? S'agit-il de prédire les caractéristiques de l'arbre moyen pour des valeurs données des variables d'entrée ? S'agit-il de prédire le volume total

1. Nous mettons en italique les termes du jargon de la théorie de l'échantillonnage; une définition en français de ces termes se trouve, par exemple, à l'annexe 2 de Bellefontaine *et al.* (2001).

du peuplement dont sont issus les arbres ayant servi à construire le tarif, ou le volume total d'un autre peuplement ? Dans ces deux derniers cas de figure, les variables d'entrée du tarif sont-elles mesurées sur tous les arbres du peuplement, ou à nouveau sur un échantillon d'arbres ? Etc. On peut ainsi construire une chaîne allant du peuplement étudié jusqu'à la grandeur que l'on cherche à prédire (figure 2.1).

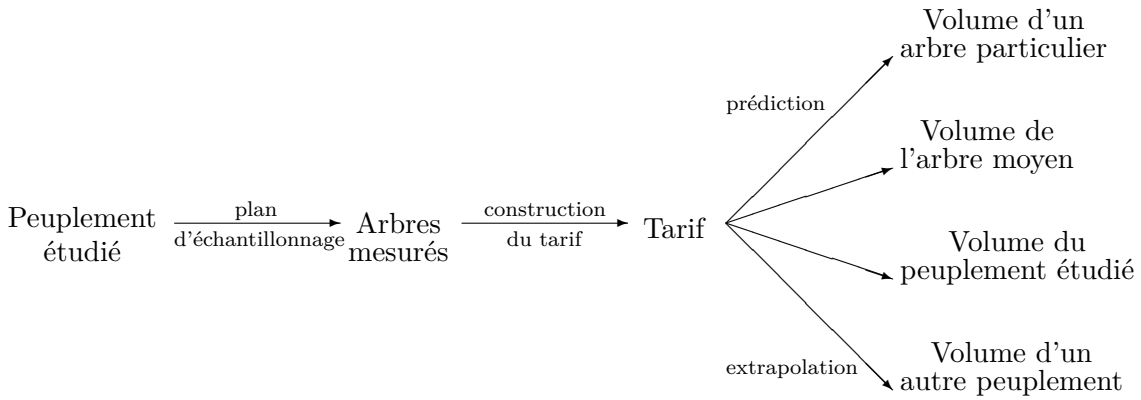


FIGURE 2.1 – *Chaîne allant du peuplement étudié jusqu'aux grandeurs que l'on cherche à prédire.*

En déroulant cette chaîne à l'envers, la précision sur la grandeur prédite dépend de la précision sur les paramètres du tarif, qui dépend elle-même du plan d'échantillonnage (nombre et choix des arbres mesurés) et de la variabilité au sein du peuplement étudié (Cunia, 1987*b*). On peut se fixer un objectif de précision à atteindre sur les prédictions, ce qui, par effet rétroactif, pour un type de tarif donné et un type d'échantillonnage donné, implique le nombre minimal d'arbres à mesurer. On peut aussi suivre une procédure d'optimisation afin de déterminer, pour une précision à atteindre donnée et un type de tarif donné, la méthode d'échantillonnage qui minimise le temps (ou le coût) des mesures (Cunia, 1987*c*,*d*). Dans certains cas, le coût des mesures est le facteur limitant. C'est en particulier le cas pour mesurer la biomasse de systèmes racinaires. Dans ce cas on ne cherche pas tant à atteindre une précision donnée pour les prédictions qu'à rester dans des limites raisonnables de coût. On peut alors rechercher, à coût de mesure donné et pour un type de tarif donné, la méthode d'échantillonnage qui maximise la précision des estimations.

Ce raisonnement, souvent trop complexe pour être suivi de manière rigoureuse, doit être fait au cas par cas puisqu'il dépend (*i*) de ce que l'on cherche à prédire, (*ii*) du type de tarif utilisé et (*iii*) du type d'échantillonnage adopté. Le fait d'utiliser un tarif est déjà en soi un choix sur la méthode d'échantillonnage: le volume total d'un peuplement pourrait être estimé à partir du cubage d'un échantillon d'arbres, sans passer par un tarif de cubage. En utilisant un tarif de cubage pour estimer le volume total d'un peuplement, on s'est déjà restreint à un type d'*estimateur* du volume total.

Qui plus est, le raisonnement qui permet de déterminer le plan d'échantillonnage en fonction de la précision à atteindre sur les prédictions suppose que l'on connaisse la relation entre la précision des prédictions et la précision des paramètres du tarif, la relation entre la précision des paramètres du tarif et la taille d'échantillon, etc. Dans certains cas simples, ces relations sont connues explicitement. Mais le plus souvent, dès que le tarif prend une forme un peu compliquée, ces relations ne sont pas explicites. On ne peut alors plus dérouler simplement le raisonnement.

Le coup de grâce est porté en pratique à ce raisonnement quand on se rend compte que: (*i*) la finalité du tarif est en général multiple, voire non précisée et (*ii*) la forme du tarif n'est en général pas connue à l'avance. On souhaite en effet le plus souvent pouvoir utiliser

un tarif à différentes fins: pour évaluer le volume d'un arbre particulier, d'un arbre moyen, d'un peuplement entier, etc. La construction d'un tarif est au bout du compte une fin en soi, sans référence à une quantité à prédire. De plus, le choix de la forme du tarif résulte le plus souvent d'une analyse exploratoire des données et n'est donc pas connue à l'avance. Certes, certaines relations, comme la fonction puissance ou les polynômes de degré deux, reviennent souvent mais on ne peut pas se fixer de règle *a priori*. Il est dès lors illusoire de chercher à optimiser un plan d'échantillonnage.

Au bout du compte, l'échantillonnage dans le cas de la construction de tarifs se ramène le plus souvent à des considérations empiriques sur le plan d'échantillonnage, le choix de l'*estimateur*, qui se ramène en fait au choix du tarif, étant raisonné *a posteriori*, en fonction des données collectées et indépendamment du plan d'échantillonnage.

2.1 Échantillonnage pour une régression linéaire simple

Nous commençons par un exemple simple qui permettra d'illustrer les idées développées précédemment. On suppose que les arbres du peuplement sont décrits par leur diamètre D, leur hauteur H et leur volume V. On utilise un tarif de cubage qui prédit le volume V en fonction de la variable D^2H. Le modèle de super-population que l'on adopte pour décrire le peuplement suppose que la relation entre V et D^2H est linéaire avec un bruit blanc ε de variance σ^2:

$$V = \alpha + \beta\, D^2H + \varepsilon \tag{2.1}$$

où ε suit une loi normale d'espérance nulle et d'écart-type σ. On suppose de plus que la quantité D^2H est distribuée selon une loi normale de moyenne μ et d'écart-type τ. Le bruit blanc ε incorpore tous les facteurs qui font que deux arbres de même diamètre et de même hauteur n'ont pas forcément le même volume. Les paramètres α et β sont inconnus. Pour les estimer, on va mesurer n arbres; on obtient ainsi un échantillon de n doublets $(D_1^2H_1, V_1)$, $\ldots$, $(D_n^2H_n, V_n)$, puis on fait la régression linéaire suivante:

$$V_i = a + b\, D_i^2H_i + \varepsilon_i \tag{2.2}$$

Dans le jargon de la théorie de l'échantillonnage, les variables d'entrées du tarif (diamètre, hauteur...) sont appelées des variables *auxiliaires*. Il faut bien distinguer ces variables, qui sont relatives à l'arbre, de variables telles que l'âge qui sont relatives au peuplement. Ces dernières sont considérées comme des paramètres (Pardé et Bouchon, 1988, p.106). Par ailleurs l'*unité* d'échantillonnage est l'arbre. Voyons à présent comment définir le plan d'échantillonnage en fonction de l'objectif fixé.

2.1.1 Prédiction du volume d'un arbre particulier

Supposons que l'objectif soit de prédire le volume d'un arbre du peuplement de diamètre D^* et de hauteur H^*. Le volume prédit est bien entendu:

$$V^* = a + b\, D^{*2}H^*$$

Le modèle de super-population stipule que, du fait du bruit blanc ε, deux arbres pris au hasard et ayant le même diamètre D et la même hauteur H n'ont pas forcément le même volume. Il en résulte une variabilité intrinsèque quand on mesure un arbre particulier, qui est égale à σ^2. À cette variabilité intrinsèque s'ajoute, pour l'erreur de prédiction du volume, la variabilité due à l'imprécision des estimations des paramètres α et β du tarif de cubage. Nous reviendrons plus tard sur ces notions (dans le chapitre 7). Ainsi, pour une régression

linéaire, la demi-amplitude de l'intervalle de confiance au seuil α (typiquement 5 %) de V^* est égale à (Saporta, 1990, p.374):

$$t_{n-2}\,\hat{\sigma}\,\sqrt{1+\frac{1}{n}+\frac{(D^{*2}H^*-\overline{D^2H}_e)^2}{\sum_{i=1}^{n}(D_i^2H_i-\overline{D^2H}_e)^2}}$$

où t_{n-2} est le quantile $1-\alpha/2$ d'une loi de Student à $n-2$ degrés de liberté, $\overline{D^2H}_e$ est la moyenne empirique des valeurs de D^2H mesurées dans l'échantillon:

$$\overline{D^2H}_e=\frac{1}{n}\sum_{i=1}^{n}D_i^2H_i$$

et $\hat{\sigma}$ est l'estimation de l'écart-type des résidus:

$$\hat{\sigma}^2=\frac{1}{n-2}\sum_{i=1}^{n}[V_i-(a+b\,D_i^2H_i)]^2$$

La valeur minimale de cette demi-amplitude (lorsque $n\to\infty$) est $1{,}96\,\sigma$. On se donne comme objectif de précision de l'estimation un écart de E % par rapport à ce minimum incompressible, c'est-à-dire que, de façon approximative, on recherche la taille d'échantillon n telle que:

$$1+E\approx\sqrt{1+\frac{1}{n}+\frac{(D^{*2}H^*-\overline{D^2H}_e)^2}{\sum_{i=1}^{n}(D_i^2H_i-\overline{D^2H}_e)^2}} \tag{2.3}$$

Échantillonnage aléatoire

Voyons tout d'abord le cas où l'on ne cherche pas à optimiser le plan d'échantillonnage, par exemple en sélectionnant les arbres de l'échantillon au hasard. La moyenne empirique de D^2H sur l'échantillon est alors une estimation de μ, tandis que la variance empirique de D^2H sur l'échantillon est une estimation de τ^2. Ainsi:

$$(1+E)^2-1\approx\frac{1}{n}\left[1+\frac{(D^{*2}H^*-\mu)^2}{\tau^2}\right]$$

À titre d'exemple numérique, prenons $\mu=5$ m^3 pour la valeur moyenne de D^2H dans le peuplement en entier et $\tau=1$ m^3 pour son écart-type. Si on veut prédire le volume d'un arbre dont la taille D^2H est égale à 2 m^3 avec un écart de précision de $E=5\,\%$, alors il faut mesurer approximativement $n=98$ arbres. On remarquera que l'expression de n en fonction de $D^{*2}H^*$ est symétrique autour de μ et passe par un minimum pour $D^{*2}H^*=\mu$. Comme $\mu-2=3$ m^3 et $\mu+3=8$ m^3, il faut donc également $n=98$ arbres pour prédire le volume d'un arbre dont la taille D^2H est égale à 8 m^3 avec un écart de précision de 5 %. On peut ainsi interpréter la taille d'échantillon $n=98$ comme celle qui assure un écart de précision d'au plus 5 % (au seuil $\alpha=5\,\%$) pour toute prédiction dans l'intervalle 2–8 m^3.

Échantillonnage optimisé

Voyons à présent le cas où l'on cherche à optimiser le plan d'échantillonnage en fonction de la valeur de $D^{*2}H^*$. L'équation (2.3) montre que l'écart de précision E est minimum lorsque $\overline{D^2H}_e=D^{*2}H^*$. On a donc intérêt à choisir les arbres de l'échantillon de telle sorte que la moyenne empirique de leur taille D^2H soit égale à $D^{*2}H^*$. En pratique, la

moyenne empirique des D^2H de l'échantillon ne sera jamais exactement égale à $D^{*2}H^*$, donc on a intérêt également à maximiser le dénominateur $\sum_i (D_i^2 H_i - \overline{D^2 H_e})^2$, c'est-à-dire à maximiser la variance empirique des valeurs de D^2H de l'échantillon. En fin de compte, le plan d'échantillonnage qui maximise la précision de la prédiction du volume d'un arbre de D^2H égal à $D^{*2}H^*$ consiste à choisir $n/2$ arbres de D^2H égal à $D^{*2}H^* - \Delta$ et $n/2$ arbres de D^2H égal à $D^{*2}H^* + \Delta$, avec Δ aussi grand que possible (figure 2.2). Ce plan

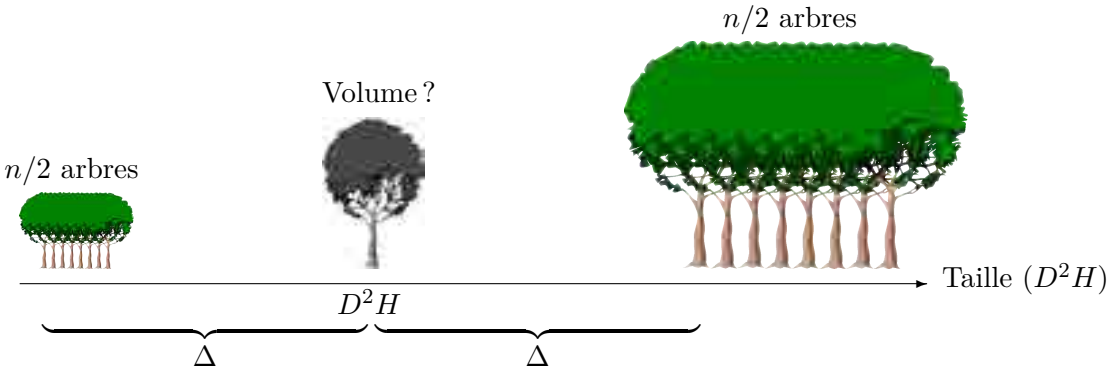


Figure 2.2 – *Plan d'échantillonnage optimisant la précision de la prédiction du volume pour un arbre particulier. L'écart de taille Δ doit être aussi grand que possible.*

d'échantillonnage permet de négliger le terme qui dépend de $D^{*2}H^*$ dans (2.3), de sorte que cette relation se simplifie en:

$$(1+E)^2 - 1 \approx \frac{1}{n}$$

Pour $E = 5\,\%$, on obtient alors $n = 10$ arbres. L'optimisation du plan d'échantillonnage a permis l'« économie » de la mesure de 88 arbres, par rapport au plan d'échantillonnage consistant à prendre les arbres au hasard. Cependant le plan d'échantillonnage optimisé est inféodé à l'estimation du volume d'un arbre de taille $D^{*2}H^*$. Il n'est pas optimisé pour l'estimation du volume d'arbre de toute autre taille. On voit donc la limite de ce raisonnement, un tarif de cubage n'étant généralement pas (si ce n'est jamais) construit pour prédire le volume d'une seule taille d'arbres.

Plus grave, le plan d'échantillonnage optimisé est également inféodé au modèle de super-population et peut conduire à des estimations erronées si ce modèle de super-population est mal vérifié dans la réalité. La figure 2.3 illustre cela. Le plan d'échantillonnage optimisé pour une taille $D^{*2}H^*$ donnée conduit à choisir pour l'échantillon des points extrêmes (en noir dans la figure 2.3). Cette situation est critique pour une régression linéaire car le fait d'avoir deux groupes de points éloignés va donner un R^2 élevé sans que l'on sache ce qui se passe réellement entre les deux. Si la relation linéaire supposée par le modèle de super-population est exacte (figure 2.3 gauche), alors il n'y a pas de problème: le volume prédit par le tarif (représentée par une étoile) sera effectivement proche du volume réel (point grisé). En revanche si on s'est trompé pour le modèle de super-population, alors le volume prédit sera erroné: c'est ce que l'on observe sur la figure 2.3 de droite (dont les points échantillons, en noir, sont exactement les mêmes que ceux de la figure 2.3 de gauche), où la relation taille-volume est en réalité parabolique et non pas linéaire. En pratique, la forme de la relation taille-volume (et donc du tarif) n'est pas connue à l'avance, et on a donc fortement intérêt à échantillonner les arbres dans tout l'intervalle de variation de la taille de manière à visualiser la nature de la relation taille-volume.

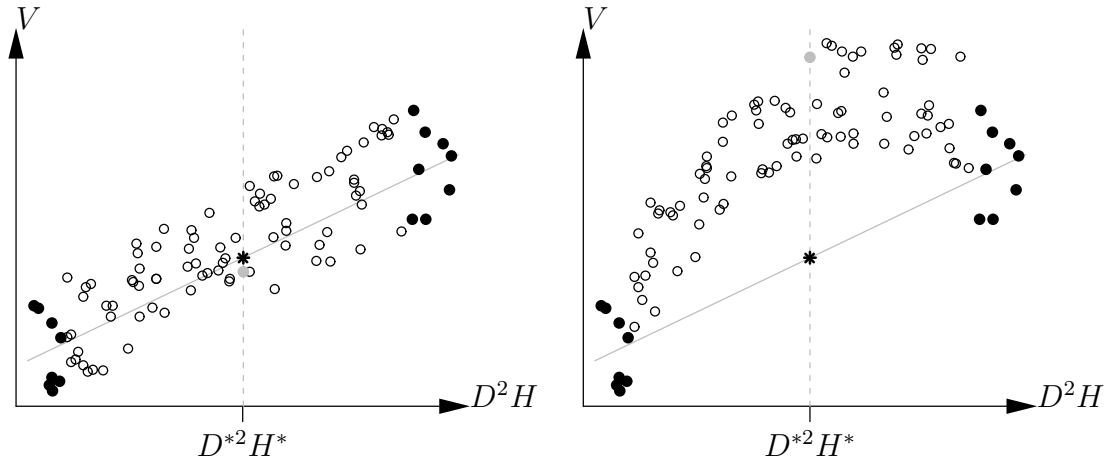


FIGURE 2.3 – *Prédiction du volume à l'aide d'une régression linéaire calée sur les points extrêmes (en noir) lorsque la relation taille-volume est effectivement linéaire (à gauche) et quand elle ne l'est pas (à droite). Les points noirs sont les mêmes dans les deux cas. L'étoile indique le volume prédit par la régression linéaire calée sur les points noirs, tandis que le point grisé indique le volume réel correspondant à* $D^{*2}H^*$.

2.1.2 Prédiction du volume du peuplement

Supposons à présent que l'objectif soit de prédire le volume de l'ensemble du peuplement. Pour ce faire, on suppose tout d'abord que l'on mesure le diamètre D et la hauteur H de *tous* les arbres du peuplement. Soit N le nombre total d'arbres dans le peuplement (y compris les n arbres de l'échantillon). Modulo une renumérotation des arbres, on dispose donc d'une mesure du volume V pour $i = 1, \ldots, n$ et d'une mesure de la taille D^2H pour $i = 1, \ldots, N$. L'estimateur du volume total du peuplement déduit du tarif de cubage est alors:

$$V_{\text{tot}} = \sum_{i=1}^{N} (a + bD_i^2 H_i)$$

ce que l'on peut encore écrire: $V_{\text{tot}} = N\bar{V}$, où: $\bar{V} = a + b\overline{D^2H}$ représente le volume moyen des arbres du peuplement, et $\overline{D^2H} = (\sum_{i=1}^{N} D_i^2 H_i)/N$ est le diamètre moyen des arbres du peuplement. Dans la mesure où le tarif de cubage est obtenu par régression linéaire de (V_1, …, V_n) par rapport à ($D_1^2H_1$, …, $D_n^2H_n$), les valeurs numériques des coefficients a et b vérifient (Saporta, 1990, p.363): $\bar{V}_e = a + b\overline{D^2H}_e$, où $\bar{V}_e = (\sum_{i=1}^{n} V_i)/n$ est le volume moyen des arbres de l'échantillon et $\overline{D^2H}_e = (\sum_{i=1}^{n} D_i^2 H_i)/n$ est la taille moyenne des arbres de l'échantillon. Par soustraction, on débouche sur le résultat suivant:

$$\bar{V} = \bar{V}_e + b(\overline{D^2H} - \overline{D^2H}_e) \tag{2.4}$$

Dans cette équation, on prendra bien garde que $\bar{V}$ et $\overline{D^2H}$ sont des moyennes sur l'ensemble du peuplement, tandis que $\bar{V}_e$ et $\overline{D^2H}_e$ sont des moyennes sur l'échantillon. De plus $\overline{D^2H}$, $\overline{D^2H}_e$ et $\bar{V}_e$ sont issus des mesures, alors que $\bar{V}$ est la quantité que l'on cherche à estimer.

Sous la forme (2.4), on reconnaît un type d'estimateurs bien connu en théorie de l'échantillonnage: les *estimateurs par régression*. La théorie des estimateurs par régression est exposée en détail dans Cochran (1977, chapitre 7) ou dans Thompson (1992, chapitre 8). Dans le cadre de la foresterie, un exposé sur les estimateurs par régression se trouve dans de Vries (1986) et dans Shiver et Borders (1996, chapitre 6) (le premier est plutôt théorique,

le deuxième plutôt appliqué). La théorie des estimateurs par régression s'applique dans le cas d'une relation linéaire entre la quantité à prédire ($\bar{V}$ dans l'exemple ci-dessus) et une variable auxiliaire ($\overline{D^2H}$ dans l'exemple ci-dessus). En revanche cette théorie est moins bien développée dans le cas de relations non-linéaires ou de régressions multiples; pourtant ces cas de figure sont fréquents pour les tarifs de cubage.

La demi-amplitude de l'intervalle de confiance au seuil α (typiquement 5 %) de $\bar{V}$ est égale à (Cochran, 1977, p.199; Thompson, 1992, p.83):

$$t_{n-2}\,\hat{\sigma}\,\sqrt{\frac{1}{n}-\frac{1}{N}+\frac{(\overline{D^2H}-\overline{D^2H}_e)^2}{\sum_{i=1}^{n}(D_i^2H_i-\overline{D^2H}_e)^2}} \tag{2.5}$$

On remarquera que le minimum de cette amplitude est zéro, qui est atteint lorsque tout le peuplement est inclus dans l'échantillon ($n = N$, ce qui entraîne $\overline{D^2H} = \overline{D^2H}_e$). Comme précédemment, le plan d'échantillonnage optimal est tel que $\overline{D^2H}_e$ soit le plus proche possible de $\overline{D^2H}$, avec une variance empirique de D^2H dans l'échantillon maximale.

Dans la dérivation de l'estimateur par régression, nous avons supposé que la taille D^2H est mesurée sur *tous* les arbres du peuplement pour aboutir à l'estimation du volume total V_{tot}. En pratique, un protocole de mesure plus réaliste est le suivant: on mesure la taille des arbres sur un échantillon de taille $n' < N$; on mesure à la fois la taille et le volume sur un sous-échantillon de taille $n < n'$ de cet échantillon. La régression du volume par rapport au diamètre (c'est-à-dire le tarif de cubage) est réalisée à partir du sous-échantillon; on en déduit une estimation du volume de l'échantillon puis, par extrapolation, de l'ensemble du peuplement. Cette stratégie d'échantillonnage est appelée *double échantillonnage*. Sa théorie est exposée dans Cochran (1977, section 12.6) ou, de manière plus pragmatique, dans Shiver et Borders (1996, chapitre 7). Son application à l'estimation de la biomasse des peuplements a été développée par Cunia (1987*b*,*c*,*d*).

Enfin, les propriétés de l'estimateur par régression (2.4) sont connues dans la théorie classique de l'échantillonnage, qui ne requiert pas l'hypothèse du modèle linéaire (2.1) mais considère que la seule source de variabilité est l'échantillonnage. Dans le cas d'un plan d'échantillonnage aléatoire simple et pour une taille d'échantillon n suffisamment grande, la variance de $\bar{V}$ est dans la théorie classique approximativement égale à (Cochran, 1977, p.195; Shiver et Borders, 1996, p.181):

$$\widehat{\text{Var}}(\bar{V}) = \frac{1-n/N}{n(n-2)}\left\{\sum_{i=1}^{n}(V_i-\bar{V}_e)^2-\frac{\left[\sum_{i=1}^{n}(V_i-\bar{V}_e)(D_i^2H_i-\overline{D^2H}_e)\right]^2}{\sum_{i=1}^{n}(D_i^2H_i-\overline{D^2H}_e)^2}\right\} \tag{2.6}$$

et la demi-amplitude de l'intervalle de confiance au seuil α (typiquement 5 %) de $\bar{V}$ est approximativement égale à (Thompson, 1992, p.80; Shiver et Borders, 1996, p.185):

$$t_{n-2}\,\sqrt{\widehat{\text{Var}}(\bar{V})}$$

Cette dernière expression est du reste considérée comme plus robuste que l'expression (2.5) lorsque la réalité s'écarte du modèle de super-population (2.1) (Thompson, 1992, p.84).

En conclusion, cet exemple simple montre à la fois l'apport et les limites de l'échantillonnage pour planifier des tarifs de cubage: apport, parce que la théorie de l'échantillonnage permet de planifier le nombre d'arbres minimal à mesurer pour atteindre une précision donnée dans les prédictions et permet d'optimiser le plan d'échantillonnage; limites, parce que le raisonnement suppose que la forme du tarif de cubage (et du modèle de super-population sous-jacent) est connue à l'avance et que le tarif sera utilisé pour une application donnée.

Aucun de ces deux pré-requis n'est vérifié en pratique. De plus, les calculs qui sont relativement simples dans le cas du modèle linéaire que nous venons d'exposer deviennent vite inextricables pour des modèles plus réalistes.

2.2 Échantillonnage pour la construction d'un tarif

Dans un premier temps, considérons le problème de la prédiction du volume ou de la biomasse d'un arbre particulier à l'aide d'un tarif. Combien d'arbres faut-il mesurer pour construire ce tarif (§ 2.2.1) ? Comment choisir ces arbres dans le peuplement ? Cette deuxième question sous-entend: comment ventiler les arbres de l'échantillon en fonction des variables d'entrée du tarif, à commencer par leur taille (§ 2.2.2) ? Comment, le cas échéant, stratifier l'échantillon (§ 2.2.3) ? Vaut-il mieux sélectionner des individus de l'échantillon de façon éparpillée dans la forêt, ou au contraire inventorier tous les arbres d'une parcelle donnée (§ 2.2.4) ?

2.2.1 Nombre d'arbres

À cause des limites de la théorie de l'échantillonnage, le nombre d'arbres cubés ou pesés (en d'autres termes, la taille d'échantillon) est généralement choisi de manière empirique, à partir de règles issues de l'expérience. Un principe général est que, à précision égale, la taille d'échantillon doit être d'autant plus élevée que le matériel est variable: des effectifs moins élevés seront nécessaires pour une plantation de clones que pour une forêt tropicale naturelle, pour une espèce donnée que pour un groupe d'espèces, ou pour une parcelle de 10 ha que pour une région naturelle. Dans certains cas, comme pour la biomasse racinaire, c'est le coût de la mesure qui guide le choix de la taille d'échantillon plutôt que la précision escomptée des prédictions: on choisira un nombre d'arbres qui génère une quantité de travail acceptable pour la mesure. À titre indicatif, pour la construction d'un tarif de cubage, le mémento du forestier (CTFT, 1989, p.256) recommande la mesure d'environ 100 arbres « dans le cas d'un ou plusieurs peuplements de plantation récente sur une surface restreinte (du type parcelles de recherche sylvicole) ». Pardé et Bouchon (1988, p.108), quant à eux, recommandent les effectifs donnés dans le tableau 2.1, en fonction de surface de la zone sur laquelle on veut utiliser le tarif. Des compilations de tarifs de cubage et de biomasse ont été faites par Zianis *et al.* (2005) pour l'Europe et par Henry *et al.* (2011) pour l'Afrique sub-saharienne. Les tailles d'échantillon reportées pour les tarifs listés dans ces revues de la littérature permettent de se faire une idée de l'effort d'échantillonnage consenti. Chave *et al.* (2004) ont montré qu'en utilisant 300 arbres pour construire un tarif de biomasse, l'estimation de la biomasse d'un peuplement tropical humide (Barro Colorado Island au Panama) qui en résultait avait un coefficient de variation d'à peine 3,1 %. Ce coefficient de variation passait au-dessus de 10 % dès lors que le nombre d'arbres utilisés pour construire le tarif de biomasse passait en-dessous de 50, avec une décroissance du coefficient de variation approximativement proportionnelle à $1/\sqrt{n}$ (Chave *et al.*, 2004, figure 3). Van Breugel *et al.* (2011) ont trouvé le même type de décroissance de la précision d'estimation avec la taille d'échantillon utilisé pour construire le tarif de biomasse, pour n compris entre 49 et 195 arbres.

Plus l'acquisition d'une observation est coûteuse en termes de temps et d'effort de mesure, plus le plan d'échantillonnage tend à être piloté par l'effort d'échantillonnage que l'on est prêt à consentir plutôt que par la précision d'estimation escomptée. La biomasse épigée d'un arbre étant plus difficile à mesurer que le volume de sa tige, les tarifs de biomasse tendent ainsi à être construits à partir de moins d'observations que les tarifs de cubage.

Table 2.1 – *Nombre d'arbres à mesurer pour l'établissement d'un tarif de cubage en fonction de la superficie sur laquelle on veut utiliser le tarif: recommandations de Pardé et Bouchon (1988).*

Zone	n
Peuplement unique et homogène	30
Parcelle de 15 ha	100
Forêt de 1000 ha	400
Région naturelle	800
Aire de l'espèce	2000 à 3000

Certains tarifs de biomasse sont construits à partir d'une dizaine de mesures d'arbres seulement (8 arbres pour Brown *et al.*, 1995 au Brésil, 12 arbres pour Ebuy Alipade *et al.*, 2011 en République Démocratique du Congo, 14 arbres pour Deans *et al.*, 1996, 15 arbres pour Russell, 1983 au Brésil). Les tarifs racinaires, qui nécessitent un effort de mesure encore plus élevé, reposent bien souvent sur des tailles d'échantillon encore plus faibles. Les tarifs construits sur des échantillons aussi petits sont le plus souvent peu fiable, et n'ont de toute façon qu'une validité très locale. Cependant, ces petits jeux de données peuvent ensuite être regroupés en jeux de données plus imposants, qui eux ont un intérêt pour l'ajustement de tarifs (sous réserve que l'on sache contrôler par des covariables explicatives – âge, densité du bois... – ou par des facteurs de stratification – espèce, type de formation végétale... – la variabilité induite par la réunion des données).

2.2.2 Ventilation des arbres

La ventilation des arbres de l'échantillon en fonction de leur taille (et plus généralement, en fonction des variables utilisées en entrée du tarif) peut en principe être optimisée. Dans le cas d'une régression linéaire, par exemple, la demi-amplitude de l'intervalle de confiance au seuil α de la pente de régression est (Saporta, 1990, p.367):

$$t_{n-2}\,\frac{\hat{\sigma}}{S_X\sqrt{n}}$$

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 est la taille d'échantillon et S_X est l'écart-type empirique de la variable d'entrée X au sein de l'échantillon:

$$S_X^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2 \quad \text{avec} \quad \bar{X} = \frac{1}{n}\sum_{i=1}^{n}X_i$$

La précision de l'estimation de la pente est donc d'autant meilleure que S_X est élevé ce qui, à taille d'échantillon fixée, nous ramène à une ventilation des arbres semblable à celle montrée dans la figure 2.2. On a déjà vu les limites de ce raisonnement: le plan d'échantillonnage consistant à prendre des arbres aux deux extrémités du gradient de taille, s'il est optimal lorsque l'hypothèse d'une relation linéaire est bien vérifiée, conduit à des estimations erronées lorsque cette hypothèse n'est pas vérifiée (figure 2.3). On a donc en pratique intérêt à échantillonner les arbres tout le long du gradient de taille de manière à s'assurer de la forme de la relation entre leur volume (ou leur masse) et leur taille.

La théorie des surfaces de réponse (Box et Draper, 1987; Goupy, 1999; Myers et Montgomery, 2002) permet d'optimiser la ventilation des arbres en fonction de leur diamètre (et plus généralement, en fonction des variables utilisées en entrée du tarif). Nous n'entrerons pas ici dans les détails de cette théorie et nous contenterons de quelques principes généraux. Le premier principe est d'étaler au maximum le gradient de taille des arbres de l'échantillon.

Si la variance du volume (ou de la masse) est constante quelle que soit la taille de l'arbre, la règle est de mesurer autant d'arbres dans chaque classe de taille (Pardé et Bouchon, 1988, p.108; CTFT, 1989, p.256). Prendre pour l'échantillon un nombre d'arbres par classe de taille proportionnel aux effectifs dans le peuplement par classe de taille (en d'autres termes, tirer les arbres au hasard) serait une erreur. Cependant la variance du volume est rarement constante; généralement elle augmente avec la taille (hétéroscédasticité des résidus). Dans ce cas, la règle est d'augmenter l'intensité d'échantillonnage des classes les plus variables, de manière à assurer la meilleure précision. En théorie, l'idéal est de mesurer dans une classe de taille donnée un effectif d'arbres proportionnel à l'écart-type du volume des arbres de cette classe (CTFT, 1989, p.256). En pratique, lorsque la variable d'entrée est le diamètre, une règle empirique consiste à prendre des effectifs d'arbres constants par classe de surface terrière, ce qui assure une meilleure représentation des arbres de fort diamètre (CTFT, 1989, p.256–257).

Le raisonnement s'étend à d'autres variables explicatives. Si la variable d'entrée du tarif est D^2H, on ventilera les arbres selon des classes de D^2H. Pour les tarifs de biomasse pluri-spécifiques, la densité du bois ρ est souvent utilisée comme une variable d'entrée du tarif (avec la spécificité qu'elle intervient au niveau de l'espèce et non pas au niveau de l'arbre). Pour un tarif pluri-spécifique utilisant le diamètre D et la densité du bois spécifique ρ comme variable d'entrée, une ventilation adéquate des arbres de l'échantillon consisterait à les répartir de manière uniforme par classe de diamètre et par classe de densité du bois.

2.2.3 Stratification

On a déjà vu, pour la ventilation des arbres dans les classes de taille, que tirer les arbres au hasard, en donnant une *probabilité d'inclusion* égale à tous les arbres, constitue un plan d'échantillonnage sous-optimal. La stratification vise de même à tenir compte d'informations exogènes pour définir des *strates* d'échantillonnage homogènes, de façon à améliorer la précision des estimations. Le principe est, comme précédemment, d'augmenter l'intensité d'échantillonnage des strates les plus variables (relativement aux autres strates). Pour reprendre l'exemple du paragraphe 2.1, la variance de l'estimateur par régression $\bar{V}$ devient, dans le cas d'un échantillonnage stratifié (Cochran, 1977, p.202):

$$\begin{aligned}\widehat{\mathrm{Var}}(\bar{V}) \quad = \quad & \sum_h \left(\frac{N_h}{N}\right)^2 \frac{1-n_h/N_h}{n_h(n_h-2)} \left\{ \sum_{i=1}^{n_h} (V_{hi}-\bar{V}_{eh})^2 \right. \\ & \left. - \frac{\left[\sum_{i=1}^{n_h}(V_{hi}-\bar{V}_{eh})(D_{hi}^2 H_{hi} - \overline{D^2H}_{eh})\right]^2}{\sum_{i=1}^{n_h}(D_{hi}^2 H_{hi} - \overline{D^2H}_{eh})^2} \right\} \end{aligned} \tag{2.7}$$

où h désigne la strate, N_h est le nombre d'individus du peuplement appartenant à la strate h, n_h est le nombre d'individus de l'échantillon appartenant à la strate h, V_{ih} est le volume du i^{e} individu de la strate h au sein de l'échantillon, $\bar{V}_{eh}$ est la moyenne empirique du volume dans la strate h de l'échantillon, etc. Cette formule vient en remplacement de (2.6). Illustrons le gain de précision apporté par la stratification à l'aide d'un petit exemple numérique. Supposons pour simplifier qu'il y a deux strates, que chacune correspond à 50 % du peuplement (de sorte que $N_1/N = N_2/N = 0{,}5$), et que l'échantillonnage au sein de chaque

strate est choisi de sorte que le second terme entre accolades de (2.7) soit négligeable. On suppose de plus que $n_1 \ll N_1$ et $n_2 \ll N_2$. La variance de l'estimateur par régression est alors approximativement proportionnelle à:

$$\widehat{\mathrm{Var}}(\bar{V}) \propto \frac{1}{n_1 - 2}\left\{\frac{1}{n_1}\sum_{i=1}^{n_1}(V_{1i} - \bar{V}_{e1})^2\right\} + \frac{1}{n_2 - 2}\left\{\frac{1}{n_2}\sum_{i=1}^{n_2}(V_{2i} - \bar{V}_{e2})^2\right\}$$

Les termes entre accolades représentent les variances du volume intra-strates. Supposons que l'écart-type du volume soit de 4 m^3 dans la première strate et de 2 m^3 dans la seconde. La taille totale de l'échantillon est fixée à $n_1 + n_2 = 60$ individus. Si on ne tient pas compte de la stratification, c'est-à-dire si on tire le nombre d'arbres dans chaque strate proportionnellement à la fréquence N_h/N de la strate dans le peuplement, alors on a dans le cas de figure présent autant d'arbres dans chaque strate de l'échantillon: $n_1 = n_2 = 30$ individus. La variance de l'estimateur par régression est alors approximativement:

$$\frac{4^2}{30 - 2} + \frac{2^2}{30 - 2} = 0{,}71 \text{ m}^6$$

En revanche si on fixe le nombre d'arbres dans chaque strate proportionnellement à l'écart-type du volume dans la strate, alors: $n_1 = 2n_2$, d'où $n_1 = 40$ individus et $n_2 = 20$ individus. La variance de l'estimateur par régression est alors approximativement:

$$\frac{4^2}{40 - 2} + \frac{2^2}{20 - 2} = 0{,}64 \text{ m}^6$$

On voit ainsi que du point de vue de la variance de l'estimateur, $30 + 30$ n'est pas égal à $40 + 20$. On pourra d'ailleurs vérifier que le minimum de la fonction qui à n_1 associe $16/(n_1 - 2) + 4/(58 - n_1)$ est obtenu pour $n_1 = 39{,}333$.

Du point de vue de la théorie de l'échantillonnage, la stratification a pour but d'augmenter la précision de l'estimation en ajustant le plan d'échantillonnage à la variabilité au sein de chaque strate. Mais du point de vue de la construction d'un tarif de cubage, la stratification a un second objectif tout aussi important que le premier: vérifier que la relation entre le volume (ou la biomasse) et la taille des arbres est la même au sein de chaque strate et, le cas échéant, décliner le tarif en autant de relations que nécessaire. Ce second point est implicite dans la formule (2.7) qui repose sur un ajustement d'une pente b différente (cf. équation 2.2) au sein de chaque strate.

En résumé, la stratification vise à explorer la variabilité au sein de la zone d'étude afin (*i*) de faire varier, le cas échéant, la forme du tarif en fonction des strates et (*ii*) d'adapter le plan d'échantillonnage à la variabilité au sein des strates. Souvent, pour la construction d'un tarif de cubage, le point (*i*) prédomine sur le point (*ii*), alors que c'est le contraire en théorie de l'échantillonnage. La figure 2.4 illustre ces deux objectifs.

Facteurs de stratification

Tout facteur susceptible d'expliquer la variabilité au sein de la zone d'étude peut être envisagé: âge du peuplement (surtout dans le cas de plantations), fertilité, station, traitement sylvicole, variété ou espèce, altitude, profondeur de la nappe phréatique, etc. (Pardé et Bouchon, 1988, p.106; CTFT, 1989, p.255). Les facteurs de stratification peuvent être emboîtés: stratification selon la région morpho-pédologique, puis selon la fertilité au sein de chaque région, puis selon l'âge au sein de chaque classe de fertilité, puis selon la densité au sein de chaque classe d'âge. La « finesse » des facteurs de stratification doit également être adapté au contexte. Les facteurs de stratification ne seront pas les mêmes selon que

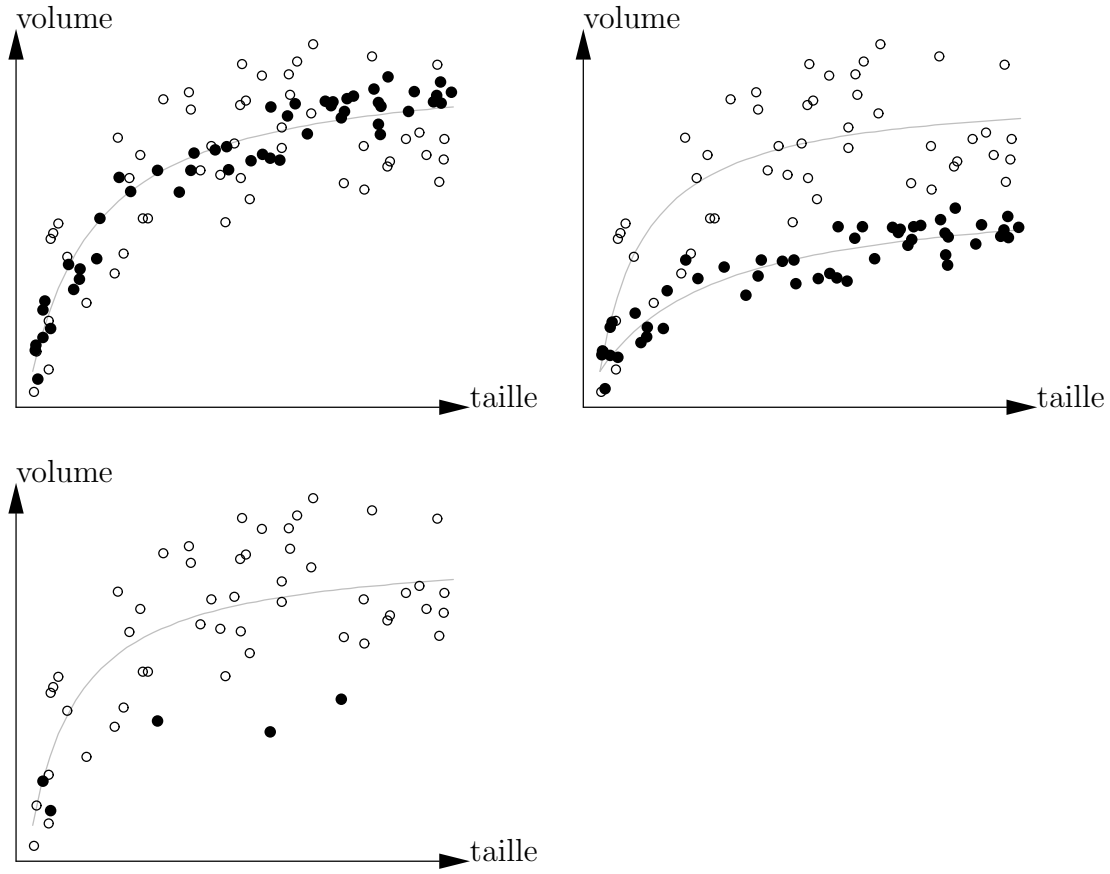


FIGURE 2.4 – *Prédiction du volume en fonction de la taille pour deux strates (correspondant aux points noirs et blancs): en haut à gauche, les deux strates correspondent à deux variances des résidus (variance plus élevée pour les points blancs que pour les points noirs) mais la relation est la même; en haut à droite, à la fois la variance et la relation varient entre les strates; en bas, la situation est la même qu'en haut à droite mais la seconde strate a été sous-échantillonnée, de sorte que l'on peut croire qu'on a affaire à la même relation dans les deux strates.*

l'on raisonne à l'échelle globale comme Brown (1997), à l'échelle d'un paysage comme van Breugel *et al.* (2011), ou à l'échelle d'une plantation de clones comme Saint-André *et al.* (2005). Brown (1997) propose des tarifs toutes essences pour des zones climatiques (forêts sèches, forêts humides). À l'autre extrême, sur une parcelle « eddy-correlation », avec pour objectif de comparer les estimations de NEP, la stratification pourra se faire en fonction de l'âge de la parcelle, de la saison et du « foot-print » de la tour à flux.

L'espèce comme facteur de stratification

Pour les formations naturelles renfermant plusieurs essences, l'essence peut également être considérée comme un facteur de stratification. Il est habituel pour les formations pluri-spécifiques de construire un tarif de cubage séparément pour chaque essence (ou du moins les plus abondantes), puis de tenter des regroupements soit au niveau du genre, soit en rassemblant toutes les essences (tarif « toutes essences »). En fusionnant les données, on augmente la taille de l'échantillon, ce qui a un intérêt si cela compense l'augmentation de la variabilité liée au mélange des différentes essences. Comparé à un modèle mono-spécifique, l'utilisation d'un modèle pluri-spécifique revient à introduire un biais de prédiction, que l'on peut voir comme de la variabilité inter-espèce. Van Breugel *et al.* (2011) ont ainsi quantifié le biais de prédiction résultant de l'agrégation de plusieurs essences. Ainsi, fusionner les données relatives à plusieurs espèces a un intérêt si le gain de variabilité intra-espèce apporté par cette fusion compense la variabilité inter-espèce introduite. Il faut toutefois s'assurer que (*i*) ce mélange a un sens et que (*ii*) les tailles d'échantillon relatives aux différentes essences sont comparables (figure 2.4). Lorsque d'emblée on cherche à construire un tarif « toutes essences » (ce qui est souvent le cas pour les tarifs de peuplement naturel), il faut prendre garde à ce que le choix des individus qui feront partie de l'échantillon soit indépendant de leur essence, de manière à ne pas biaiser le tarif en faveur d'une essence particulière.

Allocation entre strates

Une fois les strates identifiées, le plan d'échantillonnage y sera adapté selon des règles empiriques. Si on a une estimation *a priori* de la variabilité du volume (pour un tarif de cubage) ou de la biomasse (pour un tarif de biomasse) au sein de chaque strate, une règle empirique est de prendre une intensité d'échantillonnage proportionnelle à l'écart-type au sein de chaque strate; si on n'a pas d'estimation *a priori* de la variabilité, on tâchera de prendre une intensité d'échantillonnage constante au sein de chaque strate (ce qui ne correspond pas à un échantillonnage au hasard dès que les strates n'ont pas la même fréquence dans le peuplement).

Tarif paramétré

L'information relative aux strates sera ensuite incorporée dans le tarif de cubage en établissant un tarif différent pour chaque strate. On pourra tester si les tarifs calés pour deux strates sont significativement différents et, le cas échéant, fusionner les deux jeux de données pour construire un seul tarif. On pourra aussi élaborer un tarif *paramétré* à partir de plusieurs tarifs relatifs à différentes strates suivant le principe du modèle mixte: les paramètres du tarif deviennent eux-mêmes des fonctions des variables définissant les strates. Ces différents points seront développés dans les paragraphes suivants consacrés à la construction proprement dite des tarifs. À titre d'exemple, Ketterings *et al.* (2001) ont développé des tarifs de biomasse individuels à une entrée de la forme puissance:

$$B = aD^b$$

où D est le diamètre des arbres et B leur biomasse, pour différents arbres de différentes essences en différents sites de la province de Jambi à Sumatra, en Indonésie. Le facteur site a ensuite été pris en compte dans le paramètre b qui a été écrit comme $b = 2 + c$, où c est le paramètre de l'équation allométrique liant la hauteur au diamètre en chaque site: $H = kD^c$. Le facteur essence a quant à lui été pris en compte dans le paramètre a qui a été écrit comme $a = r\rho$, où ρ est la densité du bois de l'essence et r un paramètre constant. Le tarif finalement obtenu, et valable pour toutes les essences en tous les sites, est un tarif paramétré:

$$B = r\rho D^{2+c}$$

2.2.4 Sélection des arbres

Une fois la composition de l'échantillon définie, il faut identifier les arbres à mesurer sur le terrain. Dans la mesure où il s'agit de mesures lourdes et, en ce qui concerne la biomasse, destructrices, le choix des arbres doit être fait soigneusement. Une stratégie qui a été adoptée par certains pour la construction de tarifs de biomasse consiste à abattre tous les arbres dans un périmètre donné (par exemple dans un demi-hectare). Cette stratégie a l'avantage de faire d'une pierre deux coups, puisqu'elle fournit en même temps une estimation de la biomasse du peuplement, et des observations individuelles pour la construction du tarif. De façon pratique, l'espace qui est dégagé par l'abattage des premiers arbres facilite ensuite l'abattage des suivants. Mais cette stratégie présente un inconvénient majeur: la distribution des tailles des arbres dans le peuplement n'ayant que très peu de chance de coïncider avec la ventilation souhaitée des arbres de l'échantillon par classe de taille, elle aboutira à une distribution des tailles des arbres de l'échantillon qui n'est pas optimale. Il en sera de même pour tout facteur structurant l'échantillon (classes de densité du bois, strates, etc.). De plus, la perturbation du peuplement à cette échelle a parfois des conséquences inattendues. Djomo *et al.* (2010) font ainsi état d'une parcelle ayant été envahie par des fourmis après l'abattage des arbres, à tel point que les mesures de biomasse des arbres ont dû être abandonnées. Cette stratégie de choix des arbres sera donc à éviter dans les zones infestées par *Wasmannia*, genre myrmicole dont les attaques sont redoutables.

Plutôt que de choisir tous les arbres au sein d'un périmètre donné, on préférera donc sélectionner les tiges à mesurer pied par pied, en fonction des besoins identifiés pour constituer l'échantillon. Cette stratégie est peut-être plus longue à mettre en œuvre puisqu'elle nécessite un repérage individuel des arbres. Compte-tenu des contraintes engendrées par la mesure de la biomasse des arbres (cf. chapitre 3), on favorisera, parmi tous les arbres qui satisfont les critères du plan d'échantillonnage, ceux qui sont le plus facilement accessibles.

2.3 Échantillonnage pour l'estimation d'un peuplement

Considérons à présent le problème de la prédiction du volume ou de la biomasse d'un peuplement. De manière statistiquement rigoureuse, il faudrait considérer toute la chaîne de propagation des erreurs telle que définie dans la figure 2.1 (Parresol, 1999). Cela débouche sur des questions de double échantillonnage et d'estimateur par régression, telles que définies dans le paragraphe 2.1.2. Cunia (1987*b*,*c*,*d*), Chave *et al.* (2004) et van Breugel *et al.* (2011) sont de rares exemples où l'ensemble de la chaîne de propagation des erreurs a effectivement été prise en compte, et où l'erreur d'estimation de la biomasse d'un peuplement a été reliée à la taille de l'échantillon d'arbres utilisé pour construire le tarif de biomasse nécessaire à cette estimation. Dans la pratique, on simplifie généralement le problème en considérant le tarif comme exact et ne souffrant d'aucune erreur de prédiction. Cette approximation,

qui revient à déconnecter l'échantillonnage du peuplement pour prédire son volume ou sa biomasse de l'échantillonnage des arbres pour construire le tarif, ramène le premier a un problème d'inventaire forestier classique.

Nous ne nous attarderons pas sur cette question de l'inventaire forestier, d'une part parce qu'elle reste marginale par rapport à l'objectif central de ce manuel, et d'autre part parce que des ouvrages entiers y ont déjà été consacrés (Loetsch et Haller, 1973; Lanly, 1981; de Vries, 1986; Schreuder *et al.*, 1993; Shiver et Borders, 1996; West, 2009). Nous présenterons cependant quelques développements relatifs à l'estimation de la biomasse d'un peuplement.

2.3.1 Unité d'échantillonnage

Autant, pour la construction d'un tarif de biomasse, il convient de sélectionner les arbres à inclure dans l'échantillon de manière individuelle, autant cette stratégie d'échantillonnage est irréaliste quand il s'agit d'estimer la biomasse d'un peuplement. Dans ce cas, on va plutôt opter pour la stratégie consistant à mesurer tous les arbres au sein d'un périmètre donné, quitte à répéter ce périmètre pour étendre la taille de l'échantillon. Ce périmètre, ou placette, devient alors l'unité d'échantillonnage. Soit n le nombre de placette inventoriée, N_i le nombre d'arbres trouvés dans la i^{e} placette ($i = 1, \ldots, n$), et B_{ij} la biomasse du j^{e} arbre de la i^{e} parcelle ($j = 1, \ldots, N_i$), calculée à l'aide du tarif de biomasse et des caractéristiques mesurées de l'arbre. Le nombre N_i est aléatoire mais, pour un arbre donné, la prédiction de B_{ij} est considérée comme déterministe. La biomasse de la i^{e} placette est alors: $B_i = \sum_{j=1}^{N_i} B_{ij}$.

Soit A la surface d'une placette d'échantillonnage et $\mathcal{A}$ la surface du peuplement. Dans un modèle de super-population, la biomasse du peuplement est alors estimée par: $(\mathcal{A}/A)\,\bar{B}$, où $\bar{B} = (\sum_{i=1}^{n} B_i)/n$ est la biomasse moyenne d'une placette. On considère généralement que A et $\mathcal{A}$ sont connues de manière exacte. L'erreur d'estimation de la biomasse du peuplement découle alors de celle de la biomasse moyenne $\bar{B}$.

2.3.2 Relation entre le coefficient de variation et la taille des placettes

D'après le théorème central limite, l'intervalle de confiance au seuil α de l'espérance de la biomasse d'une parcelle est approximativement (cette expression étant exacte lorsque la biomasse suit une loi normale, ou dans la limite où le nombre de parcelles tend vers l'infini) (Saporta, 1990, p.304):

$$\bar{B} \pm t_{n-1} \frac{S_B}{\sqrt{n-1}}$$

où t_{n-1} est le quantile $1 - \alpha/2$ d'une loi de Student à $n-1$ degrés de liberté, et S_B est l'écart-type empirique de la biomasse d'une parcelle:

$$S_B^2 = \frac{1}{n-1} \sum_{i=1}^{n} (B_i - \bar{B})^2$$

Par définition, la précision d'estimation E au seuil α est le ratio de la demi-amplitude de l'intervalle de confiance au seuil α sur la biomasse moyenne:

$$E = t_{n-1} \frac{S_B}{\bar{B}\sqrt{n-1}} = t_{n-1} \frac{\mathrm{CV}_B}{\sqrt{n-1}} \tag{2.8}$$

où $\mathrm{CV}_B = S_B/\bar{B}$ est le coefficient de variation de la biomasse. En arrondissant t_{n-1} à 2, la taille d'échantillon n requise pour atteindre une précision d'estimation donnée de E est

donc:

$$n \simeq \left(\frac{2\mathrm{CV}_B}{E}\right)^2 + 1$$

Le coefficient de variation de la biomasse d'une parcelle de surface A est donc l'élément essentiel pour construire le plan d'échantillonnage. De plus, la surface A des placettes n'étant *a priori* pas connue, il faut en fait connaître la relation entre le coefficient de variation de la biomasse et la surface A des placettes.

La dérivation exacte de la relation entre A et CV_B nécessite de spécifier un modèle pour la répartition spatiale des arbres. La théorie des processus ponctuels répond à ce besoin (Cressie, 1993; Stoyan et Stoyan, 1994). Le calcul exact de la relation entre A et CV_B dans le cadre d'un processus ponctuel est faisable mais compliqué (Picard *et al.*, 2004; Picard et Bar-Hen, 2007). Le calcul exact permet de se rendre compte de deux choses:

1. la *forme* des placettes, bien qu'ayant un effet sur le coefficient de variation (ce qui est connu empiriquement, cf. Johnson et Hixon, 1952; Bormann, 1953), a un effet négligeable par rapport à la taille des parcelles;
2. la relation entre A et CV_B peut être approchée par une relation puissance (Fairfield Smith, 1938; Picard et Favier, 2011):

$$\mathrm{CV}_B = kA^{-c}$$

En pratique, c'est cette relation puissance qui est le plus souvent spécifiée. Intuitivement, la valeur $c = 0{,}5$ correspond à une répartition spatiale aléatoire de la biomasse au sein du peuplement; une valeur $0 < c < 0{,}5$ correspond à une répartition spatiale agrégée de la biomasse; et une valeur $c > 0{,}5$ correspond à une répartition spatiale régulière de la biomasse (CTFT, 1989, p.284). En utilisant les données de biomasse d'une parcelle de grande taille à Paracou en Guyane française, Wagner *et al.* (2010) ont ainsi trouvé:

$$\mathrm{CV}_B = 557 \times A^{-0{,}430} \qquad (A \text{ en m}^2,\ \mathrm{CV}_B \text{ en \%})$$

Selon l'interprétation précédente, cela correspond à une répartition spatiale légèrement agrégée de la biomasse. En Amazonie brésilienne, Keller *et al.* (2001) ont trouvé la relation suivante (ajustée aux données de leur figure 4 avec $R^2 = 0{,}993$ sur les données log-transformées):

$$\mathrm{CV}_B = 706 \times A^{-0{,}350} \qquad (A \text{ en m}^2,\ \mathrm{CV}_B \text{ en \%})$$

La valeur plus faible (en valeur absolue) de l'exposant traduit une répartition spatiale de la biomasse plus fortement agrégée qu'en Guyane. Une étude semblable a été menée par Chave *et al.* (2003) en utilisant les données de la parcelle de 50 ha de Barro Colorado Island au Panama. Chave *et al.* (2003) ont reporté dans leur tableau 5 les valeurs de l'amplitude de l'intervalle de confiance à 95 % non pas de l'espérance de la biomasse d'une parcelle, mais de l'espérance de la biomasse par unité de surface. L'amplitude de l'intervalle de confiance à 95 % de l'espérance de la biomasse d'une parcelle correspond donc à l'amplitude reportée par Chave *et al.* (2003) fois la surface de la placette, soit:

$$2t_{n-1}\frac{S_B}{\sqrt{n-1}} = \Delta \times A$$

où Δ est l'amplitude de l'intervalle de confiance à 95 % reportée par Chave *et al.* (2003) dans leur tableau 5. On en déduit:

$$\mathrm{CV}_B = \frac{S_B}{\bar{B}} = \frac{\Delta A\sqrt{n-1}}{2t_{n-1}\bar{B}} = \frac{\Delta\sqrt{n-1}}{2t_{n-1}\mu}$$

Table 2.2 – *Coefficient de variation de la biomasse d'une parcelle en fonction de sa taille: données complétées du tableau 5 de Chave et al. (2003) pour Barro Colorado Island au Panama.*

A (m^2)	n	Δ ($Mg\,ha^{-1}$)	CV_B (%)
100	5000	17,4	114,5
200	2500	18,7	87,0
400	1250	20,0	65,7
1000	500	21,4	44,4
2500	200	20,1	26,2
5000	100	22,4	20,5
10000	50	23,5	14,9

où μ est la biomasse moyenne par unité de surface, égale à 274 $Mg\,ha^{-1}$ dans l'étude de Chave *et al.* (2003). Le tableau 2.2 complète le tableau 5 de Chave *et al.* (2003) avec la valeur de CV_B ainsi calculée. Les valeurs de CV_B données dans le tableau 2.2 s'ajustent très bien ($R^2 = 0{,}998$ sur les données log-transformées) à la relation puissance suivante en fonction de la taille des placettes:

$$CV_B = 942 \times A^{-0{,}450} \qquad (A \text{ en } m^2,\ CV_B \text{ en } \%)$$

La variabilité de la biomasse (traduite par la valeur du coefficient multiplicateur $k = 942$) est plus forte qu'à Paracou, mais la structuration spatiale de la biomasse (traduite par l'exposant $c = 0{,}45$) est assez semblable à celle observée à Paracou par Wagner *et al.* (2010). De plus, le fait que c s'approche de la valeur 0,5 traduit une faible agrégation spatiale de la biomasse. Chave *et al.* (2003) soulignent d'ailleurs qu'il n'y a pas d'autocorrélation spatiale significative de la biomasse (ce qui correspondrait à $c = 0{,}5$, ou à une valeur constante de Δ).

2.3.3 Choix de la taille des placettes

Le choix de la taille des placettes d'échantillonnage peut être fait de manière à optimiser la précision d'estimation à effort d'échantillonnage donné (Bormann, 1953; Schreuder *et al.*, 1987; Hebert *et al.*, 1988), ou de manière à minimiser l'effort d'échantillonnage à précision d'estimation donnée (Zeide, 1980; Gambill *et al.*, 1985; Cunia, 1987*c*,*d*). Ces deux points de vue sont duaux l'un de l'autre et conduisent au même optimum. L'effort d'échantillonnage peut être quantifié de manière simple par le taux d'échantillonnage $n \times A/\mathcal{A}$ ou, de manière plus réaliste, par un coût dont l'expression est plus complexe. Examinons ces deux options.

Taux d'échantillonnage fixe

À taux d'échantillonnage constant, la taille A et le nombre n de placettes d'échantillonnage sont liés par une relation inversement proportionnelle: $n \propto 1/A$. Le choix de la taille des placettes se ramène alors à la question suivante « vaut-il mieux installer peu de grandes placettes ou beaucoup de petites placettes ? », ce que l'on appelle aussi le compromis SLOSS (de l'anglais « single large or several small »; Lahti et Ranta, 1985). Si on reporte la relation $n \propto 1/A$ dans (2.8) (et en considérant que $t_{n-1}/\sqrt{n-1}$ est peu différent de $2/\sqrt{n}$):

$$E \propto 2\,CV_B\,\sqrt{A}$$

Si la répartition spatiale de la biomasse est aléatoire, $\mathrm{CV}_B \propto A^{-0,5}$ et donc la précision d'estimation E est indépendante de la taille A des placettes. Si la répartition spatiale de la biomasse est agrégée, $\mathrm{CV}_B \propto A^{-c}$ avec $c < 0,5$ et donc $E \propto A^{0,5-c}$ avec $0,5 - c > 0$: la précision d'estimation est d'autant meilleure (E petit) que la taille A des placettes est petite. Dans ce cas, à taux d'échantillonnage fixe, il vaut mieux installer beaucoup de petites parcelles que peu de grandes parcelles. C'est ce que l'on observe dans le tableau 2.2, où l'on voit que la valeur de Δ diminue lorsque A diminue (cette diminution reste légère car c est proche de 0,5). Si la répartition spatiale de la biomasse est régulière, $\mathrm{CV}_B \propto A^{-c}$ avec $c > 0,5$ et donc $E \propto A^{0,5-c}$ avec $0,5 - c < 0$: la précision d'estimation est d'autant meilleure (E petit) que la taille A des placettes est grande. Dans ce cas, à taux d'échantillonnage fixe, il vaut mieux installer peu de grandes parcelles que beaucoup de petites parcelles.

Les grandeurs mesurées en biologie ont le plus souvent une répartition spatiale agrégée ($c < 0,5$), parfois aléatoire ($c = 0,5$), rarement régulière ($c > 0,5$) (Fairfield Smith, 1938). En d'autres termes, le compromis SLOSS sera le plus souvent résolu au profit d'une multitude de petites parcelles. Si on pousse le raisonnement jusqu'au bout, on voit que la précision d'estimation sera optimale (E minimal) pour $A = 0$, c'est-à-dire en mettant en place une infinité de parcelles de taille nulle ! On voit là les limites de ce raisonnement. Quand on quantifie l'effort d'échantillonnage par le taux d'échantillonnage $nA/\mathcal{A}$, on suppose implicitement que le coût d'échantillonnage, c'est-à-dire le temps ou l'argent nécessaire à l'échantillonnage, est proportionnel à nA. Cela revient à ne considérer qu'un coût surfacique, c'est-à-dire un coût d'échantillonnage qui est proportionnel à la surface inventoriée.

Coût d'échantillonnage

En réalité, le coût surfacique n'est qu'une composante du coût d'échantillonnage. L'inventaire proprement dit des placettes d'échantillonnage, dont la durée est effectivement proportionnelle à la surface inventoriée, n'est pas la seule tâche qui prend du temps. Délimiter les placettes d'échantillonnage prend aussi du temps. Or la durée de délimitation des placettes est proportionnelle à leur périmètre cumulé: il s'agit donc d'un coût linéaire. Circuler d'une placette à l'autre prend aussi du temps. Il est plus réaliste de mesurer l'effort d'échantillonnage par un coût qui prend en compte toutes ces tâches, plutôt que simplement par le taux d'échantillonnage. Si on mesure ce coût en terme de temps et si les placettes d'échantillonnage ont une forme carré, le coût d'échantillonnage sera par exemple (Zeide, 1980; Gambill *et al.*, 1985):

$$C = \alpha\, nA + \beta \times 4n\sqrt{A} + \gamma\, d(n, A)$$

où α est le temps d'inventaire par unité de surface, β est le temps de délimitation par unité de longueur ($4\sqrt{A}$ représente le périmètre d'une parcelle carré de surface A), γ est la vitesse de déplacement, et $d(n, A)$ est la longueur du chemin reliant les n placettes d'échantillonnage. On peut compléter l'expression du coût d'échantillonnage pour tenir compte d'autres tâches. Le raisonnement suivi dans le paragraphe précédent revenait à poser $\beta = \gamma = 0$. Avec $\beta > 0$ et $\gamma > 0$, la solution du compromis SLOSS ne sera plus $A = 0$ dans le cas d'une répartition spatiale agrégée de la biomasse ($c < 0,5$).

Autres contraintes

Limiter la question de l'échantillonnage de la biomasse d'un peuplement à une question de précision d'estimation reste cependant trop restrictif. Souvent, la question ne se limite pas à l'estimation de la biomasse du peuplement, mais des objectifs multiples sont poursuivis simultanément. Par exemple, il s'agira d'estimer non seulement la biomasse du peuplement,

mais aussi ses variations dans le temps. Compte des processus de mortalité à prendre en compte dans ce cas, les surfaces à inventorier peuvent alors être bien supérieures (Chave *et al.*, 2003, 2004; Rutishauser *et al.*, 2010; Wagner *et al.*, 2010). Ou alors, il s'agira d'estimer la biomasse sur des placettes de manière à caler une relation avec des indices issus d'images satellite, de manière à extrapoler l'estimation de la biomasse à l'échelle d'un paysage. Dans ce cas, la surface des placettes d'échantillonnage est aussi contrainte par la résolution des images satellite et par la surface minimale nécessaire pour calculer les indices satellitaires.

De plus, le type de plan d'échantillonnage implicitement considéré ici, à savoir un plan aléatoire simple à l'aide de placettes de taille fixe, est rarement le plus efficient (*i.e.* avec la meilleure précision d'estimation à coût d'échantillon donné). À l'échelle d'un paysage comportant différents types forestiers, d'autres stratégies d'échantillonnage peuvent s'avérer plus efficientes (Whraton et Cunia, 1987; van Breugel *et al.*, 2011). Comparé à un plan aléatoire simple de même taille d'échantillon, un plan d'échantillonnage stratifié induirait un coût d'échantillonnage supplémentaire (car la stratification a un coût) mais fournirait une meilleure précision d'estimation; un plan par grappes induirait un coût moindre (car moins de déplacements à effectuer) mais fournirait une moins bonne précision d'estimation. Les techniques d'inventaire spécifiques à la foresterie telles que l'inventaire par distances (Magnussen *et al.*, 2008*a*,*b*; Picard et Bar-Hen, 2007; Picard *et al.*, 2005) ou l'inventaire au relascope de Bitterlich (Schreuder *et al.*, 1993; West, 2009), qui ont en commun de reposer sur des placettes de taille variable, peuvent aussi être des alternatives plus efficientes aux approches par placettes de taille fixe.

3
Terrain

L'étape de terrain est la plus cruciale car elle peut générer des erreurs de mesure qui ne peuvent pas être corrigées. Cette phase doit être régie par trois principes clés: (*i*) il est préférable de tout peser sur le terrain, plutôt que de calculer un volume et de le multiplier ensuite par une mesure de densité (cf. chapitre 1, et les variations de forme des tiges et de densité du bois dans les arbres); (*ii*) dès lors qu'il y a prélèvement d'une aliquote, il faut peser le total puis l'aliquote de façon à garantir le suivi de la perte d'humidité; enfin (*iii*) une campagne de biomasse étant très lourde à réaliser et très chère, d'autres mesures peuvent être réalisées en même temps pour éviter de revenir ensuite sur le terrain (profil des tiges, échantillonnage pour la minéralomasse par exemple).

La sélection des arbres à mesurer sur le terrain (voir chapitre 2), qu'elle soit par individu ou exhaustive sur une surface donnée, nécessite un marquage des arbres à la peinture, une mesure de circonférence si possible à 1,30 m (en effectuant un cercle de peinture à cette hauteur) et une mesure de hauteur. Celles-ci permettent d'une part de vérifier que l'arbre sélectionné correspond bien au plan d'échantillonnage choisi (cas d'une sélection par individu) et d'autre part de donner des mesures de contrôle une fois l'arbre abattu. Il est pratique également de prendre une photo de l'individu sélectionné et d'en faire un schéma synthétique sur la fiche terrain. Cela facilite l'interprétation des données et la vérification des résultats obtenus. En général, les arbres trop particuliers (cime cassée, tige noueuse ou sinueuse) ne sont pas sélectionnés sauf si ces tiges représentent une proportion significative du peuplement ou si l'objectif est de quantifier un accident (exemple: bris de cime suite à un gel). De même, les arbres situés dans un environnement non représentatif sont à exclure (lisière de forêt, clairière, forêt dégradée, etc.). En effet, leur architecture est souvent différente des autres arbres du peuplement. Enfin, il n'est pas rare que les contraintes de terrain (pente, accès, peuplement non conforme à la strate, etc.) remettent en cause l'échantillonnage initial.

La base générale des mesures de biomasse, et *a fortiori* de minéralomasse, réside dans une règle de trois entre la biomasse fraîche mesurée sur le terrain, la biomasse fraîche de l'aliquote et la biomasse sèche de l'aliquote. Comme les différents organes d'un arbre n'ont pas le même taux d'humidité ou la même densité, il est préférable de procéder par compartimentation des arbres pour prendre en compte les variations de densité et d'humidité dans l'arbre (et de concentration en éléments minéraux pour la minéralomasse). L'estimation de la biomasse

sera d'autant plus précise que la stratification sera fine mais cela demande plus de travail. Un compromis est à trouver entre la précision de la mesure et la rapidité du travail sur le terrain. Traditionnellement, les compartiments sont ainsi définis: le tronc en distinguant le bois de l'écorce et qu'il convient de billonner pour bien prendre en compte les variations de densité et d'humidité en fonction du diamètre des sections; les branches généralement échantillonnées par classes de diamètre en distinguant ou non le bois de l'écorce; les plus petits rameaux incluent généralement les bourgeons; les feuilles; les fruits; les fleurs; et enfin les racines par classes de diamètre. Un exemple d'une telle compartimentation est donné dans la figure 3.1 pour le hêtre.

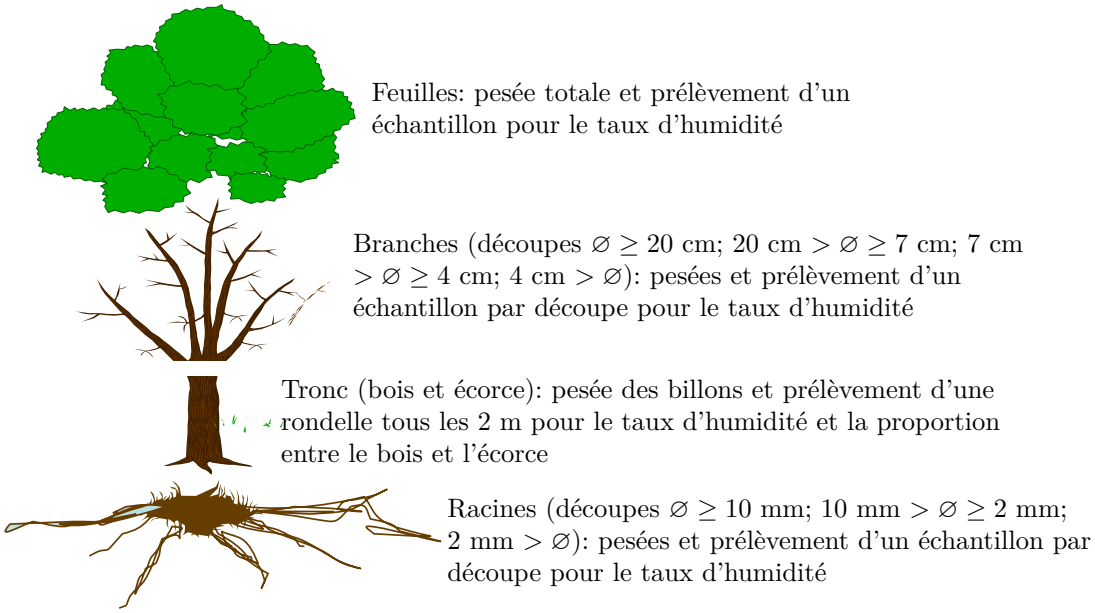


FIGURE 3.1 – *Exemple de compartimentation des arbres pour une campagne de biomasse et de minéralomasse sur le hêtre en France.*

Il est possible de profiter de chantiers d'abattage existants pour prélever les arbres nécessaires à l'établissement du tarif. En effet, l'accès et les prélèvements d'arbres sont souvent réglementés en forêt et l'exploitation sylvicole fournit un des seuls moyens d'avoir accès aux arbres recherchés. Cependant, cette méthode risque d'induire un biais dans la sélection des arbres puisque les essences abattues seront principalement commerciales. Les autres essences ne sont abattues que si elles gênent le prélèvement d'un arbre sélectionné par la compagnie forestière ou si elles sont présentes sur les chemins de débardage ou les zones de stockage. De plus, les arbres abattus pour motif commercial ne peuvent pas nécessairement être tronçonnés en billons de taille raisonnable pour pouvoir être pesés sur le terrain. Cela dépend de la capacité des pesons qui sont disponibles et de la longueur des billons. Ces contraintes impliquent alors une sélection attentive des individus et de combiner deux méthodes: (1) pesées intégrales des sections non-commerciales des arbres, notamment les branches; (2) mesures de volume et de densité du bois pour le tronc.

Il n'existe donc pas de méthode de terrain standardisée car chacun devra s'adapter en fonction des situations. Par contre, il est possible de donner dans le cadre de ce manuel trois cas typiques qui donnent les bases pour mener ensuite n'importe quelle campagne de terrain. Le premier concerne des forêts régulières (issues de régénération ou plantées), le second une forêt sèche et le troisième une forêt tropicale humide. Dans le premier cas, tous

les compartiments sont pesés directement sur le terrain. Dans le second cas, les arbres ne peuvent pas être abattus et les mesures sont semi-destructives. Le troisième cas concerne des arbres de dimensions trop grandes pour une pesée intégrale sur le terrain. L'obtention d'une mesure résulte de trois phases qui sont décrites ci-après: le terrain, le laboratoire et le calcul sur ordinateur. Le terrain et les calculs étant spécifiques à chaque méthode, ils sont présentés pour chacun des cas. Les procédures de laboratoire sont généralement les mêmes.

3.1 Pesées directes sur le terrain de tous les compartiments

Le premier cas que nous envisageons est le plus fréquent. Il s'agit de peser directement sur le terrain tous les compartiments. Le mode opératoire proposé est le résultat de plusieurs campagnes de terrain effectuées aussi bien dans des peuplements en climat tempéré que tropical. Nous l'illustrons par des exemples pris dans différents peuplements réguliers: plantations d'eucalyptus au Congo (Saint-André *et al.*, 2005), d'hévéa en Thaïlande, futaies de hêtre et chêne en France (Genet *et al.*, 2011). Un exemple de réalisation de cette méthodologie, avec un complément de mesures sur le défilement des grosses branches et le prélèvement d'échantillons pour la minéralomasse est donné par Rivoire *et al.* (2009).

3.1.1 Sur le terrain

Le chantier est une étape complexe dont l'organisation doit être fluide pour que toutes les équipes puissent travailler sans temps mort (voir le détail de ces équipes en 3.6). Il est préparé en amont par le responsable du chantier qui a présélectionné les arbres et les a localisés sur le terrain. Il s'ensuit un travail au laboratoire pour (*i*) préparer le matériel nécessaire (voir le détail en 3.5), (*ii*) préparer les feuilles de saisie (pesée des différents compartiments, mesures connexes), (*iii*) préparer les sacs qui vont accueillir les différentes aliquotes prélevées sur les arbres (voir figure 3.1), (*iv*) expliquer aux différents intervenants comment le chantier est organisé pour que tous trouvent leur poste sur le terrain. La figure 3.2 propose une organisation efficace dans différentes campagnes de biomasse, avec sept postes qui fonctionnent en même temps.

Compte tenu du fait que le temps d'ébranchage est le plus long, il est utile de commencer le chantier par un arbre de grande taille (photo 3.3). Le responsable de chantier accompagne les bûcherons et dépose au pied de l'arbre les sacs destinés à recueillir les échantillons (poste 1). La dimension des sacs doit être adaptée en fonction de la taille des échantillons à prélever. Ils doivent systématiquement porter la référence du compartiment, de l'arbre et de la parcelle. Après abattage, la première équipe à intervenir est celle qui mesure les profils de tiges (poste 2). Lorsque celle-ci a terminé, elle va sur le second arbre, abattu entre-temps par les bûcherons, tandis que les équipes d'ébrancheurs commencent le travail sur le premier arbre (poste 3). Il faut compter environ une demi-journée pour un arbre de 12 tonnes (environ 90–100 cm de diamètre). Lorsque les ébrancheurs terminent le premier arbre, les bûcherons ont eu le temps d'abattre suffisamment d'arbres en avance pour que l'équipe des profils ait suffisamment d'arbre à mesurer toute la journée. Les bûcherons peuvent ensuite revenir sur le premier arbre pour le billonner et prélever les rondelles (poste 4). Une fois le billonnage et le prélèvement des rondelles réalisés sur ce premier arbre, les bûcherons partent sur le second arbre qui a été ébranché entre-temps. Sur le premier arbre, les pesées des feuilles, des billons et des fagots de branches sont effectués (poste 5) tandis que le responsable de chantier prélève les échantillons de feuilles et de branches (poste 6). L'ensemble des échantillons, dont les rondelles, est apporté dans la zone de pesée des échantillons (poste 7). Lorsque l'équipe des profils de tiges a terminé tous les arbres du jour,

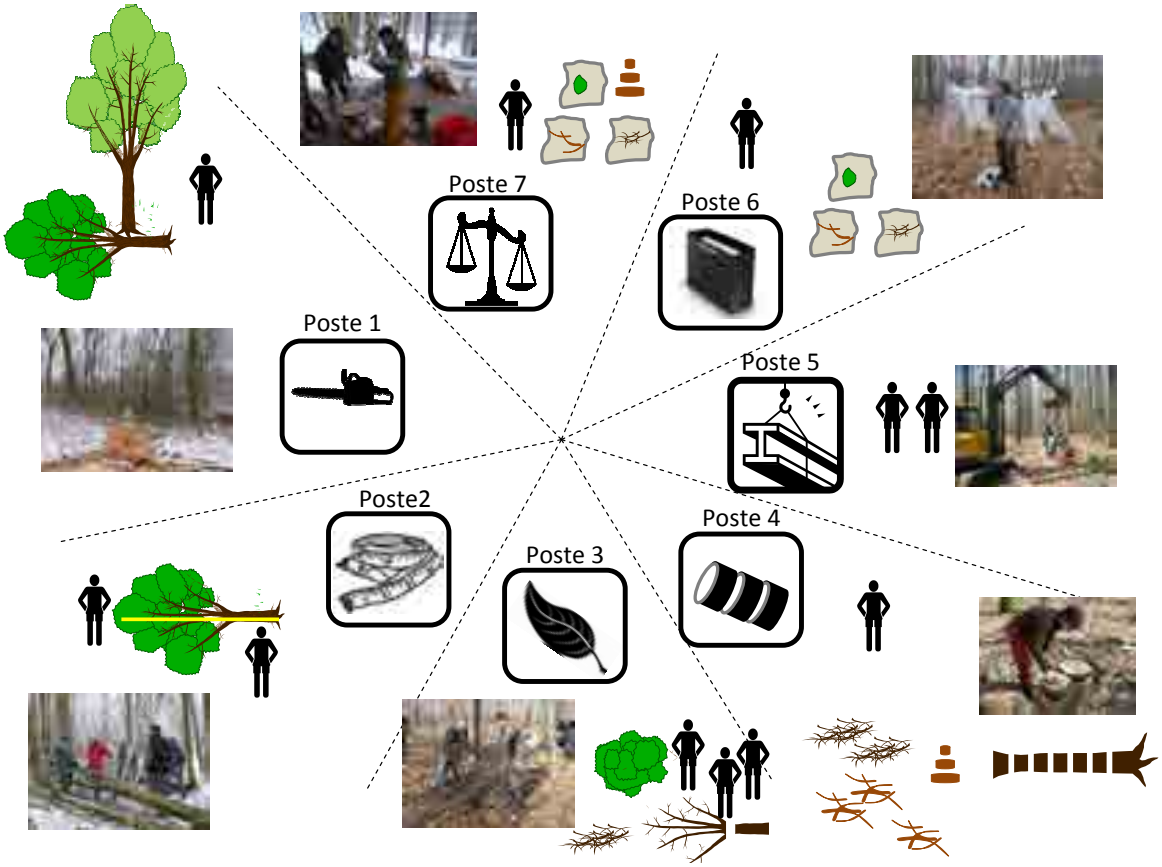


FIGURE 3.2 – *Organisation d'un chantier de biomasse avec 7 postes. Poste 1, préparation du terrain et abattage des arbres (photo: L. Saint-André); poste 2, mesure sur arbres abattus: profils de tige, position des billons (photo: M. Rivoire); poste 3, effeuillage et ébranchage (photo: R. D'Annunzio et M. Rivoire); poste 4, billonnage et étiquetage des rondelles (photo: C. Nys); poste 5, pesée des billons et des fagots (photo: J.-F. Picard); poste 6, échantillonnage des branches (photo: M. Rivoire); poste 7, zone de pesées des échantillons (photo: M. Rivoire).*

elle intervient sur cette zone pour clore les pesées.

Photo 3.3 – *Campagne de mesure dans un taillis sous futaie en France. À gauche, arrivée sur le chantier et déploiement du matériel (photo: M. Rivoire); à droite, abattage du premier arbre (photo: L. Saint-André).*

Ce schéma chronologique est valable lorsque les conditions climatiques sont tempérées. Sous climat tropical, il n'est pas possible d'attendre la fin de la journée pour effectuer la pesée des échantillons. Aussi, la mesure des échantillons doit se faire en même temps que les billons et les fagots. Si cette pesée n'est pas possible *in situ*, il faut la faire au laboratoire mais après transport des échantillons dans une boîte hermétique pour limiter au maximum l'évaporation de l'eau contenue dans les échantillons. Ceci doit être fait en dernier recours et il est plus fiable de faire toutes les pesées sur le terrain.

L'abattage (poste 1)

Le bûcheron prépare l'arbre sélectionné pendant que les techniciens abattent les petites tiges qui peuvent gêner la chute de l'arbre et dégagent les lieux avant abattage. On peut envisager d'étendre une bâche au sol pour ne pas perdre les feuilles lors de l'abattage (photo 3.4). La chute de l'arbre pouvant entraîner d'autres houppiers, les techniciens séparent les branches appartenant à l'arbre sélectionné des branches des autres arbres.

Les mesures sur l'arbre (poste 2)

Les profils de tige sont ensuite mesurés (photo 3.5). Le tronçonnage n'étant jamais effectué au niveau du sol, il est indispensable de marquer la hauteur à 1,30 m avec de la peinture sur le tronc avant abattage et de positionner le mètre ruban en mettant la graduation de 1,30 m du ruban sur la marque de peinture une fois l'arbre abattu. Cela permet d'éviter l'introduction d'un biais dans la localisation des sections (le décalage qui est induit par la hauteur d'abattage). Les circonférences sont en général mesurées tous les mètres ou, plus utile pour construire des modèles de profils de tiges, par pourcentage de la hauteur totale. Cette méthode est toutefois beaucoup plus difficile à réaliser sur le terrain. Lorsque la mesure de circonférence n'est pas possible car le tronc repose complètement sur le sol, il faut effectuer cette mesure au compas forestier avec la prise de deux diamètres perpendiculairement l'un à l'autre. Lors de ce passage le long du tronc, les découpes fixées en accord avec l'organisme en charge des forêts (ou celui qui a acheté les bois) sont matérialisées sur le tronc avec de la peinture ou une griffe forestière.

PHOTO 3.4 – *Campagne de biomasse au Congo dans une plantation d'eucalyptus. À gauche, effeuillage d'un arbre sur une bâche (photo: R. D'Annunzio). À droite, fin du chantier pour un arbre avec sur la zone de pesée les sacs contenant les feuilles, les billons, les fagots de branches (photo: L. Saint-André).*

Dans le cas d'arbres rectilignes avec une tige principale clairement identifiée, il n'est pas nécessaire de choisir l'axe principal. Par contre, dans le cas de tiges très sinueuses ou branchues (houppier des arbres feuillus), il est nécessaire de bien identifier l'axe principal. Celui-ci peut être différencié par un trait de peinture par exemple. L'axe principal se différencie des autres de par son diamètre plus important à chaque division du tronc. Tous les axes branchés sur la tige principale sont considérés comme des branches. Dans le cas des arbres multicaules, il est possible d'inclure chaque brin dans la tige principale (photo 3.6), ou alors de considérer chaque brin comme un individu et il faudra alors tracer l'axe principal sur chacun d'entre eux.

PHOTO 3.5 – *À gauche, campagne de biomasse au Ghana dans une futaie de teck: mesure de branchaison (photo: S. Adu-Bredu). À droite, campagne de biomasse en France dans un taillis sous futaie: mesure de profils de tiges (photo: M. Rivoire).*

La longueur du tronc ainsi que la position de la première branche vivante et des grosses fourches sont ensuite déterminées. Une mesure de la hauteur peut être faite pour différentes

Photo 3.6 – *Campagne de biomasse dans des plantations d'hévéa en Thaïlande. À gauche, cas d'un arbre abattu multicaule (3 tiges sur la même souche): ébranchage et effeuillage. À droite, mélange des feuilles avant de prendre une aliquote (photos: L. Saint-André).*

découpes, par exemple: hauteur du fin bout < 1 cm, hauteur au diamètre de découpe 4 cm, et hauteur au diamètre de découpe 7 cm. Les mesures effectuées sur l'arbre abattu peuvent être ensuite confrontées aux mesures effectuées pendant l'inventaire forestier sur les arbres sur pied. Ceci permet de vérifier la cohérence des jeux de données et de corriger éventuellement des données aberrantes tout en sachant que des différences peuvent exister du fait de l'imprécision de la mesure de hauteur avant abattage (en général 1 m), ou inversement de la sinuosité de la tige ou des casses lors des mesures de longueur après abattage.

La découpe (postes 3 et 4)

L'idéal est de pouvoir tronçonner l'arbre en billons de 2 m de long pour pouvoir tenir compte des variations de densité du bois et d'humidité dans la tige. Une fois l'arbre préparé, les branches sont séparées du tronc (ainsi que les feuilles si nécessaire). Les branches sont ensuite redécoupées pour faire des fagots par classes de diamètre fin bout. Dans le cas d'un peuplement feuillu tempéré les découpes pratiquées sont en général par classe de diamètre > 20 cm, 20–7 cm, 7–4 cm, < 4 cm. Dans le cas de l'eucalyptus au Congo, les branches ont été divisées en deux groupes: < 2 cm et > 2 cm. Les fagots sont réalisés avec des cadres en fer et deux ficelles solides (voir partie 3.5 et photo 3.14). Lorsque les branches sont feuillues, il convient de séparer les feuilles des brindilles. Pour cela, il est nécessaire d'utiliser des bâches pour ne pas perdre de feuilles. Si les feuilles ne se détachent pas bien des axes ligneux (exemple: chêne vert ou résineux), il convient alors d'adopter une stratégie de sous-échantillonnage (voir l'exemple suivant au Cameroun). Les feuilles sont mises dans des grands sacs en plastique pour leur pesée. L'ébranchage et l'effeuillage sont des activités longues et les capacités humaines adéquates (nombre d'équipes suffisant) doivent être allouées pour ne pas ralentir le travail des bûcherons. Pour les branches maîtresses d'un arbre qui sont souvent de très gros diamètre (> 20 cm), il convient de procéder comme pour le tronc par billonnage et prélèvement de rondelles.

Le billonnage est réalisé lorsque les branches ont été séparées de la tige principale. Une rondelle d'environ 3–5 cm d'épaisseur est prélevée au niveau de la souche, puis tous les x

mètres (photo 3.7). La longueur x des billons est dépendante de la dimension de l'arbre et des dispositions prises avec l'organisme en charge des forêts ou l'exploitant forestier. Ce travail de terrain étant fastidieux et long, il faut absolument en profiter pour réaliser des prélèvements multiples (par exemple prélever une rondelle supplémentaire pour des mesures plus détaillées de densité du bois ou de minéralomasse — voir par exemple Saint-André *et al.*, 2002*b*, pour les concentrations en éléments minéraux dans les tiges d'eucalyptus). Il est important d'indiquer la position de chaque rondelle prélevée. Les rondelles doivent être pesées *in situ* le jour même du traitement de l'arbre afin de minimiser les pertes d'humidité (cela nécessite une équipe de deux personnes — en général c'est l'équipe du profil de tige qui réalise cette tâche en interrompant son travail un peu plus tôt pour réaliser les pesées des rondelles, voir figure 3.2).

Pesée des billons et des fagots (poste 5)

Les pesées des billons et des fagots sont réalisées sur le terrain (photo 3.7) et dans le même intervalle de temps afin de s'assurer que les mesures pour un arbre donné ont été réalisées au même taux d'humidité. Il est très pratique d'effectuer ces pesées à l'aide d'un montage pesons-pelleteuse. Les fagots sont accrochés aux pesons et la masse fraîche est mesurée. Les ficelles et la bâche des fagots sont récupérées pour leur réutilisation.

PHOTO 3.7 – *Campagne de biomasse dans une chênaie. À gauche les rondelles prélevées pour un arbre et positionnées dans leur big-bag avant transport vers la zone de pesée des échantillons; au milieu zone de pesée des échantillons; à droite mise en œuvre de la pelleteuse pour la pesée des billons (photos: C. Nys).*

Prélèvement des aliquotes (postes 6 et 7)

Lorsque les fagots sont mesurés, des aliquotes pour chaque fagot sont prélevées pour estimer le taux d'humidité des branches. Il est préférable de prélever des échantillons de différents diamètres dans différentes branches de façon à être représentatifs de l'architecture d'une branche type. En effet, le prélèvement dans une seule branche peut conduire à des biais si elle était plus humide ou plus sèche que les autres. Les branches sont différentiées en quatre groupes en fonction de leur diamètre (classe 1: $0 < \varnothing \leq 4$ cm, classe 2: $4 < \varnothing \leq 7$ cm, classe 3: $7 < \varnothing \leq 20$ cm , et classe 4: $\varnothing > 20$ cm). Pour les branches de classe 1, des échantillons d'environ 10 cm de long sont prélevés. Pour les autres classes, le principe est similaire mais leur diamètre étant plus élevé, des rondelles sont prélevées au lieu des morceaux de 10 cm de long. Environ 9, 6 et 3 rondelles sont prélevées pour les classes 2, 3 et 4. Ces chiffres sont indicatifs mais résultent d'une synthèse des différentes campagnes effectuées dans différents écosystèmes. Les aliquotes sont mises dans des sacs en papiers préparés à cet effet (et qui avaient été déposés au préalable au pied de l'arbre, voir la première étape). Puis, les sacs

en papiers sont disposés dans un sac en plastique pour un arbre donné afin d'assurer que les échantillons ne sont pas mélangés entre arbres.

Afin d'éviter le biais d'échantillonnage, il est important que ce soit toujours la même personne qui échantillonne et qu'elle le fasse de façon systématique et représentative de la variabilité dans chaque classe de taille de branches. Afin de minimiser le biais lié à la mesure du taux d'humidité, les échantillons seront apportés sur l'aire de pesage (même endroit que les rondelles) et pesés dans leur sac en papier avant leur traitement au laboratoire. Si la pesée des échantillons n'est pas possible sur le terrain (ce qui n'est pas recommandé), il s'agira de limiter au maximum les pertes d'humidité et donc l'usage d'une glacière est très fortement recommandé. Pour le prélèvement des feuilles, il conviendra de bien mélanger les échantillons, puis de piocher aléatoirement au milieu du tas ainsi constitué. Il est recommandé de réaliser cette opération mélange-prélèvement cinq ou six fois pour chaque arbre (photo 3.6). Les échantillons pour chaque arbre sont mis dans le même sac (la quantité est à adapter selon la taille des feuilles et de leur hétérogénéité, notamment la proportion de feuilles vertes et de feuilles sénescentes — en général, un sac plastique classique convient très bien).

3.1.2 Au laboratoire

Si les rondelles de tronc ne peuvent pas être mesurées immédiatement, elles devront être stockées à l'air libre et placées sur des tasseaux afin de laisser circuler l'air entre elles (risque de moisissure). Elles peuvent sécher librement dès lors que la pesée de la biomasse fraîche a été réalisée sur le terrain. Par contre, si la pesée n'a pas pu se faire sur le terrain, il convient de les peser tout de suite en arrivant.

Pour les aliquotes pesées dans un sac sur le terrain, il sera nécessaire de réaliser une tare avec un sac vide (si possible mesure de chaque sac, ou s'il est trop détérioré, prendre un lot de 10 à 20 sacs et appliquer un poids moyen correctif). Cette mesure est à soustraire des valeurs mesurées sur le terrain. En cas de remplacement du sac pour faire sécher les aliquotes, il est indispensable de reporter toutes les informations nécessaires.

Le passage à l'étuve se fera à 70°C pour les feuilles, les fleurs, les fruits ou 65°C si des analyses chimiques doivent suivre sur les aliquotes. Pour des opérations de biomasse et pour le bois seulement, la température sera de 105°C. Pour toutes les catégories d'échantillons, un minimum de trois témoins seront pesés tous les jours jusqu'à stabilisation du poids. Cela évite de sortir tous les échantillons à chaque contrôle journalier. La stabilisation prend en général deux jours pour les feuilles et environ une semaine pour les éléments ligneux en fonction des la taille des échantillons.

La figure 3.3 représente le mode opératoire à adopter pour la mesure des échantillons. Les mesures en laboratoire débutent par la mesure de la masse des échantillons humides avec leur sac (mesure de contrôle par rapport à la pesée de terrain). Dans le cas des rondelles de bois, si elles sont trop grosses, il est possible de sous-échantillonner. Il est alors impératif de repeser la rondelle complète, puis le morceau échantillonné. La perte d'humidité entre le terrain et la mesure de la rondelle en laboratoire sera ajoutée à celle mesurée au laboratoire après le passage à l'étuve du morceau échantillonné. Si l'intervalle de temps entre la phase de terrain et la phase au laboratoire est important, le fait d'oublier cette étape du protocole peut générer des erreurs très fortes — jusqu'à 60–70% — sur la biomasse sèche. L'écorçage est en général réalisé à l'aide d'un couteau à écorcer ou d'un ciseau à bois (photo 3.8) — le fait de passer les rondelles au congélateur lorsqu'elles sont encore humides peut parfois faciliter cette opération (par exemple sur le chêne). Les échantillons d'écorce et de bois sont ensuite mesurés et les échantillons sont séchés dans l'étuve (éviter la multiplication des sacs

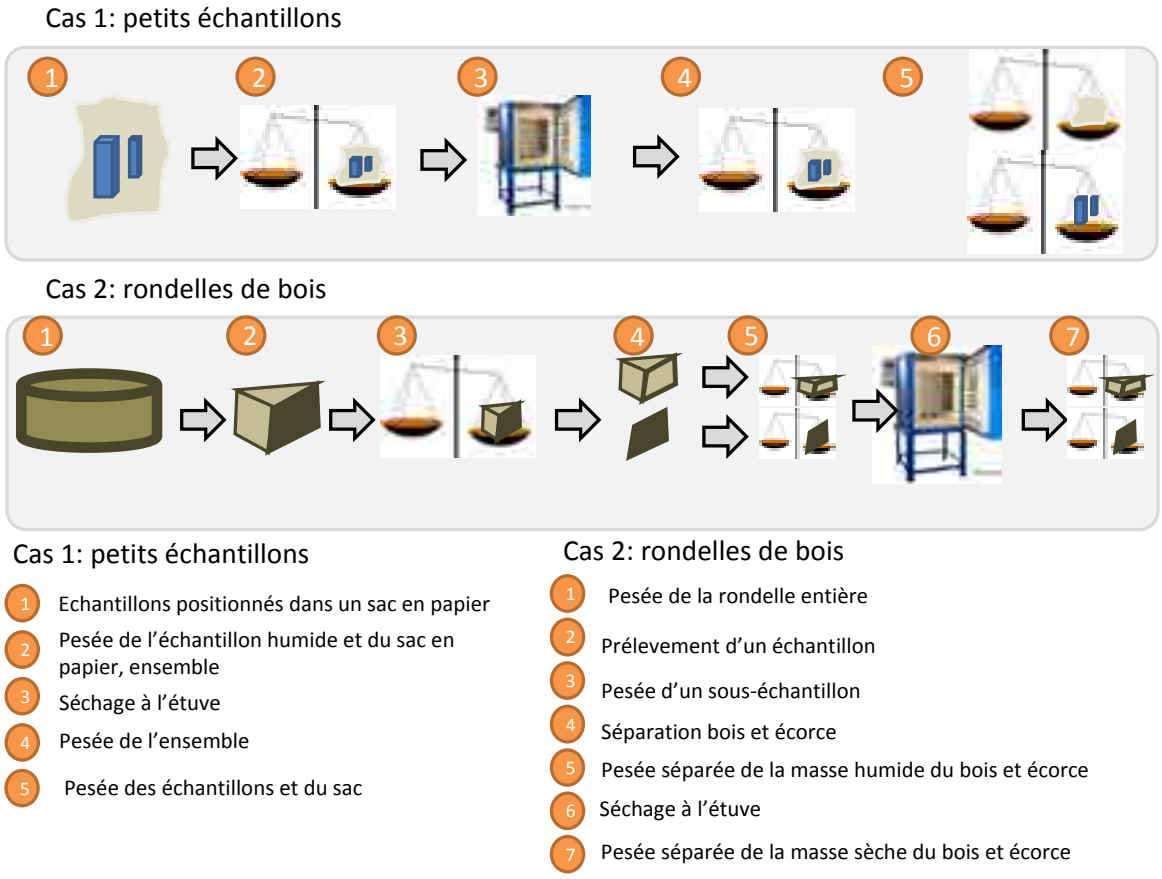


FIGURE 3.3 – *Mode opératoire pour les pesées des échantillons lorsqu'ils arrivent au laboratoire.*

dans l'étuve).

3.1.3 Les calculs

Calcul de la biomasse du tronc

Pour chaque billon i, la mesure de la circonférence aux deux extrémités a été faite: la circonférence C_{1i} au fin bout est la circonférence de la rondelle qui a été prélevée au fin bout et la circonférence C_{2i} au gros bout est la circonférence de la rondelle qui a été prélevée au gros bout. Ceci permet de calculer le volume du billon frais selon la formule du tronçon de cône (ou formule de Newton):

$$V_{\text{frais},i} = L_i \times \frac{\pi}{3} \times (R_{1i}^2 + R_{1i}R_{2i} + R_{2i}^2) \tag{3.1}$$

où L_i est la longueur du billon i, et $R_{1i} = C_{1i}/(2\pi)$ et $R_{2i} = C_{2i}/(2\pi)$ sont les rayons du billon i à ses deux extrémités. Ce volume peut être calculé sur écorce (avec les circonférences mesurées sur le terrain) ou sous écorce (avec les circonférences mesurées sur les rondelles après écorçage au laboratoire). Le volume frais sur écorce est très employé pour les ventes de bois, tandis que la seconde mesure permet de contrôler la cohérence des données en permettant le calcul de la densité du bois dans l'arbre.

À noter qu'il existe d'autres formules pour calculer le volume d'un billon. Les plus utilisées sont la formule de Huber (basée sur la circonférence mesurée au milieu du billon) et

Photo 3.8 – *Mesures en laboratoire: (A) écorçage des rondelles, (B) pesée du bois, (C) pesée de l'écorce (photos: L. Saint-André), (D) mise à l'étuve des échantillons, (E) pesées régulières jusqu'à stabilisation du poids (photos: M. Henry).*

celle de Smalian (basée sur la moyenne quadratique des circonférences mesurées en haut et en bas du billon). Mais dans le cas où la longueur des billons est faible (1 ou 2 m), la forme du tronc s'écarte peu du tronc de cône avec un défilement très peu marqué et la différence entre ces formules est faible.

De plus, pour chaque échantillon prélevé dans le billon i on calcule:

– la proportion en biomasse fraîche du bois (hors écorce):

$$\omega_{\text{bois frais},i} = \frac{B^{\text{aliquote}}_{\text{bois frais},i}}{B^{\text{aliquote}}_{\text{bois frais},i} + B^{\text{aliquote}}_{\text{écorce fraîche},i}}$$

où $B^{\text{aliquote}}_{\text{bois frais},i}$ est la biomasse fraîche du bois (sans écorce) de l'échantillon du billon i, et $B^{\text{aliquote}}_{\text{écorce fraîche},i}$ est la biomasse fraîche de l'écorce de l'échantillon du billon i;

– le taux d'humidité du bois (hors écorce):

$$\chi_{\text{bois},i} = \frac{B^{\text{aliquote}}_{\text{bois sec},i}}{B^{\text{aliquote}}_{\text{bois frais},i}} \tag{3.2}$$

où $B^{\text{aliquote}}_{\text{bois sec},i}$ est la biomasse sèche du bois (sans écorce) de l'échantillon du billon i;

– la proportion en biomasse fraîche de l'écorce:

$$\omega_{\text{écorce fraîche},i} = 1 - \omega_{\text{bois frais},i}$$

– le taux d'humidité de l'écorce:

$$\chi_{\text{écorce},i} = \frac{B^{\text{aliquote}}_{\text{écorce sèche},i}}{B^{\text{aliquote}}_{\text{écorce fraîche},i}}$$

où $B^{\text{aliquote}}_{\text{écorce sèche},i}$ est la biomasse sèche de l'écorce de l'échantillon du billon i.
Puis on extrapole les mesures faites sur l'échantillon du billon i au billon i tout entier par des règles de trois:

– la biomasse sèche du bois (hors écorce) du billon i est:

$$B_{\text{bois sec},i} = B_{\text{frais},i} \times \omega_{\text{bois frais},i} \times \chi_{\text{bois},i}$$

où $B_{\text{frais},i}$ est la biomasse fraîche (écorce comprise) du billon i;

– la biomasse sèche de l'écorce du billon i est:

$$B_{\text{écorce sèche},i} = B_{\text{frais},i} \times \omega_{\text{écorce fraîche},i} \times \chi_{\text{écorce},i}$$

– la densité du bois du billon i est:

$$\rho_i = \frac{B_{\text{bois sec},i}}{V_{\text{frais},i}}$$

où $V_{\text{frais},i}$ est le volume frais *sous* écorce donné par l'équation (3.1).
On fait ensuite la somme des poids secs de tous les billons afin d'obtenir le poids sec du tronc:

– la biomasse sèche du bois (hors écorce) du tronc est:

$$B_{\text{bois sec tronc}} = \sum_i B_{\text{bois sec},i}$$

où la somme porte sur tous les billons i qui composent le tronc;

– la biomasse sèche de l'écorce du tronc est:

$$B_{\text{écorce sèche tronc}} = \sum_i B_{\text{écorce sèche},i}$$

La densité du bois ρ_i qui intervient dans le calcul de la biomasse sèche doit être la densité anhydre (en anglais: « ovendry wood density »), c'est-à-dire le rapport de la biomasse *sèche* (séchage en étuve jusqu'à stabilisation du poids sec) sur le volume *frais* du bois. On prendra garde à ne pas confondre cette densité du bois avec la masse volumique du bois, qui est le rapport masse sur volume, à même teneur en humidité pour la masse et le volume (c'est-à-dire masse sèche sur volume sec, ou masse fraîche sur volume frais). Toutefois, la norme AFNOR (1985) définit autrement la densité du bois, comme le rapport de la biomasse séchée à l'air libre sur le volume du bois à 12 % d'humidité (Fournier-Djimbi, 1998). La densité anhydre du bois peut être calculée à partir de la densité du bois à 12 % d'humidité par la relation (Gourlet-Fleury *et al.*, 2011):

$$\rho_\chi = \frac{\rho\,(1+\chi)}{1-\eta\,(\chi_0-\chi)}$$

où ρ_χ est le ratio de la biomasse séchée à l'air libre sur le volume du bois à χ % d'humidité (en $\mathrm{g\,cm^{-3}}$), ρ est le ratio de la biomasse séchée à l'étuve sur le volume frais du bois (en $\mathrm{g\,cm^{-3}}$), η est le coefficient de retrait volumique (nombre sans dimension) et χ_0 est le point de saturation des fibres. Les coefficients η et χ_0 varient d'une espèce à l'autre et requièrent une connaissance des propriétés technologiques du bois des espèces. En utilisant les données de ρ et $\rho_{12\%}$ de 379 arbres, Reyes *et al.* (1992) ont par ailleurs établi une relation empirique entre la densité anhydre ρ et la densité à 12 % d'humidité $\rho_{12\%}$: $\rho = 0{,}0134 + 0{,}800\rho_{12\%}$ avec un coefficient de détermination $R^2 = 0{,}988$.

Calcul de la biomasse des feuilles

Pour chaque échantillon i de feuillage prélevé, on calcule le taux d'humidité du feuillage:

$$\chi_{\text{feuille},i} = \frac{B^{\text{aliquote}}_{\text{feuille sèche},i}}{B^{\text{aliquote}}_{\text{feuille fraîche},i}}$$

où $B^{\text{aliquote}}_{\text{feuille sèche},i}$ est la biomasse sèche du feuillage de l'échantillon i, et $B^{\text{aliquote}}_{\text{feuille fraîche},i}$ est la biomasse fraîche du feuillage de l'échantillon i. Puis on extrapole par une règle de trois l'échantillon i au compartiment i dont cet échantillon est extrait:

$$B_{\text{feuille sèche},i} = B_{\text{feuille fraîche},i} \times \chi_{\text{feuille},i}$$

où $B_{\text{feuille sèche},i}$ est la biomasse sèche (calculée) du feuillage du compartiment i, et $B_{\text{feuille fraîche},i}$ est la biomasse fraîche (mesurée) du feuillage du compartiment i. Souvent il n'y a qu'un seul compartiment correspondant au houppier tout entier. Mais lorsque le houppier a été compartimenté (par exemple par tiers successifs), le poids sec total des feuilles s'obtient en sommant sur les compartiments i:

$$B_{\text{feuille sèche}} = \sum_i B_{\text{feuille sèche},i}$$

Calcul de la biomasse des branches

Dans le cas des très grosses branches (par exemple > 20 cm de diamètre), il faut procéder comme pour le tronc, tandis que pour les fagots, il faut procéder comme pour les feuilles.

Calcul de la biomasse des fruits, fleurs

Le calcul est identique à celui des feuilles.

3.2 Pesées directes pour certaines compartiments et mesures de volume et de densité pour d'autres compartiments

Le second cas que nous envisageons est celui avec contraintes d'abattage, amenant à faire des mesures semi-destructives qui combinent des pesées directes pour certaines parties de l'arbre, et des mesures de volume et de densité pour d'autres parties. Nous illustrons ce cas par le développement d'une équation allométrique pour des forêts sèches dans le nord du Cameroun. L'évaluation de la biomasse des forêts sèches est particulièrement difficile du fait de la complexité de l'architecture des arbres. Dans les zones sèches, l'intervention humaine est particulièrement significative de par la rareté de la ressource forestière et l'importance de la demande bioénergétique. Celle-ci se reflète par des pratiques d'émondage, de taille et d'entretien des arbres souvent situés dans des forêts claires, des parcs agro-forestiers ou des haies (photo 3.9).

Dans la plupart des zones sèches les arbres sont protégés car la régénération de la ressource ligneuse est particulièrement lente et mise en péril par les activités humaines. Les mesures de biomasse sont préférablement non-destructives et profitent de la pratique de l'émondage pour mesurer la biomasse des compartiments émondés. Les activités de pâturage limitent la régénération et les petits arbres sont souvent peu représentés. Aussi, cette partie du manuel tend à considérer seulement les arbres matures.

PHOTO 3.9 – *Émondage des karités (Vitellaria paradoxa) dans le nord du Cameroun (photo: R. Peltier).*

3.2.1 Sur le terrain: cas de mesures semi-destructives

Généralement, le tronc et les grosses branches ne sont pas émondés. Ce sont les petites branches qui sont concernées. La mesure de la biomasse fraîche (en kg) peut être divisée en deux parties: mesure de la biomasse fraîche émondée et mesure de la biomasse fraîche non émondée (figure 3.4A).

Biomasse fraîche émondée

Les branches peuvent être émondées en suivant les pratiques locales (souvent à la machette). Le diamètre à la base de chaque branche émondée est mesuré à l'aide d'un mètre ruban. Puis les feuilles et le bois des branches émondées sont séparés. La biomasse fraîche des feuilles des branches émondées ($B_{\text{feuille fraîche émondée}}$) et la biomasse fraîche du bois des branches émondées ($B_{\text{bois frais émondé}}$) sont pesées séparément. La mesure du poids est effectuée à l'aide de pesons adaptés. Lorsque la masse des feuilles est inférieure à deux kilogrammes, il est possible de mesurer leur poids à l'aide d'une balance électronique de terrain.

Une aliquote de feuilles est prélevée au hasard parmi les feuilles des branches émondées. Un minimum de trois échantillons de feuilles provenant de trois branches différentes est en général requis pour former l'aliquote. Sa masse fraîche ($B^{\text{aliquote}}_{\text{feuille fraîche}}$ en g) est mesurée. Une aliquote de bois est également prélevée au hasard dans le bois des branches émondées, sans enlever l'écorce. Sa masse fraîche ($B^{\text{aliquote}}_{\text{bois frais}}$ en g) est mesurée sur le terrain aussitôt après la coupe. Les aliquotes sont mis dans des sacs plastique numérotés et amenés laboratoire. Le volume frais de l'aliquote de bois sera ultérieurement mesuré en laboratoire (cf. § 3.2.2), ce qui permettra de déterminer la densité moyenne du bois $\bar{\rho}$.

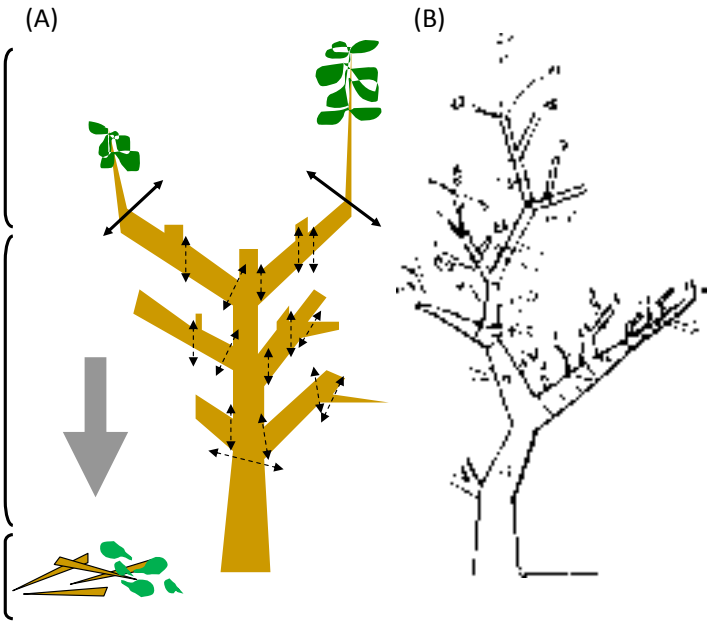


Figure 3.4 – *Détermination de la biomasse fraîche totale. (A) Séparation et mesure de la biomasse émondée et non-émondée, (B) numérotations des tronçons et des branches mesurées sur l'arbre émondé.*

Biomasse fraîche non émondée

La mesure de la biomasse non émondée est indirecte puisque non destructive. Les différentes ramifications de l'arbre émondé sont identifiées et les branches numérotées (figure 3.4B). Les petites branches non émondées sont traitées différemment des grosses branches et du tronc (figure 3.4A). Pour les petites branches non émondées, seul le diamètre à la base est mesuré. La biomasse des petites branches non émondées sera estimée à partir d'une relation entre leur diamètre à la base et leur masse, comme expliqué dans la section 3.2.3.

La biomasse du tronc et des grosses branches est estimée à partir de mesures de volumes (V_i en cm^3) et de la densité moyenne du bois ($\bar{\rho}$ en $\mathrm{g\,cm}^{-3}$). Les grosses branches et le tronc de l'arbre émondé sont virtuellement divisés en tronçons matérialisés par des marques sur l'arbre. Le volume V_i de chaque tronçon i est obtenu à partir de la mesure de son diamètre (ou de sa circonférence) et de sa longueur. Une longueur de tronçon d'environ un mètre est souhaitable afin de pouvoir mieux considérer les variations de diamètre le long du tronc et des branches.

3.2.2 Au laboratoire

On mesure en premier lieu le volume ($V^{\text{aliquote}}_{\text{bois frais}}$) de l'aliquote de bois extrait des compartiments émondés. Le volume de bois peut être mesuré de différentes façons (Maniatis *et al.*, 2011). La méthode la plus couramment utilisée consiste à mesurer le déplacement du volume d'eau provoqué par l'immersion de l'échantillon dans l'eau. La mesure du volume d'eau peut être faite à l'aide d'une éprouvette adaptée à la taille de l'échantillon (figure 3.5). Une autre méthode consiste à découper les échantillons pour leur donner une forme dont le volume peut être mesuré le plus précisément possible. Cette méthode nécessite des outils de précision et du personnel formé pour la découpe du bois.

Les aliquotes de bois et de feuille sont ensuite soumises aux mêmes mesures en laboratoire (séchage à l'étuve, pesée du poids sec, etc.) que celles décrites dans le paragraphe 3.1.2.

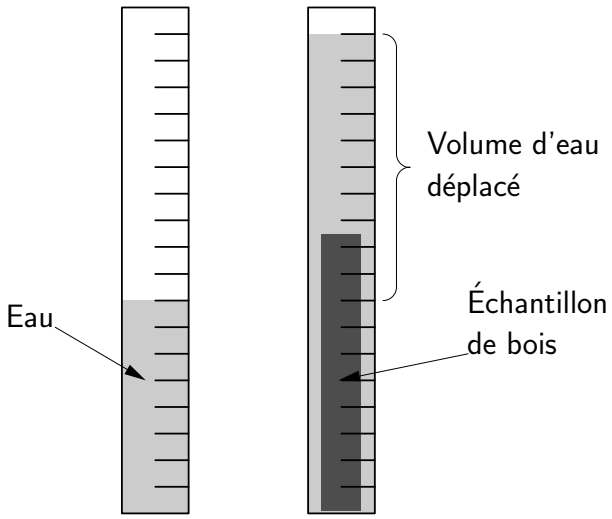


FIGURE 3.5 – *Mesure du volume des échantillons par déplacement du volume d'eau.*

3.2.3 Les calculs

La biomasse sèche de l'arbre s'obtient comme la somme de la biomasse sèche émondée et de la biomasse sèche non émondée:

$$B_{\text{sec}} = B_{\text{sec émondé}} + B_{\text{sec non-émondé}}$$

Calcul de la biomasse émondée

À partir de la biomasse fraîche $B^{\text{aliquote}}_{\text{bois frais}}$ de l'aliquote de bois et de sa biomasse sèche $B^{\text{aliquote}}_{\text{bois sec}}$, on calcule comme précédemment (cf. équation 3.2) le taux d'humidité du bois (écorce comprise):

$$\chi_{\text{bois}} = \frac{B^{\text{aliquote}}_{\text{bois sec}}}{B^{\text{aliquote}}_{\text{bois frais}}}$$

De même, on calcule le taux d'humidité du feuillage à partir de la biomasse fraîche $B^{\text{aliquote}}_{\text{feuille fraîche}}$ de l'aliquote de feuille et de sa biomasse sèche $B^{\text{aliquote}}_{\text{feuille sèche}}$:

$$\chi_{\text{feuille}} = \frac{B^{\text{aliquote}}_{\text{feuille sèche}}}{B^{\text{aliquote}}_{\text{feuille fraîche}}}$$

On peut alors calculer la biomasse sèche émondée:

$$B_{\text{sec émondé}} = B_{\text{bois frais émondé}} \times \chi_{\text{bois}} + B_{\text{feuille fraîche émondée}} \times \chi_{\text{feuille}}$$

où $B_{\text{feuille fraîche émondée}}$ est la biomasse fraîche des feuilles des branches émondées et $B_{\text{bois frais émondé}}$ est la biomasse fraîche du bois des branches émondées.

Calcul de la biomasse non émondée

Pour la biomasse sèche de la partie non émondée (celle qui reste sur pied), deux calculs sont faits: un calcul pour les petites branches, et un autre pour les grosses branches et le tronc. La biomasse non émondée est la résultante des deux:

$$B_{\text{sec non-émondé}} = B_{\text{branche sèche non-émondée}} + B_{\text{sec tronçon}}$$

Chaque tronçon i du tronc et des grosses branches peut être considéré comme un cylindre dont le volume est (formule de Smalian):

$$V_i = \frac{\pi}{8} L_i \, (D_{1i}^2 + D_{2i}^2) \tag{3.3}$$

où V_i est le volume du i^{e} tronçon, L_i sa longueur, et D_{1i} et D_{2i} les diamètres des deux extrémités du tronçon i. La formule du tronçon de cône (voir équation 3.1) peut également être utilisée à la place de la formule (3.3) du cylindre, mais les différences seront faibles entre les deux calculs car le défilement sur un mètre de long n'est pas très prononcé pour les arbres.

La biomasse sèche des grosses branches et du tronc s'obtient comme le produit de la densité moyenne du bois et du volume total des grosses branches et du tronc:

$$B_{\text{sec tronçon}} = \bar{\rho} \times \sum_i V_i \tag{3.4}$$

où la somme porte sur l'ensemble des tronçons qui composent les grosses branches et le tronc (figure 3.4B), et où la densité moyenne du bois se calcule par:

$$\bar{\rho} = \frac{B_{\text{bois sec}}^{\text{aliquote}}}{V_{\text{bois frais}}^{\text{aliquote}}}$$

On prendra garde à ce que les unités de mesure soient cohérentes entre elles. Par exemple, si la densité moyenne du bois $\bar{\rho}$ dans (3.4) s'exprime en $\text{g}\,\text{cm}^{-3}$, alors le volume V_i doit s'exprimer en cm^3, ce qui amène à exprimer à la fois la longueur L_i et les diamètres D_{1i} et D_{2i} en cm. La biomasse est alors exprimée en g.

La biomasse sèche des petites branches non émondées est calculée à l'aide d'un modèle entre la biomasse sèche et le diamètre à la base. Pour cela, le modèle est établi en suivant la même procédure que pour le développement d'un modèle allométrique (voir chapitres 4 à 7 du manuel). Les équations utilisées sont souvent de type puissance:

$$B_{\text{branche sèche}} = a + bD^c$$

où a, b et c sont les paramètres du modèle et D le diamètre à la base de la branche, mais d'autres régressions peuvent être testées et plus adaptées (cf. tableau 5.1). Avec un modèle de ce type, la biomasse sèche des petites branches non émondées serait:

$$B_{\text{branche sèche non-émondée}} = \sum_j (a + bD_j^c)$$

où la somme porte sur l'ensemble des petites branches non émondées et D_j est le diamètre à la base de la j^{e} branche.

3.3 Pesées partielles sur le terrain

Le troisième cas que nous envisageons est celui d'arbres de dimensions trop grandes pour une pesée intégrale à la main. Nous l'illustrons par l'établissement d'une équation allométrique pour estimer la biomasse aérienne des arbres d'une forêt tropicale humide par mesure destructive. Il est important que la méthode proposée soit adaptée aux circonstances locales et aux moyens mis à disposition. La valeur commerciale et la demande en bois sont deux facteurs à prendre en compte dans le cas des mesures dans les concessions forestières.

Les arbres sélectionnés sont abattus en suivant des pratiques d'abattage adaptées. Une fois l'arbre abattu, les variables telles que la hauteur totale et la hauteur des contreforts (lorsque l'arbre en possède) peuvent être mesurées à l'aide d'un mètre ruban. Ensuite, l'architecture de l'arbre est analysée (figure 3.6). L'approche proposée différentie les arbres qui peuvent être mesurés par pesée manuelle sur le terrain (par exemple les arbres de diamètre ≤ 20 cm) et ceux nécessitant des moyens techniques plus conséquents (les arbres de diamètre > 20 cm).

3.3.1 Arbres ayant un diamètre inférieur à 20 cm

Pour les arbres de diamètre ≤ 20 cm, la démarche est similaire à celle décrite dans le premier exemple (§ 3.1). Les branches et le tronc sont séparés. Les biomasses fraîches du tronc ($B_{\text{tronc frais}}$) et des branches ($B_{\text{branche fraîche}}$, bois et feuilles ensemble) sont mesurées à l'aide de pesons adaptés. Pour mesurer la biomasse des feuilles, un nombre limité de branches sont sélectionnées au hasard pour chaque individu. Les feuilles et le bois de cet échantillon de branches sont séparés. La biomasse fraîche des feuilles ($B^{\text{échantillon}}_{\text{feuille fraîche}}$) et la biomasse fraîche du bois ($B^{\text{échantillon}}_{\text{bois frais}}$) de cet échantillon de branches sont mesurées séparément à l'aide de pesons. La proportion foliaire des branches est alors calculée:

$$\omega_{\text{feuille}} = \frac{B^{\text{échantillon}}_{\text{feuille fraîche}}}{B^{\text{échantillon}}_{\text{feuille fraîche}} + B^{\text{échantillon}}_{\text{bois frais}}}$$

Les biomasses fraîches foliaire ($B_{\text{feuille fraîche}}$) et ligneuse ($B_{\text{bois frais}}$) des branches sont ensuite calculées à partir de cette proportion moyenne de feuille:

$$\begin{aligned} B_{\text{feuille fraîche}} &= \omega_{\text{feuille}} \times B_{\text{branche fraîche}} \\ B_{\text{bois frais}} &= (1 - \omega_{\text{feuille}}) \times B_{\text{branche fraîche}} \end{aligned}$$

Des aliquotes de feuille et de bois sont ensuite prélevées à différent niveaux dans les branches et le long du tronc. La biomasse fraîche ($B^{\text{aliquote}}_{\text{feuille fraîche}}$ et $B^{\text{aliquote}}_{\text{bois frais}}$) des aliquotes est mesurée à l'aide d'une balance électronique sur le terrain. Les aliquotes sont amenées au laboratoire, séchées et pesées, selon le même protocole que celui décrit dans le premier exemple (§ 3.1.2). La biomasse sèche ($B^{\text{aliquote}}_{\text{feuille sèche}}$ et $B^{\text{aliquote}}_{\text{bois sec}}$) des aliquotes permet de calculer le taux d'humidité des feuilles et du bois:

$$\chi_{\text{feuille}} = \frac{B^{\text{aliquote}}_{\text{feuille sèche}}}{B^{\text{aliquote}}_{\text{feuille fraîche}}}, \qquad \chi_{\text{bois}} = \frac{B^{\text{aliquote}}_{\text{bois sec}}}{B^{\text{aliquote}}_{\text{bois frais}}}$$

Les biomasses sèches folaire et ligneuse sont finalement obtenues à partir de leur biomasse fraîche et des taux d'humidité calculés à partir des aliquotes. Pour la biomasse ligneuse, on ajoutera la biomasse fraîche du bois des branches et celle du tronc:

$$\begin{aligned} B_{\text{feuille sèche}} &= \chi_{\text{feuille}} \times B_{\text{feuille fraîche}} \\ B_{\text{ligneux sec}} &= \chi_{\text{bois}} \times (B_{\text{bois frais}} + B_{\text{tronc frais}}) \end{aligned}$$

La masse sèche totale est finalement obtenue comme la somme de la biomasse sèche foliaire et de la biomasse sèche ligneuse:

$$B_{\text{sec}} = B_{\text{feuille sèche}} + B_{\text{ligneux sec}}$$

3.3.2 Arbres ayant un diamètre supérieur à 20 cm

Il n'est pas pratique de séparer les branches du tronc lorsque les arbres deviennent trop grands, à cause de la quantité de branchage et de feuillage. La méthode alternative proposée ici consiste à traiter différemment le tronc et les grosses branches (de diamètre basal supérieur à, mettons, 10 cm) d'une part, et les petites branches (de diamètre basal inférieur à 10 cm) d'autre part. Tandis que les grosses branches de diamètre basal > 10 cm ne sont constituées que de bois, les petites branches de diamètre basal ≤ 10 cm peuvent comporter également du feuillage. Les grosses branches de diamètre basal > 10 cm sont traitées de la même façon que le tronc. La première étape consiste à les diviser en sections de bois. Alors que la biomasse des sections de diamètre supérieur à 10 cm est déduite de leur volume mesuré ($V_{\text{billon},i}$) et de la densité moyenne du bois ($\bar{\rho}$), la biomasse des branches de diamètre basal ≤ 10 cm est estimée à partir d'une régression entre leur diamètre à la base et la biomasse qu'elles portent.

Mesure du volume des sections de diamètre supérieur à 10 cm (tronc ou branche)

Une fois le tronc et les branches de diamètre basal > 10 cm divisés en sections, le volume des sections est calculé à partir de leur longueur et de leurs diamètres (ou de leurs circonférences) aux deux extrémités (D_{1i} et D_{2i}). Il est possible de fixer un intervalle fixe (par exemple tous les deux mètres) pour mesurer le diamètre de chacune des sections (photo 3.10A). Par endroits, une longueur de section plus courte que la longueur fixée doit être utilisée car une ramification empêche de donner une forme cylindrique au billon. Le technicien note alors la longueur et les diamètres de chacune des sections. Il établit un schéma représentant l'architecture de l'arbre (figure 3.6). Ce schéma est particulièrement utile pour l'analyse des résultats et leur interprétation.

Les arbres de diamètre > 20 cm présentent souvent des contreforts. Le volume des contreforts peut être estimé en faisant l'hypothèse que la forme des contreforts correspond à une pyramide dont l'arête supérieure est un quart d'ellipse (encart de la figure 3.6; Henry *et al.*, 2010). Pour chaque contrefort j, on mesure sa hauteur H_j, sa largeur l_j et sa longueur L_j (encart de la figure 3.6).

Des aliquotes de bois sont ensuite prélevées dans les différentes sections de diamètre supérieur à 10 cm (tronc, branches, et contreforts le cas échéant; photo 3.10B). Les aliquotes de bois frais sont disposés dans des sacs hermétiques et transportés jusqu'au laboratoire. En laboratoire, leur volume ($V^{\text{aliquote}}_{\text{bois frais}}$) est mesuré selon le protocole décrit au paragraphe 3.2.2. Les aliquotes de bois sont ensuite séchées et pesées comme décrit au paragraphe 3.1.2, ce qui permet d'obtenir leur biomasse sèche ($B^{\text{aliquote}}_{\text{bois sec}}$).

Calcul de la biomasse des sections de diamètre supérieur à 10 cm (tronc ou branche)

Comme précédemment (cf. équation 3.3), le volume $V_{\text{billon},i}$ de la i^{e} section (tronc ou branche de diamètre basal > 10 cm) est calculé par la formule de Smalian:

$$V_{\text{billon},i} = \frac{\pi \times L_i}{8}(D_{1i}^2 + D_{2i}^2)$$

où L_i est la longueur de i^{e} section, D_{1i} est le diamètre d'une de ses extrémités et D_{2i} est le diamètre de son autre extrémité. Compte-tenu de sa forme pyramidale, une formule différente est utilisée pour calculé le volume $V_{\text{contrefort},j}$ du j^{e} contrefort:

$$V_{\text{contrefort},j} = \left(1 - \frac{\pi}{4}\right)\frac{L_j H_j l_j}{3}$$

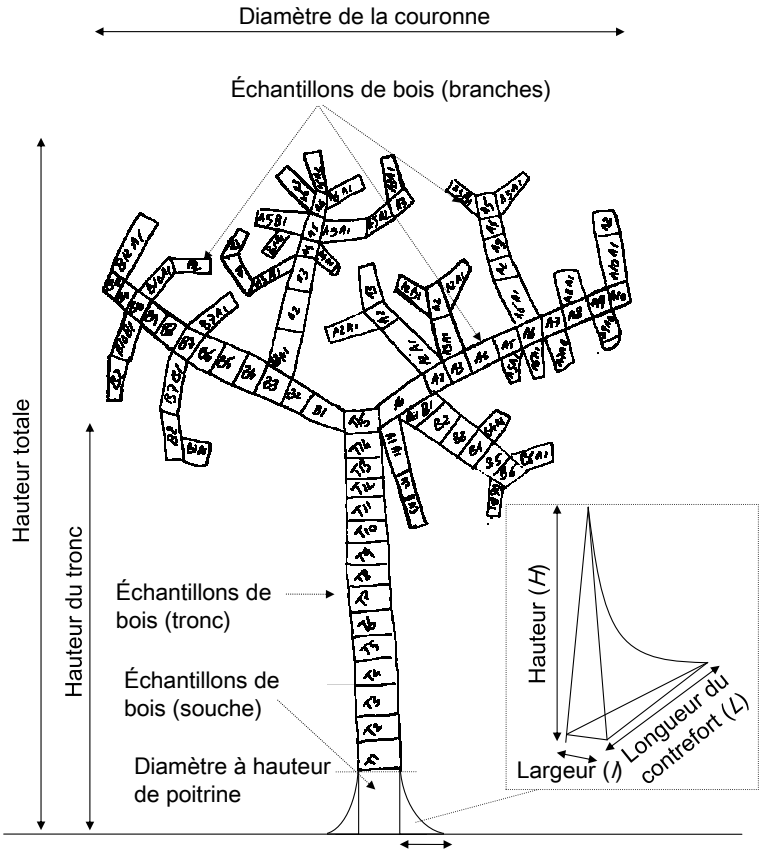


FIGURE 3.6 – *Schéma représentant les différentes sections d'un arbre pour le calcul de son volume.*

où l_j est la largeur du j^{e} contrefort, L_j sa longueur et H_j sa hauteur.

Par ailleurs, à partir de la biomasse et sèche et du volume frais des aliquotes de bois, on peut calculer la densité moyenne du bois:

$$\bar{\rho} = \frac{B^{\text{aliquote}}_{\text{bois sec}}}{V^{\text{aliquote}}_{\text{bois frais}}}$$

La biomasse sèche cumulée des sections (tronc et branches de diamètre basal > 10 cm) est alors:

$$B_{\text{sèche sections}} = \bar{\rho} \times \sum_i V_{\text{billon},i}$$

où la somme porte sur l'ensemble des sections, tandis que la biomasse sèche des contreforts est:

$$B_{\text{sèche contreforts}} = \bar{\rho} \times \sum_j V_{\text{contrefort},j}$$

où la somme porte sur l'ensemble des contreforts. Alternativement à la densité moyenne du bois, une densité de bois spécifique à chaque compartiment (tronc, branches, contreforts) pourra être utilisée. Dans ce cas, la densité moyenne du bois $\bar{\rho}$ sera remplacée dans les formules ci-dessus par la densité spécifique au compartiment idoine.

Photo 3.10 – *Mesures sur le terrain d'un grand arbre: (A) mesure du volume d'un arbre de diamètre* > 20 *cm, (B) prélèvement d'aliquotes de bois au niveau du tronc (photos: M. Henry).*

Mesure des branches de diamètre inférieur à 10 cm

Pour toutes les branches de diamètre basal ≤ 10 cm, le diamètre à la base est mesuré. Leur biomasse sèche sera estimée à partir d'une régression entre le diamètre à la base de la branche et la masse sèche qu'elle porte. Cette régression est établie à partir d'un échantillon de branches sélectionnées dans l'arbre afin de représenter les différentes classes de diamètres à leur base. Pour chaque branche de cet échantillon, les feuilles et le bois sont séparés. La biomasse fraîche des feuilles ($B^{\text{échantillon}}_{\text{feuille fraîche},i}$ pour la i^{e} branche) et la biomasse fraîche du bois ($B^{\text{échantillon}}_{\text{bois frais},i}$ pour la i^{e} branche) de chaque branche de l'échantillon sont pesées séparément sur le terrain.

Il est possible que certaines branches aient des malformations et qu'elles n'aboutissent pas à une architecture ramifiée. Dans ce cas, le volume peut être mesuré et l'anomalie est enregistrée sur les feuilles de terrain.

Des aliquotes de bois et de feuilles sont ensuite prélevées et leur biomasse fraîche ($B^{\text{aliquote}}_{\text{bois frais}}$ et $B^{\text{aliquote}}_{\text{feuille fraîche}}$) est aussitôt pesées sur le terrain. Les aliquotes sont placées dans des sacs plastiques hermétiques, amenées au laboratoire où elles sont séchées et pesées selon le protocole indiqué dans le paragraphe 3.1.2. On obtient ainsi leur biomasse sèche ($B^{\text{aliquote}}_{\text{bois sec}}$ et $B^{\text{aliquote}}_{\text{feuille sèche}}$)

Calcul de la biomasse des branches de diamètre inférieur à 10 cm

La biomasse fraîche et sèche des aliquotes sert à déterminer la teneur en humidité des feuilles et du bois:

$$\chi_{\text{feuille}} = \frac{B^{\text{aliquote}}_{\text{feuille sèche}}}{B^{\text{aliquote}}_{\text{feuille fraîche}}}, \qquad \chi_{\text{bois}} = \frac{B^{\text{aliquote}}_{\text{bois sec}}}{B^{\text{aliquote}}_{\text{bois frais}}}$$

On en déduit, pour chaque branche i de l'échantillon de branches, la biomasse sèche des feuilles, la biomasse sèche du bois, puis la biomasse sèche totale de la branche i:

$$\begin{aligned} B^{\text{échantillon}}_{\text{feuille sèche},i} &= \chi_{\text{feuille}} \times B^{\text{échantillon}}_{\text{feuille fraîche},i} \\ B^{\text{échantillon}}_{\text{bois sec},i} &= \chi_{\text{bois}} \times B^{\text{échantillon}}_{\text{bois frais},i} \\ B^{\text{échantillon}}_{\text{branche sèche},i} &= B^{\text{échantillon}}_{\text{feuille sèche},i} + B^{\text{échantillon}}_{\text{bois sec},i} \end{aligned}$$

Comme dans le paragraphe 3.2.3, un tarif de biomasse pour les branches est ensuite ajusté aux données ($B^{\text{échantillon}}_{\text{branche sèche},i}$, $D^{\text{échantillon}}_{i}$), où $D^{\text{échantillon}}_{i}$ est le diamètre à la base de la i^{e} branche de l'échantillon. Le tarif de biomasse pour les branches est établi en suivant la même procédure que pour le développement d'une équation allométrique (voir chapitre 4 à 7 du manuel). Afin d'augmenter la taille d'échantillon, le tarif pourra être établi à partir de toutes les branches mesurées pour tous les arbres de la même espèce ou par groupes fonctionnels d'espèces (Hawthorne, 1995).

En utilisant le tarif de biomasse pour branche ainsi établi, on peut calculer la biomasse sèche des branches de diamètre basal ≤ 10 cm:

$$B_{\text{branche sèche}} = \sum_i f(D_i)$$

où la somme porte sur l'ensemble des branches de diamètre basal ≤ 10 cm, D_i est le diamètre basal de la i^{e} branche, et f est le tarif de biomasse qui prédit la biomasse sèche d'une branche en fonction de son diamètre basal.

Calcul de la biomasse de l'arbre

La biomasse sèche de l'arbre s'obtient en sommant la biomasse sèche des sections (tronc et branches de diamètre basal > 10 cm), la biomasse sèche des contreforts et la biomasse sèche des branches de diamètre ≤ 10 cm:

$$B_{\text{sec}} = B_{\text{sèche sections}} + B_{\text{sèche contreforts}} + B_{\text{branche sèche}}$$

3.4 Cas des mesures racinaires

Les mesures de biomasse racinaires sont beaucoup plus difficiles à effectuer que les biomasses aériennes. Les méthodes que nous proposons ici sont le résultat de campagnes réalisées dans différents écosystèmes et ont fait l'objet d'une étude comparative au Congo (Levillain *et al.*, 2011).

La première étape, quel que soit l'écosystème, consiste à tracer le diagramme de Voronoï [1] autour de l'arbre sélectionné. La figure 3.7 indique la démarche à suivre: (*i*) tracer les segments qui relient l'arbre sélectionné à chacun de ses voisins; (*ii*) tracer les médiatrices

1. Un diagramme de Voronoï (aussi appelé décomposition de Voronoï, partition de Voronoï, polygones de Voronoï) représente une décomposition particulière d'un espace métrique déterminée par les distances à un ensemble discret d'objets de l'espace, en général un ensemble discret de points.

de chaque segment, (*iii*) relier les médiatrices entre elles pour délimiter un espace autour de l'arbre; (*iv*) cet espace peut ensuite être divisé en triangles jointifs, la surface de chaque zone étant facile à calculer en utilisant la formule du triangle et en connaissant les longueurs des trois cotés (a, b et c):

$$A = \sqrt{p(p-2a)(p-2b)(p-2c)}$$

où $p = a + b + c$ est le périmètre du triangle et A sa surface. La figure 3.8 illustre cette démarche pour des plantations de cocotier au Vanuatu (Navarro *et al.*, 2008).

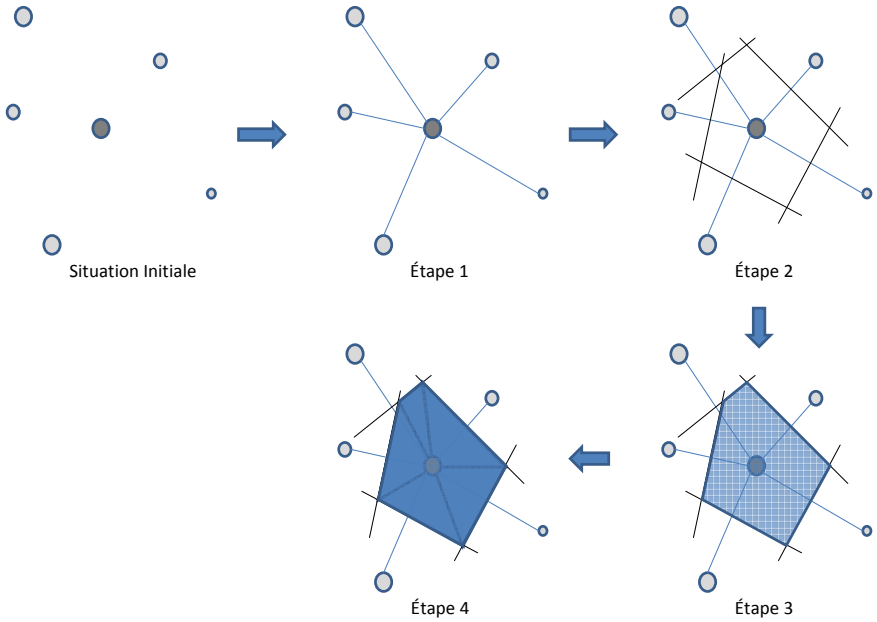


Figure 3.7 – *Méthode pour tracer un espace de Voronoï et ses subdivisions autour d'un arbre et dans une situation de voisinage quelconque.*

L'espace ainsi délimité ne constitue pas une matérialisation de l'espace « vital » de l'arbre. Il s'agit juste d'une façon de découper l'espace en zones jointives pour ensuite faciliter l'échantillonnage de la biomasse souterraine. L'hypothèse majeure est que les racines d'un autre arbre qui viennent coloniser cet espace compensent celles qui en sortent et qui appartiennent à l'arbre sélectionné.

Dans le cas de peuplements plurispécifiques ou agro-forestiers, il est parfois difficile, voire impossible, de séparer les racines des différentes espèces. Dans ce cas, il sera très hasardeux de faire des modèles individuels (biomasse racinaire reliée à l'arbre échantillonné) mais les estimations de biomasse racinaire à l'hectare, sans distinction des essences, restent tout à fait valables.

Les méthodes d'échantillonnage varient ensuite selon la grosseur des racines. Levillain *et al.* (2011) ont réalisé une étude qui compare différentes méthodes sur un même arbre (photo 3.11). Ils montrent qu'il est plus rentable en termes de coût–précision d'échantillonner les racines fines par carottage, tandis que les racines moyennes nécessitent une excavation partielle et les grosses racines une excavation totale de l'espace de Voronoï.

Le nombre de carottes et la dimension de la fosse à excaver varient d'un écosystème à un autre. Au Congo, dans les plantations d'eucalyptus, le nombre optimum de carottes pour obtenir une précision de 10 % est de l'ordre de 300 en surface (0–10 cm) et d'une centaine

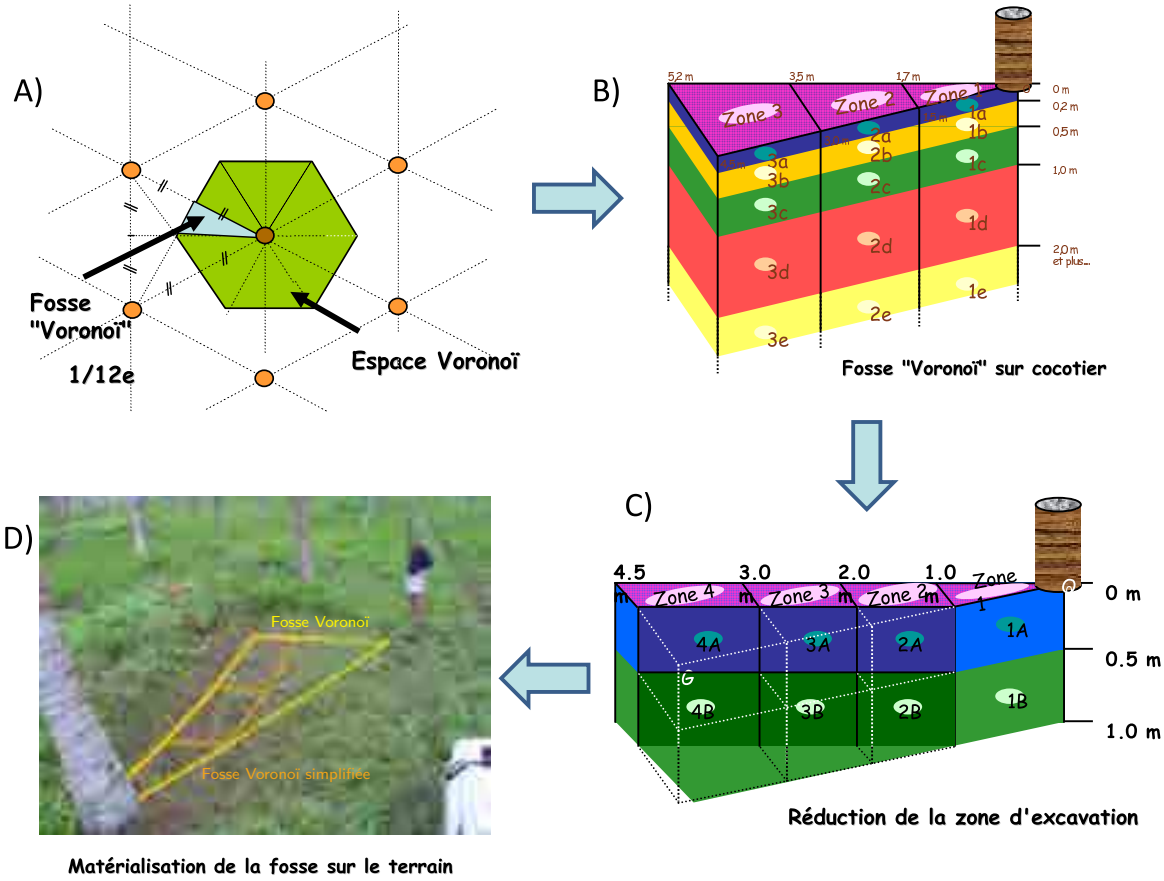


FIGURE 3.8 – *Exemple de tracé de l'espace de Voronoï pour l'échantillonnage des racines dans une cocoteraie au Vanuatu (photo: C. Jourdan). (A) Tracé de l'espace de Voronoï et choix de travailler sur 1/12^{e} de cet espace; (B) coupe schématique des fosses réalisées; (C) simplification du protocole compte tenu de la variabilité observée sur un premier échantillonnage; (D) matérialisation des tracés sur un cas réel.*

pour les horizons en profondeur (10–50 et 50–100 cm). Pour obtenir cette précision sur 1 m de profondeur, cela nécessite 36 hommes·jours de travail. En revanche, si la précision souhaitée est de 30 % seulement, le temps nécessaire à l'échantillonnage est diminué de 75 %. Cet exemple illustre parfaitement l'utilité de faire un pré-échantillonnage (cf. chapitre 2) afin d'évaluer la variabilité dans l'écosystème étudié et ensuite d'adapter le protocole en fonction des objectifs et de la précision souhaitée.

Une fois le sol prélevé avec les racines, le tri peut se faire au laboratoire pour les carottes qui contiennent les racines fines. En revanche, pour les racines moyennes et grosses, il est nécessaire de faire ce tri sur le terrain compte tenu du volume et du poids de terre excavée. Au laboratoire, le sol est lavé en faisant attention à mettre un filtre, les racines pouvant ensuite être récupérées par flottaison et/ou par tri sur tamis. *In situ*, le sol est trié manuellement sur des bâches. Pour les grosses racines et les racines moyennes, il est possible de se servir d'un couteau à air qui permet d'excaver complètement le système racinaire tout en conservant son architecture. Cette méthode, intéressante sur sols sableux, permet de répondre à deux objectifs (biomasse et architecture), mais nécessite cependant d'avoir sur le chantier un compresseur mobile (photo 3.12).

Une fois les racines triées et récoltées, elles sont mises à sécher à l'étuve en suivant les mêmes principes que pour la biomasse aérienne. Les racines fines vont nécessiter globalement

Photo 3.11 – *À gauche, superposition des méthodes d'échantillonnage (carottes, excavations par cubes, excavation partielle du Voronoï, excavation totale du Voronoï), d'après Levillain et al. (2011) (photo: C. Jourdan). À droite, excavation manuelle des grosses racines dans une plantation d'hévéa en Thaïlande (photo: L. Saint-André).*

Photo 3.12 – *Déploiement d'un couteau à air au Congo pour l'extraction des systèmes racinaires (grosses et moyennes racines) des eucalyptus. À gauche, l'opérateur avec les équipements de sécurité (poussière, bruit); au milieu, le compresseur et un zoom sur le cadran indiquant la pression d'air (environ 8 bar); à droite, le résultat (photos: C. Jourdan).*

le même temps de séchage que les feuilles, tandis que les racines moyennes et grosses vont plutôt nécessiter des temps équivalents à ceux des branches.

Pour la souche, il faudra prélever un sous-échantillon, de préférence vertical pour mieux tenir compte des variations de densité du bois dans cette partie de l'arbre, et suivre les mêmes procédures que pour les rondelles de tronc.

Les calculs à effectuer sont ensuite les mêmes que pour la biomasse aérienne.

3.5 Recommandation pour le matériel à utiliser

3.5.1 Matériel lourd et véhicules

- Voitures, camions, remorque: transport des personnes, du matériel et des échantillons depuis/jusqu'au laboratoire.
- Quad (si possible): transport du matériel encombrant et des échantillons sur le terrain.
- Pelleteuse pour la pesée des fagots.

3.5.2 Matériel général

- Boîte à outils avec outillage de base.
- Caisses en plastique (stockage et transport du matériel – environ 10).
- Sacs en plastique (compter un ou deux grand sac par arbre) pour regrouper les échantillons d'un arbre et éviter les pertes d'humidité. Sacs en papier (compter un sac en papier par compartiment et par arbre) pour mettre les échantillons juste après le prélèvement. Idéalement, les sacs sont déjà marqués avec les numéros d'arbre et de compartiment (ce qui permet de gagner du temps sur le terrain). Mais prévoir des lots non marqués et un feutre noir à encre indélébile pour pallier d'éventuelles erreurs ou permettre de prélever des échantillons supplémentaires.
- Grandes bâches pour le houppier (soit découpées pour le fagotage des branches, soit étalées au sol pour récupérer les feuilles arrachées des arbres).
- Étiquettes (à agrafer sur les rondelles), agrafes et agrafeuse; ou crayon fuschine (si les échantillons doivent ensuite être conservés pour mesurer la minéralomasse, il faut éviter la fuschine) (photo 3.13).
- Cutters, machettes, haches, cisailles, scies (photo 3.13).
- Cadres pour le façonnage des fagots (photo 3.14) ou alors des poubelles de différentes tailles.
- Tronçonneuses (idéalement une tronçonneuse adaptée à l'abattage des arbres et une autre, plus petite et maniable pour découper les branches – photo 3.13).
- Ficelles résistantes pour faire les fagots de branches (réutilisées au fur et à mesure de la campagne, il est donc nécessaire de faire des nœuds réversibles).
- Grands sacs très résistants (type big-bag à grain ou sable – photo 3.15) pour le transport des rondelles et des échantillons du terrain aux véhicules (si ceux-ci sont éloignés du chantier).

3.5.3 Saisie des données de terrain

- Pocket PC (avec chargeur de batterie et câbles) ou fiches terrain sur papier pluie ou papier cartonné, si possible reliées en cahiers avec couverture plastifiée dessus-dessous.
- Flore ou clé de détermination des essences pour les chantiers en forêt tropicale humide.
- Crayons à papier 2B, gommes, taille crayons.

Photo 3.13 – *Matériel de terrain. À gauche, matériel pour découper les aliquotes et étiquetage; au milieu, exemples de gabarits pour les découpes de branches; à droite tronçonneuse et équipement de sécurité (photo: A. Genet).*

Photo 3.14 – *Façonnage des fagots. À gauche, cadre en fer, bâche plastique et ficelle pour la mise en fagot des branches (photo: A. Genet); au milieu opération de mise en fagot sur le terrain (photo: M. Rivoire); à droite, le fagot terminé prêt à peser (photo: M. Rivoire).*

Photo 3.15 – *Transports des rondelles et des aliquotes dans un « big-bag » à sable ou à grain (photo: J.-F. Picard).*

- Balances de terrain ou pesons (avec 2 batteries et chargeur) pour la pesée des échantillons (précision de 1 g), l'idéal étant d'avoir une gamme complète adaptée aux poids des échantillons (un billon de 1 ou 2 m pouvant faire plusieurs centaines de kg, tandis que les rondelles de bois vont de quelques dizaines de g à plusieurs dizaines de kg). L'utilisation d'une pelleteuse permet de faciliter la pesée sur le terrain des gros billons. Il faut alors prévoir des sangles pour attacher deux pesons sur le godet et des griffes autobloquantes pour accrocher le billon.
- Décamètre pour la mesure des hauteurs le long du tronc (profils de tiges).
- Compas forestier et ruban forestier pour les mesures de circonférence.
- Bombe de marquage forestier (marquage des arbres sur pied et marquage de la tige principale dans les houppiers très développés).
- Griffe forestière pour indiquer les endroits où les rondelles seront prélevées (ou marquage avec la bombe de peinture).

3.5.4 Matériel au laboratoire

- Étuves.
- Éprouvette d'un demi litre minimum.
- Couteau à écorcer.
- Sécateur.
- Balance de capacité 2 à 2000 g (précision de 0,1 g à 1 g).
- Scie à ruban.

3.6 Recommandation pour la composition des équipes de terrain

Équipe abattage: un bûcheron, deux aides bûcherons, deux personnes (déblayage de la zone avant abattage). Compter deux jours pour l'abattage de 40 arbres pour des circonférences allant de 31 à 290 cm (moyenne 140 cm). Il est possible d'abattre tous les arbres au début de la campagne pour ensuite libérer cette équipe qui s'occupera du billonnage des arbres. Pour qu'ils puissent être opérationnels sans temps morts, il faut qu'une dizaine d'arbres (d'une vingtaine de mètre de haut — soit environ 10 à 20 rondelles par arbre) soient prêts à être débités en permanence sur le chantier.

Équipe profils de tige: deux personnes (un pointeur et un mesureur). Elle intervient dès que l'arbre est abattu et suit l'équipe de bûcherons. En général ces deux équipes sont assez synchrones. L'équipe de profils de tige n'est jamais en situation d'attente par rapport à celle des bûcherons, sauf en cas très rares de soucis à l'abattage (par exemple pour les grosses tiges, ou alors pour des tiges encrouées dans d'autres arbres et qu'il faut dégager).

Équipe ébranchage: trois personnes par unité de traitement. Chaque unité comprend un tronçonneur (avec la tronçonneuse maniable) et deux fagoteurs. Ces équipes peuvent être doublées ou triplées selon la dimension du houppier à traiter. Un ordre de grandeur est le suivant: au-dessus de 200 cm de circonférence à 1,30 m, il faut trois unités; entre 80 et 200 cm de circonférence, il faut deux unités; et en dessous de 80 cm de circonférence, une unité suffit. Ces ordres de grandeur tiennent compte du fait que les unité ne doivent pas se perturber l'une l'autre dans leur travail. À trois unités sur le même arbre, il en faut une en bas de l'arbre qui remonte le long de la tige, tandis que les deux autres sont positionnées de part et d'autre de l'axe principal et partent du milieu du houppier environ pour remonter vers le haut de l'arbre.

Équipe billonnage: elle comprend un bûcheron (découpe des tiges) et une personne (étiquetage des rondelles). En général, c'est l'équipe d'abattage qui prend le relais. Une fois tous les arbres abattus, le bûcheron vient dans cette équipe de billonnage et les aides bûcherons viennent renforcer les unités d'ébranchage.

Équipe pesage: trois personnes (conducteur pour la pelleteuse et deux autres personnes pour la manutention des billons et des fagots.).

Équipe échantillonnage des branches: une à deux personnes.

4

Saisie et mise en forme des données

Après la phase de mesures sur le terrain et avant la phase d'analyse des données se situe la phase de structuration des données, qui inclut la saisie des données, l'apurement des données et la mise en forme des données.

4.1 Saisie des données

La saisie consiste à transférer dans un fichier informatique les données présentes sur les fiches de terrain. Il faut au préalable avoir choisi un logiciel. Pour un petit jeu de données, un tableur comme Microsoft Excel ou OpenOffice Calc pourra faire l'affaire. Pour des campagnes de mesure plus conséquentes, il faudra utiliser un logiciel de gestion de base de données, comme par exemple Microsoft Access ou MySQL (www.mysql.com).

4.1.1 Les erreurs de saisie

La saisie doit être faite aussi soigneusement que possible pour limiter les erreurs de saisie. Une façon de minimiser les erreurs de saisie est de faire une double saisie: un premier opérateur fait la saisie; un deuxième opérateur (si possible différent du premier) refait la saisie de manière totalement indépendante du premier. Il suffit alors de comparer les deux fichiers de saisie pour déceler les erreurs de saisie. Comme il est peu probable que deux opérateurs fassent la même erreur de saisie, cette méthode assure une bonne qualité de la saisie. En revanche elle est coûteuse en temps et fastidieuse.

Lors de la saisie des données, il faut également s'attacher à un certain nombre de détails qui ont leur importance. Il faut d'abord bien différencier les nombres des chaînes de caractères. Pour le logiciel de statistiques qui traitera ensuite les données, un nombre n'a pas le même rôle qu'une chaîne de caractères, donc il est important de faire cette distinction dès la saisie. Un nombre sera interprété comme la valeur d'une variable numérique, tandis qu'une chaîne de caractère sera interprétée comme la modalité d'une variable qualitative. La différence entre les deux est en général bien claire, mais pas toujours. Prenons le cas des latitudes et des longitudes. Si l'on souhaite calculer la corrélation entre la latitude ou la longitude et une autre variable (pour identifier un gradient nord-sud ou est-ouest), il faut faire en sorte que le logiciel perçoive les coordonnées géographiques comme des nombres. Il

ne faut donc pas saisir les coordonnées géographiques comme par exemple « 7°28'55,1" » ou « 13°41'25,9" ». Ces coordonnées seraient interprétées comme des variables qualitatives, et aucun calcul ne serait alors possible. Une solution possible consiste à convertir les coordonnées géographiques en valeurs décimales. Une autre solution consiste à saisir les coordonnées géographiques sur trois colonnes (une colonne pour les degrés, une autre pour les minutes, une troisième pour les secondes).

Lorsque l'on saisit des variables qualitatives, il faut éviter de saisir des chaînes de caractères de grande longueur car cela multiplie des risques d'erreur de saisie. Il faut mieux saisir un code abrégé et préciser dans la méta-information (cf. ci-dessous) la signification de ce code.

Un autre détail qui a son importance est le symbole décimal utilisé. Pratiquement tous les logiciels de statistique permettent de basculer de la virgule (symbole utilisé en français) au point (symbole utilisé en anglais), donc l'usage de l'un ou de l'autre est indifférent. En revanche, une fois que l'on a choisi la virgule ou le point comme symbole décimal, il faut s'y tenir dans l'ensemble de la saisie. Si on utilise tantôt l'un, tantôt l'autre, une partie des données normalement numériques seront interprétés par le logiciel de statistiques comme des chaînes de caractères.

4.1.2 La méta-information

Au cours de la saisie des données, il faut songer à la méta-information. La méta-information est l'information qui accompagne les données, sans être elle-même une donnée mesurée. La méta-information donnera par exemple la date à laquelle les mesures ont été effectuées, et par qui. Si des codes sont utilisés dans la saisie, la méta-information précisera la signification de ces codes. Il n'est pas rare par exemple que les noms des espèces soient saisis en abrégé. Un code espèce comme ANO dans une savane sèche d'Afrique de l'ouest, par exemple, est ambiguë: il peut s'agir d'*Annona senegalensis* ou d'*Anogeissus leiocarpus*. La méta-information est là pour lever cette ambiguïté. La méta-information doit aussi préciser la nature des variables mesurées. Par exemple si on mesure les circonférences d'arbres, noter « circonférence » dans le tableau de données est en soi insuffisant. La méta-information doit préciser à quelle hauteur la circonférence a été mesurée (à la base, à 20 cm, à 1,30 m...) et, point extrêmement important, dans quelle unité la circonférence est exprimée (en cm, en dm...). Insistons sur le fait que l'unité de mesure de chacune des variables doit être précisée dans la méta-information. Trop souvent on a des tableaux de données qui ne précisent pas les unités dans lesquelles ces données sont exprimées, ce qui donne alors lieu à un jeu de devinettes risqué.

Pour la personne qui a conçu le dispositif de mesure et encadré les mesures, les informations contenues dans la méta-information sont souvent si évidentes qu'elle ne voit pas la nécessité de passer du temps à les renseigner. Il faut cependant s'imaginer une personne étrangère à la mesure retrouvant le jeu de données dix ans plus tard. Si la méta-information a été bien faite, cette personne doit pouvoir travailler sur le jeu de données comme si elle l'avait constitué elle-même.

4.1.3 Niveaux emboîtés

Les données sont saisies dans des tableaux, avec une ligne par individu. Si les données comportent plusieurs niveaux emboîtés, il doit y avoir autant de tableaux qu'il y a de niveaux. Imaginons que l'on cherche à construire un tarif pour des formations de taillis sous futaie à l'échelle d'une région. Pour les individus multicaules de l'échantillon, chaque tige

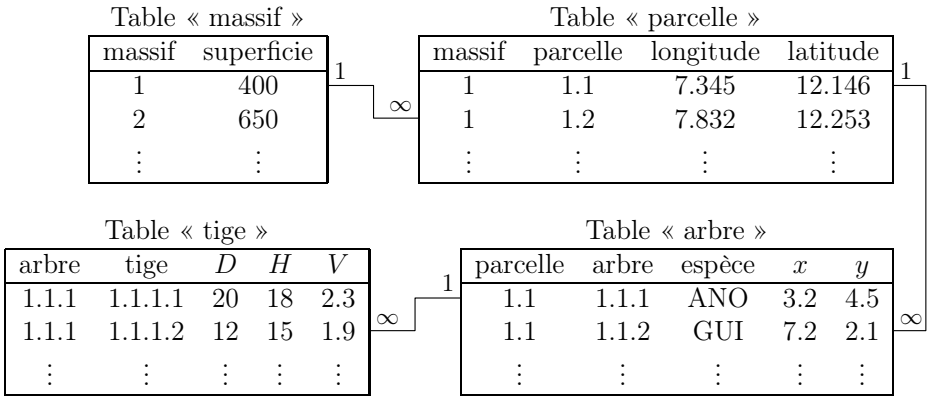


Figure 4.1 – *Exemple de quatre tableaux de données pour quatre niveaux emboîtés.*

Table 4.1 – *Saisie des données avec quatre niveaux emboîtés dans un seul tableau.*

massif	superficie	parcelle	longitude	latitude	arbre	espèce	x	y	tige	D	H	V
1	400	1.1	7.345	12.146	1.1.1	ANO	3.2	4.5	1.1.1.1	20	18	2.3
1	40*1*	1.1	7.345	12.146	1.1.1	ANO	3.2	4.5	1.1.1.2	12	15	1.9
1	400	1.1	7.345	12.146	1.1.2	GUI	7.2	2.1	⋮	⋮	⋮	⋮
1	400	1.2	7.832	12.253	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2	650	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

est cubée séparément. Les individus sont sélectionnés au sein de parcelles, elles-mêmes sélectionnées au sein de massifs forestiers répartis dans la zone d'étude. Il y a dans ce cas quatre niveaux emboîtés: le massif, qui comporte plusieurs parcelles; la parcelle, qui comporte plusieurs arbres; l'arbre, qui comporte plusieurs tiges; et enfin la tige. Il devra donc y avoir dans ce cas quatre tableaux de données (figure 4.1). Chaque tableau renseignera les variables décrivant les individus du niveau correspondant, avec une ligne de tableau par individu. Par exemple, le premier tableau donnera la superficie de chacun des massifs forestiers. Le deuxième tableau donnera les coordonnées géographiques de chacune des parcelles. Le troisième tableau donnera l'espèce et les coordonnées au sein de la parcelle de chaque arbre. Enfin le quatrième tableau donnera le volume et la taille de chaque tige. À chaque ligne d'un tableau correspond plusieurs lignes dans le tableau du niveau inférieur. Un identifiant doit permettre de faire la correspondance entre les lignes des différents tableaux. Ainsi le numéro du massif sera répété dans les tableaux « massif » et « parcelle », le numéro de la parcelle sera répété dans les tableaux « parcelle » et « arbre », et le numéro de l'arbre sera répété dans les tableaux « arbre » et « tige » (figure 4.1).

Cette structuration des données minimise la répétition de l'information, donc les erreurs de saisie. Une alternative consisterait à saisir toutes les données dans le même tableau, comme indiqué pour l'exemple précédent dans le tableau 4.1. Cette alternative n'est pas recommandable car elle répète inutilement de l'information, donc multiplie les risques d'erreur de saisie. Par exemple dans le tableau 4.1 nous avons volontairement introduit une erreur de saisie dans la deuxième ligne du tableau, où la superficie du massif 1, normalement égale à 400 ha, est ici de 401 ha. En répétant inutilement de l'information, on se retrouve avec

une multiplication de ce genre d'incohérences qu'il faut ensuite corriger.

Une bonne façon de résoudre ces problèmes de niveaux emboîtés est de construire une base de données relationnelle. Les bases de données relationnelles sont précisément faites pour gérer différents tableaux avec des liens entre eux. Elles permettent d'éliminer toute incohérence comme celle illustrée dans le tableau 4.1 en vérifiant systématiquement l'intégrité des relations entre les tableaux. Cependant la construction d'une base de données relationnelles est une étape technique qui nécessite éventuellement le recours à une personne compétente dans ce domaine.

Pour récapituler, au cours de la saisie des données, il est préférable de:
- éviter la répétition de l'information,
- favoriser les bases de données relationnelles,
- donner des informations supplémentaires (méta-information),
- faire attention aux unités,
- faire la différence entre l'information qualitative et l'information quantitative,
- vérifier les données,
- réduire ou corriger les données absentes.

4.2 Apurement des données

L'apurement nécessite un aller-retour entre les fiches de mesure et le logiciel de statistiques (ou éventuellement un logiciel d'apurement des données spécialement conçu pour cela). Cette étape vise à éliminer toute incohérence dans les données. Éventuellement, si le dispositif de mesure est encore en place, elle nécessitera de refaire certaines mesures. L'apurement permettra d'éliminer:
- les données aberrantes. Par exemple un arbre de 50 mètres de diamètre.
- les données incohérentes. Par exemple un arbre ayant une biomasse du tronc de 755 kg et une biomasse totale de 440 kg, ou alors un arbre de 5 cm de diamètre et de 40 m de haut.
- les fausses modalités des variables qualitatives. Par exemple un logiciel faisant une différence entre majuscules et minuscules interprétera « oui » et « Oui » comme deux modalités différentes, alors qu'il s'agit de la même modalité.

La difficulté pour repérer les données aberrantes vient du choix du seuil entre ce qui est une mesure normale et ce qui est une mesure aberrante. Il arrive que les données aberrantes soit le résultat d'un changement d'unité au cours de la saisie. Si la fiche de terrain mentionne 1,2 kg puis 900 g pour des mesures de biomasse foliaire, il faudra prendre garde à saisir 1,2 et 0,9 (en kg), ou alors 1200 et 900 (en g), mais en aucun cas il ne faudra saisir 1,2 et 900. Les données incohérentes sont plus difficiles à repérer car elles nécessitent de confronter plusieurs variables entre elles. Dans l'exemple précédent, un arbre avec une biomasse du tronc de 755 kg n'a rien d'anormal et un arbre avec une biomasse totale de 440 kg n'a rien d'anormal non plus, mais bien entendu les deux mesures ne peuvent pas être correctes simultanément pour le même arbre. De même un arbre avec un diamètre de 5 cm n'a rien d'anormal, pas plus qu'un arbre de 40 m de haut, mais un arbre avec un diamètre de 5 cm et une hauteur de 40 m est anormal.

Le dépistage des données aberrantes et incohérentes pourra être fait à l'aide de statistiques descriptives et de graphiques deux à deux: l'examen des moyennes, quantiles, valeurs maximales et minimales permet le plus souvent de détecter les données aberrantes; les graphiques des variables deux à deux permet de détecter les données incohérentes. Dans l'exemple précédent, on pourra faire le graphique de la biomasse totale en fonction de la biomasse du tronc, et vérifier que tous les points se situent au-dessus de la droite $y = x$.

Les graphiques hauteur en fonction du diamètre, volume en fonction du diamètre, etc., permettent également de repérer des données anormales. Les modalités des variables qualitatives pourront être inspectées en faisant le décompte du nombre d'observations par modalité. Deux variables qualitatives pourront être croisées en construisant le tableau de contingence correspondant. On s'assurera lors de cette inspection des données que le logiciel de statistiques a bien interprété les données numériques comme des données numériques, et les données qualitatives comme des données qualitatives.

Les fausses modalités résultent souvent d'un manque de rigueur dans la saisie. Les fautes d'orthographe involontaires, fréquentes par exemple lorsque les noms des essences sont saisies en plein, génèrent de fausses modalités. Ces fausses modalités peuvent être très ambiguës et difficiles à corriger. Prenons l'exemple d'un jeu de données sur des arbres de forêt tropicale humide d'Afrique centrale, où figurent, entre autres, les deux essences alombi (*Julbernardia seretii*) et ilomba (*Picnanthus angolensis*). Mettons que, par erreur, la fausse modalité « alomba » ait été saisie. Cette fausse modalité ne diffère des vraies modalités alombi et ilomba que d'une lettre. Comment savoir alors quelle est la vraie modalité ? Les accents sont souvent sources aussi de fausses modalités, selon que le texte ait été saisi avec ou sans accents. Prenons l'exemple de la saisie d'une couleur: pour la personne qui saisit les données, il est peut-être clair que « vert foncé » et « vert fonce » désignent la même modalité, mais le logiciel considérera cela comme deux modalités différentes. Les accords en genre peuvent aussi poser problème. Les modalités « feuille vert clair » et « feuille verte clair » seront comprises par le logiciel comme deux modalités différentes. Une fausse modalité couramment trouvée est celle liée à l'espace. La modalité « vert » (sans espace) et la modalité « vert␣ » (avec espace, ici matérialisé par le symbole « ␣ ») sera comprise par le logiciel comme deux modalités différentes. Cette fausse modalité est particulièrement déroutante car l'espace n'est pas visible à l'écran, de sorte que l'utilisateur a vraiment l'impression qu'il s'agit de la même modalité. Tous les caractères invisibles (retour chariot, tabulation, etc.), ou les caractères qui apparaissent de la même façon à l'écran bien qu'ayant des codes ASCII différents, peuvent générer le même genre d'erreur déroutante.

Les fausses modalités peuvent être évitées en utilisant des masques de saisie, qui ne laissent le choix pour la saisie des variables qualitatives que parmi une liste de modalités admissibles. L'utilisation de scripts automatiques d'apurement des données, qui vont supprimer les espaces intempestifs, vérifier les accents ou la casse des lettres, vérifier que les variables qualitatives prennent leur valeur parmi une liste de modalités admissibles, est une nécessité pour les gros jeux de données.

4.3 Mise en forme des données

La mise en forme consiste à organiser les données dans un format permettant l'exécution des calculs nécessaires à la construction du tarif. Typiquement il s'agit d'un tableau comportant une ligne par individu statistique (un arbre pour un tarif individuel, une parcelle pour un tarif peuplement) et autant de colonnes que de variables descriptives (aussi bien variables à prédire: biomasse, volume..., que variables explicatives: diamètre, hauteur...). Cette phase de mise en forme peut requérir des manipulations assez avancées des données. Dans certains cas il faudra agréger les données d'un niveau de description à un autre. Par exemple, si on veut construire un tarif individuel pour un taillis et que les mesures portent sur les tiges, il faudra agréger les données relatives aux tiges d'une même souche: additionner les volumes et les masses, calculer le diamètre équivalent (c'est-à-dire la moyenne quadratique) de la souche à partir des diamètres de ses tiges. Un autre exemple est la construction d'un tarif de peuplement à partir de mesures sur des arbres. Il faudra alors agréger les

données relatives aux arbres en données caractérisant le peuplement (volume à l'hectare, hauteur dominante, etc.). Dans d'autres cas au contraire, il faut diviser le jeu de données. Par exemple on a cubé des arbres au hasard dans un peuplement pluri-spécifique, et on veut construire un tarif séparé pour les cinq essences dominantes: il faut alors diviser le jeu de données en fonction de l'essence des arbres.

La mise en forme des données sera d'autant plus aisée que la saisie des données aura été effectuée dans un format adéquat. Les bases de données relationnelles ont l'avantage d'offrir un langage de requêtes qui permet de construire facilement de tels tableaux synthétiques. Dans le tableur Microsoft Excel, la notion de « tableau croisé dynamique » sera utilement mise à profit pour mettre en forme les données.

Jeu de données du fil rouge

Pour illustrer ce manuel, nous utiliserons le jeu de données collectées au Ghana par Henry *et al.* (2010). Ce jeu de données donne la biomasse sèche de 42 arbres appartenant à 16 espèces d'une forêt tropicale humide. Pour chaque arbre, on a mesuré son diamètre à hauteur de poitrine, sa hauteur, le diamètre de sa couronne, la densité moyenne de son bois, son volume, et sa biomasse sèche dans cinq compartiments: biomasse des branches, biomasse des feuilles, biomasse du tronc, biomasse des contreforts, et biomasse totale.

Le tableau 4.2 présente les données de Henry *et al.* (2010) telles qu'elles devraient être mises en forme dans un tableur. Le tableau des données se présente dans le tableur sous la forme d'un rectangle de données; il ne doit y avoir ni ligne blanche, ni colonne blanche, ni aucune mise en page qui dévie de cette présentation matricielle des données. Les effets décoratifs, les indentations, les cellules laissées vides pour « aérer » la présentation doivent être proscrites, car le logiciel de statistiques ne sera pas à même de lire un jeu de données qui dévie du format matriciel. Les libellés des colonnes sont réduits à des mots courts, voire des abréviations. L'information sur la signification de ces variables et leur unité est saisie à part, dans la méta-information.

Si des informations devaient être saisies sur les espèces, celles-ci seraient saisies dans un second tableau, puisqu'il y a deux niveaux emboîtés: le niveau espèce, avec plusieurs arbres par espèce; et le niveau arbre, emboîté dans le niveau espèce. Ainsi, si on voulait saisir la guilde écologique et le nom vernaculaire des espèces, on obtiendrait un second tableau 4.3 propre à l'espèce, le nom scientifique de l'espèce étant ici l'identifiant permettant de faire le lien entre le tableau 4.2 et le tableau 4.3.

Lecture des données. Supposons que les données, mises en forme au format matriciel, sont enregistrées dans un fichier Excel `Henry_et_al2010.xls`, dont la première feuille intitulée `biomasse` contient le tableau 4.2. Dans le logiciel R, la lecture des données s'effectue à l'aide des commandes:

```
library(RODBC)
ch <- odbcConnectExcel("Henry_et_al2010.xls")
dat <- sqlFetch(ch,"biomasse")
odbcClose(ch)
```

Les données sont alors stockées dans l'objet `dat`.

Apurement des données. Quelques vérifications peuvent être faites pour vérifier la qualité des données. Dans R, la commande `summary` donne les statistiques descriptives de base des variables d'un tableau de données:

```
summary(dat)
```

En particulier pour le diamètre le résultat est:

```
      dbh
 Min.   :  2.60
 1st Qu.: 15.03
 Median : 59.25
 Mean   : 58.59
 3rd Qu.: 89.75
 Max.   :180.00
```

Ainsi le diamètre des arbres mesurés va de 2,6 cm à 180 cm, avec une moyenne de 58,59 cm et un diamètre médian de 59,25 cm. Les statistiques descriptives de base pour la biomasse sèche totale sont:

```
      Btot
 Min.   : 0.0000
 1st Qu.: 0.1375
 Median : 3.1500
 Mean   : 6.8155
 3rd Qu.: 9.6075
 Max.   :70.2400
```

La biomasse sèche totale du plus gros arbre est de 70,24 tonnes. La biomasse sèche du plus petit arbre prend une valeur nulle dans le jeu de données. Les biomasses étant exprimées en tonnes avec deux chiffres significatifs, cette valeur nulle n'est pas une donnée aberrante mais signifie simplement que la biomasse sèche de cet arbre est inférieure à 0,01 tonnes = 10 kg. Cette valeur nulle posera cependant problème par la suite quand on voudra transformer les données par le logarithme.

On peut enfin s'assurer que la biomasse sèche totale est bien la somme des biomasses des quatre autres compartiments:

```
max(abs(dat$Btot-rowSums(dat[,c("Bbran","Bfol","Btronc","Bctf")])))
```

La plus grande différence en valeur absolue est égale à 0,01 tonnes, ce qui correspond bien à la précision des données (deux chiffres significatifs). Il n'y a donc pas d'incohérence à ce niveau dans le jeu de données.

TABLE 4.2 – *Données de biomasse d'arbres de Henry et al. (2010) au Ghana.* `dbh` *est le diamètre en cm,* `haut` *est la hauteur en m,* `houp` *est le diamètre de la couronne en m,* `dens` *est la densité moyenne du bois en* $g\,cm^{-3}$, `volume` *est le volume en* m^3, `Bbran` *est la biomasse sèche des branches en tonnes,* `Bfol` *est la biomasse foliaire sèche en tonnes,* `Btronc` *est la biomasse sèche du tronc en tonnes,* `Bctf` *est la biomasse sèches des contreforts en tonnes, et* `Btot` *est la biomasse sèche totale en tonnes.*

espèce	dbh	haut	houp	dens	volume	Bbran	Bfol	Btronc	Bctf	Btot
Heritiera utilis	7,3	5,1	3,7	0,58	0,03	0,02	0	0	0	0,02
Heritiera utilis	12,4	12	5	0,62	0,11	0,02	0	0,05	0	0,07
Heritiera utilis	31	22	9	0,61	1,34	0,1	0,01	0,71	0,02	0,83
Heritiera utilis	32,5	27,5	7,1	0,61	1,12	0,07	0,01	0,61	0,01	0,7
Heritiera utilis	48,1	35,6	7,9	0,61	3,83	0,24	0,01	2,07	0,01	2,33
Heritiera utilis	56,5	35,1	8	0,6	5,43	0,85	0,03	2,28	0,14	3,31
Heritiera utilis	62	40,4	11,1	0,6	6,84	0,68	0,04	3,28	0,15	4,15
Heritiera utilis	71,9	42,3	20	0,6	9,84	1,34	0,05	4,43	0,11	5,93
Heritiera utilis	83	39,4	15,9	0,6	11,89	2,2	0,09	4,83	0,04	7,16
Heritiera utilis	100	50,5	19,1	0,58	31,71	8,71	0,11	8,39	1,4	18,61
Heritiera utilis	105	50,5	19,2	0,58	35,36	8,81	0,13	11,18	0,65	20,76
Heritiera utilis	6,5	8,1	1,5	0,78	0,01	0,01	0	0	0	0,01
Tieghemella heckelii	12	17	4,7	0,78	0,15	0,12	0,01	0	0	0,13
Tieghemella heckelii	73,5	45	11,1	0,66	11,08	1,27	0,04	5,91	0,14	7,36
Tieghemella heckelii	80,5	50,7	13	0,66	12,25	1,54	0,05	6,45	0,09	8,13
Tieghemella heckelii	93	45	17	0,66	17,79	3,66	0,06	7,8	0,21	11,73
Tieghemella heckelii	180	61	41	0,62	112,81	27,28	0,74	35,07	7,16	70,24
Piptadeniastrum africanum	70	39,7	10,5	0,58	10,98	2,97	0,06	3,29	0,07	6,39
Piptadeniastrum africanum	89	50	18,8	0,57	15,72	3,69	0,05	5,16	0,16	9,06
Piptadeniastrum africanum	90	50,2	16	0,57	22,34	5,73	0,38	6,23	0,74	13,08
Aubrevillea kerstingii	65	32,5	9	0,62	4,79	1,52	0,02	1,45	0	2,99
Afzelia bella	83,6	40	13,5	0,67	14,57	3,17	0,03	6	0,58	9,79
Cecropia peltata	7,8	2,3	2,5	0,17	0,07	0	0	0,01	0	0,01
Cecropia peltata	20,5	21,2	6,2	0,23	0,44	0,03	0	0,07	0	0,11
Cecropia peltata	29,3	22,5	8,9	0,27	1,11	0,13	0,01	0,16	0	0,31
Cecropia peltata	35,5	12	7,3	0,26	1,39	0,12	0,02	0,25	0	0,38
Ceiba pentandra	132	45	16	0,54	28,55	1,53	0,04	13,37	0,44	15,39
Ceiba pentandra	170	51	27,1	0,26	64,84	3,2	0,1	11,87	1,88	17,05
Nauclea diderrichii	2,6	4,9	8,4	0,76	0	0	0	0	0	0
Nauclea diderrichii	94,6	50,5	12	0,5	17,19	1,06	0,02	7,49	0,06	8,64
Nauclea diderrichii	110	58,8	14,1	0,4	28,71	3,47	0,06	7,9	0,07	11,49
Nauclea diderrichii	112	40	13,2	0,47	22,74	3,41	0,1	7,19	0,13	10,82
Daniellia thurifera	9	9,3	8	0,42	0,11	0,05	0,01	0	0	0,05
Guarea cedrata	12,8	13	3,1	0,62	0,12	0,08	0,01	0	0	0,08
Guarea cedrata	71,5	45,5	14	0,5	10,12	0,65	0,02	4,3	0,13	5,1
Strombosia glaucescens	7,6	11,3	3,9	0,66	0,07	0,05	0,01	0	0	0,05
Strombosia glaucescens	26,5	26	12,2	0,73	1,09	0,2	0,01	0,58	0	0,8
Garcinia epunctata	7,1	5,7	3,8	0,65	0,08	0,05	0,01	0	0	0,06
Drypetes chevalieri	13,2	15,7	5	0,65	0,22	0,15	0,02	0	0	0,16
Cola nitida	23,6	23,4	6,3	0,56	0,68	0,09	0,01	0,28	0	0,39
Nesogordonia papaverifera	24,3	30,2	6,5	0,69	0,73	0,12	0,01	0,36	0,02	0,51
Dialium aubrevilliei	98	43,7	98	0,65	18,49	2,55	0,05	9,07	0,4	12,07

Table 4.3 – *Données sur les espèces échantillonnées par Henry et al. (2010) au Ghana.*

guilde	espèce	vernaculaire
héliophile non pionnière	Heritiera utilis	Nyankom
héliophile non pionnière	Tieghemella heckelii	Makore
héliophile non pionnière	Piptadeniastrum africanum	Dahoma
héliophile non pionnière	Aubrevillea kerstingii	Dahomanua
héliophile non pionnière	Afzelia bella	Papao-nua
pionnière	Cecropia peltata	Odwuma
pionnière	Ceiba pentandra	Onyina
pionnière	Nauclea diderrichii	Kusia
pionnière	Daniellia thurifera	Sopi
tolérante à l'ombre	Guarea cedrata	Kwabohoro
tolérante à l'ombre	Strombosia glaucescens	Afena
tolérante à l'ombre	Garcinia epunctata	Nsokonua
tolérante à l'ombre	Drypetes chevalieri	Katreka
tolérante à l'ombre	Cola nitida	Bese
tolérante à l'ombre	Nesogordonia papaverifera	Danta
tolérante à l'ombre	Dialium aubrevilliei	Dua bankye

5

Exploration graphique des données

L'exploration graphique des données est la première étape de l'analyse des données. Elle consiste à étudier visuellement les relations entre variables afin de se faire une idée du type de modèle à ajuster. Concrètement, on projette sur un graphique un nuage de points dont les coordonnées correspondent à deux variables: variable explicative sur l'axe des x et variable à expliquer sur l'axe des y. Un graphique ne peut être construit que pour deux variables simultanément, au plus (les graphiques en trois dimensions sont en pratique inexploitables visuellement). Pour explorer graphiquement les relations entre p variables (avec $p > 1$), on fera donc $p(p-1)/2$ graphiques de variables deux à deux et/ou on cherchera à construire des variables explicatives synthétiques à partir de plusieurs variables explicatives (nous reviendrons sur ce point dans le § 5.1.1).

Nous supposons désormais que nous avons une variable à expliquer notée Y (le volume, la biomasse...) et p variables explicatives notées X_1, X_2, ..., X_p (le diamètre, la hauteur...). Le but de l'exploration graphique n'est pas de sélectionner parmi les p variables explicatives celles qui seront effectivement retenues pour le modèle: la sélection de variables suppose que l'on sache tester le caractère significatif d'une variable, ce qui renvoie à la phase suivante de l'ajustement du modèle. Les p variables explicatives sont donc considérées comme fixées, et on cherche la forme du modèle qui relie au mieux la variable Y aux variables X_1 à X_p. Un modèle se compose de deux termes: la moyenne et l'erreur (ou résidu). L'exploration graphique vise à préciser à la fois la forme de la relation moyenne et celle de l'erreur, mais sans se soucier de la valeur des paramètres du modèle (ce sera l'étape suivante d'ajustement du modèle). La relation moyenne peut être linéaire ou non linéaire, linéarisable ou non; l'erreur résiduelle peut être additive ou multiplicative, de variance constante (homoscédasticité) ou non (hétéroscédasticité). À titre d'exemple, la figure 5.1 présente quatre cas de figure selon que la relation est linéaire ou non et la variance des résidus constante ou non.

La phase d'exploration graphique des données est également nécessaire pour éviter de tomber dans les pièges de l'ajustement « à l'aveugle »: on peut en effet avoir l'impression que l'ajustement d'un modèle à des données est de bonne qualité alors qu'en fait il s'agit d'un artefact. Cela est illustré par la figure 5.2 dans le cas de la relation linéaire. Dans les quatre cas montrés dans cette figure, le R^2 de la régression linéaire de Y par rapport à X est élevé, alors qu'en réalité la relation linéaire $Y = a + bX + \varepsilon$ n'est pas adaptée aux données. Dans la figure 5.2A, le nuage de points se structure en trois sous-ensembles

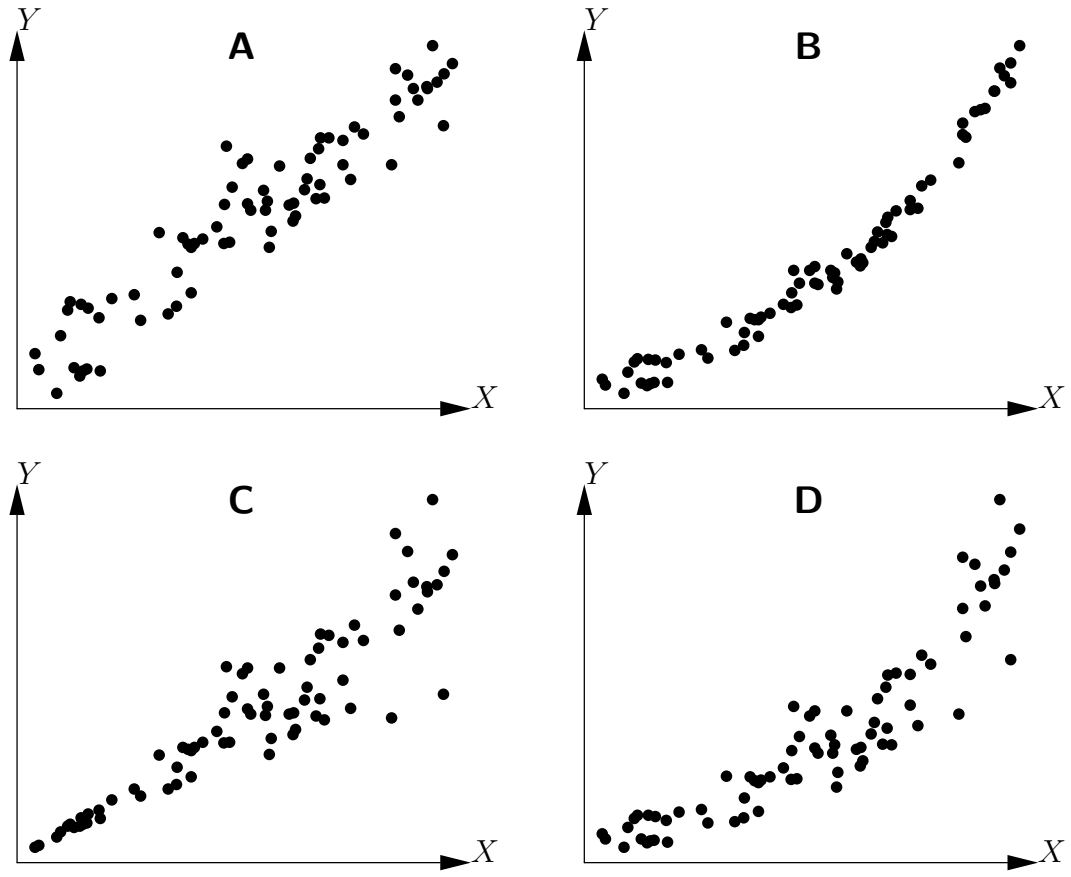


FIGURE 5.1 – *Exemple de relations entre deux variables X et Y: (A) relation linéaire et variance des résidus constante, (B) relation non linéaire et variance des résidus constante, (C) relation linéaire et variance des résidus non constante, (D) relation non linéaire et variance des résidus non constante.*

au sein desquels la relation entre Y et X est linéaire avec un coefficient de corrélation négatif. Mais ces trois sous-ensembles s'organisent le long d'une droite de pente positive, qui est la droite renvoyée par la régression linéaire. En 5.2B, le nuage de points, à l'exception d'une seule donnée excentrée (vraisemblablement une donnée aberrante), ne présente aucune relation entre Y et X. Mais la donnée excentrée suffit à laisser croire qu'il existe une relation positive entre Y et X. En 5.2C, la relation entre Y et X est parabolique. Enfin en 5.2D, le nuage de points, à l'exception d'une seule donnée excentrée, se structure le long d'une droite de pente positive. Dans ce cas une relation linéaire entre Y et X serait adaptée pour décrire les données amputées de la donnée excentrée. Cette donnée excentrée fait décroître artificiellement la valeur du R^2 (par opposition au graphique 5.2B où la donnée excentrée gonfle artificiellement le R^2).

Comme son nom l'indique, la phase d'exploration graphique procède plus de l'exploration que d'une méthode systématique. Même si un certain nombre de conseils peuvent être donnés pour arriver à déceler le bon modèle, elle requiert de l'expérience et de l'intuition.

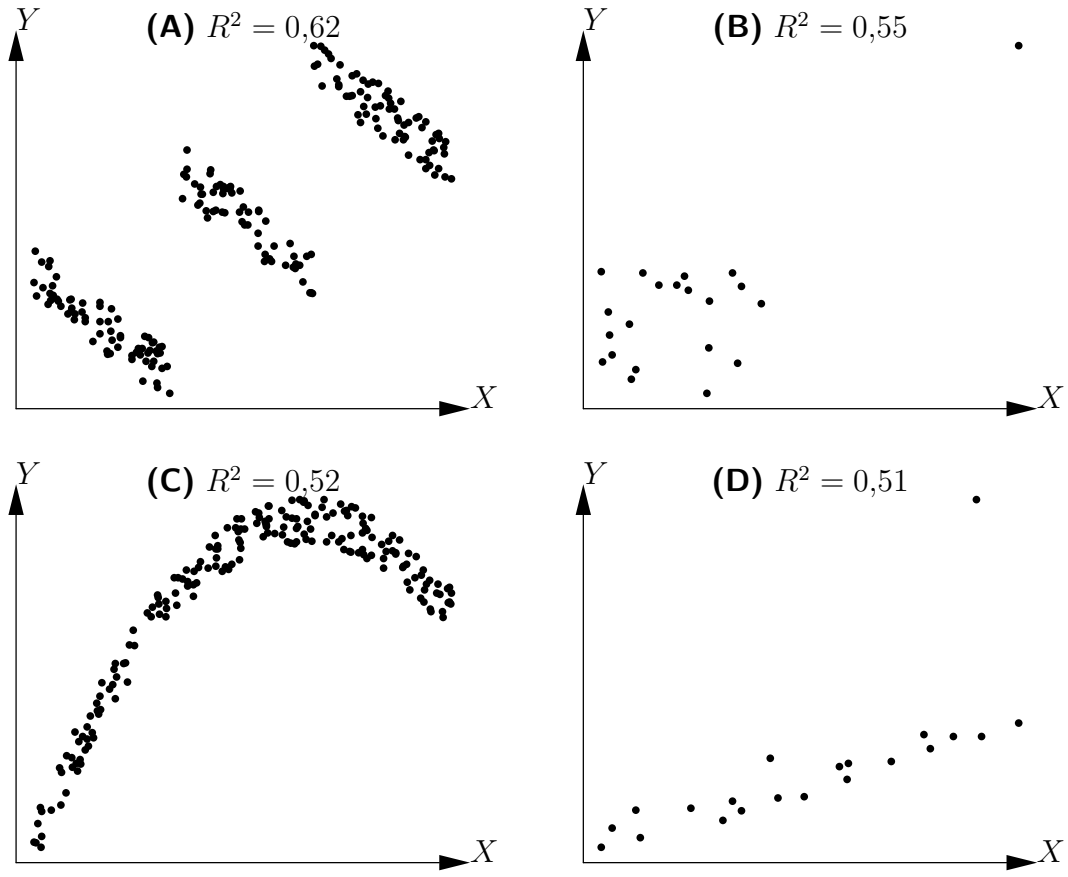


Figure 5.2 – *Coefficients de détermination (R^2) de régressions linéaires réalisées sur des nuages de points ne présentant pas de relations linéaires.*

5.1 Exploration de la relation moyenne

On s'intéresse dans cette section à la façon graphique de déterminer la nature de la relation moyenne entre deux variables X et Y, c'est-à-dire à trouver la forme de la fonction f (si elle existe !) telle que $\mathrm{E}(Y) = f(X)$. Lorsqu'il n'y a qu'une seule variable explicative X, l'exploration graphique consiste à tracer le nuage des points de Y en fonction de X.

②

Exploration de la relation biomasse–diamètre

Pour voir la forme de la relation entre la biomasse sèche totale et le diamètre des arbres, on trace le nuage de points de la biomasse en fonction du diamètre. Le jeu de données étant lu (cf. fil rouge n° 1), la commande pour tracer le nuage de points est:

```
plot(dat$dbh,dat$Btot,xlab="Diamètre (cm)",ylab="Biomasse (t)")
```

Le nuage de points résultant est montré dans la figure 5.3. Ce nuage de points est du même type que le graphe de la figure 5.1D: la relation entre la biomasse et le diamètre n'est pas linéaire et la variance de la biomasse augmente lorsque le diamètre augmente.

La méthode graphique du nuage de points ne pouvant être utilisée que pour une seule

variable explicative, on tâchera de se ramener à ce cas quand il y a plusieurs variables explicatives. Explicitons tout d'abord ce dernier point.

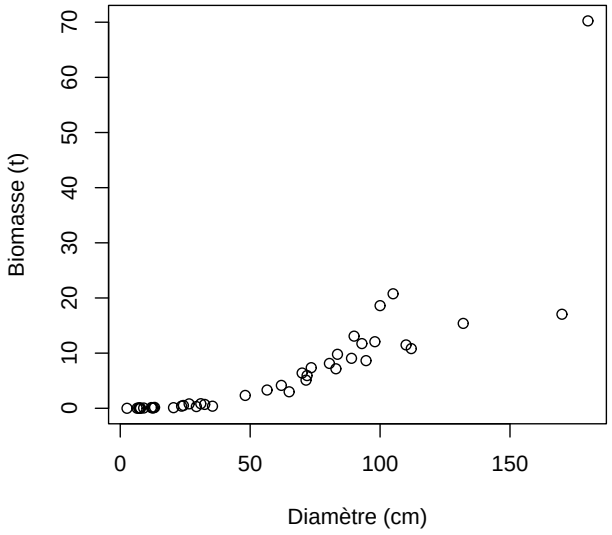


FIGURE 5.3 – *Nuage de points de la biomasse sèche totale (tonnes) en fonction du diamètre à hauteur de poitrine (cm) pour les 42 arbres mesurés au Ghana par Henry et al. (2010).*

5.1.1 Quand il y a plus d'une variable explicative

La première chose est de voir s'il n'est pas possible de former à partir de plusieurs variables explicatives une seule variable explicative synthétique. Par exemple si on cherche à prédire le volume du tronc à partir de son diamètre D et de sa hauteur H, on peut parier que la nouvelle variable D^2H sera un prédicteur efficace. Dans ce cas on a formé à partir des deux variables explicatives D et H une nouvelle (et unique !) variable explicative D^2H. Par exemple Louppe *et al.* (1994) ont créé le tarif de cubage individuel suivant pour *Afzelia africana* dans la forêt classée de Badénou en Côte d'Ivoire:

$$V = -0{,}0019 + 0{,}04846C^2H$$

où V est le volume total en m^3, C la circonférence à 1,30 m en m et H la hauteur en m. Même s'il s'agit d'un tarif à deux entrées (la circonférence et la hauteur), il n'y a en fait qu'une seule variable explicative: C^2H. Un autre exemple est le tarif de cubage de peuplement établi par Fonweban et Houllier (1997) au Cameroun pour *Eucalyptus saligna*:

$$V = \beta_1 G^{\beta_2} \left(\frac{H_0}{N}\right)^{\beta_3}$$

où V est le volume du peuplement en $m^3\,ha^{-1}$, G est la surface terrière en $m^2\,ha^{-1}$, H_0 est la hauteur dominante du peuplement, N est la densité du peuplement (nombre de tiges par hectare) et les β sont des paramètres constants. Même s'il s'agit d'un tarif à trois entrées (la surface terrière, la hauteur dominante et la densité), il n'y a en fait que deux variables explicatives: G et le rapport H_0/N.

③ **Exploration de la relation biomasse–D^2H**

Pour un tarif de biomasse à deux entrées par rapport au diamètre D et la hauteur H, la quantité D^2H constitue une approximation du volume du tronc (au coefficient de forme près) et peut donc être utilisée comme variable explicative synthétique. Le nuage de points de la biomasse en fonction de D^2H s'obtient par la commande:

```
with(dat,plot(dbh^2*haut,Btot,xlab="D2H (cm2.m)",ylab="Biomasse (t)"))
```

et le résultat est représenté dans la figure 5.4. Ce nuage de points est du même type que le graphe de la figure 5.1C: la relation entre la biomasse et D^2H est linéaire mais la variance de la biomasse augmente lorsque D^2H augmente.

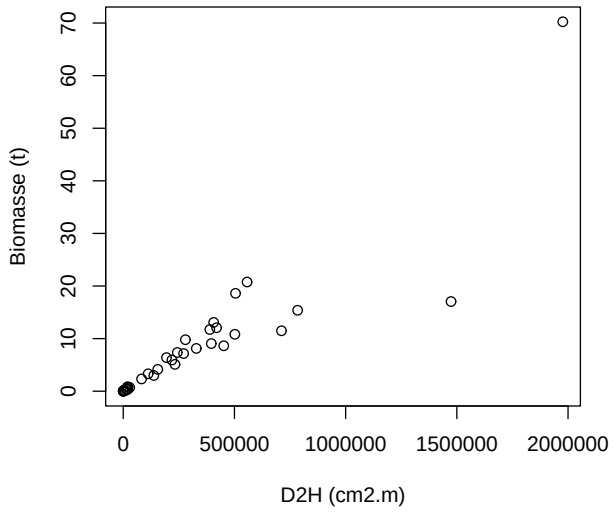


Figure 5.4 – *Nuage de points de la biomasse sèche totale (tonnes) en fonction de D^2H, où D est le diamètre à hauteur de poitrine (cm) et H la hauteur (m) pour les 42 arbres mesurés au Ghana par Henry et al. (2010).*

À supposer que, après cette phase d'agrégation des variables explicatives, il reste encore p variables explicatives $X_1, \ldots, X_p$ (avec $p > 1$), on peut d'abord explorer les p relations entre Y et chacune des p variables explicatives. Il s'agit bien de relations entre deux variables et les méthodes graphiques que l'on présentera après s'appliquent donc bien. Cependant cette approche s'avère en fait souvent peu informative, car la relation entre Y et p variables ne se ramène pas aux p relations entre Y et chacune des p variables séparément. Un exemple didactique simple peut illustrer ce propos: supposons que la variable Y soit (aux erreurs près) la somme de deux variables explicatives:

$$Y = X_1 + X_2 + \varepsilon \tag{5.1}$$

où ε est une erreur d'espérance nulle, et que les variables X_1 et X_2 soient liées de sorte que X_1 varie entre 0 et $-X_{\max}$ et que, à X_1 donné, X_2 varie entre $\max(0, -X1)$ et $\min(X_{\max}, 1 - X_1)$. La figure 5.5 montre les deux graphiques de Y en fonction de chacune des variables explicatives X_1 et X_2 pour des données simulées selon ce modèle (avec $X_{\max} = 5$). Aucune structure n'est apparente dans les nuages de points et on ne peut donc pas déceler le modèle générateur $\mathrm{E}(Y) = X_1 + X_2$.

Une façon de s'en sortir quand on a deux variables explicatives passe par le conditionnement. Il consiste à regarder la relation entre la variable à expliquer Y et l'une des variables

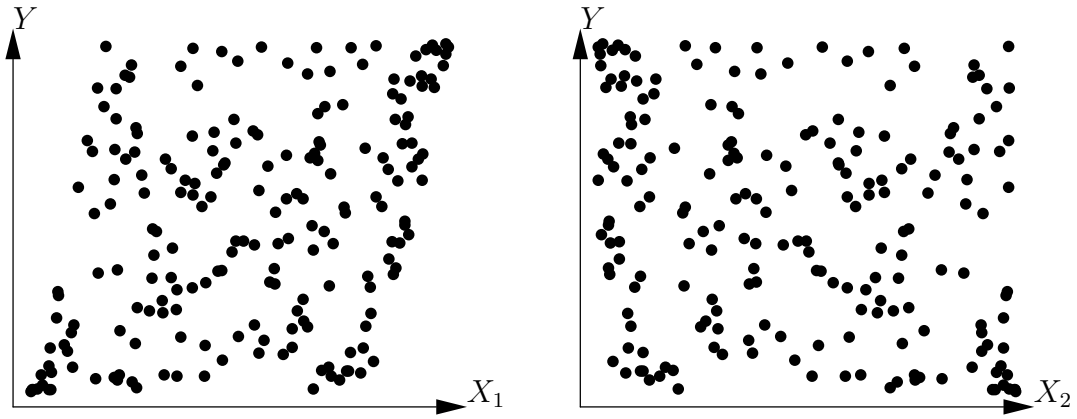


FIGURE 5.5 – *Graphes d'une variable Y en fonction de chacune de deux variables explicatives X_1 et X_2 telles que $E(Y) = X_1 + X_2$.*

explicatives (mettons X_2) conditionnellement aux valeurs de l'autre variable explicative (dans ce cas X_1). Concrètement, on découpe le jeu de données selon des classes de valeurs de X_1, puis on explore la relation entre Y et X_2 au sein de chaque sous-jeu de données. Pour poursuivre l'exemple précédent, on a découpé les valeurs de X_1 en 12 intervalles larges de 0,5 unités: le premier va de -5 à $-4{,}5$, le second de $-4{,}5$ à -4, etc., jusqu'au dernier intervalle qui va de 0,5 à 1. Le jeu de données représenté dans la figure 5.5 a été découpé en 12 sous-jeux de données en fonction des 12 classes de valeurs de X_1, puis on a tracé les 12 graphiques de Y en fonction de X_2 pour ces 12 sous-jeux de données. Le résultat est représenté dans la figure 5.6. La superposition des graphiques de la figure 5.6 redonnerait le graphique de droite de la figure 5.5. Ces graphiques révèlent que, pour une valeur donnée de X_1, la relation entre Y et X_2 est bien linéaire. De plus on peut s'apercevoir que la pente de la droite reliant Y à X_2 pour X_1 donné est constante quelle que soit la valeur de X_1. Cette exploration graphique révèle donc que le modèle est du type:

$$\mathrm{E}(Y) = f(X_1) + aX_2$$

où a est une coefficient constant (en l'occurrence égal à 1, mais l'exploration graphique ne se préoccupe pas de la valeur des paramètres), et $f(X_1)$ représente l'ordonnée à l'origine de la droite reliant Y à X_2 pour X_1 donné. Cette ordonnée à l'origine varie potentiellement en fonction de X_1, selon une fonction f qui reste à déterminer.

Pour explorer la forme de la fonction f, on peut ajuster par régression linéaire une droite à chacun des 12 sous-jeux de données de Y et X_2 correspondant aux 12 classes de valeurs de X_1. On retient la valeur de l'ordonnée à l'origine, notée y_0, de ces 12 droites, et on trace le graphique de y_0 en fonction du milieu de chaque classe de valeurs de X_1. Ce graphique est représenté sur la figure 5.7 pour les mêmes données simulées que précédemment. Cette exploration graphique révèle que la relation entre y_0 et X_1 est linéaire, c'est-à-dire: $f(X_1) = bX_1 + c$. Au final, l'exploration graphique fondée sur le conditionnement par rapport à X_1 a révélé qu'un modèle adéquat était bien:

$$\mathrm{E}(Y) = aX_2 + bX_1 + c$$

Dans la mesure où les variables X_1 et X_2 jouent un rôle symétrique dans le modèle (5.1), le conditionnement est également symétrique vis-à-vis de ces deux variables. Nous avons ici étudié la relation entre Y et X_2 conditionnellement à X_1, mais nous aurions abouti de la

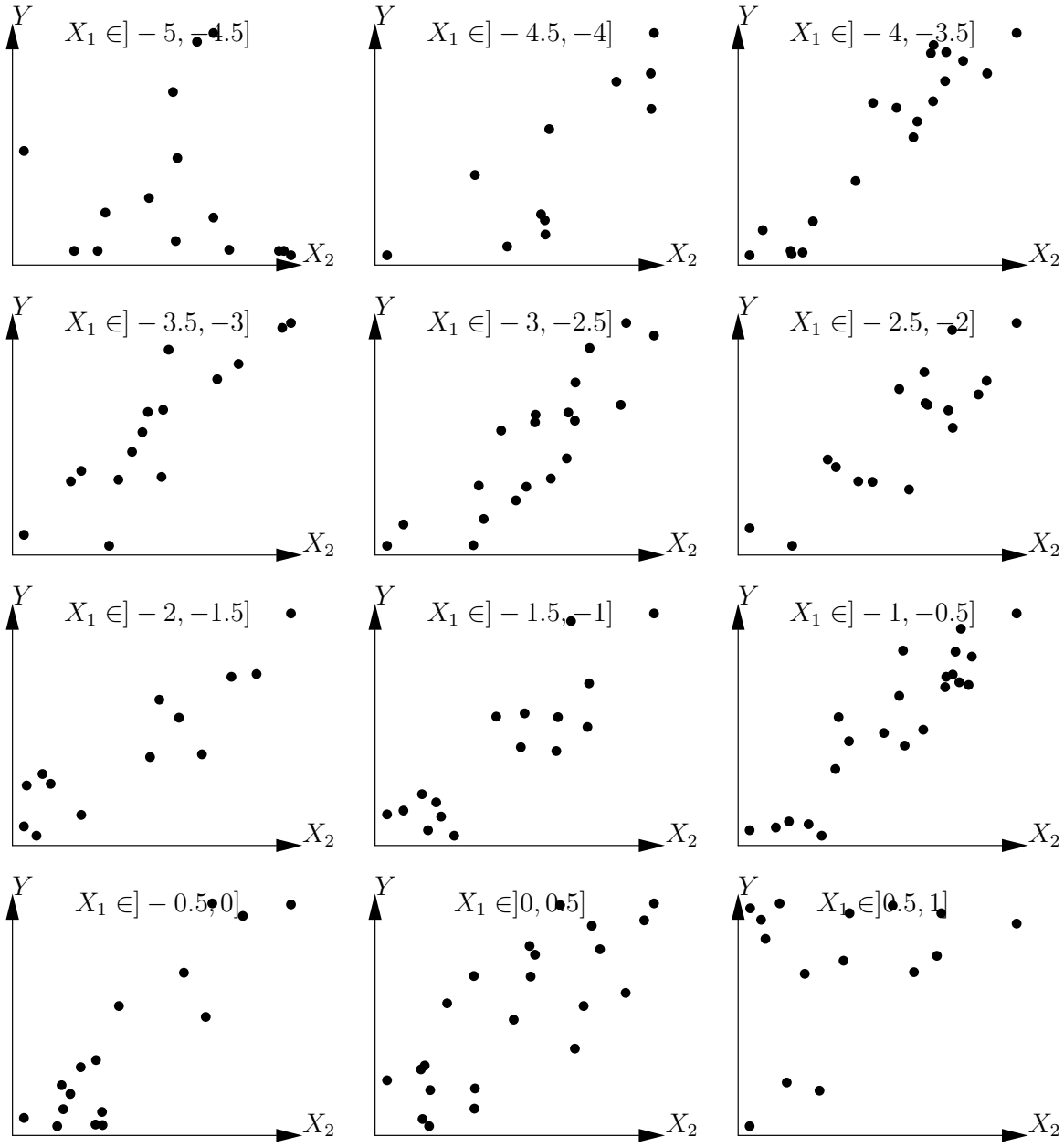


Figure 5.6 – *Graphes d'une variable Y en fonction d'une variable explicative X_2 pour chacun des sous-jeux de données définis par des classes de valeurs d'une autre variable explicative X_1, avec $E(Y) = X_1 + X_2$.*

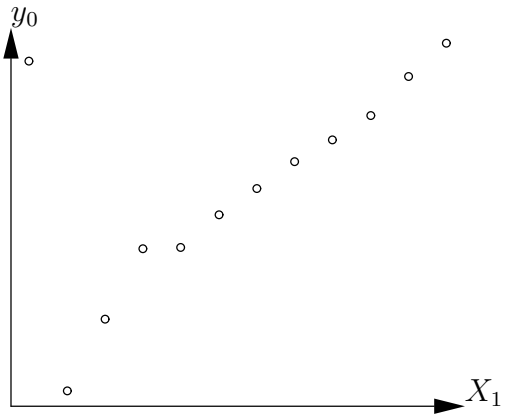


FIGURE 5.7 – *Graphe de l'ordonnée à l'origine de la régression linéaire de Y par rapport à X_2 pour un sous-jeu de données correspondant à une classe de valeurs de X_1 en fonction du milieu de ces classes, pour des données simulées selon le modèle $Y = X_1 + X_2 + \varepsilon$.*

même manière au même modèle en explorant la relation entre Y et X_1 conditionnellement à X_2.

Dans cet exemple, la relation entre Y et X_2 pour X_1 donné est une droite dont la pente est indépendante de X_2: on dit qu'il n'y a pas d'interaction entre X_1 et X_2. Un modèle avec interaction serait par exemple: $\mathrm{E}(Y) = X_1 + X_2 + X_1X_2$. Dans ce cas, la relation entre Y et X_2 à X_1 donné est une droite dont la pente, égale à $1 + X_1$, dépend bien de X_1. Le conditionnement permet sans plus de difficulté d'explorer la forme de modèles avec interactions entre les variables explicatives.

Le conditionnement s'étend en principe à un nombre quelconque de variables explicatives. Pour trois variables explicatives X_1, X_2, X_3 par exemple, on pourra explorer la relation entre Y et X_3 pour X_1 et X_2 fixés; notons f la fonction définissant cette relation et $\theta(X_1, X_2)$ les paramètres de f (qui dépendent potentiellement de X_1 et X_2):

$$\mathrm{E}(Y) = f[X_3; \theta(X_1, X_2)]$$

Il s'agit ensuite d'explorer la relation entre θ et les deux variables X_1 et X_2. À nouveau on conditionne en explorant la relation entre θ et X_2 pour X_1 fixé; notons g la fonction définissant cette relation et $\phi(X_1)$ les paramètres de g (qui dépendent potentiellement de X_1):

$$\theta(X_1, X_2) = g[X_2; \phi(X_1)]$$

Enfin, on explore la relation entre ϕ et X_1; soit h la fonction définissant cette relation. Au bout du compte, le modèle décrivant les données sera:

$$\mathrm{E}(Y) = f\{X_3; g[X_2; h(X_1)]\}$$

Ce raisonnement s'étend en principe à un nombre quelconque de variables explicatives, mais on voit bien en pratique qu'il devient difficile à mettre en œuvre pour $p > 3$. Le conditionnement requiert par ailleurs des données abondantes puisque chaque sous-jeu de données, défini par les classes de valeurs des variables conditionnelles, doit comporter suffisamment de données pour pouvoir explorer graphiquement les relations entre variables. Dans le cas de trois variables explicatives, les sous-jeux de données sont définis par le croisement des classes de valeurs de X_1 et X_2 (par exemple). Si le jeu de données complet comporte n observations, si X_1 et X_2 sont divisés en 10 classes de valeurs, et si les données se répartissent

également selon ces classes, alors chaque sous-jeu de données ne comporte plus que $n/100$ observations ! En pratique, à moins d'avoir un jeu de données particulièrement grand, il est difficile d'utiliser le principe du conditionnement pour plus de deux variables explicatives.

Pour l'ajustement de tarifs de biomasse ou de cubage, le nombre d'entrées du tarif est le plus souvent limité (deux ou trois entrées au plus), de sorte qu'on n'est généralement pas confronté au problème de l'exploration graphique avec un grand nombre de variables explicatives. Si cela venait à se produire, des analyses multivariées telles que l'analyse en composantes principales peuvent être utilement mises à profit (Philippeau, 1986; Härdle et Simar, 2003). Ces analyses visent à projeter les observations sur un sous-espace de dimension réduite (le plus souvent deux ou trois), construit à partir de combinaisons linéaires des variables explicatives et de sorte à maximiser la variabilité des observations dans ce sous-espace. En d'autres termes, ces analyses multivariées permettent de visualiser dans le plan les relations entre variables en perdant le moins possible d'information, ce qui est bien l'objectif recherché de l'exploration graphique.

④

Conditionnement sur la densité du bois

Explorons à présent la relation entre la biomasse, D^2H et la densité du bois ρ. On définit `n` classes de densité du bois de sorte que chaque classe contienne approximativement le même nombre d'observations:

```
d <- quantile(dat$dens,(0:n)/n)
i <- findInterval(dat$dens,d,rightmost.closed=TRUE)
```

L'objet `d` définit les bornes des classes de densité, tandis que l'objet `i` contient le numéro de la classe de densité à laquelle appartient chaque observation. Le graphique de la biomasse en fonction de D^2H en coordonnées logarithmiques, avec différents symboles et couleurs selon la classe de densité s'obtient par la commande:

```
with(dat,plot(dbh^2*haut,Btot,xlab="D2H (cm2m)",ylab="Biomasse (t)",log="xy",pch=i,
col=i))
```

et est représenté sur la figure 5.8 pour `n` $= 4$ classes de densité du bois. En anticipant sur le chapitre 6, on ajuste une régression linéaire entre $\ln(B)$ et $\ln(D^2H)$ pour chaque sous-ensemble d'observations correspondant à chaque classe de densité du bois:

```
m <- as.data.frame(lapply(split(dat,i), function(x)
coef(lm(log(Btot)~I(log(dbh^2*haut)),data=x[x$Btot>0,]))))
```

Quand on trace les graphiques de l'ordonnée à l'origine de la régression et de sa pente en fonction de la densité médiane de la classe:

```
dmid <- (d[-1]+d[-n])/2
plot(dmid,m[1,],xlab="Densité du bois (g/cm3)",ylab="Ordonnée à l'origine")
plot(dmid,m[2,],xlab="Densité du bois (g/cm3)",ylab="Pente")
```

on n'observe à première vue aucune relation particulière (figure 5.9).

5.1.2 Comment détecter qu'une relation est adéquate ?

Désormais, on suppose qu'on a une seule variable explicative X et que l'on cherche à explorer la relation entre X et la variable à expliquer Y. La première étape est de projeter

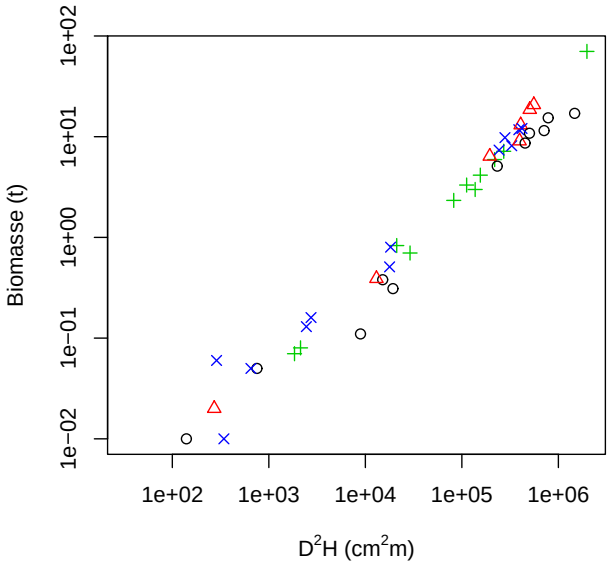


FIGURE 5.8 – *Nuage de points (données log-transformées) de la biomasse sèche totale (tonnes) en fonction de D^2H, où D est le diamètre à hauteur de poitrine (cm) et H la hauteur (m) pour les 42 arbres mesurés au Ghana par Henry et al. (2010), avec différents symboles selon les classes de densité du bois: rond noir, $0{,}170 \leq \rho < 0{,}545$ g cm^{-3}; triangle rouge, $0{,}545 \leq \rho < 0{,}600$ g cm^{-3}; signe plus vert, $0{,}600 \leq \rho < 0{,}650$ g cm^{-3}; croix bleue, $0{,}650 \leq \rho < 0{,}780$ g cm^{-3}.*

sur un graphique avec X en abscisse et Y en ordonnée le nuage de points correspondant aux données. Il s'agit ensuite de deviner visuellement la fonction qui passe au milieu de ce nuage de points en épousant ses formes. Il se trouve que l'œil humain est assez peu doué pour discriminer des formes proches. À titre d'exemple, la figure 5.10 présente trois nuages de points correspondants dans le désordre aux trois modèles suivants (on a mis ici le terme d'erreur à zéro):

modèle puissance:	$Y = aX^b$
modèle exponentiel:	$Y = a\exp(bX)$
modèle polynômial:	$Y = a + bX + cX^2 + dX^3$

Les trois nuages de points ont des allures semblables et bien malin qui pourrait dire à quel modèle correspond chacun de ces trois nuages de points.

En revanche l'œil humain est doué pour déceler si une relation est linéaire ou pas. Pour détecter visuellement si la forme d'un nuage de point s'ajuste ou non à une fonction, on a donc fortement intérêt, lorsque cela est possible, à utiliser une transformation de variables qui rend la relation linéaire. Dans le cas du modèle puissance par exemple, $Y = aX^b$ implique $\ln Y = \ln a + b \ln X$, donc la transformation de variables:

$$\begin{cases} X' = \ln X \\ Y' = \ln Y \end{cases} \tag{5.2}$$

rend la relation linéaire. Dans le cas du modèle exponentiel, $Y = a\exp(bX)$ implique $\ln Y = \ln a + bX$, donc la transformation de variables:

$$\begin{cases} X' = X \\ Y' = \ln Y \end{cases} \tag{5.3}$$

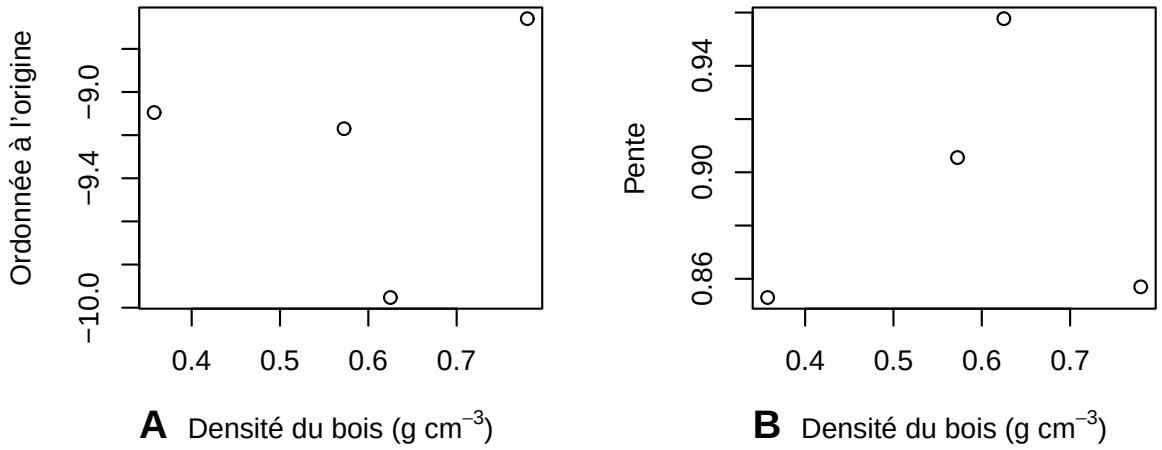


Figure 5.9 – *Ordonnée à l'origine a (A) et pente b (B) de la régression linéaire* $\ln(B) = a + b\ln(D^2H)$ *conditionnelle à la classe de densité du bois, en fonction de la densité du bois médiane des classes. Les régressions sont ajustées aux données des 42 arbres mesurés par Henry et al. (2010) au Ghana.*

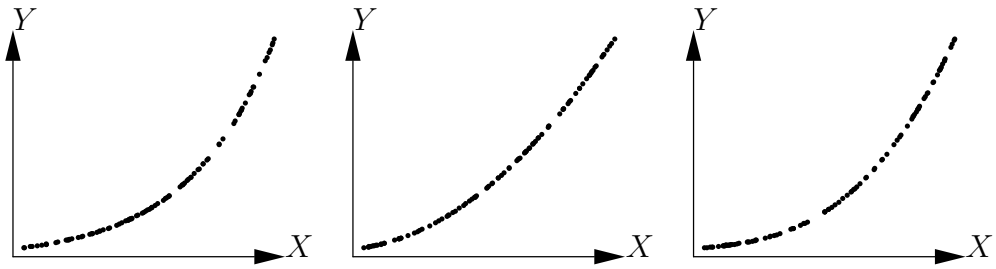


Figure 5.10 – *Trois nuages de points correspondants dans le désordre à trois modèles: modèle puissance, modèle exponentiel et modèle polynômial.*

rend la relation linéaire. En revanche aucune de ces deux transformations ne linéarise le modèle polynômial. En appliquant ces transformations de variable aux données représentées sur la figure 5.10, on va donc être en mesure de déceler lequel des nuages de points correspond à chacun des modèles. La figure 5.11 représente les trois nuages de points après application de la transformation de variables (5.3). Le premier nuage de point adopte la forme d'une droite tandis que les deux autres gardent une forme courbe. Le nuage de points le plus à gauche de la figure 5.10 correspond ainsi au modèle exponentiel.

La figure 5.12 représente les trois nuages de points après application de la transformation de variables (5.2). Le second nuage de point adopte la forme d'une droite tandis que les deux autres gardent une forme courbe. Le nuage de points au centre de la figure 5.10 correspond ainsi au modèle puissance. Par déduction, le nuage de points le plus à droite de la figure 5.10 correspond au modèle polynômial.

Il n'est pas toujours possible de trouver une transformation de variable qui linéarise la relation. C'est par exemple le cas du modèle polynômial $Y = a + bX + cX^2 + dX^3$: on ne peut pas trouver de transformation de X en X' et de Y en Y' telle que la relation entre Y' et X' soit une droite, quels que soient les coefficients a, b, c et d. Il doit être clair également que la linéarité dont il est question ici est celle de la relation entre la variable à expliquer Y et la variable explicative X. Ce n'est pas la linéarité au sens du modèle linéaire, qui est

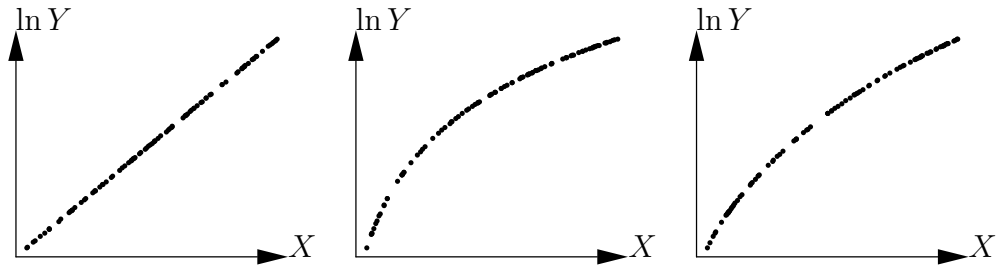


FIGURE 5.11 – *Application de la transformation de variables $X \to X$, $Y \to \ln Y$ aux nuages de points représentés dans la figure 5.10.*

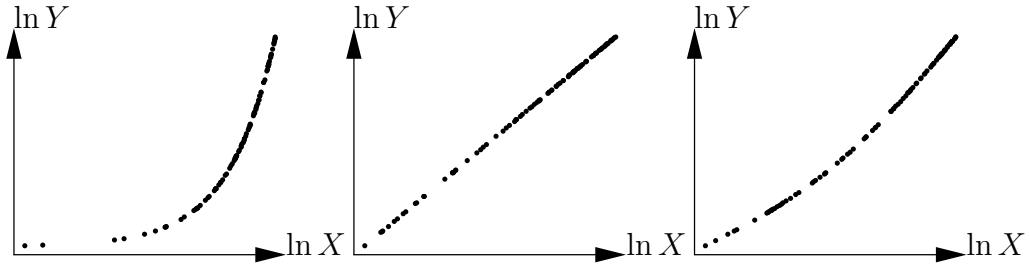


FIGURE 5.12 – *Application de la transformation de variables $X \to \ln X$, $Y \to \ln Y$ aux nuages de points représentés dans la figure 5.10.*

une linéarité vis-à-vis des coefficients du modèle (ainsi le modèle $Y = a + bX^2$ est linéaire au sens du modèle linéaire alors que ce modèle définit une relation non linéaire entre X et Y).

Lorsqu'aucune transformation de variable ne permet de linéariser la relation entre X et Y, le mieux est d'ajuster le modèle et de regarder si la courbe ajustée passe au milieu du nuage de points en s'adaptant à ses formes. On aura d'ailleurs intérêt dans ce cas à regarder le graphique des résidus en fonction des valeurs prédites.

⑤

Exploration de la relation biomasse–diamètre: transformation des variables

Utilisons la transformation logarithmique pour transformer simultanément le diamètre et la biomasse. Le nuage de points des données log-transformées s'obtient comme suit:

```
plot(dat$dbh,dat$Btot,xlab="Diamètre (cm)",ylab="Biomasse (t)",log="xy")
```

Le nuage de points résultant est montré dans la figure 5.13. La transformation logarithmique a linéarisé la relation entre la biomasse et le diamètre: la relation entre $\ln(D)$ et $\ln(B)$ a la forme d'une droite et la variance de $\ln(B)$ ne varie pas avec le diamètre (comme dans la figure 5.1A).

⑥

Exploration de la relation biomasse–D^2H: transformation des variables

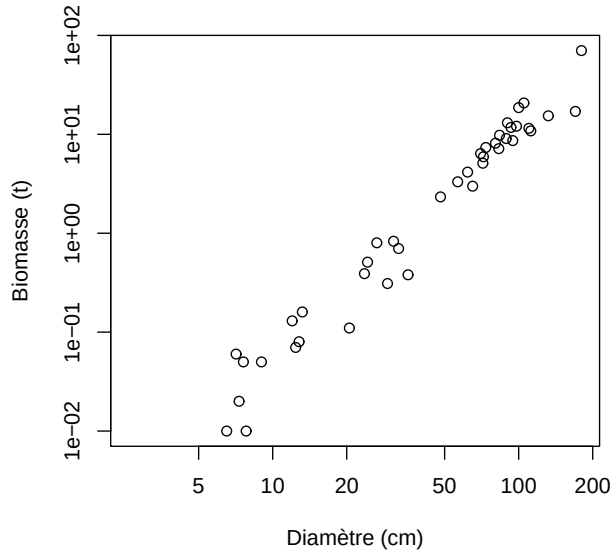


Figure 5.13 – *Nuage de points (données log-transformées) de la biomasse sèche totale (tonnes) en fonction du diamètre à hauteur de poitrine (cm) pour les 42 arbres mesurés au Ghana par Henry et al. (2010).*

Utilisons la transformation logarithmique pour transformer simultanément D^2H et la biomasse. Le nuage de points des données log-transformées s'obtient comme suit:

```
with(dat,plot(dbh^2*haut,Btot,log="xy",xlab="D2H (cm2.m)",ylab="Biomasse (t)"))
```

Le nuage de points résultant est montré dans la figure 5.14. La transformation logarithmique a linéarisé la relation entre la biomasse et D^2H: la relation entre $\ln(D^2H)$ et $\ln(B)$ a la forme d'une droite et la variance de $\ln(B)$ ne varie pas avec D^2H (comme dans la figure 5.1A).

5.1.3 Catalogue de primitives

Les synthèses des tarifs réalisées par Zianis *et al.* (2005) pour l'Europe, par Henry *et al.* (2011) pour l'Afrique sub-saharienne ou plus spécifiquement par Hofstad (2005) pour l'Afrique australe permettront de se faire une idée de la forme des modèles les plus fréquemment trouvés dans la littérature sur les tarifs de biomasse et de cubage. Deux types de modèles ressortent le plus fréquemment: le modèle puissance et le modèle polynômial (de degré deux ou, au maximum, trois). Ces deux types de modèles seront donc le point de départ de l'exploration graphique des données pour la construction d'un tarif de cubage ou de biomasse. Le modèle puissance $Y = aX^b$ est également appelé relation allométrique et un certain nombre d'interprétation biologique de ce modèle ont été fournies (Gould, 1979; Franc *et al.*, 2000, § 1.1.5). En particulier, la « metabolic scaling theory » (Enquist *et al.*, 1998, 1999; West *et al.*, 1997, 1999) prédit de manière théorique, en s'appuyant sur une description fractale de la structure interne des arbres, que la biomasse d'un arbre est liée à son diamètre par une relation puissance avec un exposant égal à $8/3 \approx 2{,}67$:

$$B \propto \rho D^{8/3}$$

où ρ est la densité spécifique du bois. Même si la « metabolic scaling theory » a été largement remise en cause (Muller-Landau *et al.*, 2006), elle a au moins le mérite de donner une base

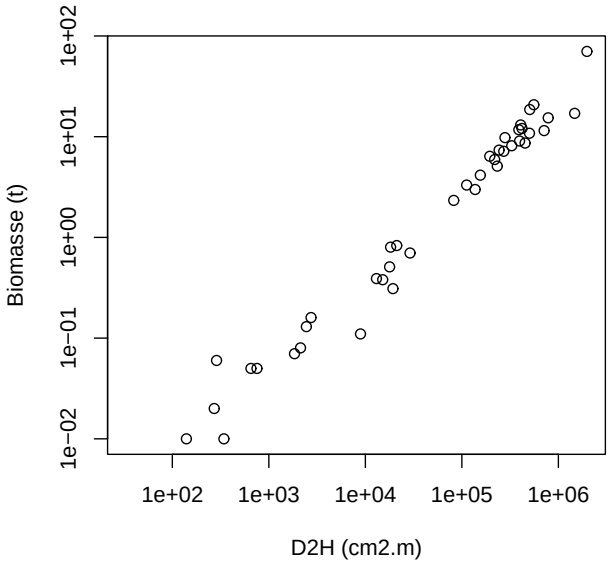


FIGURE 5.14 – *Nuage de points (données log-transformées) de la biomasse sèche totale (tonnes) en fonction de D^2H, où D est le diamètre à hauteur de poitrine (cm) et H la hauteur (m) pour les 42 arbres mesurés au Ghana par Henry et al. (2010).*

biologique à la relation puissance qui est souvent observée.

Outre le modèle puissance $B = aD^b$ et le modèle polynômial du second degré $B = a_0 + a_1 D + a_2 D^2$, et sans prétention à l'exhaustivité, les tarifs de biomasse suivants sont fréquemment trouvés (Yamakura *et al.*, 1986; Brown *et al.*, 1989; Brown, 1997; Martinez-Yrizar *et al.*, 1992; Araújo *et al.*, 1999; Nelson *et al.*, 1999; Ketterings *et al.*, 2001; Chave *et al.*, 2001, 2005; Nogueira *et al.*, 2008; Basuki *et al.*, 2009; Návar, 2009; Djomo *et al.*, 2010; Henry *et al.*, 2010):

1. tarif à deux entrées sous la forme puissance par rapport à la variable D^2H: $B = a(\rho D^2 H)^b$
2. tarif à deux entrées: $\ln(B) = a_0 + a_1 \ln(D) + a_2 \ln(H) + a_3 \ln(\rho)$
3. tarif à une entrée: $\ln(B) = a_0 + a_1 \ln(D) + a_2[\ln(D)]^2 + a_3[\ln(D)]^3 + a_4 \ln(\rho)$,

où ρ est la densité spécifique du bois. Au facteur de forme près, la variable D^2H représente le volume du tronc, ce qui explique qu'elle soit fréquemment utilisée comme variable explicative. La deuxième équation peut être vue comme une généralisation de la première. En effet, en appliquant la transformation logarithmique, la première équation équivaut à: $\ln(B) = \ln(a) + 2b \ln(D) + b \ln(H) + b \ln(\rho)$. La première équation est donc équivalente à la seconde dans le cas particulier où $a_2 = a_3 = a_1/2$. Enfin la troisième équation peut être vue comme une généralisation du modèle puissance $B = aD^b$.

Plus généralement, le tableau 5.1 récapitule un certain nombre de fonctions pouvant modéliser la relation entre deux variables. Quand elle existe, on indique la transformation de variables qui linéarise la relation entre X et Y. On remarquera que le modèle puissance modifié n'est qu'une réécriture du modèle exponentiel, et que le modèle racine n'est qu'une réécriture du modèle exponentiel modifié. On remarquera également qu'une bonne partie de ces modèles ne sont que des cas particuliers de modèles plus complexes (et comportant davantage de paramètres). Par exemple le modèle linéaire n'est qu'un cas particulier du modèle polynômial, le modèle de Gompertz n'est qu'un cas particulier du modèle de Sloboda, etc.

Table 5.1 – *Quelques modèles reliant deux variables.*

Nom	Équation	Transformation
Modèles polynômiaux		
linéaire	$Y = a + bX$	identité
parabolique ou quadratique	$Y = a + bX + cX^2$	
polynômial d'ordre p	$Y = a_0 + a_1X + a_2X^2 + \ldots + a_pX^p$	
Modèles exponentiels		
exponentiel ou de Malthus	$Y = a\exp(bX)$	$Y' = \ln Y$, $X' = X$
exponentiel modifié	$Y = a\exp(b/X)$	$Y' = \ln Y$, $X' = 1/X$
logarithme	$Y = a + b\ln X$	$Y' = Y$, $X' = \ln X$
log réciproque	$Y = 1/(a + b\ln X)$	$Y' = 1/Y$, $X' = \ln X$
pression de vapeur	$Y = \exp(a + b/X + c\ln X)$	
Modèles en loi de puissance		
puissance	$Y = aX^b$	$Y' = \ln Y$, $X' = \ln X$
puissance modifié	$Y = ab^X$	$Y' = \ln Y$, $X' = X$
puissance décalé	$Y = a(X - b)^c$	
géométrique	$Y = aX^{bX}$	$Y' = \ln Y$, $X' = X\ln X$
géométrique modifié	$Y = aX^{b/X}$	$Y' = \ln Y$, $X' = (\ln X)/X$
racine	$Y = ab^{1/X}$	$Y' = \ln Y$, $X' = 1/X$
de Hoerl	$Y = ab^X X^c$	
de Hoerl modifié	$Y = ab^{1/X} X^c$	
Modèles de production–densité		
inverse	$Y = 1/(a + bX)$	$Y' = 1/Y$, $X' = X$
inverse quadratique	$Y = 1/(a + bX + cX^2)$	
de Bleasdale	$Y = (a + bX)^{-1/c}$	
de Harris	$Y = 1/(a + bX^c)$	
Modèles de croissance		
de croissance saturée	$Y = aX/(b + X)$	$Y' = X/Y$, $X' = X$
mononucléaire ou de Mitscherlisch	$Y = a[b - \exp(-cX)]$	
Modèles sigmoïdes		
de Gompertz	$Y = a\exp[-b\exp(-cX)]$	
de Sloboda	$Y = a\exp[-b\exp(-cX^d)]$	
logistique ou de Verhulst	$Y = a/[1 + b\exp(-cX)]$	
de Nelder	$Y = a/[1 + b\exp(-cX)]^{1/d}$	
de von Bertalanffy	$Y = a[1 - b\exp(-cX)]^3$	
de Chapman-Richards	$Y = a[1 - b\exp(-cX)]^d$	
de Hossfeld	$Y = a/[1 + b(1 + cX)^{-1/d}]$	
de Levakovic	$Y = a/[1 + b(1 + cX)^{-1/d}]^{1/e}$	
du facteur multiplicatif multiple	$Y = (ab + cX^d)/(b + X^d)$	
de Johnson-Schumacher	$Y = a\exp[-1/(b + cX)]$	
de Lundqvist-Matérn ou de Korf	$Y = a\exp[-(b + cX)^d]$	
de Weibull	$Y = a - b\exp(-cX^d)$	
Modèles divers		
hyperbolique	$Y = a + b/X$	$Y' = Y$, $X' = 1/X$
sinusoïdal	$Y = a + b\cos(cX + d)$	
de capacité de chaleur	$Y = a + bX + c/X^2$	
gaussien	$Y = a\exp[-(X - b)^2/(2c^2)]$	
de fraction rationnelle	$Y = (a + bX)/(1 + cX + dX^2)$	

Le modèle polynômial d'ordre p doit être manié avec prudence car les polynômes sont capables de s'ajuster à n'importe quelle forme pourvu que le degré p soit suffisamment élevé (les fonctions usuelles sont toutes décomposables dans une base de polynômes: c'est le principe du développement limité). Concrètement, on peut avoir un polynôme qui épouse très bien la forme du nuage de points dans le domaine de valeurs des données disponibles, mais qui prend une forme très invraisemblable en dehors de ce domaine. En d'autres termes, le modèle polynômial peut présenter des dangers d'extrapolation, d'autant plus grands que le degré p est important. En pratique, on évitera fortement d'ajuster des polynômes de degré supérieur à 3.

5.2 Exploration de la variance

Considérons à présent le terme d'erreur ε du modèle reliant la variable à expliquer Y à une variable explicative X. L'exploration de la forme de la variance revient basiquement à répondre à la question: la variance des résidus est-elle constante (homoscédasticité) ou non (hétéroscédasticité) ? La réponse à cette question dépend implicitement de la forme précise de la relation qui sera utilisée pour ajuster le modèle. À titre d'exemple, pour la relation puissance $Y = aX^b$, on peut

- soit ajuster le modèle non linéaire $Y = aX^b + \varepsilon$, ce qui revient à estimer directement les paramètres a et b;
- soit ajuster le modèle linéaire $Y' = a' + bX' + \varepsilon$ sur les données transformées $Y' = \ln Y$ et $X' = \ln X$, ce qui revient à estimer les paramètres $a' = \ln a$ et b.

Les deux options ne sont bien sûr pas interchangeables puisque le terme d'erreur ε (que l'on supposera suivre une loi normale d'écart-type constant) ne joue pas le même rôle dans les deux cas. Dans le premier cas on a une erreur additive par rapport au modèle puissance. Dans le second cas on a erreur additive par rapport au modèle linéarisé donc, si l'on revient au modèle puissance:

$$Y = \exp(Y') = aX^b \exp(\varepsilon) = aX^b \varepsilon'$$

ce qui correspond à une erreur multiplicative par rapport au modèle puissance, où ε' suit une loi lognormale. La différence entre ces deux options est illustrée dans la figure 5.15. L'erreur additive se traduit par une variance constante dans le graphique (a) de Y en fonction de X et par une variance décroissante avec X dans le graphique (c) de ces mêmes données transformées par le logarithme. L'erreur multiplicative se traduit par une variance croissante avec X dans le graphique (b) de Y en fonction de X et par une variance constante dans le graphique (d) de ces mêmes données transformées par le logarithme.

Ainsi le processus de linéarisation du modèle reliant Y à X par une transformation de variables affecte aussi bien la forme de la relation moyenne que celle du terme d'erreur. Cette propriété peut du reste être exploitée pour stabiliser des résidus qui varient avec X afin de les rendre constants, mais c'est un point qui sera abordé au chapitre suivant. Pour l'instant, nous cherchons à explorer la forme de l'erreur $Y - \mathrm{E}(Y)$ en fonction de X, sans chercher à transformer les variables X et Y.

La forme de la relation moyenne $\mathrm{E}(Y) = f(X)$ ayant préalablement été déterminée de façon graphique, il suffit d'examiner visuellement sur le graphique de Y en fonction de X si les points se répartissent également de chaque côté de la courbe f quelle que soit la valeur de X. Les graphiques (a) et (b) de la figure 5.1, par exemple, illustrent le cas de résidus de variance constante quel que soit X, tandis que les graphiques (c) et (d) de cette même figure montrent le cas de résidus dont la variance augmente avec X. Des relations plus complexes, comme celle illustrée par la figure 5.16, peuvent également être imaginées. Dans le cas de

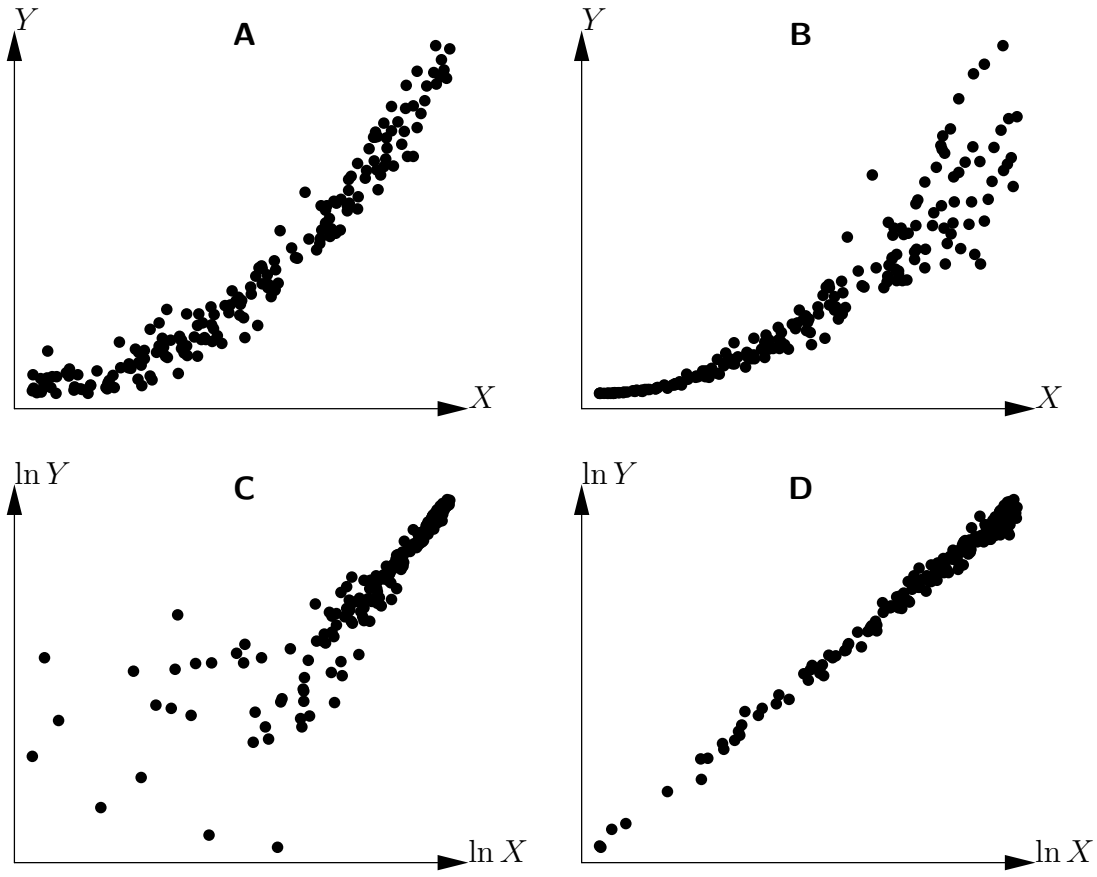


FIGURE 5.15 – *Modèle puissance avec erreur additive (A et C) ou multiplicative (B et D). Le graphique (C) (respectivement D) résulte du graphique (A) (respectivement B) par la transformation de variables* $X \to \ln X$ *et* $Y \to \ln Y$.

figure illustrée par la figure 5.16, la variance des résidus fluctue de façon périodique avec X. De telles situations ont peu de chances d'être rencontrées en pratique dans le contexte des tarifs de biomasse ou de cubage. Presque systématiquement, on aura à choisir entre deux situations: la variance des résidus est constante ou elle augmente avec X. Dans le premier cas, il n'y a rien de plus à faire. Dans le second cas, on ne cherchera pas à préciser la forme exacte de la relation entre X et la variance des résidus et on adoptera d'emblée un modèle puissance pour relier la variance des résidus ε à X:

$$\mathrm{Var}(\varepsilon) = \alpha X^{\beta}$$

Les valeurs des coefficients α et β seront estimés en même temps que celles des autres coefficients du modèle lors de la phase d'ajustement du modèle, qui sera traitée dans le prochain chapitre.

5.3 L'exploration n'est pas une sélection

En guise de conclusion, nous tenons à préciser que l'exploration graphique ne vise pas à sélectionner une seule forme de modèle, mais plutôt à faire le tri entre les modèles qui sont acceptables pour décrire le jeu de données et ceux qui ne le sont pas. Loin de chercher à sélectionner « le » modèle censé être le meilleur pour décrire les données, il faut plutôt

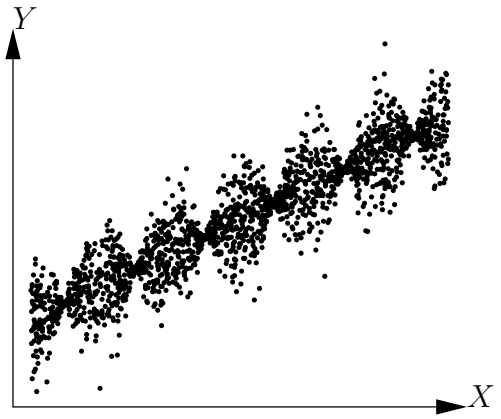


FIGURE 5.16 – *Graphique d'un nuage de points générés par le modèle* $Y = a + bX + \varepsilon$*, où* ε *suit une loi normale de moyenne nulle et d'écart-type proportionnel au cosinus de* X*.*

chercher à sélectionner trois ou quatre modèles candidats, susceptibles de décrire les données. Le choix final entre ces trois ou quatre modèles identifiés lors de l'exploration graphique se fera après la phase d'ajustement aux données que nous allons voir dans le chapitre suivant.

6

Ajustement du tarif

L'ajustement d'un modèle consiste à estimer les paramètres de ce modèle à partir de données. Cela suppose donc d'une part que les données sont déjà disponibles et mises en forme, et d'autre part que l'expression mathématique du modèle à ajuster est connue. Par exemple, ajuster le modèle puissance $B = aD^b$ consiste à estimer les coefficients a et b à partir d'un jeu de données donnant les valeurs B_i et D_i de la biomasse et du diamètre de n arbres ($i = 1, \ldots, n$). La variable réponse (aussi appelée dans la littérature variable de sortie, variable d'intérêt, variable dépendante) du modèle est la variable qui est prédite par le modèle. Il n'y en a qu'une. Dans le cadre de ce manuel, la variable réponse sera toujours un volume ou une biomasse. Les variables explicatives sont les variables utilisées pour prédire la variable réponse. Il peut y en avoir plusieurs, et leur nombre est dénoté p. Il ne faut pas confondre les variables explicatives et les entrées du tarif. Le modèle $B = a(D^2H)^b$ comporte une seule variable explicative (à savoir D^2H) mais deux entrées (le diamètre D et la hauteur H). À l'inverse, le modèle $B = a_0 + a_1D + a_2D^2$ comporte deux variables explicatives (D et D^2) mais une seule entrée (le diamètre D). À chaque variable explicative est associée un coefficient à estimer. À cela s'ajoute le cas échéant une ordonnée à l'origine ou un coefficient multiplicateur, de sorte que le nombre total de coefficients à estimer dans un modèle à p variables explicatives sera p ou $p + 1$.

Une observation consiste en la donnée de la variable réponse (volume ou biomasse) et des variables explicatives pour un arbre. Pour reprendre l'exemple du modèle $B = aD^b$, une observation consiste en le doublet (B_i, D_i). Le nombre d'observations est donc n. Une observation découle d'une mesure sur le terrain. La prédiction du modèle est la valeur de la variable réponse prédite par le modèle étant données les variables explicatives. Une prédiction découle d'un calcul. Par exemple, la prédiction du modèle $B = aD^b$ pour un arbre de diamètre D_i est $\hat{B}_i = aD_i^b$. Il y autant de prédictions qu'il y a d'observations. Un concept clé de l'ajustement des modèles est le résidu. Le résidu, ou erreur résiduelle, est l'écart entre la valeur observée de la variable réponse et sa prédiction. Toujours pour le même exemple, le résidu de la i^{e} observation est: $\varepsilon_i = B_i - \hat{B}_i = B_i - aD_i^b$. Il y autant de résidus qu'il y a d'observations. L'ajustement d'un modèle sera d'autant meilleur que ses résidus seront faibles. De plus, les propriétés statistiques du modèle découleront des propriétés que les résidus seront supposés vérifier *a priori*, en particulier leur loi de distribution. Le type d'ajustement du modèle dépendra directement des propriétés de ses résidus.

Dans la totalité des modèles que nous verrons, les observations seront supposées être *indépendantes* ou, ce qui revient au même, les résidus seront supposés être indépendants: pour tout $i \neq j$, ε_i est supposé être indépendant de ε_j. Cette propriété d'indépendance est relativement facile à assurer *via* le protocole d'échantillonnage. Typiquement, il faudra s'assurer que les caractéristiques d'un arbre mesuré à un endroit n'influencent pas les caractéristiques d'un autre arbre de l'échantillon. Sélectionner pour l'échantillon des arbres suffisamment éloignés les uns des autres suffit en général à assurer cette propriété d'indépendance. Si les résidus ne sont pas indépendants, on peut modifier le modèle pour en tenir compte. Par exemple, on pourra introduire une structure de dépendance spatiale dans les résidus pour tenir compte d'une auto-corrélation spatiale des mesures. Nous n'aborderons pas ces modèles qui sont beaucoup plus complexes à mettre en œuvre.

Dans tous les modèles que nous verrons, on supposera de plus que les résidus ont une distribution *normale* d'espérance nulle. La moyenne nulle des résidus est en fait une propriété qui découle automatiquement de l'ajustement du modèle, et qui assure que les prédictions ne sont pas biaisées. Ce sont bien les résidus qui sont supposés avoir une distribution normale, et non pas les observations. Pour des données de volume ou de biomasse, cette hypothèse n'est en fait guère contraignante. Dans l'hypothèse peu probable où la distribution des résidus s'éloignerait fortement d'une loi normale, on pourrait éventuellement envisager d'ajuster d'autres types de modèle, comme le modèle linéaire généralisé, mais cela ne sera pas abordé dans le cadre de ce manuel.

L'hypothèse d'indépendance et l'hypothèse de distribution normale des résidus sont les deux premières hypothèses qui sous-tendent l'ajustement des modèles. Nous verrons une troisième hypothèse ultérieurement. Il convient de vérifier que ces hypothèses sont effectivement vérifiées. Dans la mesure où ces hypothèses portent sur les résidus du modèle et non pas sur les observations, on ne peut pas les tester tant que les résidus n'ont pas été calculés, c'est-à-dire tant que le modèle n'a pas été ajusté. Ce sont donc des hypothèses que l'on vérifie *a posteriori*, après que le modèle a été ajusté. Les modèles que nous verrons sont par ailleurs *robustes* vis-à-vis de ces hypothèses, c'est-à-dire que les qualités de prédiction des modèles ajustés restent correctes quand bien même les hypothèses d'indépendance et de distribution normale des résidus ne sont pas tout à fait vérifiées. C'est pourquoi on ne cherchera pas à tester de manière très formelle ces deux hypothèses. En pratique, on se contentera d'une vérification visuelle fondée sur des graphiques.

6.1 Ajustement d'un modèle linéaire

Le modèle linéaire est le plus simple des modèles à ajuster. L'adjectif *linéaire* signifie ici que le modèle dépend *linéairement* de ses coefficients. Par exemple, $Y = a + bX^2$ et $Y = a + b\ln(X)$ sont des modèles linéaires car la variable réponse Y dépend linéairement des coefficients a et b, même si Y ne dépend pas linéairement de la variable explicative X. À l'inverse, $Y = aX^b$ n'est pas un modèle linéaire car Y ne dépend pas linéairement du coefficient b. Une autre propriété du modèle linéaire est que le résidu est additif. Pour faire ressortir cela, on écrit explicitement le résidu ε dans l'expression du modèle. Par exemple, pour une régression linéaire de Y par rapport à X, on écrira: $Y = a + bX + \varepsilon$.

6.1.1 Régression linéaire simple

La régression linéaire simple est le plus simple des modèles linéaires. Elle suppose (*i*) qu'il n'y qu'une seule variable explicative X, (*ii*) que la relation entre la variable réponse Y

et X a la forme d'une droite:

$$Y = a + bX + \varepsilon$$

où a est l'ordonnée à l'origine de la droite et b sa pente, et (*iii*) que les résidus ont une variance constante: $\mathrm{Var}(\varepsilon) = \sigma^2$. Par exemple, le modèle

$$\ln(B) = a + b\ln(D) + \varepsilon \tag{6.1}$$

est un exemple de régression linéaire simple, avec comme variable réponse $Y = \ln(B)$ et comme variable explicative $X = \ln(D)$. Il correspond à un modèle puissance pour la biomasse: $B = \exp(a)D^b$. Ce modèle est fréquemment utilisé pour ajuster un tarif de biomasse monospécifique. Un autre exemple est le tarif de biomasse à deux entrées:

$$\ln(B) = a + b\ln(D^2H) + \varepsilon \tag{6.2}$$

L'hypothèse de variance constante des résidus vient s'ajouter aux deux hypothèses d'indépendance et de distribution normale (on parle aussi d'homoscédasticité). On résume ces trois hypothèses en écrivant:

$$\varepsilon \underset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma)$$

où $\mathcal{N}(\mu, \sigma)$ désigne la loi normale d'espérance μ et d'écart-type σ, le tilde « $\sim$ » signifie « est distribué selon », et « i.i.d. » est l'abréviation de « indépendamment et identiquement distribué ».

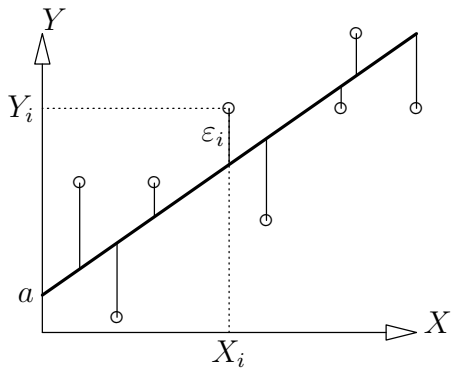


Figure 6.1 – *Schéma des observations (points), de la droite de régression (trait épais) et des résidus (traits fins).*

Estimation des coefficients

La figure 6.1 schématise les observations et la droite des valeurs prédites. Le meilleur ajustement va être celui qui minimise l'erreur résiduelle. On peut envisager plusieurs façons de quantifier cette erreur résiduelle. D'un point de vue mathématique, cela revient à choisir une norme pour mesurer ε, et diverses normes pourraient faire l'affaire. La norme classiquement utilisée est la norme L_2, ce qui revient à quantifier l'écart résiduel entre les observations et les prédictions par la somme des carrés des résidus, que l'on appelle aussi la somme des carrés des écarts (SCE):

$$\mathrm{SCE}(a, b) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - a - bX_i)^2$$

Le meilleur ajustement est donc celui qui minimise SCE. Autrement dit, les estimations $\hat{a}$ et $\hat{b}$ des coefficients a et b sont les valeurs de a et b qui minimisent la somme des carrés des écarts:

$$(\hat{a},\, \hat{b}) = \arg\min_{(a,\, b)} \mathrm{SCE}(a, b)$$

Ce minimum s'obtient en calculant les dérivées partielles de SCE par rapport à a et b, et en cherchant les valeurs de a et b qui annulent ces dérivées partielles. Des calculs simples aboutissent aux résultats suivants: $\hat{b} = \widehat{\mathrm{Cov}}(X,\, Y)/S_X^2$ et $\hat{a} = \bar{Y} - \hat{b}\bar{X}$, où $\bar{X} = (\sum_{i=1}^n X_i)/n$ est la moyenne empirique de la variable explicative, $\bar{Y} = (\sum_{i=1}^n Y_i)/n$ est la moyenne empirique de la variable réponse,

$$S_X^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

est la variance empirique de la variable explicative, et

$$\widehat{\mathrm{Cov}}(X,\, Y) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$$

est la covariance empirique entre la variable explicative et la variable réponse. L'estimation de la variance résiduelle est quant à elle:

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{a} - \hat{b}X_i)^2 = \frac{\mathrm{SCE}(\hat{a},\, \hat{b})}{n-2}$$

Parce que cette méthode d'estimation des coefficients repose sur la minimisation de la somme des carrés des écarts, on l'appelle la méthode *des moindres carrés* (on précise parfois « moindres carrés ordinaires », par opposition aux moindres carrés pondérés que l'on verra au § 6.1.3). L'avantage de cette méthode d'estimation est qu'elle fournit une expression explicite des coefficients estimés.

Interprétation des résultats d'une régression

Lorsque l'on ajuste une régression linéaire simple, plusieurs sorties sont à analyser. Le coefficient de détermination, plus généralement appelé R^2, mesure la qualité de l'ajustement. Le R^2 est directement lié à la variance résiduelle puisque:

$$R^2 = 1 - \frac{\hat{\sigma}^2(n-2)/n}{S_Y^2}$$

où $S_Y^2 = [\sum_{i=1}^n (Y_i - \bar{Y})^2]/n$ est la variance empirique de Y. La différence $S_Y^2 - \hat{\sigma}^2(n-2)/n$ entre la variance de Y et la variance résiduelle représente la variance expliquée par le modèle. Le coefficient de détermination R^2 s'interprète donc comme le ratio de la variance expliquée par le modèle sur la variance totale. Il est compris entre 0 et 1 et plus il est proche de un, meilleure est la qualité de l'ajustement. Dans le cas d'une régression linéaire simple, et uniquement dans ce cas-là, le R^2 est aussi égal au carré du coefficient de corrélation linéaire (aussi appelé coefficient de Pearson) entre X et Y. On a vu dans le chapitre 5 (en particulier dans la figure 5.2) les limites de l'interprétation du R^2.

Outre les valeurs estimées des coefficients a et b, l'ajustement du modèle fournit également les écart-types de ces estimations (c'est-à-dire les écart-types des estimateurs $\hat{a}$ et $\hat{b}$), ainsi que les résultats des tests de significativité de ces coefficients. Il y a un test pour l'ordonnée à l'origine a, qui teste l'hypothèse nulle $a = 0$, et de même un test pour la pente b, qui teste l'hypothèse nulle $b = 0$.

Enfin, le résultat du test de significativité globale du modèle est à analyser. Ce test repose sur la décomposition de la variance totale de Y comme étant la somme de la variance expliquée par le modèle et de la variance résiduelle. Comme en analyse de variance, le test est un test de Fisher utilisant comme statistique de test un ratio pondéré de la variance expliquée sur la variance résiduelle. Dans le cas de la régression linéaire simple, et uniquement dans ce cas-là, le test de significativité globale du modèle donne le même résultat que le test de l'hypothèse nulle $b = 0$. Cela se comprend intuitivement: une droite reliant X à Y n'est significative que si la pente de cette droite est non nulle.

Vérification des hypothèses

L'ajustement du modèle s'achève en vérifiant que les hypothèses posées *a priori* sur les résidus sont bien vérifiées. Nous ne reviendrons pas sur l'hypothèse d'indépendance des résidus, que l'on considère vérifiée grâce au plan d'échantillonnage adopté. Éventuellement, s'il existe un ordre naturel dans les observations, on pourra utiliser le test de Durbin-Watson pour tester que les résidus sont bien indépendants (Durbin et Watson, 1971). L'hypothèse de distribution normale des résidus se vérifie visuellement à partir du graphe quantile–quantile. Ce graphe représente les quantiles empiriques des résidus en fonction des quantiles théoriques de la loi normale centrée réduite. Si l'hypothèse de distribution normale des résidus est acceptable, les points s'alignent approximativement le long d'une droite, comme dans la figure 6.2 (graphique à droite).

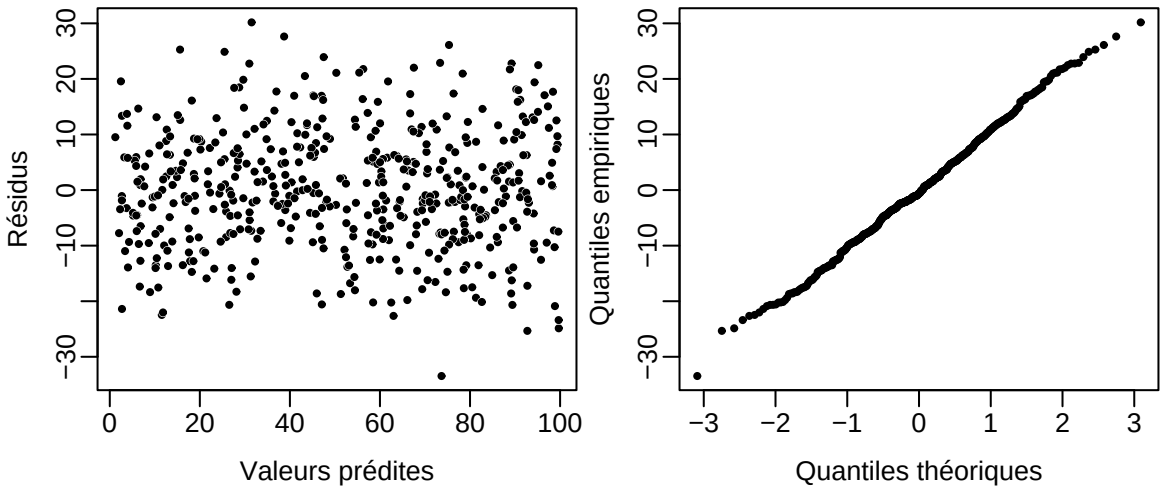


Figure 6.2 – *Allure du graphe des résidus en fonction des valeurs prédites (à gauche) et du graphe quantile–quantile (à droite) lorsque les hypothèses de distribution normale et de variance constante des résidus sont bien vérifiées.*

Dans le cas de l'ajustement de tarifs de cubage ou de biomasse, l'hypothèse la plus importante à vérifier est celle de la constance de la variance des résidus. On la vérifie visuellement en traçant le nuage de points des résidus $\varepsilon_i = Y_i - \hat{Y}_i$ en fonction des valeurs prédites $\hat{Y}_i = \hat{a} + \hat{b}X_i$. Si la variance des résidus est bien constante, ce nuage de points ne doit laisser apparaître aucune tendance, aucune structuration particulière. C'est par exemple le cas dans le graphique de gauche de la figure 6.2. En revanche, si une structuration particulière apparaît dans ce nuage de points, l'hypothèse doit être remise en cause. C'est par exemple le cas dans la figure 6.3, où le nuage de points des résidus en fonction des valeurs prédites a la forme d'un entonnoir. Cette forme est typique d'une augmentation de la variance résiduelle

avec la variable explicative (ce que l'on appelle de l'hétéroscédasticité). Si tel est le cas, un autre modèle que la régression linéaire simple doit être ajusté.

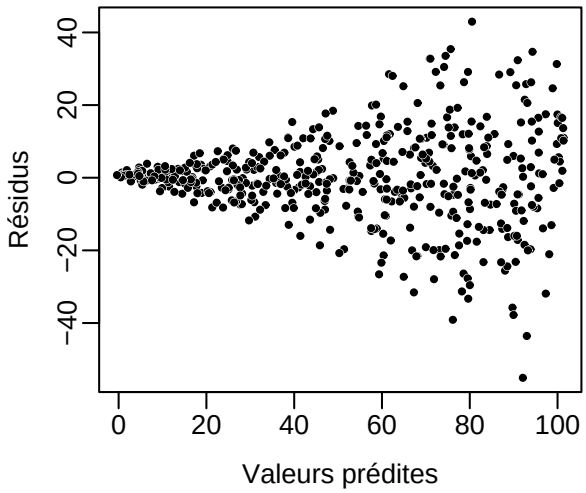


FIGURE 6.3 – *Allure du graphe des résidus en fonction des valeurs prédites lorsque les résidus ont une variance non constante (hétéroscédasticité).*

Dans le cas de données biologiques telles que le volume ou la biomasse d'arbres, l'hétéroscédasticité est la règle et l'homoscédasticité l'exception. Cela signifie simplement que la variabilité de biomasse (ou de volume) des arbres est d'autant plus grande qu'ils sont grands. Cette variabilité croissante de la biomasse des individus avec leur taille est un principe général en biologie. Ainsi, dans le cas d'ajustement de tarifs de biomasse ou de cubage, la régression linéaire simple utilisant la biomasse comme variable réponse ($Y = B$) sera généralement de peu d'utilité. La transformation logarithme (c'est-à-dire $Y = \ln(B)$) permet de résoudre ce problème, de sorte que les régressions linéaires que nous utiliserons dans le cadre de l'ajustement de tarifs seront presque toujours des régressions sur données log-transformées. Nous reviendrons longuement sur ce point qui est fondamental.

⑦

Régression linéaire simple entre $\ln(B)$ et $\ln(D)$

L'analyse exploratoire (fil rouge n° 5) a montré que la relation entre le logarithme de la biomasse et le logarithme du diamètre était linéaire, avec une variance de $\ln(B)$ qui était approximativement constante. On peut donc ajuster une régression linéaire simple pour prédire $\ln(B)$ en fonction de $\ln(D)$:

$$\ln(B) = a + b\ln(D) + \varepsilon$$

avec

$$\mathrm{Var}(\varepsilon) = \sigma^2$$

La régression est ajustée par moindres carrés ordinaires. Comme on ne peut pas appliquer la transformation logarithmique à une valeur nulle, les données de biomasse nulles (cf. fil rouge n° 1) sont préalablement retirées du jeu de données:

```
m <- lm(log(Btot)~I(log(dbh)),data=dat[dat$Btot>0,])
summary(m)
```

L'écart-type résiduel vaut $\hat{\sigma} = 0{,}462$, le R^2 est de 0,9642 et le modèle est hautement significatif (test de Fisher: $F_{1,39} = 1051$, p-value $< 2{,}2 \times 10^{-16}$). Les valeurs des coefficients sont données dans le tableau suivant:

```
             Estimate  Std. Error  t value  Pr(>|t|)
(Intercept) -8.42722     0.27915   -30.19    <2e-16  ***
I(log(dbh))  2.36104     0.07283    32.42    <2e-16  ***
```

La première colonne de ce tableau donne les valeurs des coefficients. Le modèle s'écrit donc: $\ln(B) = -8{,}42722 + 2{,}36104 \ln(D)$. La deuxième colonne donne les écart-types des estimateurs des coefficients. La troisième colonne donne la valeur de la statistique de test de l'hypothèse nulle « le coefficient est nul ». Enfin la quatrième colonne donne la p-value de ce test. Dans le cas présent, à la fois la pente et l'ordonnée à l'origine sont significativement différents de zéro.

Il reste à vérifier graphiquement que les hypothèses de la régression linéaire sont vérifiées:

```
plot(m,which=1:2)
```

Le résultat est représenté dans la figure 6.4. Même si le graphe quantile–quantile des résidus semble légèrement structuré, on considérera que les hypothèses de la régression linéaire simple sont convenablement respectées.

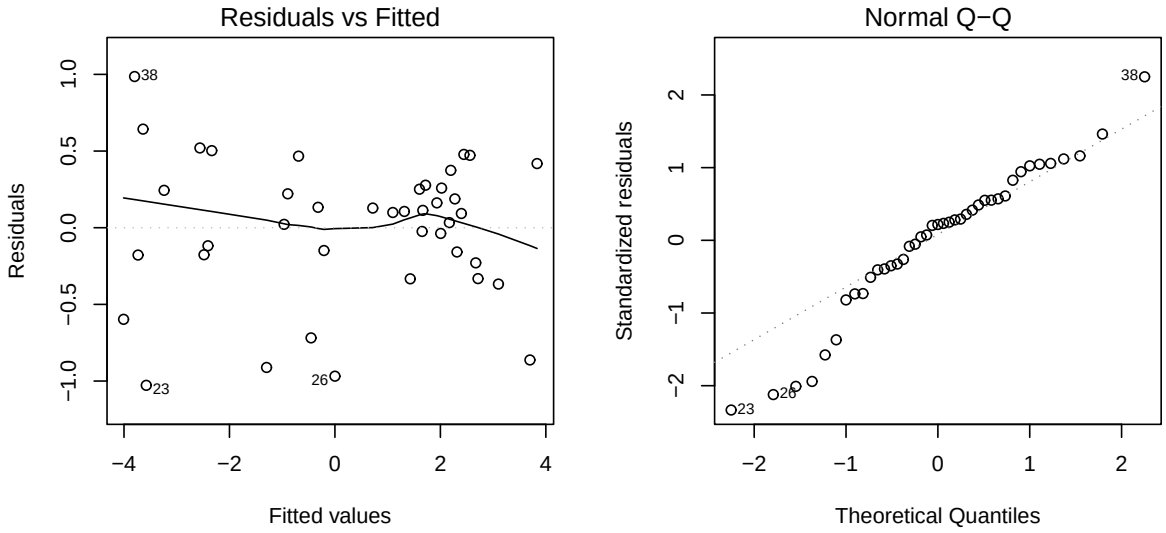


Figure 6.4 – *Graphique des résidus en fonction des valeurs prédites (gauche) et graphe quantile–quantile (droite) des résidus de la régression linéaire simple de* $\ln(B)$ *par rapport à* $\ln(D)$ *ajustée aux 42 arbres mesurés par Henry et al. (2010) au Ghana.*

⑧ **Régression linéaire simple entre** $\ln(B)$ **et** $\ln(D^2H)$

L'analyse exploratoire (fil rouge n° 6) a montré que la relation entre le logarithme de la biomasse et le logarithme de D^2H était linéaire, avec une variance de $\ln(B)$ qui était approximativement constante. On peut donc ajuster une régression linéaire simple pour prédire $\ln(B)$ en fonction de $\ln(D^2H)$:

$$\ln(B) = a + b \ln(D^2H) + \varepsilon$$

avec

$$\mathrm{Var}(\varepsilon) = \sigma^2$$

La régression est ajustée par moindres carrés ordinaires. Comme on ne peut pas appliquer la transformation logarithmique à une valeur nulle, les données de biomasse nulles (cf. fil rouge n° 1) sont préalablement retirées du jeu de données:

```
m <- lm(log(Btot)~I(log(dbh^2*haut)),data=dat[dat$Btot>0,])
summary(m)
```

L'écart-type résiduel vaut $\hat{\sigma} = 0{,}4084$, le R^2 est de 0,972 et le modèle est hautement significatif (test de Fisher: $F_{1,39} = 1356$, p-value $< 2{,}2 \times 10^{-16}$). Les valeurs des coefficients sont les suivantes:

```
                  Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)       -8.99427     0.26078   -34.49    <2e-16  ***
I(log(dbh^2*haut))  0.87238     0.02369    36.82    <2e-16  ***
```

La première colonne de ce tableau donne les valeurs des coefficients. Le modèle s'écrit donc: $\ln(B) = -8{,}99427 + 0{,}87238 \ln(D^2H)$. La deuxième colonne donne les écart-types des estimateurs des coefficients. La troisième colonne donne la valeur de la statistique de test de l'hypothèse nulle « le coefficient est nul ». Enfin la quatrième colonne donne la p-value de ce test. Dans le cas présent, à la fois la pente et l'ordonnée à l'origine sont significativement différents de zéro.

Il reste à vérifier graphiquement que les hypothèses de la régression linéaire sont vérifiées:

```
plot(m,which=1:2)
```

Le résultat est représenté dans la figure 6.5. Même si le graphe des résidus en fonction des valeurs prédites semble légèrement structuré, on considérera que les hypothèses de la régression linéaire simple sont convenablement respectées.

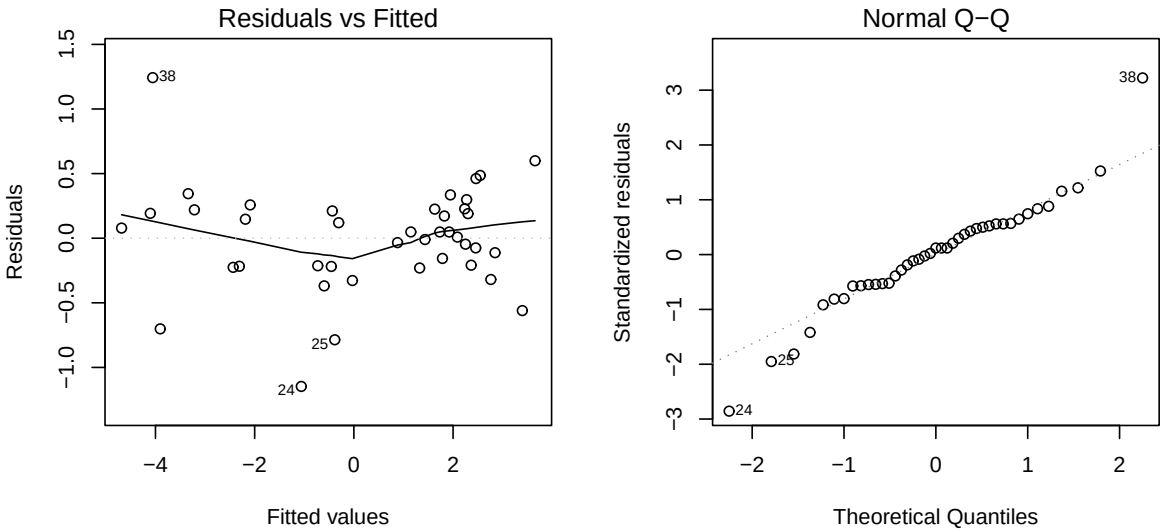


FIGURE 6.5 – *Graphique des résidus en fonction des valeurs prédites (gauche) et graphe quantile–quantile (droite) des résidus de la régression linéaire simple de* $\ln(B)$ *par rapport à* $\ln(D^2H)$ *ajustée aux 42 arbres mesurés par Henry et al. (2010) au Ghana.*

6.1.2 Régression multiple

La régression multiple est l'extension de la régression linéaire simple au cas où il y a plusieurs variables explicatives, et s'écrit:

$$Y = a_0 + a_1 X_1 + a_2 X_2 + \ldots + a_p X_p + \varepsilon \tag{6.3}$$

où Y est la variable réponse, X_1, ..., X_p les p variables explicatives, a_0, ..., a_p sont les coefficients à estimer, et ε est l'erreur résiduelle. En comptant l'ordonnée à l'origine a_0, il y a $p+1$ coefficients à estimer. Comme pour la régression linéaire simple, la variance des résidus est supposée constante, égale à σ^2:

$$\varepsilon \underset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma)$$

Les tarifs de biomasse suivants sont des exemples de régression multiple:

$$\ln(B) = a_0 + a_1 \ln(D^2 H) + a_2 \ln(\rho) + \varepsilon \tag{6.4}$$
$$\ln(B) = a_0 + a_1 \ln(D) + a_2 \ln(H) + \varepsilon \tag{6.5}$$
$$\ln(B) = a_0 + a_1 \ln(D) + a_2 \ln(H) + a_3 \ln(\rho) + \varepsilon \tag{6.6}$$
$$\ln(B) = a_0 + a_1 \ln(D) + a_2 [\ln(D)]^2 + a_3 [\ln(D)]^3 + \varepsilon \tag{6.7}$$
$$\ln(B) = a_0 + a_1 \ln(D) + a_2 [\ln(D)]^2 + a_3 [\ln(D)]^3 + a_4 \ln(\rho) + \varepsilon \tag{6.8}$$

où ρ désigne la densité du bois. Dans tous ces exemples, la variable réponse est le logarithme de la biomasse: $Y = \ln(B)$. Le modèle (6.4) généralise (6.2) en ajoutant la dépendance vis-à-vis de la densité spécifique du bois: on préférera typiquement (6.4) à (6.2) quand le jeu de données est plurispécifique. Le modèle (6.5) généralise (6.2) en considérant que l'exposant associé à la hauteur H n'est pas forcément égal à la moitié de l'exposant associé au diamètre D. Il introduit donc un peu plus de souplesse dans la forme de la relation entre la biomasse et D^2H. Le modèle (6.6) généralise (6.2) en considérant à la fois qu'il y a plusieurs espèces et que la biomasse n'est pas tout à fait une puissance de D^2H. Le modèle (6.7) généralise (6.1) en considérant que la relation entre $\ln(B)$ et $\ln(D)$ n'est pas exactement linéaire. Il offre donc un peu plus de souplesse dans la forme de cette relation. Le modèle (6.8) est une extension de (6.7) pour tenir compte de la présence de plusieurs essences dans le jeu de données.

Estimation des coefficients

De la même façon que pour la régression linéaire simple, l'estimation des coefficients de la régression multiple est fondée sur la méthode des moindres carrés. Les estimateurs $\hat{a}_0$, $\hat{a}_1$, ..., $\hat{a}_p$ sont les valeurs des coefficients a_0, a_1, ..., a_p qui minimisent la somme des carrés des écarts:

$$\text{SCE}(a_0, a_1, \ldots, a_p) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - a_0 - a_1 X_{i1} - \ldots - a_p X_{ip})^2$$

où X_{ij} est la valeur de la j^{e} variable explicative pour la i^{e} observation ($i = 1, \ldots, n$ et $j = 1 \ldots, p$). À nouveau, les estimations des coefficients s'obtiennent en calculant les dérivées partielles de SCE par rapport aux coefficients, et en cherchant les valeurs des coefficients qui annulent ces dérivées partielles. Les calculs ne sont guère plus compliqués que pour la régression linéaire simple, à condition de les poser sous forme matricielle. Soit $\mathbf{X}$ la matrice

à n lignes et p colonnes, appelée matrice du plan, qui rassemble les valeurs observées des variables explicatives:

$$\mathbf{X} = \begin{bmatrix} 1 & X_{11} & \cdots & X_{1p} \\ \vdots & \vdots & & \vdots \\ 1 & X_{n1} & \cdots & X_{np} \end{bmatrix}$$

Soit $\mathbf{Y} = {}^{\mathrm{t}}[Y_1, \ldots, Y_n]$ le vecteur des n valeurs observées de la variable réponse, et $\mathbf{a} = {}^{\mathrm{t}}[a_0, \ldots, a_p]$ le vecteur des $p+1$ coefficients à estimer. Ainsi

$$\mathbf{Xa} = \begin{bmatrix} a_0 + a_1 X_{11} + \ldots + a_p X_{1p} \\ \vdots \\ a_0 + a_1 X_{n1} + \ldots + a_p X_{np} \end{bmatrix}$$

n'est autre que le vecteur $\hat{\mathbf{Y}}$ des n valeurs prédites par le modèle de la variable réponse. En utilisant ces notations matricielles, la somme des carrés des écarts s'écrit:

$$\mathrm{SCE}(\mathbf{a}) = {}^{\mathrm{t}}(\mathbf{Y} - \hat{\mathbf{Y}})(\mathbf{Y} - \hat{\mathbf{Y}}) = {}^{\mathrm{t}}(\mathbf{Y} - \mathbf{Xa})(\mathbf{Y} - \mathbf{Xa})$$

En utilisant les règles de calcul différentiel matriciel (Magnus et Neudecker, 2007), on obtient finalement:

$$\hat{\mathbf{a}} = \arg\min_{\mathbf{a}} \mathrm{SCE}(\mathbf{a}) = ({}^{\mathrm{t}}\mathbf{XX})^{-1\,\mathrm{t}}\mathbf{XY}$$

L'estimation de la variance résiduelle est quant à elle:

$$\hat{\sigma}^2 = \frac{\mathrm{SCE}(\hat{\mathbf{a}})}{n - p - 1}$$

Comme pour la régression linéaire simple, cette méthode d'estimation a l'avantage de fournir une expression explicite des coefficients estimés. La régression linéaire simple étant un cas particulier de la régression multiple (cas où $p = 1$), on peut s'assurer que les expressions matricielles des estimations des coefficients et de $\hat{\sigma}$ redonnent bien, lorsque $p = 1$, les expressions données précédemment dans le cas de la régression linéaire simple.

Interprétation des résultats d'une régression multiple

De la même manière que pour la régression linéaire simple, l'ajustement d'une régression multiple fournit un coefficient de détermination R^2 qui représente la part de variance expliquée par le modèle; les valeurs $\hat{\mathbf{a}}$ des coefficients a_0, a_1, …, a_p du modèle; les écart-types de ces estimations; les résultats des tests de significativité des coefficients (il y en a $p+1$, un pour chaque coefficient, d'hypothèses nulles $a_i = 0$ pour $i = 0, \ldots, p$); et le résultat du test de significativité globale du modèle.

Comme précédemment, la valeur du R^2 est comprise entre 0 et 1. Il est d'autant plus élevée que la qualité d'ajustement du modèle est meilleure. Il faut toutefois prendre garde que la valeur du R^2 augmente automatiquement avec le nombre de variables explicatives utilisées. Ainsi, si on prédit Y par un polynôme de degré p en X,

$$Y = a_0 + a_1 X + a_2 X^2 + \ldots + a_p X^p$$

le R^2 sera automatiquement une fonction croissante du degré p. Cela peut donner l'illusion qu'une régression polynômiale sera d'autant meilleure que le degré p du polynôme sera élevé. Cela n'est bien entendu pas le cas. Une valeur trop élevée du degré p entraînera une surparamétrisation du modèle. Autrement dit, le R^2 n'est pas un critère valable pour faire de la sélection de modèle. Nous reviendrons sur ce point dans le paragraphe 6.3.

Vérification des hypothèses

Comme la régression linéaire simple, la régression multiple repose sur trois hypothèses: indépendance des résidus; distribution normale des résidus; variance constante des résidus. Ces hypothèses se vérifient exactement de la même façon que pour la régression linéaire simple. Pour vérifier la distribution normale des résidus, on fera un graphe quantile–quantile et on s'assurera visuellement que le nuage de points forme une droite. Pour vérifier la variance constante des résidus, on fera le graphe des résidus en fonction des valeurs prédites et on s'assurera visuellement que le nuage de points ne présente aucune tendance particulière.

La même restriction que pour la régression linéaire simple s'applique aux données biologiques de volume ou de biomasse, qui présentent presque toujours (si ce n'est toujours...) de l'hétéroscédasticité. De ce fait, la régression multiple ne sera généralement applicable pour l'ajustement de tarifs que sur des données log-transformées.

⑨ **Régression polynômiale entre $\ln(B)$ et $\ln(D)$**

L'analyse exploratoire (fil rouge n° 5) a montré que la relation entre le logarithme de la biomasse et le logarithme du diamètre correspondait à une relation linéaire. On peut se demander si cette relation est réellement linéaire, ou si elle n'a pas une forme plus complexe. Pour cela, on peut faire une régression polynômiale d'ordre p, c'est-à-dire une régression multiple de $\ln(B)$ par rapport à $\ln(D)$, $[\ln(D)]^2$, ..., $[\ln(D)]^p$:

$$\ln(B) = a_0 + a_1 \ln(D) + a_2[\ln(D)]^2 + \ldots + a_p[\ln(D)]^p + \varepsilon$$

avec

$$\mathrm{Var}(\varepsilon) = \sigma^2$$

La régression est ajustée par moindres carrés ordinaires. Comme la transformation logarithme stabilise la variance résiduelle, les hypothèses de la régression multiple sont *a priori* vérifiées. Pour un polynôme d'ordre 2, la régression polynômiale est ajustée par la ligne de code suivante:

```
m2 <- lm(log(Btot)~I(log(dbh))+I(log(dbh)^2),data=dat[dat$Btot>0,])
print(summary(m2))
```

On obtient:

```
                Estimate  Std. Error  t value  Pr(>|t|)
 (Intercept)   -8.322190    1.031359   -8.069  9.25e-10  ***
 I(log(dbh))    2.294456    0.633072    3.624  0.000846  ***
 I(log(dbh)^2)  0.009631    0.090954    0.106  0.916225
```

avec $R^2 = 0{,}9642$. Quant à régression polynômiale d'ordre 3:

```
m3 <- lm(log(Btot)~I(log(dbh))+I(log(dbh)^2)+I(log(dbh)^3),data=dat[dat$Btot>0,])
print(summary(m3))
```

elle donne:

```
               Estimate  Std. Error  t value  Pr(>|t|)
 (Intercept)   -5.46413    3.80855   -1.435     0.160
 I(log(dbh))   -0.42448    3.54394   -0.120     0.905
 I(log(dbh)^2)  0.82073    1.04404    0.786     0.437
 I(log(dbh)^3) -0.07693    0.09865   -0.780     0.440
```

avec $R^2 = 0{,}9648$. Enfin la régression polynômiale d'ordre 4:

```
m4 <- lm(log(Btot)~I(log(dbh))+I(log(dbh)^2)+I(log(dbh)^3)+I(log(dbh)^4),data=dat[
dat$Btot>0,])
print(summary(m4))
```

donne:

```
               Estimate  Std. Error  t value  Pr(>|t|)
 (Intercept)   -26.7953     15.7399   -1.702    0.0973  .
 I(log(dbh))    26.3990     19.5353    1.351    0.1850
 I(log(dbh)^2) -11.2782      8.7301   -1.292    0.2046
 I(log(dbh)^3)   2.2543      1.6732    1.347    0.1863
 I(log(dbh)^4)  -0.1628      0.1166   -1.396    0.1714
```

avec $R^2 = 0{,}9666$. L'ajout de termes de degré supérieur à 1 n'apporte rien au modèle. Les coefficients associés à ces termes ne sont pas significativement différents de zéro. Cependant le R^2 du modèle ne cesse d'augmenter avec l'ordre p du polynôme. Le R^2 n'est donc pas un bon critère pour sélectionner l'ordre du polynôme. On peut superposer au nuage de points biomasse–diamètre les courbes prédites par ces différents polynômes: l'objet `m` désignant la régression linéaire de $\ln(B)$ par rapport à $\ln(D)$ ajustée dans le fil rouge n° 7,

```
with(dat,plot(dbh,Btot,xlab="Diamètre (cm)",ylab="Biomasse (t)",log="xy"))
D <- 10^seq(par("usr")[1],par("usr")[2],length=200)
lines(D,exp(predict(m,newdata=data.frame(dbh=D))))
lines(D,exp(predict(m2,newdata=data.frame(dbh=D))))
lines(D,exp(predict(m3,newdata=data.frame(dbh=D))))
lines(D,exp(predict(m4,newdata=data.frame(dbh=D))))
```

Les courbes sont montrées dans la figure 6.6: plus l'ordre du polynôme est élevé, plus la courbe se déforme pour s'ajuster aux données, avec une extrapolation en dehors du domaine des données qui est de plus en plus irréaliste (ce qui est typique d'un sur-paramétrage du modèle).

10

Régression multiple entre $\ln(B)$, $\ln(D)$ et $\ln(H)$

L'exploration graphique (fils rouges n^{os} 3 et 6) a montré que la variable synthétique D^2H était liée à la biomasse par une relation puissance (soit une relation linéaire en coordonnées logarithmiques): $B = a(D^2H)^b$. On peut se demander toutefois si les variables D^2 et H ont bien le même exposant b, ou bien si elles auraient des exposants différents: $B = a \times (D^2)^{b_1} H^{b_2}$. En travaillant sur les données log-transformées (ce qui au passage stabilise la variance résiduelle), cela revient à ajuster une régression multiple de $\ln(B)$ par rapport à $\ln(D)$ et $\ln(H)$:

$$\ln(B) = a + b_1 \ln(D) + b_2 \ln(H) + \varepsilon$$

avec

$$\mathrm{Var}(\varepsilon) = \sigma^2$$

La régression est ajustée par moindres carrés ordinaires. L'ajustement de cette régression multiple:

```
m <- lm(log(Btot)~I(log(dbh))+I(log(haut)),data=dat[dat$Btot>0,])
summary(m)
```

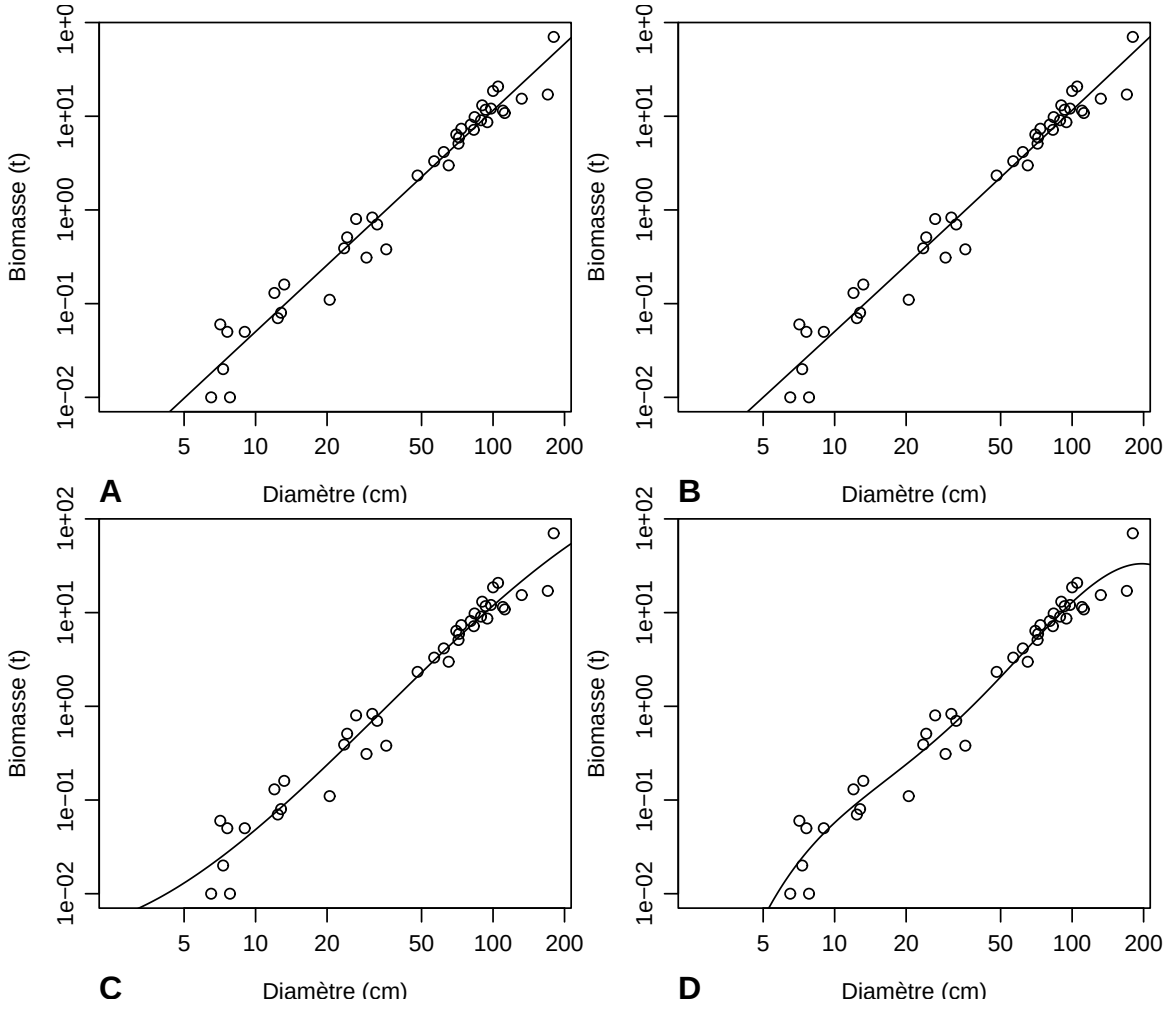


Figure 6.6 – *Biomasse en fonction du diamètre (en coordonnées logarithmiques) pour 42 arbres mesurés au Ghana par Henry et al. (2010) (points), et prédictions (courbes) par une régression polynômiale de* $\ln(B)$ *par rapport à* $\ln(D)$*: (A) polynôme d'ordre 1 (droite); (B) polynôme d'ordre 2 (parabole); (C) polynôme d'ordre 3; (D) polynôme d'ordre 4.*

donne:

```
               Estimate  Std. Error  t value  Pr(>|t|)
 (Intercept)    -8.9050      0.2855  -31.190   < 2e-16  ***
 I(log(dbh))     1.8654      0.1604   11.632  4.35e-14  ***
 I(log(haut))    0.7083      0.2097    3.378   0.00170  **
```

avec un écart-type résiduelle de 0,4104 et $R^2 = 0{,}9725$. Le modèle est hautement significatif (test de Fisher: $F_{2,38} = 671{,}5$, p-value $< 2{,}2 \times 10^{-16}$). Le modèle, dont tous les coefficients sont significativement différents de zéro, s'écrit: $\ln(B) = -8{,}9050 + 1{,}8654 \ln(D) + 0{,}7083 \ln(H)$. En appliquant la fonction exponentielle pour revenir aux données de départ, le modèle devient: $B = 1{,}357 \times 10^{-4} D^{1{,}8654} H^{0{,}7083}$. L'exposant associé à la hauteur vaut un peu moins de la moitié de celui associé au diamètre, et est un peu plus faible que l'exposant 0,87238 qui avait été trouvé pour la variable synthétique D^2H (cf. fil rouge n° 8). L'examen des résidus:

```
plot(m,which=1:2)
```

ne révèle rien de particulier (figure 6.7).

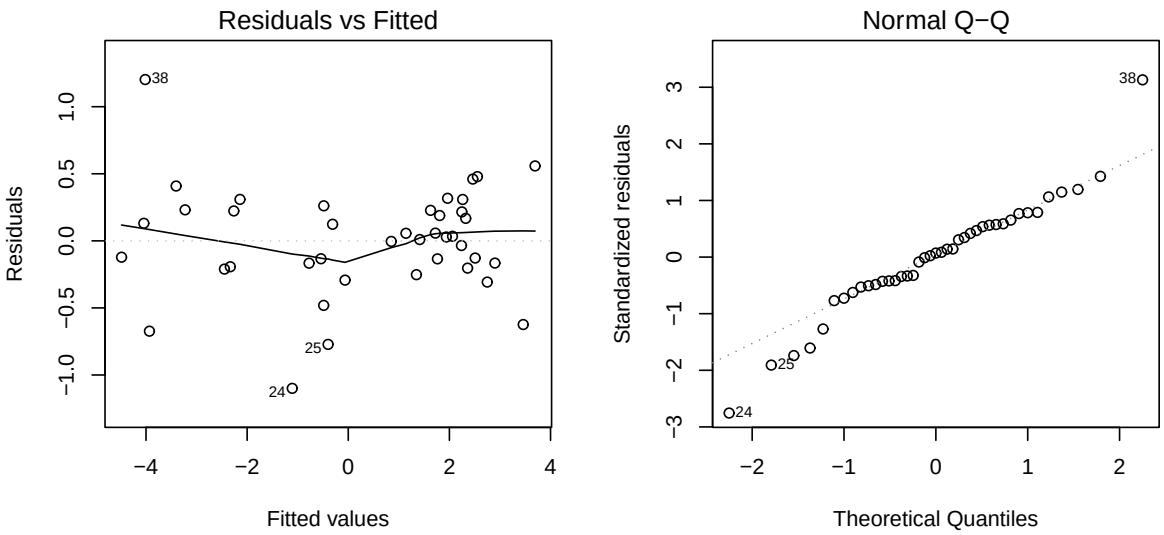


FIGURE 6.7 – *Graphique des résidus en fonction des valeurs prédites (gauche) et graphe quantile–quantile (droite) des résidus de la régression multiple de* $\ln(B)$ *par rapport à* $\ln(D)$ *et* $\ln(H)$ *ajustée aux 42 arbres mesurés par Henry et al. (2010) au Ghana.*

6.1.3 Régression pondérée

Supposons à présent que l'on veuille ajuster directement sur la biomasse B un modèle polynômial par rapport au diamètre D. Par exemple pour un polynôme de degré 2:

$$B = a_0 + a_1 D + a_2 D^2 + \varepsilon \tag{6.9}$$

Comme évoqué précédemment, la biomasse a presque toujours (si ce n'est toujours...) une variabilité qui augmente avec le diamètre D de l'arbre. Autrement dit, la variance de ε augmente avec D, en contradiction avec l'hypothèse d'homoscédasticité requise par la régression multiple. On ne pourrait donc pas ajuster le modèle (6.9) en faisant une régression

multiple. La transformation logarithmique permet de stabiliser la variance résiduelle (nous y reviendrons dans le paragraphe 6.1.5). En prenant $\ln(B)$ comme variable réponse, le modèle à ajuster devient:

$$\ln(B) = \ln(a_0 + a_1 D + a_2 D^2) + \varepsilon \tag{6.10}$$

Il est raisonnable de supposer que la variance des résidus d'un tel modèle est bien constante. Mais malheureusement, ce n'est plus un modèle linéaire puisque la dépendance de la variance réponse vis-à-vis des coefficients a_0, a_1 et a_2 n'est pas linéaire. On ne peut donc pas ajuster le modèle (6.10) à l'aide d'un modèle linéaire. Nous verrons plus loin (§ 6.2) comment ajuster ce modèle non-linéaire.

La régression pondérée permet d'ajuster un modèle tel que (6.9) dont la variance des résidus n'est pas constante, tout en s'appuyant sur le formalisme du modèle linéaire. On peut la voir comme une extension de la régression multiple au cas où la variance des résidus n'est pas constante. La régression pondérée s'est développée en foresterie à partir des années 1960 et jusque dans les années 1980, en particulier grâce aux travaux de Cunia (1964, 1987*a*). Elle a été largement utilisée pour ajuster les tarifs linéaires (Whraton et Cunia, 1987; Brown *et al.*, 1989; Parresol, 1999), avant d'être remplacée par des méthodes d'ajustement plus efficaces que nous verrons au paragraphe 6.1.4.

La régression pondérée s'écrit de façon identique à la régression multiple (6.3):

$$Y = a_0 + a_1 X_1 + a_2 X_2 + \ldots + a_p X_p + \varepsilon$$

à cela près que l'on ne suppose plus que la variance des résidus est constante. Chaque observation a à présent sa propre variance résiduelle σ_i^2:

$$\varepsilon_i \sim \mathcal{N}(0,\, \sigma_i)$$

À chaque observation est associé un *poids* positif w_i (d'où l'adjectif « pondéré » qualifiant cette régression), qui est inversement proportionnel à la variance résiduelle:

$$w_i \propto 1/\sigma_i^2$$

Le coefficient de proportionnalité entre w_i et $1/\sigma_i^2$ n'a pas à être précisé, car la méthode est en fait insensible à toute renormalisation des poids (comme on le verra au paragraphe suivant). Le fait d'associer à chaque observation un poids qui est inversement proportionnel à sa variance est assez naturel. Une observation qui a une forte variance résiduelle s'interprète comme une observation qui a une forte variabilité intrinsèque, et il est donc naturel qu'elle ait moins de poids dans l'ajustement du modèle. Comme on ne peut pas estimer n poids à partir de n observations, il faut modéliser la pondération. Pour des données biologiques telles que la biomasse ou le volume, l'hétéroscédasticité des résidus correspond presque toujours à une relation puissance entre la variance résiduelle et la taille des arbres. On supposera donc que, parmi les p variables explicatives de la régression pondérée, il y en a une (typiquement, le diamètre des arbres) telle que σ_i est une fonction puissance de cette variable. Sans perte de généralité, on peut poser que cette variable est X_1, de sorte que:

$$\sigma_i = k\, X_{i1}^c$$

avec $k > 0$ et $c \geq 0$. Par conséquent:

$$w_i \propto X_{i1}^{-2c}$$

L'exposant c ne peut pas être estimé au même titre que a_0, a_1, …, a_p, mais doit être fixé *a priori*. C'est le principal inconvénient de cette méthode d'ajustement. Nous verrons plus loin comment choisir la valeur de l'exposant c. En revanche le coefficient multiplicateur k n'est pas à estimer puisque les poids w_i ne sont définis qu'à un facteur multiplicateur près. En pratique, on pourra donc poser $w_i = X_{i1}^{-2c}$.

Estimation des coefficients

La méthode des moindres carrés est ajustée de manière à tenir compte de la pondération des observations. On parle alors de la méthode des moindres carrés pondérés. Pour un exposant c fixé, les estimations des coefficients $a_0, \ldots, a_p$ sont les valeurs qui minimisent la somme pondérée des carrés des écarts:

$$\mathrm{SCE}(a_0, a_1, \ldots, a_p) = \sum_{i=1}^{n} w_i \, \varepsilon_i^2 = \sum_{i=1}^{n} w_i (Y_i - a_0 - a_1 X_{i1} - \ldots - a_p X_{ip})^2$$

soit en écriture matricielle:

$$\mathrm{SCE}(\mathbf{a}) = {}^{\mathrm{t}}(\mathbf{Y} - \hat{\mathbf{Y}})\mathbf{W}(\mathbf{Y} - \hat{\mathbf{Y}}) = {}^{\mathrm{t}}(\mathbf{Y} - \mathbf{X}\mathbf{a})\mathbf{W}(\mathbf{Y} - \mathbf{X}\mathbf{a})$$

où $\mathbf{W}$ est la matrice diagonale $n \times n$ ayant w_i sur sa diagonale:

$$\mathbf{W} = \begin{bmatrix} w_1 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & w_n \end{bmatrix}$$

Le minimum de SCE est obtenu pour (Magnus et Neudecker, 2007):

$$\hat{\mathbf{a}} = \arg\min_{\mathbf{a}} \mathrm{SCE}(\mathbf{a}) = ({}^{\mathrm{t}}\mathbf{X}\mathbf{W}\mathbf{X})^{-1\,\mathrm{t}}\mathbf{X}\mathbf{W}\mathbf{Y}$$

Ce minimum ne change pas quand les poids w_i sont tous multipliés par le même scalaire, ce qui prouve bien que la méthode n'est pas sensible à la normalisation des poids. On peut s'assurer que l'estimation par la méthode des moindres carrés pondérés appliquées aux observations X_{ij} et Y_i donne le même résultat que l'estimation par la méthode des moindres carrés ordinaires appliquées aux observations $\sqrt{w_i}\, X_{ij}$ et $\sqrt{w_i}\, Y_i$. Comme précédemment, un avantage de cette méthode d'ajustement est que les estimations des coefficients ont une expression explicite.

Interprétation des résultats et vérification des hypothèses

L'interprétation des résultats de la régression pondérée se fait exactement de la même façon que ceux de la régression multiple. Quant à la vérification des hypothèses relatives aux résidus, c'est également la même chose, à cela près que les résidus sont remplacés par les résidus pondérés $\varepsilon'_i = \sqrt{w_i}\, \varepsilon_i = \varepsilon_i / X_i^c$. Il faut s'assurer que le graphe des résidus pondérés ε'_i en fonction des valeurs prédites ne présente pas de tendance particulière (comme dans la figure 6.8B). Si le nuage de points des résidus *versus* valeurs prédites présente une forme d'entonnoir qui s'ouvre vers la droite (comme dans la figure 6.8A), c'est que la valeur de l'exposant c est trop faible (la valeur la plus faible possible étant zéro). Si le nuage de points présente une forme d'entonnoir qui se ferme vers la droite (comme dans la figure 6.8C), c'est que la valeur de l'exposant c est trop forte.

Choix de la pondération

Un point crucial de la régression pondérée est le choix *a priori* de la valeur de l'exposant c qui définit la pondération. Plusieurs méthodes peuvent être utilisées pour déterminer c. Une première méthode consiste à procéder par tâtonnement, en fonction de l'allure du graphe des résidus pondérés en fonction des valeurs prédites. Puisque l'allure du graphe renseigne sur la pertinence de la valeur de c (figure 6.8), il suffit de tester plusieurs valeurs de c

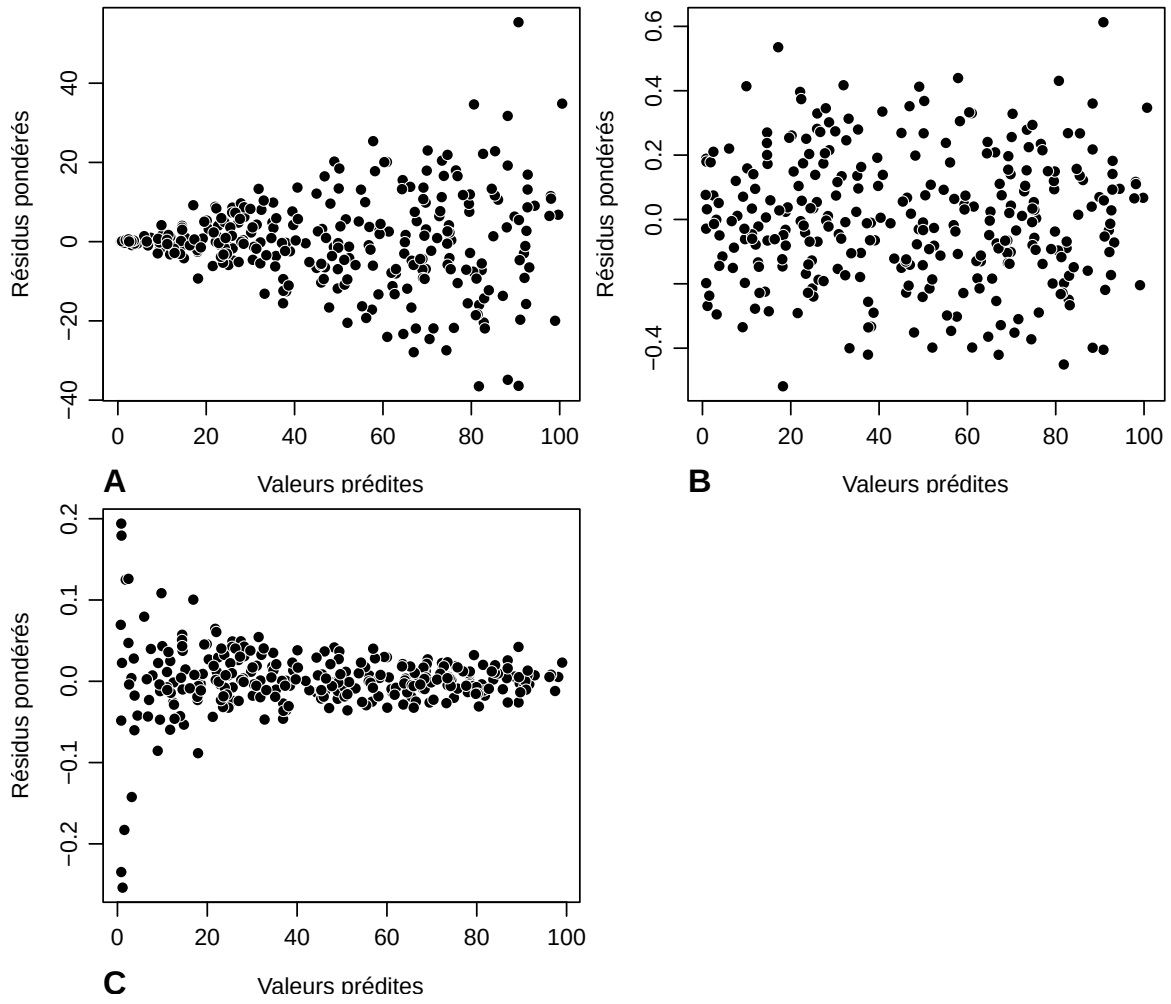


Figure 6.8 – *Graphe des résidus pondérés en fonction des valeurs prédites pour une régression pondérée: (A) la valeur de l'exposant c pour la pondération est trop faible; (B) la valeur de l'exposant c est appropriée; (C) la valeur de l'exposant c est trop forte. On remarquera qu'au fur et à mesure que l'exposant c augmente, le rang de valeurs des résidus pondérés* ε/X^c *diminue.*

jusqu'à ce que le nuage de points des résidus pondérés en fonction des valeurs prédites ne présente plus de tendance particulière. Comme la régression linéaire est robuste vis-à-vis de l'hypothèse de variance constante des résidus, il n'est pas nécessaire de déterminer c avec une grande précision. Le plus souvent, il suffit simplement de tester les valeurs entières de c. Concrètement, on pourra ajuster la régression pondérée pour c valant 0, 1, 2, 3 ou 4 (il est rarement utile d'aller au-delà de 4), et retenir la valeur entière qui assure la meilleure allure au nuage de points des résidus pondérés en fonction des valeurs prédites. Cette méthode simple est le plus souvent amplement suffisante.

Si l'on veut obtenir une valeur plus précise de l'exposant c, on peut procéder en calculant approximativement la variance conditionnelle de la variable réponse Y sachant X_1:

1. découper X_1 en K classes centrées sur X_{1k} ($k = 1, \ldots, K$);
2. calculer la variance empirique, σ_k^2, de Y pour les observations appartenant à la classe k (avec $k = 1, \ldots, K$);
3. faire une régression linéaire de $\ln(\sigma_k)$ par rapport à $\ln(X_{1k})$.

La pente de cette régression est une estimation de c.

Une troisième façon d'estimer c consiste à chercher la valeur de c qui minimise l'indice de Furnival (1961). Cet indice est défini page 160.

11

Régression linéaire pondérée entre B et D^2H

L'analyse exploratoire de la relation entre la biomasse et D^2H a montré (fil rouge n° 3) que cette relation était linéaire avec une variance de la biomasse qui augmentait avec D^2H. On peut donc ajuster une régression pondérée de la biomasse B par rapport à D^2H:

$$B = a + bD^2H + \varepsilon$$

avec

$$\mathrm{Var}(\varepsilon) \propto D^{2c}$$

La régression linéaire est ajustée par moindres carrés pondérés, ce qui nécessite de connaître *a priori* la valeur de l'exposant c.

Cherchons tout d'abord à estimer le coefficient c pour la pondération des observations. Pour cela, on va répartir les observations en classes de diamètre et calculer l'écart-type de la biomasse dans chaque classe de diamètre:

```
D <- quantile(dat$dbh,(0:5)/5)
i <- findInterval(dat$dbh,D,rightmost.closed=TRUE)
sdB <- data.frame(D=(D[-1]+D[-6])/2,sdB=tapply(dat$Btot,i,sd))
```

L'objet `D` contient les bornes des classes de diamètre, calculées de sorte à avoir 5 classes contenant approximativement le même nombre d'observations. L'objet `i` contient le numéro de la classe de diamètre à laquelle appartient chaque observation. La figure 6.9, obtenue par la commande:

```
with(sdB,plot(D,sdB,log="xy",xlab="Diamètre (cm)",ylab=
"Ecart-type de la biomasse (t)"))
```

montre l'écart-type de la biomasse en fonction du diamètre médian de chaque classe de diamètre, en coordonnées logarithmiques. Les points s'alignent à peu près le long d'une droite, ce qui confirme que le modèle puissance est approprié pour modéliser la variance

résiduelle. La régression linéaire du logarithme de l'écart-type de la biomasse par rapport au logarithme du diamètre médian des classes, ajustée par la commande:

```
summary(lm(log(sdB)~I(log(D)),data=sdB))
```

donne:

```
             Estimate  Std. Error  t value  Pr(>|t|)
 (Intercept)  -7.3487      0.7567   -9.712   0.00232  **
 I(log(D))     2.0042      0.1981   10.117   0.00206  **
```

La pente de la régression est égale à $c = 2$. Ainsi l'écart-type σ de la biomasse est approximativement proportionnel à D^2, et on prendra une pondération des observations inversement proportionnelle à D^4.

L'ajustement de la régression pondérée de la biomasse B par rapport à D^2H avec cette pondération, obtenue par la commande:

```
m <- lm(Btot~I(dbh^2*haut),data=dat,weights=1/dat$dbh^4)
summary(m)
```

donne:

```
                Estimate  Std. Error  t value  Pr(>|t|)
 (Intercept)   1.181e-03   2.288e-03    0.516     0.608
 I(dbh^2*haut) 2.742e-05   1.527e-06   17.957    <2e-16  ***
```

Un examen du résultat de cet ajustement montre que l'ordonnée à l'origine n'est pas significativement différente de zéro. On est donc amené à ajuster une nouvelle régression pondérée de la biomasse B par rapport à D^2H sans ordonnée à l'origine:

```
m <- lm(Btot~-1+I(dbh^2*haut),data=dat,weights=1/dat$dbh^4)
summary(m)
```

ce qui donne:

```
                Estimate  Std. Error  t value  Pr(>|t|)
 I(dbh^2*haut) 2.747e-05   1.511e-06    18.19    <2e-16  ***
```

Le modèle s'écrit donc: $B = 2{,}747 \times 10^{-5} D^2 H$, avec un R^2 de 0,8897 et un écart-type résiduel de $k = 0{,}0003513$ tonnes cm^{-2}. Le modèle est hautement significatif (test de Fisher: $F_{1,41} = 330{,}8$, p-value $< 2{,}2 \times 10^{-16}$). Comme ce modèle a été ajusté directement sur les données non-transformées, on remarquera qu'il n'a pas été nécessaire de retirer les observations avec une biomasse nulle (contrairement au fil rouge n° 8). Le graphique 6.10A, obtenu par la commande:

```
plot(fitted(m),residuals(m)/dat$dbh^2,xlab="Valeurs prédites",ylab=
"Résidus pondérés")
```

montre les résidus pondérés en fonction des valeurs prédites. À titre de comparaison, la figure 6.10B montre les résidus pondérés en fonction des valeurs prédites si la pondération avait été trop faible (avec des poids inversement proportionnel à D^2):

```
m <- lm(Btot~-1+I(dbh^2*haut),data=dat,weights=1/dat$dbh^2)
plot(fitted(m),residuals(m)/dat$dbh,xlab="Valeurs prédites",ylab="Résidus pondérés")
```

tandis que la figure 6.10C montre les résidus pondérés en fonction des valeurs prédites si la pondération avait été trop forte (avec des poids inversement proportionnel à D^5):

```
m <- lm(Btot~-1+I(dbh^2*haut),data=dat,weights=1/dat$dbh^5)
plot(fitted(m),residuals(m)/dat$dbh^2.5,xlab="Valeurs prédites",ylab=
"Résidus pondérés")
```

Ainsi le coefficient $c = 2$ pour la pondération s'avère bien être celui qui convient.

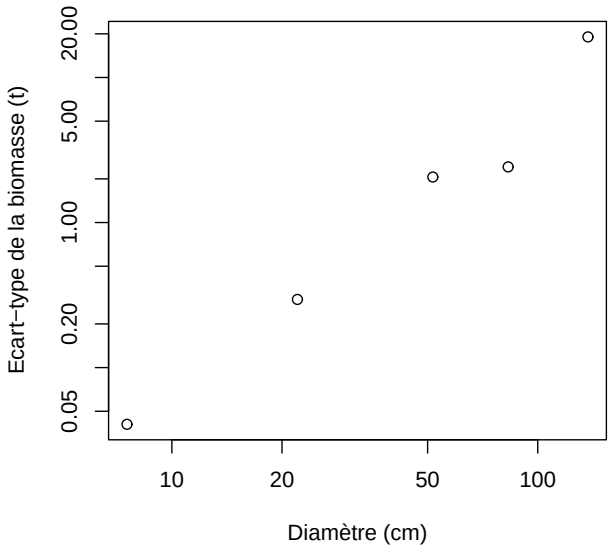


FIGURE 6.9 – *Écart-type de la biomasse calculé dans cinq classes de diamètre en fonction du diamètre médian de la classe (en coordonnées logarithmiques), pour 42 arbres mesurés au Ghana par Henry et al. (2010).*

12

Régression polynômiale pondérée entre B et D

L'analyse exploratoire (fil rouge n° 2) a montré que la relation entre la biomasse et le diamètre avait une forme parabolique, avec une augmentation de la variance de la biomasse avec le diamètre. La transformation logarithmique permet de linéariser la relation entre la biomasse et le diamètre, mais on peut également chercher à modéliser directement la relation entre la biomasse et le diamètre par une fonction parabolique:

$$B = a_0 + a_1 D + a_2 D^2 + \varepsilon$$

avec

$$\mathrm{Var}(\varepsilon) \propto D^{2c}$$

Dans le fil rouge n° 11, nous avons vu que la valeur $c = 2$ de l'exposant convenait pour modéliser l'écart-type conditionnel de la biomasse sachant le diamètre. On va donc ajuster la régression multiple par moindres carrés pondérés avec une pondération des observations proportionnelle à $1/D^4$:

```
m <- lm(Btot~dbh+I(dbh^2),data=dat,weights=1/dat$dbh^4)
summary(m)
```

ce qui donne:

```
              Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)  1.127e-02   6.356e-03    1.772   0.08415  .
dbh         -7.297e-03   2.140e-03   -3.409   0.00153  **
I(dbh^2)     1.215e-03   9.014e-05   13.478  2.93e-16  ***
```

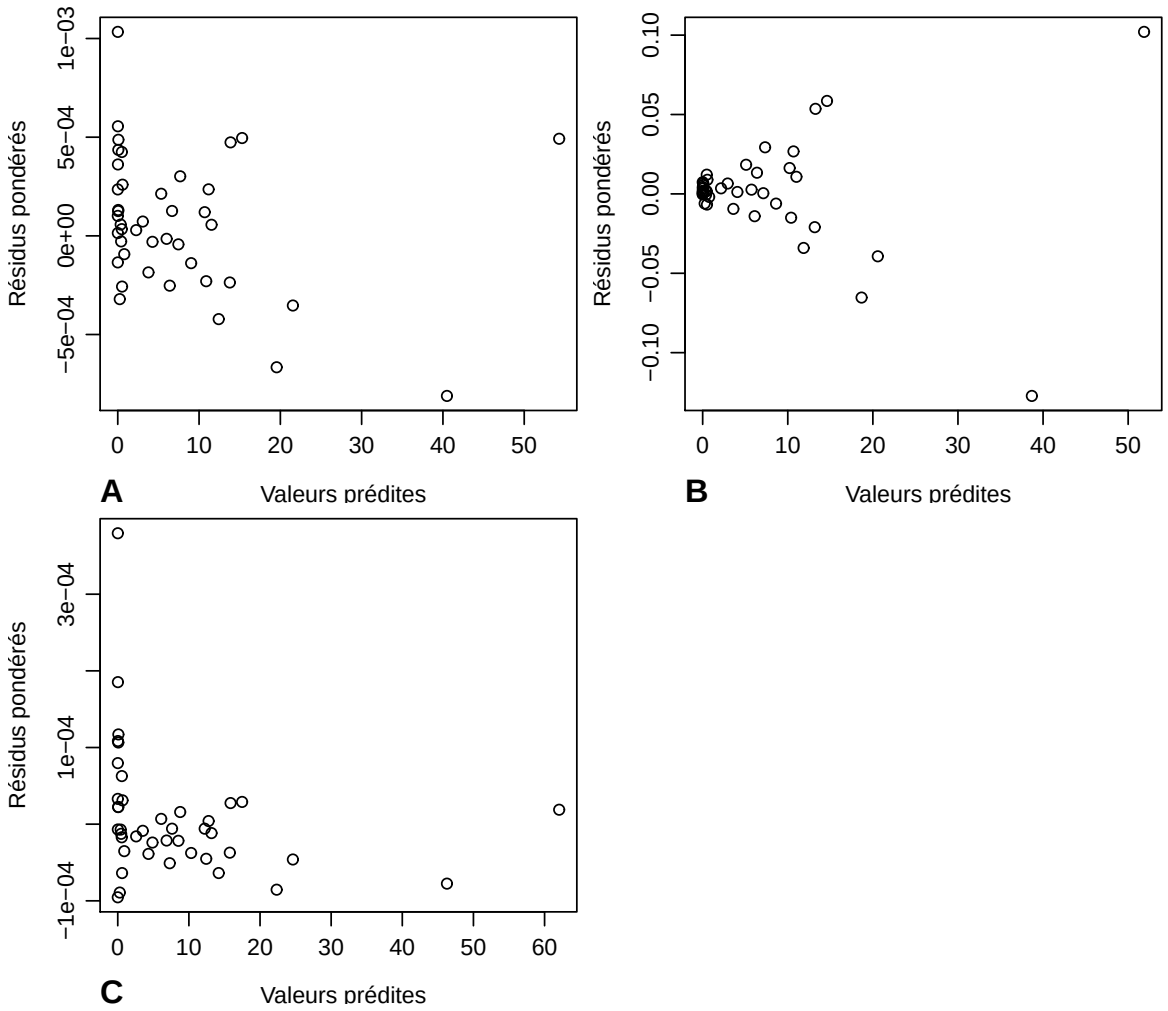


Figure 6.10 – *Graphique des résidus pondérés en fonction des valeurs prédites pour la régression pondérée de la biomasse par rapport à D^2H pour 42 arbres mesurés au Ghana par Henry et al. (2010): (A) la pondération est inversement proportionnelle à D^4; (B) la pondération est inversement proportionnelle à D^2; (C) la pondération est inversement proportionnelle à D^5.*

avec un écart-type résiduel $k = 0{,}0003882$ tonnes cm^{-2} et $R^2 = 0{,}8709$. L'ordonnée à l'origine s'avère ne pas être significativement différente de zéro. On va donc ajuster à nouveau une fonction parabolique mais sans ordonnée à l'origine:

```
m <- lm(Btot~-1+dbh+I(dbh^2),data=dat,weights=1/dat$dbh^4)
summary(m)
```

ce qui donne:

```
             Estimate  Std. Error  t value  Pr(>|t|)
dbh       -3.840e-03   9.047e-04   -4.245  0.000126  ***
I(dbh^2)   1.124e-03   7.599e-05   14.789   < 2e-16  ***
```

avec un écart-type résiduel $k = 0{,}0003985$ tonnes cm^{-2} et $R^2 = 0{,}8615$. Le modèle est hautement significatif (test de Fisher: $F_{2,40} = 124{,}4$, p-value $= 2{,}2 \times 10^{-16}$) et s'écrit: $B = -3{,}840 \times 10^{-3} D + 1{,}124 \times 10^{-3} D^2$. Le graphique 6.11 obtenu par la commande:

```
plot(fitted(m),residuals(m)/dat$dbh^2,xlab="Valeurs prédites",ylab=
"Résidus pondérés")
```

montre les résidus pondérés en fonction des valeurs prédites.

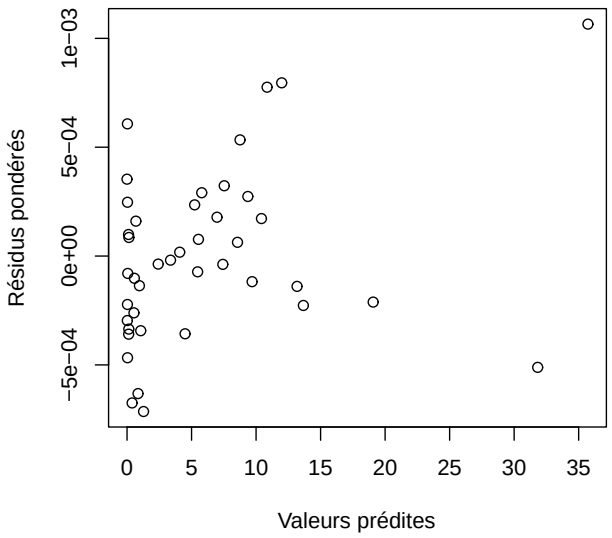


FIGURE 6.11 – *Graphique des résidus pondérés en fonction des valeurs prédites pour la régression pondérée de la biomasse par rapport à D et D^2 pour 42 arbres mesurés au Ghana par Henry et al. (2010).*

6.1.4 Régression linéaire avec modèle sur la variance

Une alternative à la régression pondérée consiste à poser explicitement un modèle pour la variance des résidus. Comme précédemment, il est réaliste de poser qu'il existe une variable explicative (sans perte de généralité, la première) telle que l'écart-type résiduel est une fonction puissance de cette variable:

$$\mathrm{Var}(\varepsilon) = (kX_1^c)^2 \tag{6.11}$$

avec $k > 0$ et $c \geq 0$. Le modèle s'écrit donc:

$$Y = a_0 + a_1 X_1 + a_2 X_2 + \ldots + a_p X_p + \varepsilon \tag{6.12}$$

avec:

$$\varepsilon \sim \mathcal{N}(0,\, kX_1^c)$$

Dans la forme, le modèle n'est guère différent de la régression pondérée. Dans le fond, il y a une différence fondamentale: les coefficients k et c sont à présent des paramètres du modèle à estimer, au même titre que les coefficients a_0, a_1, …, a_p. Du fait de ces paramètres k et c à estimer, la méthode des moindres carrés ne peut plus être utilisée pour l'estimation des coefficients du modèle. Il faut utiliser une autre méthode d'estimation, à savoir la méthode du maximum de vraisemblance. *Stricto sensu*, le modèle défini par (6.11) et (6.12) ne relève pas du modèle linéaire. Il est conceptuellement beaucoup plus proche du modèle non-linéaire que nous verrons dans la section 6.2. Nous n'irons pas plus avant ici dans la présentation du modèle non-linéaire: la méthode d'ajustement du modèle défini par (6.11) et (6.12) sera présentée comme un cas particulier du modèle non-linéaire présenté dans la section 6.2.

13

Régression linéaire entre B et D^2H avec modèle sur la variance

En anticipant sur la section 6.2, on va ajuster une régression linéaire de la biomasse par rapport à D^2H en spécifiant un modèle puissance sur la variance résiduelle:

$$B = a + bD^2H + \varepsilon$$

avec

$$\mathrm{Var}(\varepsilon) = (kD^c)^2$$

On verra plus loin (§ 6.2) que ce modèle est ajusté par maximum de vraisemblance. Cette régression est dans l'esprit très semblable à la régression pondérée de la biomasse par rapport à D^2H réalisée précédemment (fil rouge n° 11), à cela près que l'exposant c utilisé pour définir la pondération des observations est à présent un paramètre à estimer à part entière et non plus un coefficient posé *a priori*. La régression linéaire avec modèle sur la variance s'ajuste de la façon suivante:

```
library(nlme)
start <- coef(lm(Btot~I(dbh^2*haut),data=dat))
names(start) <- c("a","b")
summary(nlme(Btot~a+b*dbh^2*haut, data=cbind(dat,g="a"), fixed=a+b~1, start=start,
groups=~g, weights=varPower(form=~dbh)))
```

et donne (on reviendra dans la section 6.2 sur la signification de l'objet `start`):

```
     Value     Std.Error  DF    t-value  p-value
a 0.0012868020 0.0024211610 40  0.531481   0.598
b 0.0000273503 0.0000014999 40 18.234340   0.000
```

avec une valeur estimée de l'exposant $c = 1{,}977736$. Comme dans la régression linéaire pondérée (fil rouge n° 11), l'ordonnée à l'origine s'avère ne pas être significativement différente de zéro. On réajuste donc le modèle sans ordonnée à l'origine:

```
summary(nlme(Btot~b*dbh^2*haut, data=cbind(dat,g="a"), fixed=b~1, start=start["b"],
groups=~g, weights=varPower(form=~dbh)))
```

ce qui donne:

```
           Value    Std.Error  DF    t-value  p-value
b  2.740688e-05  1.4869e-06   41  18.43223        0
```

avec une valeur estimée de l'exposant $c = 1{,}980263$. Cette valeur est très proche de celle évaluée pour la régression linéaire pondérée ($c = 2$ dans le fil rouge n° 11). Le modèle ajusté s'écrit donc: $B = 2{,}740688 \times 10^{-5} D^2 H$, ce qui est très proche du modèle ajusté par régression linéaire pondérée (fil rouge n° 11).

14

Régression polynômiale entre B et D avec modèle sur la variance

En anticipant sur la section 6.2, on va ajuster une régression multiple de la biomasse par rapport à D et D^2 en spécifiant un modèle puissance sur la variance résiduelle:

$$B = a_0 + a_1 D + a_2 D^2 + \varepsilon$$

avec

$$\mathrm{Var}(\varepsilon) = (kD^c)^2$$

On verra plus loin (§ 6.2) que ce modèle est ajusté par maximum de vraisemblance. Cette régression est dans l'esprit très semblable à la régression polynômiale de la biomasse par rapport à D et D^2 réalisée précédemment (fil rouge n° 12), à cela près que l'exposant c utilisé pour définir la pondération des observations est à présent un paramètre à estimer à part entière et non plus un coefficient posé *a priori*. La régression linéaire avec modèle sur la variance s'ajuste de la façon suivante:

```
library(nlme)
start <- coef(lm(Btot~dbh+I(dbh^2),data=dat))
names(start) <- c("a0","a1","a2")
summary(nlme(Btot~a0+a1*dbh+a2*dbh^2,data=cbind(dat,g="a"),fixed=a0+a1+a2~1,
start=start,groups=~g,weights=varPower(form=~dbh)))
```

et donne (on reviendra dans la section 6.2 sur la signification de l'objet `start`):

```
            Value     Std.Error  DF     t-value  p-value
a0    0.009048498  0.005139129  39    1.760706   0.0861
a1   -0.006427411  0.001872346  39   -3.432812   0.0014
a2    0.001174388  0.000094063  39   12.485081   0.0000
```

avec une valeur estimée de l'exposant $c = 2{,}127509$. Comme dans la régression polynômiale pondérée (fil rouge n° 12), l'ordonnée à l'origine s'avère ne pas être significativement différente de zéro. On réajuste donc le modèle sans ordonnée à l'origine:

```
summary(nlme(Btot~a1*dbh+a2*dbh^2,data=cbind(dat,g="a"),fixed=a1+a2~1,start=start[
c("a1","a2")],groups=~g,weights=varPower(form=~dbh)))
```

ce qui donne:

```
            Value      Std.Error  DF     t-value  p-value
a1   -0.003319456  0.0006891736  40   -4.816574        0
a2    0.001067068  0.0000759745  40   14.045082        0
```

avec une valeur estimée de l'exposant $c = 2{,}139967$. Cette valeur est très proche de celle évaluée pour la régression polynômiale pondérée ($c = 2$ dans le fil rouge n° 12). Le modèle

ajusté s'écrit donc: $B = -3{,}319456 \times 10^{-3} D + 1{,}067068 \times 10^{-3} D^2$, ce qui est très proche du modèle ajusté par régression polynômiale pondérée (fil rouge n° 12).

6.1.5 Transformation de variable

Reprenons l'exemple du tarif de biomasse à une entrée (en l'occurrence le diamètre) de type puissance:

$$B = aD^b \tag{6.13}$$

On a déjà vu qu'il s'agit d'un modèle non-linéaire puisque B dépend non-linéairement des coefficients a et b. En revanche, on peut linéariser ce modèle en appliquant la transformation logarithmique. La relation (6.13) est équivalente à: $\ln(B) = \ln(a) + b\ln(D)$, que l'on peut voir comme une régression linéaire de la variable réponse $Y = \ln(B)$ par rapport à la variable explicative $X = \ln(D)$. On peut donc estimer les coefficients a et b (ou plutôt $\ln(a)$ et b) du modèle puissance (6.13) par régression linéaire sur les données log-transformées. Qu'en est-il de l'erreur résiduelle ? Si la régression linéaire sur les données log-transformées est pertinente, cela signifie que $\varepsilon = \ln(B) - \ln(a) - b\ln(D)$ suit une loi normale centrée et d'écart-type constant σ. Si on revient aux données de départ en utilisant la transformation exponentielle (qui est la transformation inverse de la transformation logarithmique), l'erreur résiduelle se retrouve en facteur:

$$B = aD^b \times \varepsilon'$$

avec $\varepsilon' = \exp(\varepsilon)$. Ainsi, on est passé d'une erreur additive sur les données log-transformées à une erreur multiplicative sur les données de départ. De plus, si ε suit une loi normale centrée et d'écart-type σ, alors, par définition, $\varepsilon' = \exp(\varepsilon)$ suit une log-normale de paramètres 0 et σ:

$$\varepsilon' \underset{\text{i.i.d.}}{\sim} \mathcal{LN}(0,\, \sigma)$$

Contrairement à ε dont la moyenne est nulle, la moyenne de ε' n'est pas nulle mais vaut: $\mathrm{E}(\varepsilon') = \exp(\sigma^2/2)$. Nous verrons au chapitre 7 les implications de cela.

Il y a deux leçons à retenir de cet exemple:

1. quand on est confronté à une relation non-linéaire entre une variable réponse et une (ou plusieurs) variable(s) explicative(s), une transformation de variable peut permettre de rendre linéaire cette relation;
2. la transformation de variable affecte non seulement la forme de la relation entre la (ou les) variable(s) explicative(s) et la variable réponse, mais aussi l'erreur résiduelle.

À propos du premier point, la transformation de variables amène à distinguer deux approches pour ajuster un modèle non-linéaire. La première approche consiste, lorsque l'on est confronté à une relation non-linéaire entre une variable réponse et des variables explicatives, à rechercher une transformation qui linéarise cette relation, pour se ramener au cas du modèle linéaire. La deuxième approche consiste à ajuster directement le modèle non-linéaire, comme nous le verrons dans la section 6.2. Chaque approche a ses avantages et inconvénients. Le modèle linéaire a l'avantage de fournir un cadre théorique relativement simple et, surtout, les estimations de ses coefficients ont des expressions explicites. L'inconvénient est que l'étape de linéarisation du modèle introduit une difficulté supplémentaire et que la transformation inverse, si on n'y prend pas garde, peut introduire un biais de prédiction (nous y reviendrons au chapitre 7). De plus, tous les modèles ne sont pas linéarisables. Par exemple, il n'existe aucune transformation de variable qui permette de linéariser le modèle suivant: $Y = a_0 + a_1 X + a_2 \exp(a_3 X)$.

À propos du second point, on sera donc désormais amené à distinguer la forme de la relation entre la variable réponse et les variables explicatives (on parle aussi de modèle pour la moyenne — sous-entendu la moyenne de la variable réponse Y), et la forme du modèle pour l'erreur résiduelle (on parle aussi de modèle pour la variance — sous-entendu la variance de Y). La transformation de variable affecte les deux simultanément. Tout l'art de la transformation de variable consiste à jouer sur les deux plans simultanément pour rendre le modèle linéaire vis-à-vis de ses coefficients et stabiliser la variance des résidus (c'est-à-dire la rendre constante).

Transformations de variable usuelles

Bien qu'il n'y ait pas de limite théorique aux transformations de variable que l'on peut utiliser, les transformations qui sont susceptibles de concerner des volumes ou des biomasses sont bien plus restreintes. La transformation qui sera le plus fréquemment utilisée pour l'ajustement de tarifs est la transformation logarithmique. Étant donné un modèle puissance:

$$Y = aX_1^{b_1}X_2^{b_2} \times \ldots \times X_p^{b_p} \times \varepsilon$$

la transformation logarithmique consiste à remplacer la variable Y par son logarithme: $Y' = \ln(Y)$, et chacune des variables réponse par son logarithme: $X'_j = \ln(X_j)$. Le modèle résultant est:

$$Y' = a' + b_1X'_1 + b_2X'_2 + \ldots + b_pX'_p + \varepsilon' \tag{6.14}$$

avec $\varepsilon' = \ln(\varepsilon)$. La transformation inverse est l'exponentielle pour l'ensemble des variables (réponse et explicatives). En termes d'erreur résiduelle, la transformation logarithmique est appropriée si ε' a une distribution normale, donc si l'erreur ε est positive et intervient de manière multiplicative. À noter que pour des variables pouvant prendre des valeurs nulles, la transformation logarithmique pose problème. Dans ce cas, on utilise la transformation $X' = \ln(X+1)$ plutôt que $X' = \ln(X)$ (ou plus généralement, $X' = \ln(X + \text{constante})$ si X peut prendre des valeurs négatives, comme un accroissement diamétrique par exemple). À titre d'exemple, les tarifs de biomasse suivants:

$$\begin{aligned} B &= aD^b \\ B &= a(D^2H)^b \\ B &= a\rho^{b_1}D^{b_2}H^{b_3} \end{aligned}$$

sont susceptibles d'être ajustés par régression linéaire suite à une transformation logarithmique des données.

Étant donné un modèle exponentiel:

$$Y = a\exp(b_1X_1 + b_2X_2 + \ldots + b_pX_p) \times \varepsilon \tag{6.15}$$

la transformation appropriée consiste à remplacer la variable Y par son logarithme: $Y' = \ln(Y)$, et à ne pas transformer les variables réponse: $X'_j = X_j$. Le modèle résultant est identique à (6.14). La transformation inverse est l'exponentielle pour la variable réponse, et pas de changement pour les variables explicatives. En termes d'erreur résiduelle, cette transformation est appropriée si ε' a une distribution normale, donc si l'erreur ε est positive et intervient de manière multiplicative. À noter que, sans perte de généralité, on peut reparamétrer les coefficients du modèle exponentiel (6.15) en posant $b'_j = \exp(b_j)$. Une écriture strictement équivalente du modèle exponentiel (6.15) est donc:

$$Y = ab_1'^{X_1}b_2'^{X_2} \times \ldots \times b_p'^{X_p} \times \varepsilon$$

À titre d'exemple, le tarif de biomasse suivant:

$$B = \exp\{a_0 + a_1 \ln(D) + a_2[\ln(D)]^2 + a_3[\ln(D)]^3\}$$

est susceptible d'être ajusté par régression linéaire suite à une transformation de variable de ce type (avec, dans cet exemple, $X_j = [\ln(D)]^j$).

La transformation de Box-Cox généralise la transformation logarithme. C'est en réalité une famille de transformations indexée par un paramètre ξ. Étant donné une variable X, sa transformée de Box-Cox X'_ξ est:

$$X'_\xi = \begin{cases} (X^\xi - 1)/\xi & (\xi \neq 0) \\ \ln(X) = \lim_{\xi \to 0}(X^\xi - 1)/\xi & (\xi = 0) \end{cases}$$

La transformation de Box-Cox permet de convertir la question du choix d'une transformation de variable en une question d'estimation d'un paramètre ξ (Hoeting *et al.*, 1999).

Une transformation de variable particulière

Les transformations de variable usuelles changent la forme de la relation entre la variable réponse et la variable explicative. Quand le nuage de points (X_i, Y_i) de la variable réponse en fonction de la variable explicative a la forme d'une droite avec de l'hétéroscédasticité, comme schématisé dans la figure 6.12, il est nécessaire d'appliquer une transformation de variable pour stabiliser la variance de Y, sans toutefois toucher à la nature linéaire de la relation entre X et Y. Le cas de figure illustré par 6.12 se retrouve assez souvent quand on ajuste une équation allométrique entre deux grandeurs qui varient proportionnellement (cf. par exemple Ngomanda *et al.*, 2012). La nature linéaire de la relation entre X et Y signifie que le modèle est de la forme:

$$Y = a + bX + \varepsilon \tag{6.16}$$

mais l'hétéroscédasticité signifie que la variance de ε n'est pas constante, ce qui empêche d'ajuster une régression linéaire. Une transformation de variable dans cas consiste à remplacer Y par $Y' = Y/X$ et X par $X' = 1/X$. En divisant chaque membre de (6.16) par X, le modèle après transformation de variable devient:

$$Y' = aX' + b + \varepsilon' \tag{6.17}$$

avec $\varepsilon' = \varepsilon/X$. Le modèle transformé correspond toujours à une relation linéaire, à cela près que l'ordonnée à l'origine a de la relation entre X et Y est devenu la pente de la relation entre X' et Y', et réciproquement la pente b de la relation entre X et Y est devenu l'ordonnée à l'origine de la relation entre X' et Y'. Le modèle (6.17) pourra être ajusté par une régression linéaire simple si la variance de ε' est constante. Comme $\mathrm{Var}(\varepsilon') = \sigma^2$ implique $\mathrm{Var}(\varepsilon) = \sigma^2 X^2$, cela sous-entend que la transformation de variable est appropriée si l'écart-type de ε est proportionnel à X.

Le modèle (6.17) étant ajusté par régression linéaire simple, sa somme des carrés des écarts vaut:

$$\mathrm{SCE}(a, b) = \sum_{i=1}^{n}(Y'_i - aX'_i - b)^2 = \sum_{i=1}^{n}(Y_i/X_i - a/X_i - b)^2 = \sum_{i=1}^{n} X_i^{-2}(Y_i - a - bX_i)^2$$

Dans la dernière expression, on reconnaît l'expression de la somme des carrés des écarts pour une régression pondérée utilisant des poids $w_i = X_i^{-2}$. Ainsi, la transformation de

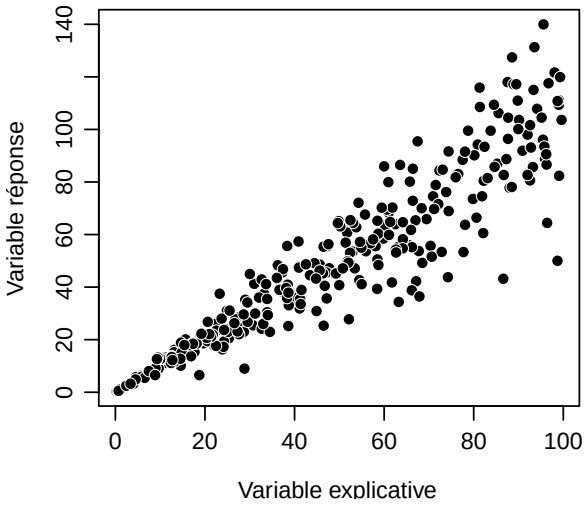


FIGURE 6.12 – *Relation linéaire entre une variable explicative (X) et une variable réponse (Y), avec accroissement de la variabilité de Y lorsque X augmente (hétéroscédasticité).*

variable $Y' = Y/X$ et $X' = 1/X$ est strictement identique à une régression pondérée de poids $w = 1/X^2$.

15

Régression linéaire entre B/D^2 et H

On a vu dans le fil rouge n° 11 qu'un tarif de biomasse à deux entrées par rapport au diamètre et la hauteur était: $B = a + bD^2H + \varepsilon$ avec $\text{Var}(\varepsilon) \propto D^4$. En divisant chaque membre de l'équation par D^2, on obtient:

$$B/D^2 = a/D^2 + bH + \varepsilon'$$

avec

$$\text{Var}(\varepsilon') = \sigma^2$$

Ainsi la régression de la variable réponse $Y = B/D^2$ par rapport aux deux variables explicatives $X_1 = 1/D^2$ et $X_2 = H$ vérifie *a priori* les hypothèses de la régression linéaire multiple. Cette régression est ajustée par moindres carrés ordinaires. L'ajustement de cette régression multiple par la commande:

```
summary(lm((Btot/dbh^2)~-1+I(1/dbh^2)+haut,data=dat))
```

donne:

```
             Estimate  Std. Error  t value  Pr(>|t|)
I(1/dbh^2)  1.181e-03   2.288e-03    0.516     0.608
haut        2.742e-05   1.527e-06   17.957    <2e-16  ***
```

d'où il ressort que le coefficient associé à $X_1 = 1/D^2$ n'est pas significativement différent de zéro. Si on revient aux données de départ, cela signifie simplement que l'ordonnée à l'origine a n'est pas significativement différente de zéro, ce que l'on avait déjà diagnostiqué dans le fil rouge n° 11. On peut donc retirer X_1 et ajuster une régression linéaire simple de $Y = B/D^2$ par rapport à $X_2 = H$:

```
with(dat,plot(haut,Btot/dbh^2,xlab="Hauteur (m)",ylab="Biomasse/carré du diamètre
(t/cm2)"))
m <- lm((Btot/dbh^2)~-1+haut,data=dat)
summary(m)
plot(m,which=1:2)
```

Le nuage de points de B/D^2 en fonction de H a effectivement la forme d'une droite avec une variance de B/D^2 qui est approximativement constante (figure 6.13). L'ajustement de la régression linéaire simple donne:

```
       Estimate  Std. Error  t value  Pr(>|t|)
haut  2.747e-05   1.511e-06    18.19    <2e-16  ***
```

avec un R^2 de 0,8897 et un écart-type résiduel de 0,0003513 tonnes cm^{-2}. Le modèle s'écrit: $B/D^2 = 2{,}747 \times 10^{-5} H$, soit en revenant aux variables de départ: $B = 2{,}747 \times 10^{-5} D^2 H$. On vérifiera que ce modèle est strictement identique à la régression pondérée de B par rapport à D^2H réalisée dans le fil rouge n° 11 avec une pondération proportionnelle à $1/D^4$. Le graphe des résidus en fonction des valeurs prédites et le graphe quantile–quantile des résidus sont représentés dans la figure 6.14.

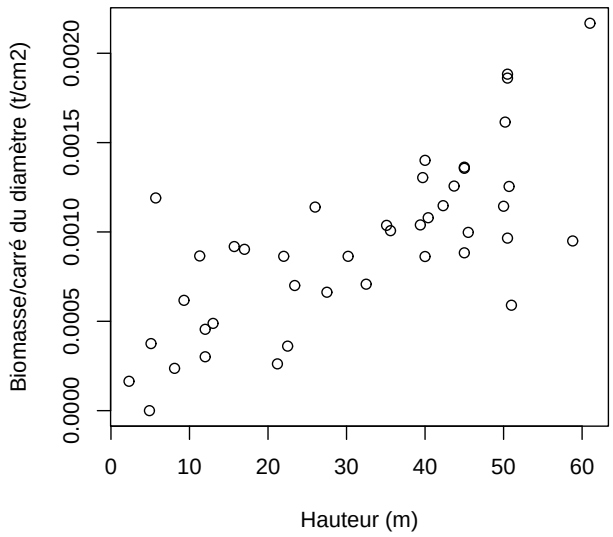


Figure 6.13 – *Nuage de points de la biomasse divisée par le carré du diamètre (tonnes cm^{-2}) en fonction de la hauteur (m) pour 42 arbres mesurés au Ghana par Henry et al. (2010).*

16

Régression linéaire entre B/D^2 et $1/D$

On a vu dans le fil rouge n° 12 qu'un tarif de biomasse polynômial par rapport au diamètre était: $B = a_0 + a_1 D + a_2 D^2 + \varepsilon$ avec $\mathrm{Var}(\varepsilon) \propto D^4$. En divisant chaque membre de l'équation par D^2, on obtient:

$$B/D^2 = a_0/D^2 + a_1/D + a_2 + \varepsilon'$$

avec

$$\mathrm{Var}(\varepsilon') = \sigma^2$$

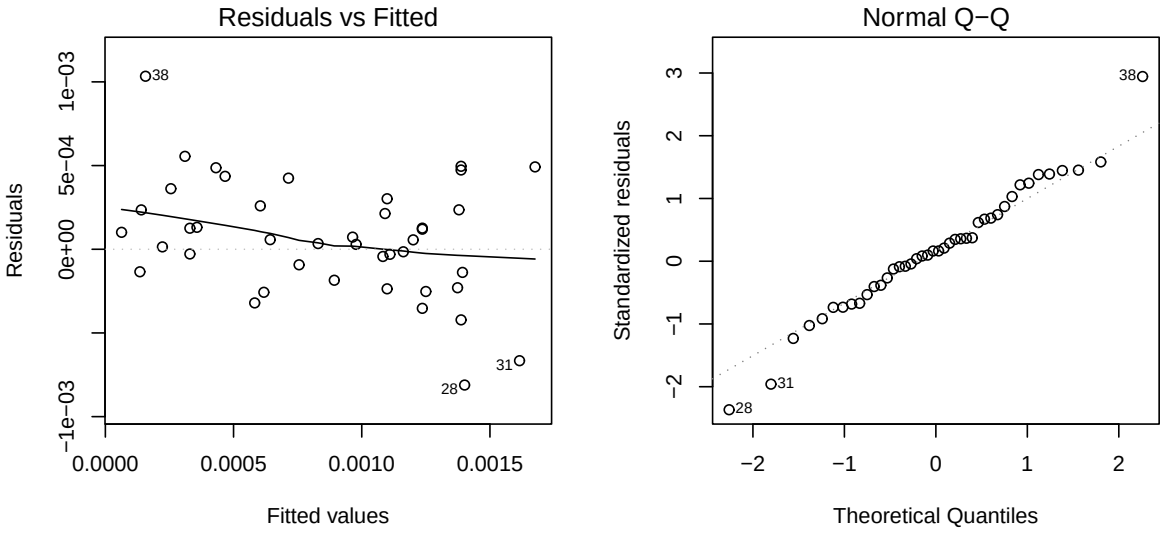


FIGURE 6.14 – *Graphique des résidus en fonction des valeurs prédites (gauche) et graphe quantile–quantile (droite) des résidus de la régression linéaire simple de B/D^2 par rapport à H ajustée aux 42 arbres mesurés par Henry et al. (2010) au Ghana.*

Ainsi la régression de la variable réponse $Y = B/D^2$ par rapport aux deux variables explicatives $X_1 = 1/D^2$ et $X_2 = 1/D$ vérifie *a priori* les hypothèses de la régression linéaire multiple. Cette régression est ajustée par moindres carrés ordinaires. L'ajustement de cette régression multiple par la commande:

```
summary(lm((Btot/dbh^2)~I(1/dbh^2)+I(1/dbh),data=dat))
```

donne:

```
               Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)   1.215e-03   9.014e-05   13.478  2.93e-16  ***
I(1/dbh^2)    1.127e-02   6.356e-03    1.772   0.08415  .
I(1/dbh)     -7.297e-03   2.140e-03   -3.409   0.00153  **
```

d'où il ressort que le coefficient associé à $X_1 = 1/D^2$ n'est pas significativement différent de zéro. Si on revient aux données de départ, cela signifie simplement que l'ordonnée à l'origine a_0 n'est pas significativement différente de zéro, ce que l'on avait déjà diagnostiqué dans le fil rouge n° 12. On peut donc retirer X_1 et ajuster une régression linéaire simple de $Y = B/D^2$ par rapport à $X_2 = 1/D$:

```
with(dat,plot(1/dbh,Btot/dbh^2,xlab="1/diamètre (/cm)",ylab="Biomasse/carré du
diamètre (t/cm2)"))
m <- lm((Btot/dbh^2)~I(1/dbh),data=dat)
summary(m)
plot(m,which=1:2)
```

Le nuage de points de B/D^2 en fonction de $1/D$ a approximativement la forme d'une droite avec une variance de B/D^2 qui est approximativement constante (figure 6.15). L'ajustement de la régression linéaire simple donne:

```
               Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)   1.124e-03   7.599e-05   14.789   < 2e-16  ***
I(1/dbh)     -3.840e-03   9.047e-04   -4.245  0.000126  ***
```

avec un R^2 de 0,3106 et un écart-type résiduel de 0,0003985 tonnes cm^{-2}. Le modèle s'écrit: $B/D^2 = 1{,}124 \times 10^{-3} - 3{,}84 \times 10^{-3} D^{-1}$, soit en revenant aux variables de départ: $B =$

$-3{,}84 \times 10^{-3} D + 1{,}124 \times 10^{-3} D^2$. On vérifiera que ce modèle est strictement identique à la régression polynômiale pondérée de B par rapport à D réalisée dans le fil rouge n° 12 avec une pondération proportionnelle à $1/D^4$. Le graphe des résidus en fonction des valeurs prédites et le graphe quantile–quantile des résidus sont représentés dans la figure 6.16.

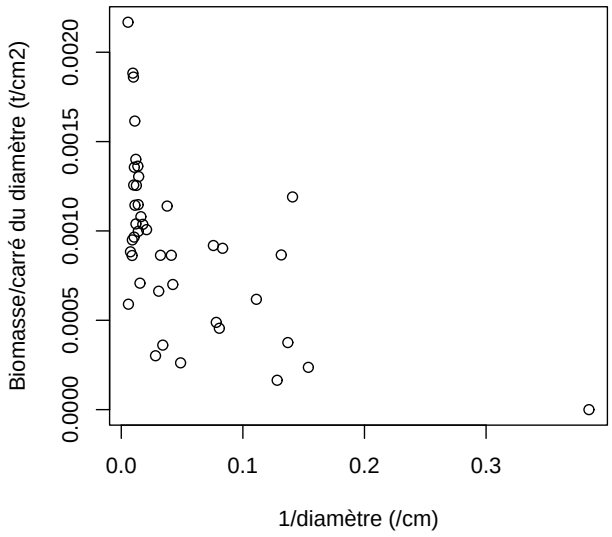


Figure 6.15 – *Nuage de points de la biomasse divisée par le carré du diamètre (tonnes cm^{-2}) en fonction de l'inverse du diamètre (cm^{-1}) pour 42 arbres mesurés au Ghana par Henry et al. (2010).*

6.2 Ajustement d'un modèle non-linéaire

Abordons à présent le cas plus général de l'ajustement d'un modèle non-linéaire. Ce modèle s'écrit:

$$Y = f(X_1, \ldots, X_p; \theta) + \varepsilon$$

où Y est la variable réponse, $X_1, \ldots, X_p$ sont les variables explicatives, θ est le vecteur de l'ensemble des coefficients du modèle, ε est l'erreur résiduelle, et f est une fonction. Si f est linéaire vis-à-vis des coefficients θ, on est ramené au modèle linéaire étudié précédemment. On ne fait désormais aucune hypothèse *a priori* sur la linéarité de la fonction f vis-à-vis des coefficients θ. Comme précédemment, on suppose les résidus indépendants et distribués selon une loi normale centrée. En revanche, on ne fait aucune hypothèse *a priori* sur leur variance. $\mathrm{E}(\varepsilon) = 0$ implique que $\mathrm{E}(Y) = f(X_1, \ldots, X_p; \theta)$. C'est pourquoi l'on dit que f définit le modèle pour la moyenne (sous-entendu: de Y). Posons:

$$\mathrm{Var}(\varepsilon) = g(X_1, \ldots, X_p; \vartheta)$$

où g est une fonction et ϑ un ensemble de paramètres. Comme $\mathrm{Var}(Y) = \mathrm{Var}(\varepsilon)$, on dit que g définit le modèle pour la variance. La fonction g peut prendre des formes diverses, mais pour des données de biomasse ou de volume, elle prend le plus souvent la forme d'une fonction puissance d'une variable caractérisant la taille de l'arbre (typiquement, son diamètre). Sans

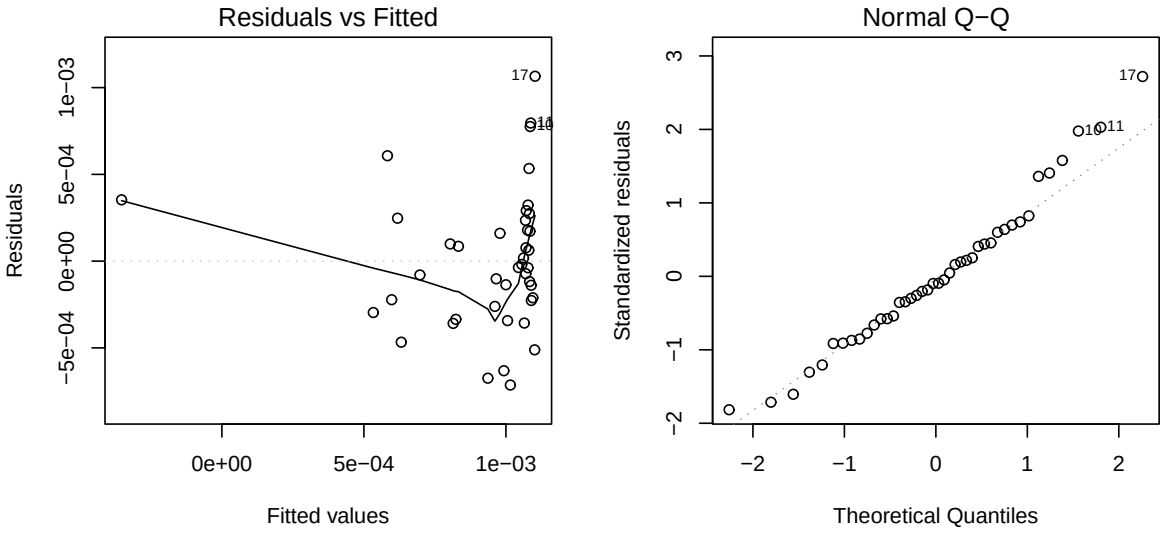


FIGURE 6.16 – *Graphique des résidus en fonction des valeurs prédites (gauche) et graphe quantile–quantile (droite) des résidus de la régression linéaire simple de B/D^2 par rapport à $1/D$ ajustée aux 42 arbres mesurés par Henry et al. (2010) au Ghana.*

perte de généralité, on posera que cette variable explicative est X_1, et donc:

$$g(X_1, \ldots, X_p; \vartheta) \equiv (kX_1^c)^2$$

avec $\vartheta \equiv (k, c)$, $k > 0$ et $c \geq 0$.

L'interprétation des résultats de l'ajustement d'un modèle non-linéaire est fondamentalement la même que pour le modèle linéaire. Outre les propriétés du modèle, la différence entre le modèle linéaire et le modèle non-linéaire est liée à la façon dont sont estimés les coefficients du modèle. Deux cas de figure sont à distinguer: (*i*) l'exposant c est fixé *a priori*; (*ii*) l'exposant c est un paramètre à estimer au même titre que les autres paramètres du modèle.

6.2.1 Exposant connu

Considérons d'abord le cas où l'exposant c du modèle pour la variance est connu *a priori*. Dans ce cas, la méthode des moindres carrés peut à nouveau être utilisée pour l'ajustement du modèle. La somme des carrés des écarts pondérés est:

$$\text{SCE}(\theta) = \sum_{i=1}^{n} w_i \, \varepsilon_i^2 = \sum_{i=1}^{n} w_i \, [Y_i - f(X_{i1}, \ldots, X_{ip}; \theta)]^2$$

où les poids sont inversement proportionnels à la variance des résidus:

$$w_i = \frac{1}{X_{i1}^{2c}} \propto \frac{1}{\text{Var}(\varepsilon_i)}$$

Comme précédemment, l'estimateur des coefficients du modèle correspond à la valeur de θ qui minimise la somme des carrés des écarts pondérés:

$$\hat{\theta} = \arg\min_{\theta} \text{SCE}(\theta) = \arg\min_{\theta} \left\{ \sum_{i=1}^{n} \frac{1}{X_{i1}^{2c}} [Y_i - f(X_{i1}, \ldots, X_{ip}; \theta)]^2 \right\}$$

Dans le cas particulier où les résidus ont une variance constante (c'est-à-dire $c = 0$), la méthode des moindres carrés pondérés se simplifie en moindres carrés ordinaires (tous les poids w_i sont égaux à 1), mais le principe des calculs reste le même. L'estimateur de θ s'obtient en résolvant

$$\frac{\partial \text{SCE}}{\partial \theta}(\hat{\theta}) = 0 \tag{6.18}$$

sous la contrainte $(\partial^2 \text{SCE}/\partial\theta^2) > 0$ qui assure qu'il s'agit bien d'un minimum et non pas d'un maximum. Dans le cas du modèle linéaire, la résolution de (6.18) avait conduit à une expression explicite pour l'estimateur $\hat{\theta}$. Dans le cas général du modèle non-linéaire, ce n'est plus le cas: il n'y a pas d'expression explicite pour $\hat{\theta}$. La minimisation de la somme des carrés des écarts doit alors être faite à l'aide d'un algorithme numérique. Nous approfondirons ce point au paragraphe 6.2.3.

Valeur *a priori* de l'exposant

La valeur *a priori* de l'exposant c s'obtient dans le cas non-linéaire de la même façon que dans le cas linéaire (cf. page 126): soit par tâtonnement, soit en découpant X_1 en classes et en estimant la variance de Y pour chaque classe, soit en minimisant l'indice de Furnival (cf. p.160).

17

Régression non-linéaire pondérée entre B et D

L'exploration graphique (fils rouges n^{os} 2 et 5) a révélé que la relation entre la biomasse B et le diamètre D était du type puissance, avec une augmentation de la variance de la biomasse avec le diamètre:

$$B = aD^b + \varepsilon$$

avec

$$\text{Var}(\varepsilon) \propto D^{2c}$$

On a vu dans le fil rouge n^o 11 que l'écart-type conditionnel de la biomasse sachant le diamètre était proportionnel au carré du diamètre: $c = 2$. On peut donc ajuster une régression non-linéaire par les moindres carrés pondérés, en utilisant une pondération inversement proportionnelle à D^4:

```
start <- coef(lm(log(Btot)~I(log(dbh)),data=dat[dat$Btot>0,]))
start[1] <- exp(start[1])
names(start) <- c("a","b")
m <- nls(Btot~a*dbh^b,data=dat,start=start,weights=1/dat$dbh^4)
summary(m)
```

L'ajustement de la régression non-linéaire est réalisé par la commande `nls`, qui demande des valeurs initiales des coefficients. Ces valeurs initiales sont contenues dans l'objet `start` et sont calculées en retransformant les coefficients de la régression linéaire sur les données log-transformées. Le résultat de l'ajustement de la régression non-linéaire par les moindres carrés pondérés est:

```
     Estimate   Std. Error   t value   Pr(>|t|)
a   2.492e-04    7.893e-05     3.157    0.00303   **
b   2.346e+00    7.373e-02    31.824    < 2e-16   ***
```

avec un écart-type résiduel $k = 0{,}0003598$ tonnes cm^{-2}. Le modèle s'écrit donc: $B = 2{,}492 \times 10^{-4} D^{2{,}346}$. Revenons à la régression linéaire ajustée aux données log-transformées (fil rouge

n° 7), qui s'écrivait: $\ln(B) = -8{,}42722 + 2{,}36104\ln(D)$. Si on revient naïvement aux données de départ en appliquant la fonction exponentielle (nous verrons au § 7.2.4 pourquoi cela est naïf), ce modèle devient: $B = \exp(-8{,}42722) \times D^{2{,}36104} = 2{,}188 \times 10^{-4} D^{2{,}36104}$. Le modèle ajusté par régression non-linéaire et le modèle ajusté par régression linéaire sur données log-transformées sont donc très proches.

18

Régression non-linéaire pondérée entre B et D^2H

On a déjà ajusté un modèle puissance $B = a(D^2H)^b$ par régression linéaire simple sur les données log-transformées (fil rouge n° 8). Ajustons à présent ce modèle directement par régression non-linéaire:

$$B = a(D^2H)^b + \varepsilon$$

avec

$$\mathrm{Var}(\varepsilon) \propto D^{2c}$$

Pour tenir compte de l'hétéroscédasticité, et considérant que l'écart-type conditionnel de la biomasse sachant le diamètre est proportionnel à D^2 (fil rouge n° 11), on peut ajuster ce modèle non-linéaire par la méthode des moindres carrés pondérés, en utilisant une pondération inversement proportionnelle à D^4:

```
start <- coef(lm(log(Btot)~I(log(dbh^2*haut)),data=dat[dat$Btot>0,]))
start[1] <- exp(start[1])
names(start) <- c("a","b")
m <- nls(Btot~a*(dbh^2*haut)^b,data=dat,start=start,weights=1/dat$dbh^4)
summary(m)
```

Comme précédemment (fil rouge n° 17), la commande `nls` demande des valeurs initiales des coefficients et celles-ci sont obtenues à partir des coefficients de la régression multiple sur les données log-transformées. Le résultat de l'ajustement est:

```
     Estimate  Std. Error  t value  Pr(>|t|)
a  7.885e-05   2.862e-05    2.755    0.0088  **
b  9.154e-01   2.957e-02   30.953    <2e-16  ***
```

avec un écart-type résiduel $k = 0{,}0003325$ tonnes cm^{-2}. Le modèle s'écrit donc: $B = 7{,}885 \times 10^{-5}(D^2H)^{0{,}9154}$. Revenons à la régression linéaire ajustée aux données log-transformées (fil rouge n° 8), qui s'écrivait: $\ln(B) = -8{,}99427 + 0{,}87238\ln(D^2H)$. Si on revient naïvement aux données de départ en appliquant la fonction exponentielle, ce modèle devient: $B = \exp(-8{,}99427) \times D^{0{,}87238} = 1{,}241 \times 10^{-4} D^{0{,}87238}$. Le modèle ajusté par régression non-linéaire et le modèle ajusté par régression linéaire sur données log-transformées sont donc relativement proches.

19

Régression non-linéaire pondérée entre B, D et H

On a déjà ajusté un modèle puissance $B = aD^{b_1}H^{b_2}$ par régression multiple sur les données log-transformées (fil rouge n° 10). Ajustons à présent ce modèle directement par régression non-linéaire:

$$B = aD^{b_1}H^{b_2} + \varepsilon$$

avec

$$\mathrm{Var}(\varepsilon) \propto D^{2c}$$

Pour tenir compte de l'hétéroscédasticité, et considérant que l'écart-type conditionnel de la biomasse sachant le diamètre est proportionnel à D^2 (fil rouge n° 11), on peut ajuster ce modèle non-linéaire par la méthode des moindres carrés pondérés, en utilisant une pondération inversement proportionnelle à D^4:

```
start <- coef(lm(log(Btot)~I(log(dbh))+I(log(haut)),data=dat[dat$Btot>0,]))
start[1] <- exp(start[1])
names(start) <- c("a","b1","b2")
m <- nls(Btot~a*dbh^b1*haut^b2,data=dat,start=start,weights=1/dat$dbh^4)
summary(m)
```

Comme précédemment (fil rouge n° 17), la commande `nls` demande des valeurs initiales des coefficients et celles-ci sont obtenues à partir des coefficients de la régression multiple sur les données log-transformées. Le résultat de l'ajustement est:

```
     Estimate  Std. Error  t value  Pr(>|t|)
a   1.003e-04   5.496e-05    1.824    0.0758  .
b1  1.923e+00   1.956e-01    9.833  4.12e-12  ***
b2  7.435e-01   3.298e-01    2.254    0.0299  *
```

avec un écart-type résiduel $k = 0{,}0003356$ tonnes cm^{-2}. Le modèle s'écrit donc: $B = 1{,}003 \times 10^{-4} D^{1{,}923} H^{0{,}7435}$. Le modèle est proche de celui qui avait été ajusté par régression multiple sur les données log-transformées (fil rouge n° 10). Le coefficient a est cependant estimé avec une moins bonne précision ici que par la régression multiple sur les données log-transformées.

6.2.2 Exposant à estimer

Considérons à présent le cas où l'exposant c est à estimer en même temps que les autres paramètres du modèle. Ce cas de figure inclut la régression linéaire avec modèle sur la variance que nous avions évoquée au paragraphe 6.1.4. La méthode des moindres carrés n'est dans ce cas plus valable. On est donc amené à utiliser une autre méthode d'ajustement: la méthode du maximum de vraisemblance. La vraisemblance d'une observation $(X_{i1}, \ldots, X_{ip}, Y_i)$ est la densité de probabilité d'observer $(X_{i1}, \ldots, X_{ip}, Y_i)$ sous le modèle spécifié. La densité de probabilité de la loi normale d'espérance μ et d'écart type σ est:

$$\phi(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

Comme Y_i est distribué selon une loi normale d'espérance $f(X_{i1}, \ldots, X_{ip}; \theta)$ et d'écart-type kX_{i1}^c, la vraisemblance de la i^{e} observation est:

$$\frac{1}{kX_{i1}^c\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{Y_i - f(X_1, \ldots, X_p; \theta)}{kX_{i1}^c}\right)^2\right]$$

Les observations étant indépendantes, leur vraisemblance conjointe est le produit des vraisemblances de chacune des observations. La vraisemblance de l'échantillon des n observations est donc:

$$\begin{aligned}\ell(\theta,\, k,\, c) &= \prod_{i=1}^{n} \frac{1}{kX_{i1}^{c}\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{Y_i - f(X_1,\, \ldots,\, X_p;\theta)}{kX_{i1}^{c}}\right)^2\right] \\ &= \frac{1}{(k\sqrt{2\pi})^n} \frac{1}{(\prod_{i=1}^{n} X_{i1})^c} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{Y_i - f(X_1,\, \ldots,\, X_p;\theta)}{kX_{i1}^{c}}\right)^2\right]\end{aligned} \tag{6.19}$$

Cette vraisemblance est considérée comme une fonction des paramètres θ, k et c.

Les valeurs des paramètres θ, k et c seront d'autant meilleures que les observations auront une probabilité élevée d'être obtenues sous le modèle correspondant à ces valeurs de paramètres. Autrement dit, les meilleures valeurs des paramètres θ, k et c sont celles qui maximisent la vraisemblance des observations. L'estimateur correspondant est, par définition, l'estimateur du maximum de vraisemblance, et s'écrit:

$$(\hat{\theta},\, \hat{k},\, \hat{c}) = \arg\max_{(\theta,\, k,\, c)} \ell(\theta,\, k,\, c) = \arg\max_{(\theta,\, k,\, c)} \ln[\ell(\theta,\, k,\, c)]$$

où la dernière égalité découle du fait qu'une fonction et son logarithmique atteignent leur maximum pour les mêmes valeurs de leur argument. Le logarithme de la vraisemblance, qu'on appelle log-vraisemblance et qu'on note $\mathcal{L}$, est plus facile à calculer que la vraisemblance, et donc, pour les calculs, c'est la log-vraisemblance que l'on cherche à maximiser. Dans le cas présent, la log-vraisemblance s'écrit:

$$\begin{aligned}\mathcal{L}(\theta,\, k,\, c) &= \ln[\ell(\theta,\, k,\, c)] \\ &= -n\ln(k\sqrt{2\pi}) - c\sum_{i=1}^{n}\ln(X_{i1}) - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{Y_i - f(X_1,\, \ldots,\, X_p;\theta)}{kX_{i1}^{c}}\right)^2 \\ &= -\frac{1}{2}\sum_{i=1}^{n}\left[\left(\frac{Y_i - f(X_1,\, \ldots,\, X_p;\theta)}{kX_{i1}^{c}}\right)^2 + \ln(2\pi) + \ln(k^2 X_i^{2c})\right]\end{aligned} \tag{6.20}$$

Pour obtenir les estimateurs du maximum de vraisemblance des paramètres, il faudrait calculer les dérivées partielles de la log-vraisemblance par rapport à ces paramètres, et rechercher les valeurs où elles s'annulent (tout en s'assurant que les dérivées secondes sont bien négatives). Dans le cas général, il n'y a pas de solution analytique à ce problème. Comme précédemment pour la somme des carrés des écarts, il faudra donc avoir recours à un algorithme numérique pour maximiser la log-vraisemblance.

On peut montrer que la méthode du maximum de vraisemblance conduit à un estimateur des coefficients qui est asymptotiquement (c'est-à-dire quand le nombre n d'observations tend vers l'infini) le meilleur. On peut montrer également que dans le cas du modèle linéaire, l'estimateur des moindres carrés et l'estimateur du maximum de vraisemblance se confondent.

20

Régression non-linéaire entre B et D avec modèle sur la variance

Reprenons la régression non-linéaire entre la biomasse et le diamètre (cf. fil rouge n° 17) mais en considérant à présent que l'exposant c du modèle pour la variance est un paramètre à estimer comme les autres. Le modèle s'écrit de la même façon que précédemment (fil rouge n° 17):

$$B = aD^b + \varepsilon$$

avec

$$\mathrm{Var}(\varepsilon) = (kD^c)^2$$

mais s'ajuste par la méthode du maximum de vraisemblance:

```
start <- coef(lm(log(Btot)~I(log(dbh)),data=dat[dat$Btot>0,]))
start[1] <- exp(start[1])
names(start) <- c("a","b")
library(nlme)
m <- nlme(Btot~a*dbh^b, data=cbind(dat,g="a"), fixed=a+b~1, start=start, groups=~g,
weights=varPower(form=~dbh))
summary(m)
```

L'ajustement est réalisé par la commande `nlme`[1], qui comme la commande `nls` (fil rouge n° 17) requiert des valeurs initiales des coefficients. Ces valeurs initiales `start` sont calculées comme dans le fil rouge n° 17. Le résultat de l'ajustement est:

```
        Value   Std.Error  DF   t-value  p-value
a   0.0002445  0.00007136  40   3.42568   0.0014
b   2.3510500  0.06947401  40  33.84071   0.0000
```

avec une valeur estimée de l'exposant $c = 2{,}090814$. Cette valeur estimée est très proche de la valeur évaluée pour la régression non-linéaire pondérée ($c = 2$, cf. fil rouge n° 11). Le modèle ajusté s'écrit donc: $B = 2{,}445 \times 10^{-4} D^{2{,}35105}$, ce qui est très proche du modèle ajusté par régression non-linéaire pondérée (fil rouge n° 17).

21

Régression non-linéaire entre B et D^2H avec modèle sur la variance

Reprenons la régression non-linéaire entre la biomasse et D^2H (cf. fil rouge n° 18) mais en considérant à présent que l'exposant c du modèle pour la variance est un paramètre à estimer comme les autres. Le modèle s'écrit de la même façon que précédemment (fil rouge n° 18):

$$B = a(D^2H)^b + \varepsilon$$

avec

$$\mathrm{Var}(\varepsilon) = (kD^c)^2$$

mais s'ajuste par la méthode du maximum de vraisemblance:

```
start <- coef(lm(log(Btot)~I(log(dbh^2*haut)),data=dat[dat$Btot>0,]))
start[1] <- exp(start[1])
names(start) <- c("a","b")
library(nlme)
m <- nlme(Btot~a*(dbh^2*haut)^b,data=cbind(dat,g="a"),fixed=a+b~1,start=start,
groups=~g,weights=varPower(form=~dbh))
summary(m)
```

1. La commande `nlme` sert en fait à ajuster les modèles non-linéaires à effet mixte. La commande `nlreg` est dédiée à l'ajustement des modèles non-linéaires avec modèle sur la variance, mais nous avons observé des résultats anormaux avec cette commande (version 3.1-96), ce qui explique que nous lui avons préféré ici la commande `nlme`, même s'il n'y a pas d'effet mixte dans les modèles considérés ici.

L'ajustement est réalisé par la commande `nlme`, qui tout comme la commande `nls` (fil rouge n° 17) requiert des valeurs initiales des coefficients. Ces valeurs initiales `start` sont calculées comme dans le fil rouge n° 17. Le résultat de l'ajustement est:

```
        Value     Std.Error  DF    t-value  p-value
a   0.0000819  0.000028528  40    2.87214   0.0065
b   0.9122144  0.028627821  40   31.86461   0.0000
```

avec une valeur estimée de l'exposant $c = 2{,}042586$. Cette valeur estimée est très proche de la valeur évaluée pour la régression non-linéaire pondérée ($c = 2$, cf. fil rouge n° 11). Le modèle ajusté s'écrit donc: $B = 8{,}19 \times 10^{-5}(D^2H)^{0{,}9122144}$, ce qui est très proche du modèle ajusté par régression non-linéaire pondérée (fil rouge n° 18).

22

Régression non-linéaire entre B, D et H avec modèle sur la variance

Reprenons la régression non-linéaire entre la biomasse, le diamètre et la hauteur (cf. fil rouge n° 19):

$$B = aD^{b_1}H^{b_2} + \varepsilon$$

avec

$$\mathrm{Var}(\varepsilon) = (kD^c)^2$$

mais en considérant à présent que l'exposant c du modèle pour la variance est un paramètre à estimer comme les autres. L'ajustement par maximum de vraisemblance:

```
library(nlme)
start <- coef(lm(log(Btot)~I(log(dbh))+I(log(haut)),data=dat[dat$Btot>0,]))
start[1] <- exp(start[1])
names(start) <- c("a","b1","b2")
m <- nlme(Btot~a*dbh^b1*haut^b2,data=cbind(dat,g="a"),fixed=a+b1+b2~1,
start=start,groups=~g,weights=varPower(form=~dbh))
summary(m)
```

requiert comme précédemment que des valeurs initiales (`start`) des coefficients soient fournies. L'ajustement donne:

```
         Value  Std.Error  DF   t-value  p-value
a   0.0001109  0.0000566  39  1.959869   0.0572
b1  1.9434876  0.1947994  39  9.976866   0.0000
b2  0.6926256  0.3211766  39  2.156526   0.0373
```

avec une valeur estimée de l'exposant $c = 2{,}055553$. Cette valeur estimée est très proche de la valeur évaluée pour la régression non-linéaire pondérée ($c = 2$, cf. fil rouge n° 11). Le modèle ajusté s'écrit donc: $B = 1{,}109 \times 10^{-4}D^{1{,}9434876}H^{0{,}6926256}$, ce qui est très proche du modèle ajusté par régression non-linéaire pondérée (fil rouge n° 19).

23

Régression non-linéaire entre B et un polynôme de $\ln(D)$

Précédemment (fil rouge n° 9), on a ajusté par régression multiple un modèle entre $\ln(B)$ et un polynôme de $\ln(D)$. Si on revient aux variable de départ, le modèle s'écrit:

$$B = \exp\{a_0 + a_1 \ln(D) + a_2[\ln(D)]^2 + \ldots + a_p[\ln(D)]^p\} + \varepsilon$$

avec

$$\mathrm{Var}(\varepsilon) = (kD^c)^2$$

On va ajuster à présent directement ce modèle non-linéaire par maximum de vraisemblance (de sorte que l'exposant c sera estimé en même temps que les autres paramètres du modèle). Pour un polynôme de degré 3, l'ajustement s'obtient par:

```
library(nlme)
start <- coef(lm(log(Btot)~I(log(dbh))+I(log(dbh)^2)+I(log(dbh)^3),data=dat[
dat$Btot>0,]))
start[1] <- exp(start[1])
names(start) <- paste("a",0:3,sep="")
m <- nlme(Btot~exp(a0+a1*log(dbh)+a2*log(dbh)^2+a3*log(dbh)^3),data=cbind(dat,
g="a"),fixed=a0+a1+a2+a3~1,start=start,groups=~g,weights=varPower(form=~dbh))
summary(m)
```

et le résultat de l'ajustement est:

```
     Value  Std.Error  DF    t-value  p-value
a0 -8.983801  2.2927006  38 -3.918436   0.0004
a1  2.939020  2.1073819  38  1.394631   0.1712
a2 -0.158585  0.6172529  38 -0.256921   0.7986
a3  0.013461  0.0581339  38  0.231547   0.8181
```

avec une valeur estimée de l'exposant $c = 2{,}099938$. On retrouve un résultat très semblable à ce qui avait été obtenu par régression multiple sur les données log-transformées (fil rouge n° 9).

6.2.3 Optimisation numérique

Que ce soit pour minimiser la somme des carrés des écarts (lorsque l'exposant c est connu) ou pour maximiser la log-vraisemblance (lorsque l'exposant c doit être estimé), il faut dans le cas du modèle non-linéaire avoir recours à un algorithme d'optimisation numérique. Maximiser la log-vraisemblance est équivalent à minimiser l'opposé de la log-vraisemblance, donc par la suite on ne considérera que le problème de minimisation d'une fonction dans un espace multidimensionel. Il existe une multitude d'algorithmes d'optimisation (Press *et al.*, 2007, chapitre 10) et l'objectif n'est pas ici de les passer en revue. Ce qui importe de savoir à ce stade, c'est que ces algorithmes sont itératifs et requièrent une valeur de départ des paramètres. À partir de cette valeur de départ et à chaque itération, l'algorithme se déplace dans l'espace des paramètres en cherchant à minimiser la fonction objectif (à savoir la somme des carrés des écarts ou moins la log-vraisemblance). On peut se représenter la fonction objectif comme une hyper-surface dans l'espace des paramètres (figure 6.17). Chaque position dans cet espace correspond à une valeur des paramètres. Une bosse de cette surface correspond à un maximum local de la fonction objectif, tandis qu'un creux de la surface correspond à un minimum local. L'objectif est de trouver le minimum global, c'est-à-dire le creux le plus profond. La position de ce creux correspond à la valeur estimée des paramètres. Si l'algorithme renvoie la position d'un creux qui n'est pas le creux le plus profond, l'estimation des paramètres est fausse.

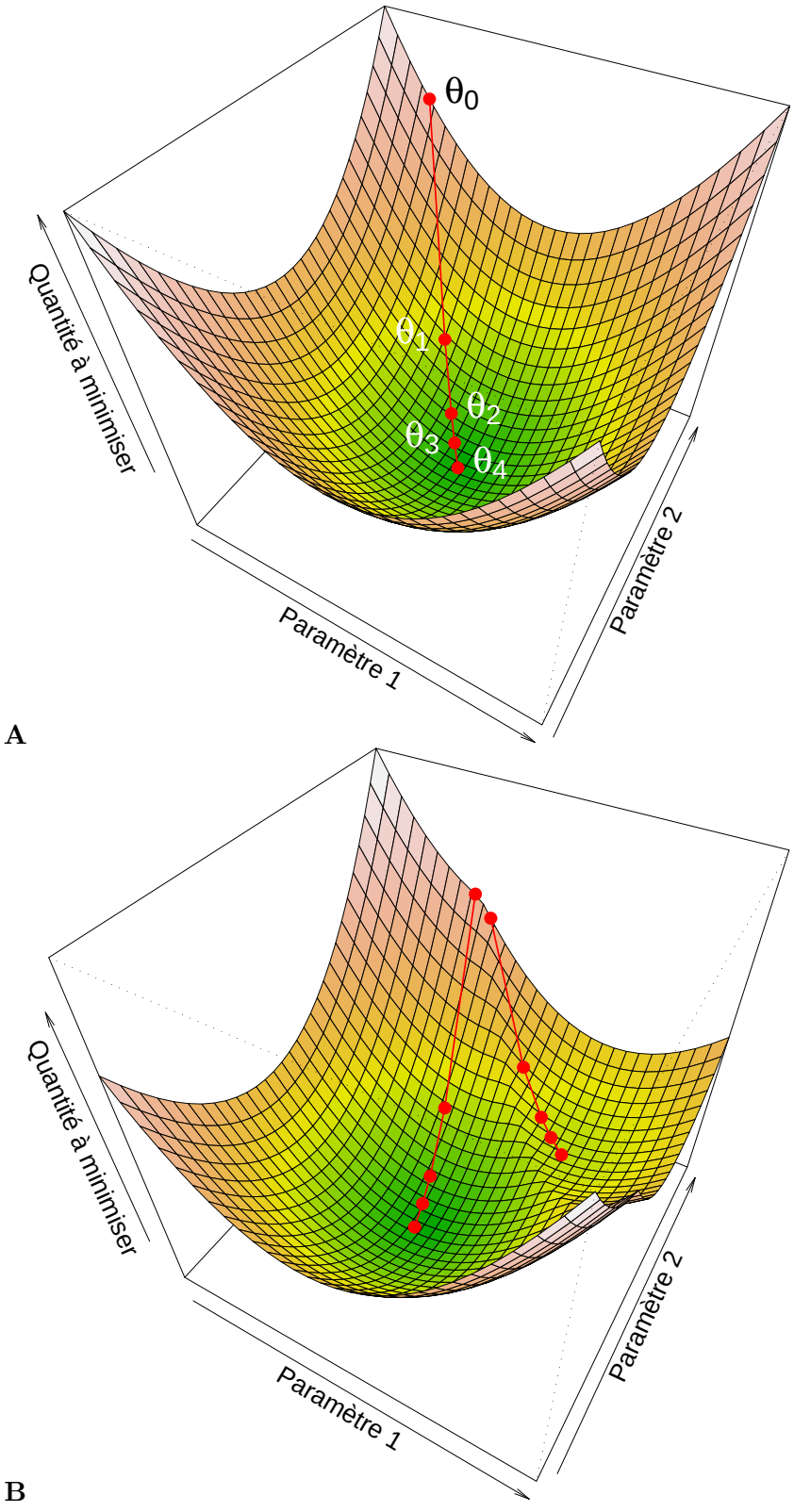


FIGURE 6.17 – *Représentation de la fonction objectif (i.e. la quantité à minimiser) comme une surface dans l'espace des paramètres. Chaque position dans cet espace correspond à une valeur des paramètres. Les valeurs successives θ_1, θ_2, ..., des paramètres sont obtenues à partir d'une valeur initiale θ_0 en descendant la surface selon sa ligne de plus forte pente. (A) La surface comporte un seul bassin versant. (B) La surface comporte plusieurs bassins versants.*

Algorithme de descente

L'algorithme d'optimisation le plus simple consiste à calculer des positions successives, c'est-à-dire des valeurs successives des paramètres, en descendant la surface définie par la fonction objectif selon sa ligne de plus forte pente (figure 6.17A). Cet algorithme conduit à un creux de la surface, mais rien ne dit que ce creux est le plus profond. En effet, la surface peut comporter plusieurs bassins versants avec plusieurs creux. Selon la position de départ, l'algorithme convergera vers un creux ou un autre (figure 6.17B). Qui plus est, deux positions initiales très proches l'une de l'autre tout en étant de part et d'autre de la ligne de crête séparant les deux bassins versants aboutiront à des creux différents, c'est-à-dire à des estimations différentes des paramètres. Le seul cas de figure où cet algorithme donne la bonne estimation des paramètres quelle que soit la valeur initiale des paramètres est lorsque la surface comporte un unique creux, c'est-à-dire lorsque la fonction objectif est convexe. C'est en particulier le cas pour le modèle linéaire, mais n'est généralement pas vrai pour le modèle non-linéaire.

Amélioration des algorithmes en cas de minima locaux

Des algorithmes plus subtils que celui de descente selon la ligne de plus forte pente ont été mis au point. On peut par exemple laisser la possibilité de ressortir d'un creux où l'algorithme a temporairement convergé pour explorer s'il n'y a pas de creux plus profond dans le voisinage. Il n'en reste pas moins qu'aucun algorithme, même le plus subtil, n'offre la certitude qu'il a bien convergé vers le creux le plus profond. Ainsi, tout algorithme d'optimisation numérique (*i*) peut être piégé par un minimum local au lieu de converger vers le minimum global, et (*ii*) est sensible à la position de départ indiquée, qui détermine en partie la position finale où l'algorithme convergera.

Si on revient au problème qui nous intéresse, cela signifie (*i*) que l'ajustement d'un modèle non-linéaire pourra donner des estimations erronées des paramètres et (*ii*) que le choix des valeurs initiales des paramètres pour l'algorithme d'optimisation est un choix sensible. C'est là le principal inconvénient de l'ajustement d'un modèle non-linéaire. Pour circonscrire cet inconvénient, il faudra choisir soigneusement la valeur initiale des paramètres, et surtout en tester plusieurs.

Choix de la valeur initiale des paramètres

Quand le modèle f pour la moyenne peut être transformé en une relation linéaire entre la variable réponse Y et les variables explicatives $X_1, \ldots, X_p$, une valeur de départ des coefficients peut être obtenue en ajustant une régression linéaire sur les variables transformées sans tenir compte de l'hétéroscédasticité éventuelle des résidus. Prenons l'exemple d'un tarif de biomasse de type puissance:

$$B = aD^{b_1}H^{b_2}\rho^{b_3} + \varepsilon \tag{6.21}$$

avec

$$\varepsilon \underset{\text{i.i.d.}}{\sim} \mathcal{N}(0,\ kD^c)$$

Le modèle puissance pour l'espérance de B peut être linéarisé en transformant les variables par le logarithme: $\ln(B) = a' + b_1 \ln(D) + b_2 \ln(H) + b_3 \ln(\rho)$. Cependant cette transformation n'est pas compatible avec l'additivité des erreurs dans le modèle (6.21). Autrement dit, la régression multiple de la variable réponse $\ln(B)$ par rapport aux variables explicatives $\ln(D)$, $\ln(H)$ et $\ln(\rho)$:

$$\ln(B) = a' + b_1 \ln(D) + b_2 \ln(H) + b_3 \ln(\rho) + \varepsilon' \tag{6.22}$$

avec $\varepsilon' \sim \mathcal{N}(0, \sigma)$, n'est pas un modèle équivalent à (6.21), quand bien même les résidus ε' de ce modèle auraient une variance constante. Même si les modèles (6.21) et (6.22) ne sont mathématiquement pas équivalents, les coefficients de (6.22) estimés par régression multiple peuvent servir de valeurs initiales pour l'algorithme numérique qui estime les coefficients de (6.21). Si on note $x^{(0)}$ la valeur initiale du paramètre x pour l'algorithme d'optimisation numérique, on aura ainsi:

$$a^{(0)} = \exp(\hat{a}'), \quad b_i^{(0)} = \hat{b}_i, \quad k^{(0)} = \hat{\sigma}, \quad c^{(0)} = 0$$

Parfois le modèle pour la moyenne n'est pas linéarisable. Par exemple, le tarif de biomasse paramétré suivant qui est utilisé pour des arbres en plantation (Saint-André *et al.*, 2005):

$$B = a + [b_0 + b_1 T + b_2 \exp(-b_3 T)] D^2 H + \varepsilon$$

où T est l'âge de la plantation et $\varepsilon \sim \mathcal{N}(0, kD^c)$, a un modèle pour la moyenne qui n'est pas linéarisable. Dans ce cas, les valeurs initiales des paramètres devront être choisies de manière empirique. Dans cet exemple précis, on pourrait prendre par exemple:

$$a^{(0)} = \hat{a}, \quad b_0^{(0)} + b_2^{(0)} = \hat{b}_0, \quad b_1^{(0)} = \hat{b}_1, \quad b_3^{(0)} = 0, \quad k^{(0)} = \hat{\sigma}, \quad c^{(0)} = 0$$

où $\hat{a}$, $\hat{b}_0$, $\hat{b}_1$ et $\hat{\sigma}$ sont les valeurs estimées des coefficients et écart-type résiduel de la régression multiple de B par rapport à D^2H et D^2HT.

Tout choix des valeurs initiales des paramètres ne dispense pas de tester plusieurs valeurs initiales. Dès lors qu'on ajuste un modèle non-linéaire avec un algorithme d'optimisation numérique, il est essentiel de tester plusieurs valeurs initiales des paramètres pour s'assurer de la stabilité des estimations.

6.3 Sélection de variables et de modèles

Quand on veut construire un tarif de cubage ou de biomasse, l'exploration graphique des données (chapitre 5) débouche généralement sur plusieurs formes possibles du modèle. On peut ajuster tous les modèles potentiellement intéressants. Mais au final, parmi tous les modèles ajustés, lequel choisir et recommander à l'utilisateur ? La sélection de variables et la sélection de modèles ont pour objectif de déterminer quelle est la « meilleure » expression possible du modèle parmi toutes celles qui ont été ajustées.

6.3.1 Sélection de variables

Prenons l'exemple d'un tarif de biomasse que l'on veut construire à partir d'un jeu de données comportant le diamètre des arbres, leur hauteur et la densité spécifique du bois. En travaillant sur les données log-transformées et selon les variables incluses dans le modèle, on pourra ajuster les modèles suivants:

$$\begin{aligned}
\ln(B) &= a_0 + a_1 \ln(D) + \varepsilon \\
\ln(B) &= a_0 + a_2 \ln(H) + \varepsilon \\
\ln(B) &= a_0 + a_3 \ln(\rho) + \varepsilon \\
\ln(B) &= a_0 + a_1 \ln(D) + a_2 \ln(H) + \varepsilon \\
\ln(B) &= a_0 + a_1 \ln(D) + a_3 \ln(\rho) + \varepsilon \\
\ln(B) &= a_0 + a_2 \ln(H) + a_3 \ln(\rho) + \varepsilon \\
\ln(B) &= a_0 + a_1 \ln(D) + a_2 \ln(H) + a_3 \ln(\rho) + \varepsilon
\end{aligned}$$

On appelle modèle *complet* le modèle qui inclut toutes les variables explicatives disponibles (le dernier dans la liste ci-dessus). Tous les autres modèles peuvent être vus comme des sous-ensembles du modèle complet, dans lesquels certaines variables explicatives ont été employées et les autres mises de côté. La sélection de variables vise à choisir, parmi les variables explicatives d'un modèle complet, celles qui sont à retenir et celles qui peuvent être mises de côté parce qu'elles apportent peu à la prédiction de la variable réponse. Autrement dit, dans cet exemple, la sélection de variables consisterait à choisir le meilleur modèle parmi les sept modèles envisagés pour $\ln(B)$.

Étant donné p variables explicatives X_1, X_2, ..., X_p, il y a $2^p - 1$ modèles incluant tout ou partie de ces variables explicatives. La sélection de variables consiste à choisir la « meilleure » combinaison de variables explicatives parmi toutes celles disponibles. Ceci signifie tout d'abord qu'il existe un critère permettant d'évaluer la qualité d'un modèle. On a déjà vu (p.120) que le R^2 est un mauvais critère pour évaluer la qualité d'un modèle par rapport à un autre puisqu'il augmente automatiquement avec le nombre de variables explicatives, que celles-ci apportent réellement de l'information pour la prédiction de la variable réponse ou non. Un meilleur critère pour sélectionner les variables explicatives est l'estimateur de la variance résiduelle, qui est lié au R^2 par la relation:

$$\hat{\sigma}^2 = \frac{n}{n-p-1}(1-R^2)\,S_Y^2$$

où S_Y^2 est la variance empirique de la variable réponse.

La recherche de la meilleure combinaison de variables explicatives peut se faire de plusieurs façons. Si p n'est pas trop élevé, on peut passer en revue les $2^p - 1$ modèles possibles de manière exhaustive. Lorsque p est trop élevé, une méthode pas à pas de sélection de variables peut être utilisée. Les méthodes pas à pas procèdent par élimination successive ou ajout successif de variables explicatives. La méthode descendante consiste à éliminer la variable la moins significative parmi les p. On recalcule alors la régression et on recommence jusqu'à ce qu'un critère d'arrêt soit satisfait (par exemple, que tous les coefficients du modèle soient significativement différents de zéro). La méthode ascendante procède en sens inverse: on part de la meilleure régression à une variable et on ajoute tour à tour la variable qui fait progresser le plus le R^2, jusqu'à ce que le critère d'arrêt soit satisfait.

La méthode dite *stepwise* est un perfectionnement de l'algorithme précédent qui consiste à effectuer en plus à chaque pas des tests de significativité du type Fisher pour ne pas introduire une variable non significative et pour éliminer éventuellement des variables déjà introduites qui ne seraient plus informatives compte tenu de la dernière variable sélectionnée. L'algorithme s'arrête quand on ne peut plus ajouter ni retrancher de variables. Les différentes méthodes de sélection pas à pas ne donnent pas forcément le même résultat, la méthode « stepwise » semblant la meilleure. Elles ne mettent cependant pas à l'abri de l'élimination intempestive de variables réellement significatives, ce qui risque de biaiser les résultats. Il faut à ce propos rappeler que si l'on sait (pour des raisons biologiques) qu'une variable doit figurer dans un modèle (la densité spécifique du bois, par exemple), ce n'est pas parce qu'un test statistique la déclare non significative qu'il faut la rejeter (à cause du risque de seconde espèce du test).

24

Sélection de variables

Faisons une sélection des variables $\ln(D)$, $[\ln(D)]^2$, $[\ln(D)]^3$, $\ln(H)$ pour prédire le logarithme de la biomasse. Le modèle complet s'écrit donc:

$$\ln(B) = a_0 + a_1\ln(D) + a_2[\ln(D)]^2 + a_3[\ln(D)]^3 + a_4\ln(H) + \varepsilon$$

avec

$$\mathrm{Var}(\varepsilon) = \sigma^2$$

La sélection de variable dans R s'effectue avec la commande `step` appliquée au modèle complet ajusté:

```
m <- lm(log(Btot)~I(log(dbh))+I(log(dbh)^2)+I(log(dbh)^3)+I(log(haut)),data=dat[
dat$Btot>0,])
summary(step(m))
```

ce qui donne:

```
                Estimate  Std. Error  t value  Pr(>|t|)
 (Intercept)    -6.50202     0.35999  -18.062   < 2e-16  ***
 I(log(dbh)^2)   0.23756     0.01972   12.044  1.53e-14  ***
 I(log(haut))    1.01874     0.17950    5.675  1.59e-06  ***
```

Les variables sélectionnées sont donc $[\ln(D)]^2$ et $\ln(H)$. Le modèle finalement retenu s'écrit: $\ln(B) = -6{,}50202 + 0{,}23756[\ln(D)]^2 + 1{,}01874\ln(H)$, avec un écart-type résiduel de 0,3994 et $R^2 = 0{,}974$.

6.3.2 Sélection de modèles

Étant donné deux modèles concurrentiels qui prédisent la même variable réponse à une transformation de variable près, lequel choisir ? Plusieurs cas de figures sont à considérer pour répondre à cette question.

Modèles emboîtés

Le cas le plus simple est lorsque les deux modèles à comparer sont emboîtés. Un modèle est *emboîté* dans un autre s'ils prédisent la même variable réponse et si on peut passer du second au premier en supprimant une ou plusieurs variables explicatives. Par exemple, le tarif de biomasse $B = a_0 + a_1 D + \varepsilon$ est emboîté dans $B = a_0 + a_1 D + a_2 D^2 H + \varepsilon$ puisque l'on passe du second au premier en supprimant D^2H des variables explicatives. De même, le modèle $B = a_0 + a_1 D^2 H + \varepsilon$ est emboîté dans $B = a_0 + a_1 D + a_2 D^2 H + \varepsilon$ puisque l'on passe du second au premier en supprimant D des variables explicatives. En revanche, le modèle $B = a_0 + a_1 D + \varepsilon$ n'est pas emboîté dans $B = a_0 + a_2 D^2 H + \varepsilon$.

Soit p le nombre de variables explicatives du modèle complet et $p' < p$ le nombre de variables explicatives du modèle emboîté. Sans perte de généralité, on peut écrire le modèle complet sous la forme:

$$Y = f(X_1, \ldots, X_{p'}, X_{p'+1}, \ldots, X_p; \theta_0, \theta_1) + \varepsilon \tag{6.23}$$

où (θ_0, θ_1) est le vecteur des coefficients associés au modèle complet et θ_0 est le vecteur des coefficients associés au modèle emboîté, qui s'obtient en posant $\theta_1 = \mathbf{0}$. En particulier dans le cas du modèle linéaire, le modèle complet s'obtient comme la somme du modèle emboîté et de termes supplémentaires:

$$\underbrace{\underbrace{Y = a_0 + a_1 X_1 + \ldots + a_{p'} X_{p'}}_{\text{modèle emboîté}} + a_{p'+1} X_{p'+1} + \ldots + a_p X_p + \varepsilon}_{\text{modèle complet}} \tag{6.24}$$

avec $\theta_0 = (a_0, \ldots, a_{p'})$ et $\theta_1 = (a_{p'+1}, \ldots, a_p)$.

Dans le cas de modèles emboîtés, on peut tester à l'aide d'un test statistique l'un des modèles contre l'autre. L'hypothèse nulle de ce test est $\theta_1 = \mathbf{0}$, c'est-à-dire: les termes supplémentaires ne sont pas significatifs, ce que l'on peut encore reformuler en: le modèle emboîté est meilleur que le modèle complet. Si la p-value de ce test s'avère être inférieure au seuil de significativité (typiquement 5 %), alors on rejette l'hypothèse nulle, c'est-à-dire que le modèle complet est le meilleur. Au contraire, si la p-value est supérieure au seuil de significativité, le modèle emboîté est considérée comme étant le meilleur.

Dans le cas du modèle linéaire (6.24), la statistique de test est un rapport de carrés moyens, qui sous l'hypothèse nulle suit une loi de Fisher. C'est du reste le même type de tests que celui utilisé pour tester le caractère significatif global d'une régression multiple, ou que celui utilisé dans la méthode « stepwise » de sélection de variables. Dans le cas général du modèle non-linéaire (6.23), la statistique de test est un rapport de vraisemblance, dont moins deux fois le logarithme suit sous l'hypothèse nulle une loi du χ^2.

25

Test de modèles emboîtés: $\ln(D)$

Dans le fil rouge n° 24, la variable $[\ln(D)]^2$ a été sélectionnée avec $\ln(H)$ comme variables explicatives de $\ln(B)$, mais pas $\ln(D)$. Pour comparer le modèle $\ln(B) = a_0 + a_1 \ln(D) + a_2[\ln(D)]^2 + a_4 \ln(H)$, qui inclut le terme additionnel $\ln(D)$, au modèle $\ln(B) = a_0 + a_2[\ln(D)]^2 + a_4 \ln(H)$, on peut réaliser un test de modèles emboîtés. La commande de R qui permet de tester un modèle emboîté est `anova`, avec comme premier argument le modèle emboîté et comme second argument le modèle complet:

```
comp <- lm(log(Btot)~I(log(dbh))+I(log(dbh)^2)+I(log(haut)),data=dat[dat$Btot>0,])
nest <- lm(log(Btot)~I(log(dbh)^2)+I(log(haut)),data=dat[dat$Btot>0,])
anova(nest,comp)
```

Le résultat du test est le suivant:

```
  Res.Df     RSS  Df  Sum of Sq      F  Pr(>F)
1     38  6.0605
2     37  5.8964   1    0.16407 1.0295  0.3169
```

La p-value vaut 0,3169 et est donc supérieure à 5 %. Le modèle emboîté (sans $\ln(D)$) est donc sélectionné au détriment du modèle complet.

26

Test de modèles emboîtés: $\ln(H)$

Dans le fil rouge n° 7, le modèle $\ln(B) = -8{,}42722 + 2{,}36104 \ln(D)$ a été obtenu tandis que dans le fil rouge n° 10, le modèle $\ln(B) = -8{,}9050 + 1{,}8654 \ln(D) + 0{,}7083 \ln(H)$ a été obtenu. Le premier étant emboîté dans le second, on peut tester lequel est le meilleur. La commande

```
comp <- lm(log(Btot)~I(log(dbh))+I(log(haut)),data=dat[dat$Btot>0,])
nest <- lm(log(Btot)~I(log(dbh)),data=dat[dat$Btot>0,])
anova(nest,comp)
```

donne:

```
  Res.Df     RSS  Df  Sum of Sq      F    Pr(>F)
1     39  8.3236
2     38  6.4014   1     1.9222 11.410  0.001698  **
```

La p-value étant inférieure à 5 %, le modèle complet (incluant $\ln(H)$ comme variable explicative) est sélectionné au détriment du modèle emboîté.

Modèles ayant la même variable réponse

Lorsque l'on veut comparer deux modèles qui ont la même variable réponse mais qui ne sont pas emboîtés, on ne peut plus faire de test statistique. Par exemple, on ne peut pas utiliser le test précédemment présenté pour comparer $B = a_0 + a_1 D + \varepsilon$ et $B = a_0 + a_2 D^2 H + \varepsilon$. Dans ce cas, on va utiliser un critère d'information (Bozdogan, 1987; Burnham et Anderson, 2002, 2004). Il en existe plusieurs, adaptés à différents contextes. Les plus utilisés sont le critère d'information bayésien (« Bayesian information criterion » ou BIC) et surtout le critère d'information d'Akaike (1974) (« Akaike information criterion » ou AIC). L'AIC a pour expression:

$$\text{AIC} = -2 \ln \ell(\hat{\theta}) + 2q$$

où $\ell(\hat{\theta})$ est la vraisemblance du modèle, c'est-à-dire la vraisemblance de l'échantillon pour les valeurs estimées des paramètres du modèle (cf. équation 6.19), et q est le nombre de paramètres libres estimés. En particulier, dans le cas d'une régression multiple par rapport à p variables explicatives, $q = p+1$ (soit les p coefficients associés aux p variables explicatives plus l'ordonnée à l'origine). Le coefficient -2 devant la log-vraisemblance dans l'expression de l'AIC est identique à celui utilisé pour la statistique du test du rapport de vraisemblance dans le cas de modèles emboîtés. Étant donné deux modèles avec le même nombre de paramètres, le modèle le meilleur sera celui avec la plus forte vraisemblance, donc celui avec le plus petit AIC. À vraisemblances égales, le modèle le meilleur sera celui qui a le moins de paramètres (selon le principe de parcimonie ou rasoir d'Occam), donc encore celui avec le plus petit AIC. En fin de compte, le meilleur modèle sera celui qui a la plus petite valeur d'AIC.

Le BIC a une expression semblable à l'AIC, mais avec un terme de pénalisation des paramètres plus fort:

$$\text{BIC} = -2 \ln \ell(\hat{\theta}) + q \ln(n)$$

où n est le nombre d'observations. Là encore, le meilleur modèle sera le modèle avec la plus petite valeur du BIC. Dans le cas de l'ajustement de tarifs de cubage ou de biomasse, on utilisera l'AIC plutôt que le BIC comme critère de sélection de modèles.

27

Sélection de modèles ayant B comme variable réponse

Les modèles suivants ayant B comme variable réponse ont été ajustés:
- fil rouge n° 12 ou 16: $B = -3{,}840 \times 10^{-3} D + 1{,}124 \times 10^{-3} D^2$
- fil rouge n° 14: $B = -3{,}319456 \times 10^{-3} D + 1{,}067068 \times 10^{-3} D^2$
- fil rouge n° 17: $B = 2{,}492 \times 10^{-4} D^{2{,}346}$
- fil rouge n° 20: $B = 2{,}445 \times 10^{-4} D^{2{,}35105}$
- fil rouge n° 11 ou 15: $B = 2{,}747 \times 10^{-5} D^2 H$
- fil rouge n° 13: $B = 2{,}740688 \times 10^{-5} D^2 H$
- fil rouge n° 18: $B = 7{,}885 \times 10^{-5} (D^2 H)^{0{,}9154}$
- fil rouge n° 21: $B = 8{,}19 \times 10^{-5} (D^2 H)^{0{,}9122144}$

– fil rouge n° 19: $B = 1{,}003 \times 10^{-4} D^{1{,}923} H^{0{,}7435}$
– fil rouge n° 22: $B = 1{,}109 \times 10^{-4} D^{1{,}9434876} H^{0{,}6926256}$

Les modèles des fils rouges n^{os} 12, 14, 11 et 13 sont ajustés par régression linéaire tandis que les autres sont ajustés par régression non-linéaire. Il y a cinq formes distinctes de modèles, et pour chacun deux modes d'ajustement: selon une régression pondérée par la méthode des moindres carrés pondérés (fils rouges n^{os} 12, 17, 11, 18 et 19) ou selon une régression avec un modèle sur la variance par la méthode du maximum de vraisemblance (fils rouges n^{os} 14, 20, 13, 21 et 22). La figure 6.18 compare les prédictions de ces différents modèles. Soit `m` l'un des modèles ajustés ayant le diamètre comme seule entrée. Le tracé des prédictions pour ce modèle s'obtient comme suit:

```
with(dat,plot(dbh,Btot,xlab="Diamètre (cm)",ylab="Biomasse (t)"))
D <- seq(par("usr")[1],par("usr")[2],length=200)
lines(D,predict(m,newdata=data.frame(dbh=D)),col="red")
```

Pour un modèle `m` ayant le diamètre et la hauteur comme entrées, les prédictions s'obtiennent comme suit:

```
D <- seq(0,180,length=20)
H <- seq(0,61,length=20)
B <- matrix(predict(m,newdata=expand.grid(dbh=D,haut=H)),length(D))
```

et le tracé de la surface de réponse de la biomasse en fonction du diamètre et de la hauteur s'obtient par:

```
M <- persp(D,H,B,xlab="Diamètre (cm)",ylab="Hauteur (m)",zlab="Biomasse (t)",
ticktype="detailed")
points(trans3d(dat$dbh,dat$haut,dat$Btot,M))
```

Étant donné un modèle ajusté `m`, son AIC se calcule par la commande:

```
AIC(m)
```

Pour les 10 modèles précédemment listés, les valeurs des AIC sont données dans le tableau 6.1. Ce tableau fait apparaître un problème que présentent plusieurs logiciels de statistiques, dont R: quand on maximise la log-vraisemblance (6.20), tout terme constant (tel que $-n \ln(2\pi)/2$) ne joue aucun rôle. La constante que l'on utilise pour calculer la log-vraisemblance, et par conséquent l'AIC, est donc affaire de convention, et différentes constantes ont été utilisées selon les calculs. Dans le tableau 6.1, on voit ainsi que les valeurs d'AIC des modèles ajustés par la commande `nls` sont très nettement supérieures à celles des autres modèles: ce n'est pas que ces modèles sont nettement moins bons que les autres; c'est simplement que la commande `nls` utilise une autre constante que les autres pour le calcul de la log-vraisemblance. On retiendra qu'avec R, il ne faut comparer des valeurs d'AIC que pour des modèles qui ont été ajustés à l'aide de la même commande.

Dans le cas présent, si on compare les deux modèles qui ont été ajustés avec la commande `lm`, le meilleur (c'est-à-dire celui avec le plus petit AIC) est celui ayant D^2H comme variable explicative (fil rouge n° 11). Si on compare les cinq modèles ajustés avec la commande `nlme`, le meilleur est à nouveau celui ayant D^2H comme variable explicative (fil rouge n° 13). Et si on compare les trois modèles ajustés avec la commande `nls`, le meilleur est encore celui ayant D^2H comme variable explicative (fil rouge n° 18). Quelle que soit la méthode d'ajustement, on peut donc conclure que c'est le tarif de biomasse utilisant D^2H comme variable explicative qui est le meilleur.

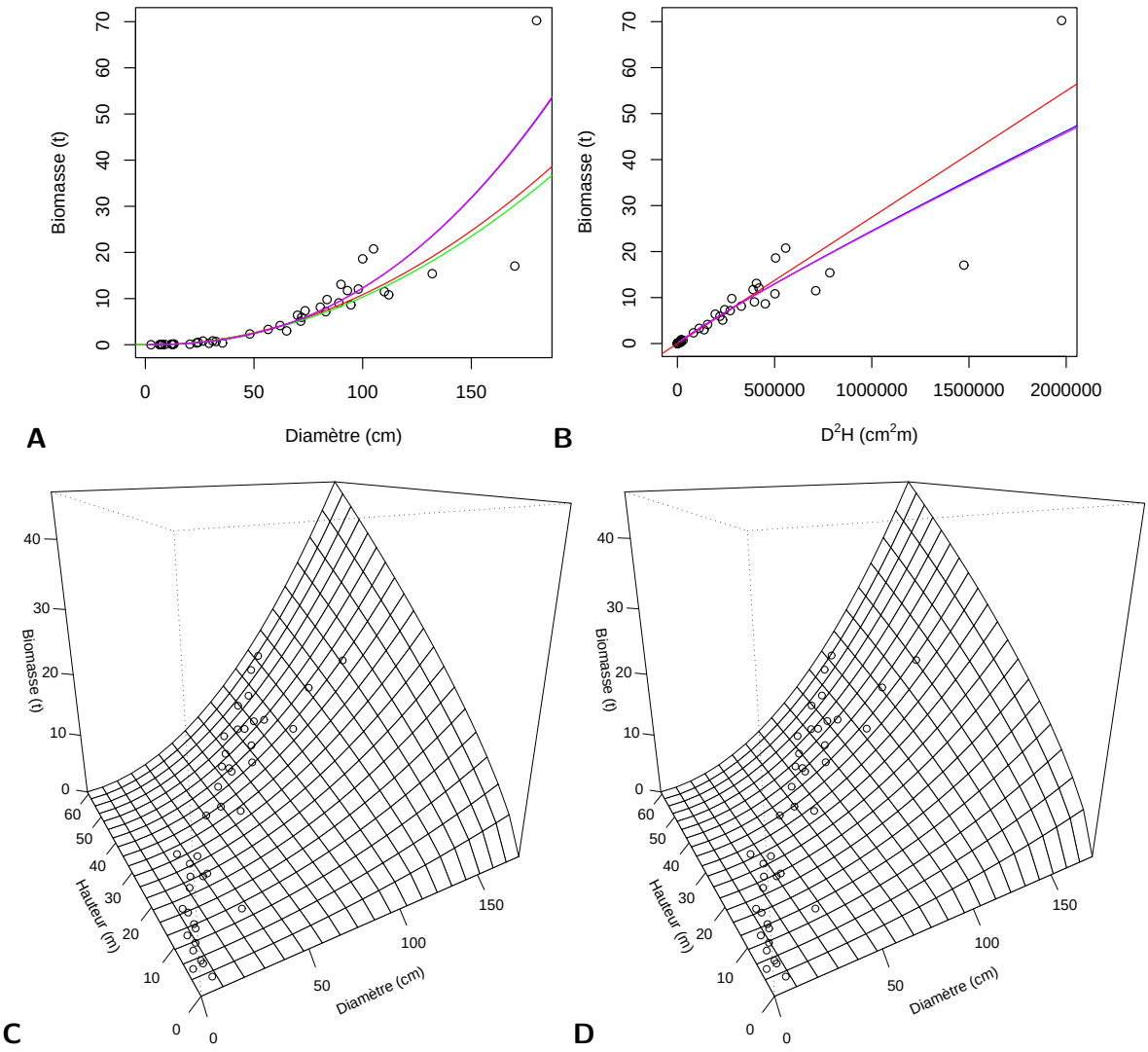


FIGURE 6.18 – *Prédictions de la biomasse par différents tarifs ajustés aux données de 42 arbres mesurés au Ghana par Henry et al. (2010). Les données sont représentées par les points. (A) Tarifs ayant le diamètre comme seule entrée, correspondant aux fils rouges n^os 12 (en rouge), 14 (vert), 17 (bleu) et 20 (magenta). (B) Tarifs ayant D^2H comme seule variable explicative, correspondant aux fils rouges n^os 11 (en rouge), 13 (rouge), 18 (bleu) et 21 (magenta). (C) Tarif correspond au fil rouge n^o 19. (D) Tarif correspond au fil rouge n^o 22.*

Table 6.1 – *Valeur de l'AIC pour 10 tarifs de biomasse ajustés aux données de 42 arbres mesurés au Ghana par Henry et al. (2010). Ces 10 tarifs prédisent directement la biomasse.*

N° fil rouge	Entrée	Méthode* ajustement	Commande R	AIC
12	D	MCP	`lm`	76,71133
14	D	MV	`nlme`	83,09157
17	D	MCP	`nls`	24 809,75727
20	D	MV	`nlme`	75,00927
11	D^2H	MCP	`lm`	65,15002
13	D^2H	MV	`nlme`	69,09644
18	D^2H	MCP	`nls`	24 797,53706
21	D^2H	MV	`nlme`	69,24482
19	D, H	MCP	`nls`	24 802,91248
22	D, H	MV	`nlme`	76,80204

*MCP = moindres carrés pondérés, MV = maximum de vraisemblance

28

Sélection de modèles ayant $\ln(B)$ comme variable réponse

Les modèles suivants ayant $\ln(B)$ comme variable réponse ont été ajustés:
- fil rouge n° 7 ou 9: $\ln(B) = -8{,}42722 + 2{,}36104\ln(D)$
- fil rouge n° 8: $\ln(B) = -8{,}99427 + 0{,}87238\ln(D^2H)$
- fil rouge n° 10: $\ln(B) = -8{,}9050 + 1{,}8654\ln(D) + 0{,}7083\ln(H)$
- fil rouge n° 24: $\ln(B) = -6{,}50202 + 0{,}23756[\ln(D)]^2 + 1{,}01874\ln(H)$

Tous ces modèles ont été ajusté selon une régression linéaire par la méthode des moindres carrés ordinaires. Le tracé des prédictions en coordonnées logarithmiques pour un modèle `m` dépendant du diamètre seulement s'obtient par la commande suivante:

```
with(dat,plot(dbh,Btot,xlab="Diamètre (cm)",ylab="Biomasse (t)",log="xy"))
D <- 10^par("usr")[1:2]
lines(D,exp(predict(m1,newdata=data.frame(dbh=D))))
```

Pour un modèle dépendant à la fois du diamètre et de la hauteur, la commande pour un graphique en coordonnées logarithmiques sera:

```
D <- exp(seq(log(1),log(180),length=20))
H <- exp(seq(log(1),log(61),length=20))
B <- matrix(predict(m,newdata=expand.grid(dbh=D,haut=H)),length(D))
M <- persp(log(D),log(H),B,xlab="log(Diamètre) (cm)",ylab="log(Hauteur) (m)",zlab=
"log(Biomasse) (t)",ticktype="detailed")
points(trans3d(log(dat$dbh),log(dat$haut),log(dat$Btot),M))
```

La figure 6.19 montre les prédiction de $\ln(B)$ selon les quatre modèles. Étant donné un modèle ajusté `m`, son AIC se calcule par la commande:

```
AIC(m)
```

Le tableau 6.2 donne l'AIC pour les quatre modèles. Les quatre modèles ayant été ajustés par la même commande `lm`, les valeurs de l'AIC sont directement comparables. Le meilleur modèle, c'est-à-dire celui avec le plus petit AIC, s'avère être le quatrième (modèle du fil rouge

n° 24). On notera également que le classement des modèles selon l'AIC est complètement cohérent avec les tests de modèles emboîtés réalisés précédemment (fils rouges n^os 25 et 26).

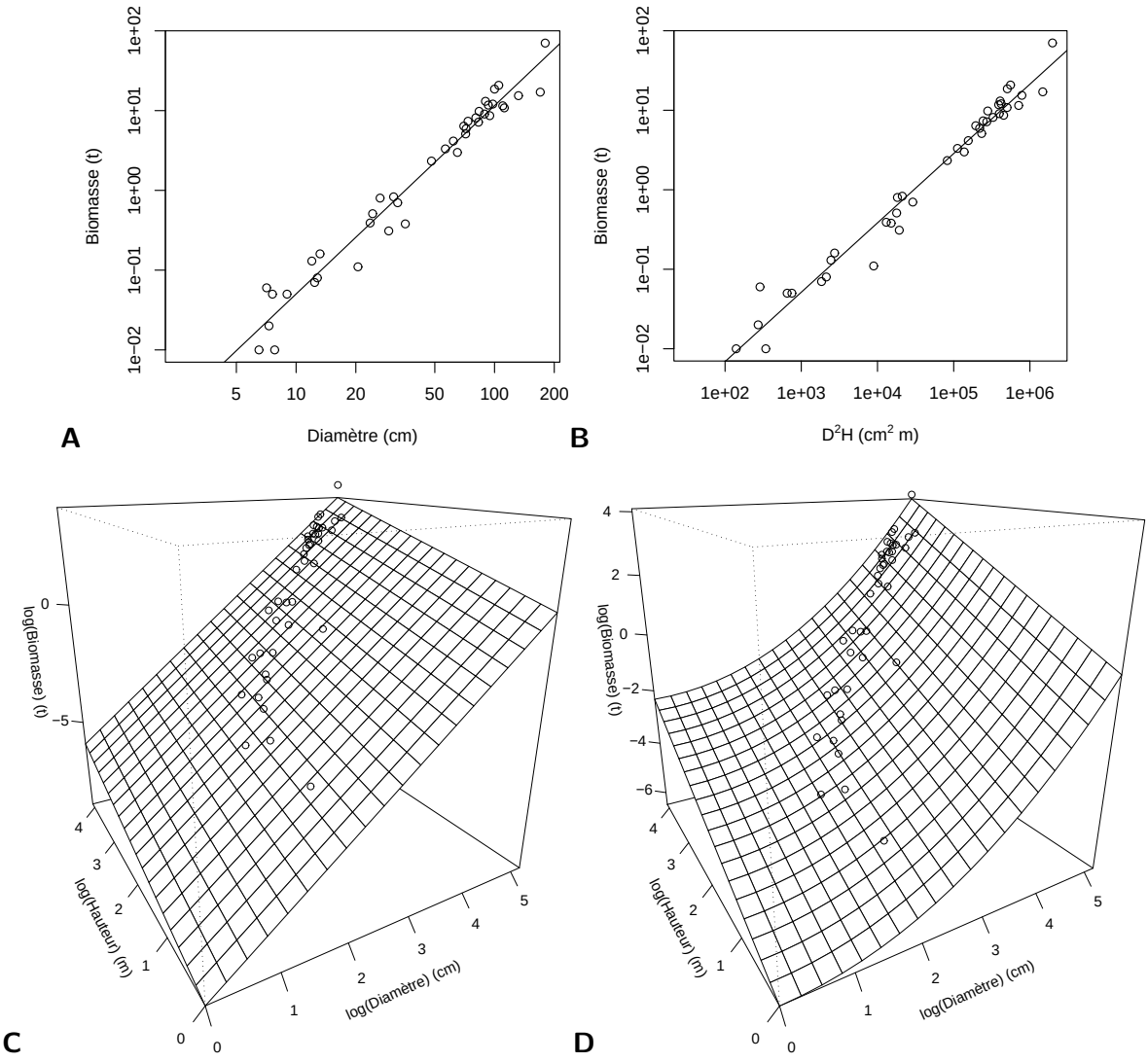


FIGURE 6.19 – *Prédictions de la biomasse par différents tarifs ajustés aux données de 42 arbres mesurés au Ghana par Henry et al. (2010). Les données sont représentées par les points. (A) Tarif du fil rouge n° 7. (B) Tarif du fil rouge n° 8. (C) Tarif du fil rouge n° 10. (D) Tarif du fil rouge n° 24.*

Modèles avec des variables réponse différentes

Le cas le plus général est lorsque l'on veut comparer deux modèles qui n'ont pas la même variable réponse, parce que l'une est une tranformée de l'autre. Par exemple, les modèles $B = aD^b + \varepsilon$ et $\ln(B) = a + b\ln(D) + \varepsilon$ prédisent tous deux la biomasse, mais la variable réponse est B dans un cas et $\ln(B)$ dans l'autre. Dans ce cas de figure, on ne peut pas utiliser

Table 6.2 – *Valeur de l'AIC pour quatre tarifs de biomasse ajustés aux données de 42 arbres mesurés au Ghana par Henry et al. (2010). Ces quatre tarifs prédisent le logarithme de la biomasse et sont tous les quatre ajustés selon une régression linéaire par la méthode des moindres carrés ordinaires (MCO).*

N° fil rouge	Entrée	Méthode ajustement	Commande R	AIC
7	D	MCO	`lm`	56,97923
8	D^2H	MCO	`lm`	46,87780
10	D, H	MCO	`lm`	48,21367
24	D, H	MCO	`lm`	45,96998

les critères d'information (AIC ou BIC) pour comparer les modèles. En revanche, l'indice de Furnival (1961) peut être utilisé dans ce cas pour comparer les modèles. Le modèle avec la plus petite valeur de l'indice de Furnival sera considéré comme le meilleur (Parresol, 1999).

L'indice de Furnival est défini uniquement pour un modèle dont l'erreur résiduelle ε a une variance supposée constante: $\mathrm{Var}(\varepsilon) = \sigma^2$. En revanche, il n'impose aucune contrainte sur la forme de la transformation de variable reliant la variable réponse Y modélisée à la variable d'intérêt (volume ou biomasse). Considérons le cas d'un tarif de biomasse (la transposition à un tarif de cubage est immédiate) et soit ψ cette transformation de variable: $Y = \psi(B)$. L'indice de Furnival est défini par:

$$F = \frac{\hat{\sigma}}{\sqrt[n]{\prod_{i=1}^{n} \psi'(B_i)}} = \exp\Big(-\frac{1}{n}\sum_{i=1}^{n} \ln[\psi'(B_i)]\Big)\,\hat{\sigma}$$

où $\hat{\sigma}$ est l'estimation de l'écart-type résiduel du modèle ajusté et B_i est la biomasse du i^{e} arbre mesuré. Lorsqu'il n'y a pas de transformation de variables, ψ est la fonction identité et l'indice de Furnival F est alors égal à l'écart-type résiduel $\hat{\sigma}$. La transformation de variables la plus fréquente est la transformation logarithmique: $\psi(B) = \ln(B)$ et $\psi'(B) = 1/B$, auquel cas l'indice de Furnival vaut:

$$F_{\ln} = \hat{\sigma}\,\sqrt[n]{\prod_{i=1}^{n} B_i} = \exp\Big(\frac{1}{n}\sum_{i=1}^{n} \ln(B_i)\Big)\,\hat{\sigma}$$

Pour les régressions linéaires dont la variance résiduelle est supposée proportionnelle à une puissance d'une variable explicative X_1, une astuce permet quand même de définir l'indice de Furnival. En effet, la régression linéaire

$$Y = a_0 + a_1X_1 + a_2X_2 + \ldots + a_pX_p + \varepsilon \tag{6.25}$$

avec $\mathrm{Var}(\varepsilon) = (kX_1^c)^2$ est strictement équivalente à la régression linéaire (cf. p.137):

$$Y' = a_0X_1^{-c} + a_1X_1^{1-c} + a_2X_2X_1^{-c} + \ldots + a_pX_pX_1^{-c} + \varepsilon' \tag{6.26}$$

avec $Y' = YX_1^{-c}$, $\varepsilon' = \varepsilon X_1^{-c}$ et $\mathrm{Var}(\varepsilon') = k^2$. Le modèle (6.26) ayant une variance résiduelle constante, son indice de Furnival est défini. Par extension, on définit l'indice de Furnival du modèle (6.25) comme étant l'indice de Furnival du modèle (6.26). Si $Y = \psi(B)$, alors $Y' = X_1^{-c}\psi(B)$, de sorte que l'indice de Furnival est alors égal à:

$$F = \frac{\hat{k}}{\sqrt[n]{\prod_{i=1}^{n} X_{i1}^{-c}\psi'(B_i)}} = \exp\Big(\frac{1}{n}\sum_{i=1}^{n}\{c\ln(X_{i1}) - \ln[\psi'(B_i)]\}\Big)\,\hat{k}$$

Ainsi, l'indice de Furnival peut aussi être utilisé pour sélectionner la valeur de l'exposant c en régression pondérée (cf. p.126).

6.3.3 Quelle méthode d'ajustement choisir ?

Revenons sur la façon d'ajuster un tarif de cubage ou de biomasse. Souvent, plusieurs solutions se présenteront pour ajuster un tarif. Considérons par exemple le tarif de biomasse

$$B = a\rho^{b_1} D^{b_2} H^{b_3} + \varepsilon$$

avec

$$\varepsilon \sim \mathcal{N}(0,\, kD^c)$$

Ce modèle pourra être ajusté comme un modèle non-linéaire (*i*) par la méthode des moindres carrés pondérés (c fixé *a priori*) ou (*ii*) par la méthode du maximum de vraisemblance (c non fixé *a priori*). Si on applique la transformation logarithmique aux données, on pourra (*iii*) ajuster la régression multiple:

$$\ln(B) = a' + b_1 \ln(\rho) + b_2 \ln(D) + b_3 \ln(H) + \varepsilon$$

avec

$$\varepsilon \sim \mathcal{N}(0,\, \sigma)$$

Ainsi, pour le même tarif qui prédit la biomasse comme une puissance des variables explicatives, on a trois méthodes d'ajustement. Certes, les méthodes (*i*)–(*ii*) et (*iii*) reposent sur des hypothèses différentes pour la structure des erreurs résiduelles: erreur additive par rapport à B dans les cas (*i*) et (*ii*), erreur multiplicative par rapport à B dans le cas (*iii*). Cependant, ces deux types d'erreur sont susceptibles de rendre compte de l'hétéroscédasticité des données, de sorte que les méthodes d'ajustement (*i*), (*ii*) et (*iii*) ont toutes trois les chances d'être valables.

Comme autre exemple, considérons le tarif de biomasse:

$$B = \exp\{a_0 + a_1 \ln(D) + a_2[\ln(D)]^2 + a_3[\ln(D)]^3 + a_4 \ln(\rho)\} + \varepsilon$$

avec

$$\varepsilon \sim \mathcal{N}(0,\, kD^c)$$

Là encore, pourra (*i*) ajuster un modèle non-linéaire par la méthode des moindres carrés (en spécifiant c *a priori*), (*ii*) ajuster un modèle non-linéaire par la méthode du maximum de vraisemblance (en estimant c), ou (*ii*) ajuster une régression multiple sur les données log-transformées:

$$\ln(B) = a_0 + a_1 \ln(D) + a_2[\ln(D)]^2 + a_3[\ln(D)]^3 + a_4 \ln(\rho) + \varepsilon$$

avec

$$\varepsilon \sim \mathcal{N}(0,\, \sigma)$$

Là encore, la structure des erreurs n'est pas la même dans les trois cas, mais toutes sont susceptibles de rendre compte correctement de l'hétéroscédasticité de la biomasse.

Le plus souvent, les différentes méthodes d'ajustement donneront des résultats très semblables en termes de prédiction. Si un doute venait à subsister sur la méthode d'ajustement la plus appropriée, on pourrait utiliser les méthodes de sélection de modèles pour trancher. Mais le choix d'une méthode d'ajustement résultera plutôt de l'importance que l'on accordera aux avantages et inconvénients respectifs de chaque méthode. La régression multiple a

l'inconvénient de poser des contraintes sur la forme des résidus, et d'avoir moins de souplesse dans la forme du modèle pour la moyenne. Comme avantage, la régression multiple offre une expression explicite des estimateurs des coefficients du modèle: il n'y a donc aucun risque d'avoir des estimations erronées des coefficients. Le modèle non-linéaire a l'avantage de ne poser aucune restriction sur la forme du modèle, que ce soit le modèle pour la moyenne ou le modèle pour la variance. Comme inconvénient, il n'y a pas d'expression explicite des estimateurs des paramètres: il y a donc un risque avec le modèle non-linéaire d'avoir des estimations erronées des paramètres.

29

Méthodes d'ajustement du modèle puissance

Nous avons vu trois façons d'ajuster le tarif puissance $B = aD^b$:

1. selon une régression linéaire simple sur les données log-transformées (fil rouge n° 7): $\ln(B) = -8{,}42722 + 2{,}36104\ln(D)$, soit $B = 2{,}18829 \times 10^{-4} D^{2{,}36104}$ si on applique « naïvement » la transformation inverse exponentielle;
2. selon une régression non-linéaire pondérée (fil rouge n° 17): $B = 2{,}492 \times 10^{-4} D^{2{,}346}$;
3. selon une régression non-linéaire avec modèle sur la variance (fil rouge n° 20): $B = 2{,}445 \times 10^{-4} D^{2{,}35105}$.

La figure 6.20 compare les prédictions de ces trois ajustements du même modèle, montrant que les différences sont minimes, bien en deçà de la précision des prédictions comme nous le verrons plus loin (§ 7.2).

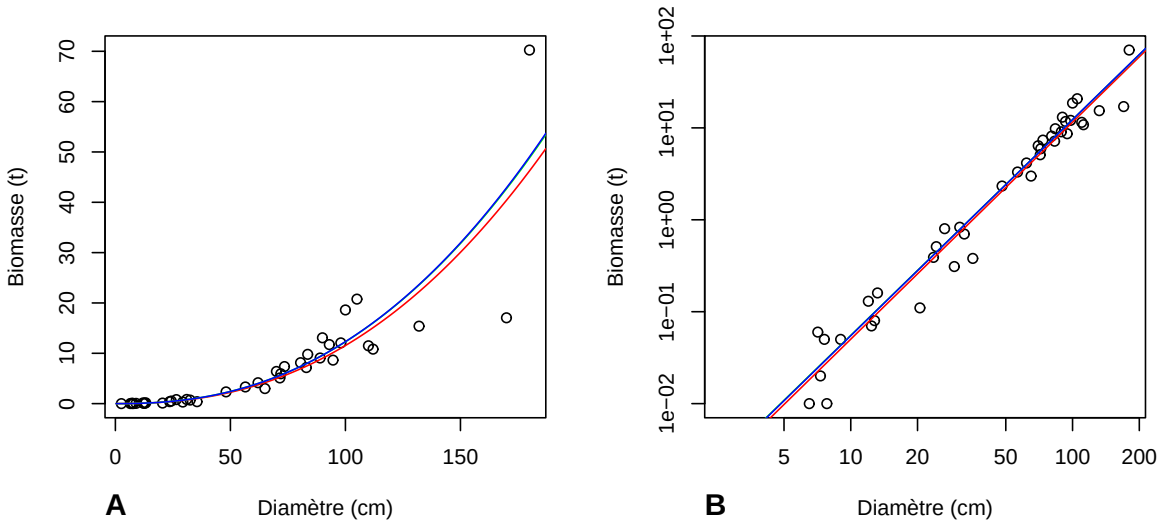


Figure 6.20 – *Prédictions de la biomasse par le même tarif puissance ajusté de trois façons différentes aux données de 42 arbres mesurés au Ghana par Henry et al. (2010). Les données sont représentées par les points. En rouge, ajustement par régression linéaire sur données log-transformées (fil rouge n° 7). En vert (pratiquement superposé avec le bleu), ajustement par régression non-linéaire pondérée (fil rouge n° 17). En bleu, ajustement par régression non-linéaire avec modèle sur la variance (fil rouge n° 20). (A) Sans transformations des données. (B) En coordonnées logarithmiques.*

6.4 Facteurs de stratification et agrégation

Jusqu'à présent, nous avons considéré que le jeu de données utilisé pour ajuster le tarif de cubage ou de biomasse était homogène. En réalité, le jeu de données peut être issu de mesures prises dans des conditions différentes, ou peut résulter de la fusion de plusieurs jeux de données distincts. Des co-variables sont généralement utilisées pour décrire cette hétérogénéité du jeu de données. Par exemple, une co-variable pourra indiquer le type de forêt dans lequel les mesures ont été faites (forêt décidue, semi-décidue, sempervirente...), ou le type de sol, ou l'année de plantation (s'il s'agit de plantation), etc. Pour les jeux de données plurispécifiques, une co-variable très importante est l'espèce de l'arbre. Dans un premier temps, toutes ces co-variables susceptibles d'expliquer l'hétérogénéité d'un jeu de données seront considérées comme des variables qualitatives (ou facteurs). Les modalités de ces facteurs définissent des strates et un jeu de données bien constitué aura été échantillonné en fonction des strates préalablement identifiées (cf. § 2.2.3). Comment prendre en compte ces co-variables qualitatives dans un tarif de cubage ou de biomasse ? Est-il valide d'analyser le jeu de données dans sa globalité, ou faut-il analyser les sous-jeux de données correspondant à chaque strate séparément ? Ce sont les questions que nous allons aborder à présent (§ 6.4.1).

Qui plus est, les mesures de biomasse sont faites séparément pour chaque compartiment de l'arbre (cf. chapitre 3). On a donc, pour chaque arbre de l'échantillon, en sus de l'estimation de sa biomasse totale, une estimation de sa biomasse foliaire, de la biomasse de son tronc, de ses grosses branches, de ses petites branches, etc. Comment tenir compte de ces différents compartiments dans l'établissement des tarifs de biomasse ? Nous aborderons aussi cette question (§ 6.4.2).

6.4.1 Stratification des données

Considérons désormais qu'il y a des co-variables qualitatives qui stratifient le jeu de données selon S strates. Chaque strate correspond à un croisement des modalités des co-variables qualitatives (dans un contexte de plans d'expérience, on parlerait de *traitement* plutôt que de strate), et nous ne considérerons pas chaque co-variable qualitative séparément. Par exemple, s'il y a une co-variable indiquant le type de forêt avec trois modalités (mettons: forêt décidue, forêt semi-décidue et forêt sempervirente) et une autre co-variable indiquant le type de sol avec trois modalités (mettons: sol sableux, sol argileux, sol limoneux), le croisement de ces deux co-variables donne $S = 3 \times 3 = 9$ strates (forêt décidue sur sol sableux, forêt décidue sur sol argileux, etc.). Nous ne chercherons pas à analyser l'effet du type de forêt séparément, pas plus que l'effet du type de sol séparément. De plus, si certaines combinaisons de modalités des co-variables ne sont pas représentées dans le jeu de données, le nombre de strates en sera diminué d'autant. Par exemple, s'il n'y a pas de forêt sempervirente sur sol limoneux, le nombre de strate sera $S = 8$ au lieu de 9.

En présence d'une stratification du jeu de données, une stratégie consisterait à ajuster un modèle séparément pour chaque strate. Dans le cas d'une régression multiple, cela s'écrirait:

$$Y_s = a_{0s} + a_{1s}X_{1s} + a_{2s}X_{2s} + \ldots + a_{ps}X_{ps} + \varepsilon_s$$

avec

$$\varepsilon_s \underset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_s)$$

où $(Y_s, X_{1s}, \ldots, X_{ps})$ désigne une observation relative à la strate s, pour $s = 1, \ldots, S$. Il y a à présent $S \times (p + 1)$ coefficients à estimer. Une stratégie alternative consiste à analyser le jeu de données dans sa globalité, en ajustant un modèle du type:

$$Y_s = a_{0s} + a_{1s}X_{1s} + a_{2s}X_{2s} + \ldots + a_{ps}X_{ps} + \varepsilon \tag{6.27}$$

avec

$$\varepsilon \underset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma)$$

L'écriture du modèle ne diffère que dans la structure de l'erreur. Ce type de modèle s'appelle une analyse de *covariance*. Elle suppose que tous les résidus ont la même variance, non seulement au sein de chaque strate, mais aussi d'une strate à l'autre. L'analyse de covariance permet de tester s'il y a un effet de la strate sur la variable réponse, seule ou en interaction avec chacune des variables explicatives $X_1, \ldots, X_p$. Tester l'effet principal de la stratification revient à tester l'hypothèse nulle $a_{01} = a_{02} = \ldots = a_{0S}$. La statistique de test est un rapport de carrés moyens qui, sous l'hypothèse nulle, suit une loi de Fisher. Tester l'effet de l'interaction entre la stratification et la j^{e} variable explicative revient à tester l'hypothèse nulle $a_{j1} = a_{j2} = \ldots = a_{jS}$. Comme précédemment, la statistique de test est un rapport de carrés moyens qui, sous l'hypothèse nulle, suit une loi de Fisher.

L'intérêt de tester ces effets est qu'à chaque fois que l'un d'entre eux ne s'avère pas significatif, on peut remplacer les S coefficients a_{j1}, a_{j2}, …, a_{jS} à estimer par un unique coefficient commun a_j. Imaginons par exemple que dans l'analyse de covariance (6.27), l'effet principal de la strate ne soit pas significatif, pas plus que l'interaction entre la strate et les p' premières variables explicatives (avec $p' < p$). Alors le modèle à ajuster s'écrit:

$$Y_s = a_0 + a_1 X_{1s} + \ldots + a_{p'} X_{p's} + a_{p'+1,s} X_{p'+1,s} + \ldots + a_{ps} X_{ps} + \varepsilon$$

avec $\varepsilon \sim \mathcal{N}(0, \sigma)$. Ce modèle ne comporte « plus » que $p'+1+(p-p')S$ coefficients à estimer, au lieu des $(p+1)S$ coefficients si on ajustait un modèle séparément pour chaque strate. L'ensemble des observations servant à estimer les coefficients communs a_0, …, $a_{p'}$, ceux-ci seront estimés avec davantage de précision que si l'on avait ajusté un modèle séparément pour chaque strate.

Ce principe d'analyse de covariance s'étend de façon immédiate au cas d'un modèle non-linéaire. Là encore, on pourra tester si les coefficients sont ou non significativement différents entre strates pour, le cas échéant, estimer un coefficient commun à toutes les strates.

30

Tarif de biomasse spécifique

Dans le fil rouge n° 8, nous avons ajusté par régression linéaire simple sur les données log-transformées un modèle puissance utilisant D^2H comme variable explicative: $\ln(B) = a + b\ln(D^2H)$. On peut à présenter intégrer l'information sur l'espèce dans ce modèle pour tester si les coefficients a et b diffèrent d'une espèce à l'autre. Le modèle correspond à une analyse de covariance:

$$\ln(B_s) = a_s + b_s \ln(D_s^2 H_s) + \varepsilon$$

avec

$$\text{Var}(\varepsilon) = \sigma^2$$

où l'indice s désigne l'espèce. L'ajustement de ce modèle s'obtient par la commande:

```
m <- lm(log(Btot)~espèce*I(log(dbh^2*haut)),data=dat[dat$Btot>0,])
```

Pour tester si les coefficients a et b diffèrent d'une espèce à l'autre, on utilise la commande:

```
anova(m)
```

qui donne:

```
                          Df   Sum Sq  Mean Sq    F value     Pr(>F)
espèce                    15  117.667    7.844    98.4396  1.647e-13  ***
I(log(dbh^2*haut))         1  112.689  112.689  1414.1228  < 2.2e-16  ***
espèce:I(log(dbh^2*haut))  7    0.942    0.135     1.6879     0.1785
Residuals                 17    1.355    0.080
```

La première ligne du tableau teste s'il y a un effet espèce, c'est-à-dire si l'ordonnée à l'origine a_s diffère d'une espèce à l'autre. L'hypothèse nulle de ce test est qu'il n'y a pas de différences entre espèces: $a_1 = a_2 = \ldots = a_S$, où $S = 16$ est le nombre d'espèces. La statistique de test est donnée dans la colonne « F value ». La p-value du test est ici inférieure à 5 %, donc on peut conclure que l'ordonnée à l'origine du modèle est significativement différente d'une espèce à l'autre. La deuxième ligne du tableau teste s'il y a un effet de la variable D^2H, c'est-à-dire si la pente moyenne associée à cette variable est significativement différente de zéro. La troisième ligne du tableau teste si l'interaction pente–espèce est significative, c'est-à-dire si la pente b_s diffère d'une espèce à l'autre. L'hypothèse nulle est qu'il n'y a pas de différences entre espèces: $b_1 = b_2 = \ldots = b_S$. La p-value vaut ici 0,1785 et est donc supérieure à 5 %: il n'y a donc pas de différence significative de pente entre les espèces.

On est donc amené à ajuster le modèle suivant:

$$\ln(B_s) = a_s + b \ln(D_s^2 H_s) + \varepsilon \tag{6.28}$$

qui considère que la pente b est la même pour toutes les espèces. La commande est:

```
m <- lm(log(Btot)~espèce+I(log(dbh^2*haut)),data=dat[dat$Btot>0,])
anova(m)
```

et donne:

```
                   Df   Sum Sq  Mean Sq  F value     Pr(>F)
espèce             15  117.667    7.844    81.99  < 2.2e-16  ***
I(log(dbh^2*haut))  1  112.689  112.689  1177.81  < 2.2e-16  ***
Residuals          24    2.296    0.096
```

Les coefficients du modèle s'obtiennent par la commande:

```
summary(m)
```

ce qui donne:

```
                                Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)                    -9.00359     0.45144  -19.944    <2e-16  ***
espèceAubrevillea kerstingii   -0.54634     0.43784   -1.248    0.2241
espèceCecropia peltata         -0.77688     0.36261   -2.142    0.0425  *
espèceCeiba pentandra          -0.70841     0.38048   -1.862    0.0749  .
espèceCola nitida              -0.46428     0.44476   -1.044    0.3069
espèceDaniellia thurifera       0.04685     0.46413    0.101    0.9204
espèceDialium aubrevilliei     -0.15626     0.43757   -0.357    0.7241
espèceDrypetes chevalieri       0.04953     0.45395    0.109    0.9140
espèceGarcinia epunctata        1.09645     0.47318    2.317    0.0293  *
espèceGuarea cedrata           -0.45255     0.38460   -1.177    0.2509
espèceHeritiera utilis         -0.26865     0.32663   -0.822    0.4189
espèceNauclea diderrichii      -0.55464     0.35759   -1.551    0.1340
espèceNesogordonia papaverifera -0.47817    0.44335   -1.079    0.2915
espècePiptadeniastrum africanum -0.17956    0.35718   -0.503    0.6197
espèceStrombosia glaucescens    0.06333     0.39597    0.160    0.8743
espèceTieghemella heckelii     -0.09104     0.33908   -0.268    0.7906
I(log(dbh^2*haut))              0.89985     0.02622   34.319    <2e-16  ***
```

La dernière ligne de ce tableau donne la valeur de la pente: $b = 0{,}89985$. Les lignes précédentes donne les valeurs des ordonnées à l'origine pour les seize espèces. Par convention, le

logiciel R procède de la façon suivante pour spécifier ces valeurs: la première ligne du tableau donne l'ordonnée à l'origine pour la première espèce selon l'ordre alphabétique. La première espèce dans l'ordre alphabétique étant *Afzelia bella*, l'ordonnée à l'origine pour *Afzelia bella* vaut donc $a_1 = -9{,}00359$. Les lignes suivantes donnent la *différence* $a_s - a_1$ entre l'ordonnée à l'origine pour l'espèce indiquée et l'ordonnée à l'origine d'*Afzelia bella*. Ainsi l'ordonnée à l'origine pour *Aubrevillea kerstingii* vaut: $a_2 = a_1 - 0{,}54634 = -9{,}00359 - 0{,}54634 = -9{,}54993$. En définitive, l'expression du tarif spécifique est:

$$\ln(B) = 0{,}89985\ln(D^2H) - \begin{cases} 9{,}00359 & \text{pour } \textit{Afzelia bella} \\ 9{,}54993 & \text{pour } \textit{Aubrevillea kerstingii} \\ 9{,}78047 & \text{pour } \textit{Cecropia peltata} \\ 9{,}71200 & \text{pour } \textit{Ceiba pentandra} \\ 9{,}46786 & \text{pour } \textit{Cola nitida} \\ 8{,}95674 & \text{pour } \textit{Daniellia thurifera} \\ 9{,}15985 & \text{pour } \textit{Dialium aubrevilliei} \\ 8{,}95406 & \text{pour } \textit{Drypetes chevalieri} \\ 7{,}90713 & \text{pour } \textit{Garcinia epunctata} \\ 9{,}45614 & \text{pour } \textit{Guarea cedrata} \\ 9{,}27223 & \text{pour } \textit{Heritiera utilis} \\ 9{,}55823 & \text{pour } \textit{Nauclea diderrichii} \\ 9{,}48176 & \text{pour } \textit{Nesogordonia papaverifera} \\ 9{,}18315 & \text{pour } \textit{Piptadeniastrum africanum} \\ 8{,}94026 & \text{pour } \textit{Strombosia glaucescens} \\ 9{,}09462 & \text{pour } \textit{Tieghemella heckelii} \end{cases}$$

À diamètre et hauteur égaux, l'espèce ayant la plus forte biomasse est *Garcinia epunctata* tandis que l'espèce ayant la plus faible biomasse est *Cecropia peltata*. L'écart-type résiduel du modèle est $\hat{\sigma} = 0{,}3093$ et $R^2 = 0{,}9901$.

Cas d'une co-variable numérique

Nous avons considéré jusqu'à présent que les co-variables définissant la stratification étaient des facteurs qualitatifs. Dans certains cas, ces co-variables peuvent être aussi interprétées comme des variables numériques. Prenons l'exemple d'un tarif de biomasse pour des plantations (Saint-André *et al.*, 2005). L'année où la plantation a été faite (ou, ce qui revient au même, l'âge des arbres) pourra être utilisée comme co-variable de stratification. Cette année ou cet âge peuvent être indifféremment vus comme des variables qualitatives (cohortes d'arbres ayant le même âge) ou comme des variables numériques. Plus généralement, toute variable numérique peut être vue comme une variable qualitative si on la découpe en classes. Dans le cas de l'âge, on pourra ainsi considérer les plantations de 0 à 5 ans comme une strate, les plantations de 5 à 10 ans comme une autre strate, les plantations de 10 à 20 ans comme une troisième strate, etc. L'avantage de découper une co-variable numérique Z en classes et de la considérer comme une variable qualitative est que cela permet de modéliser la relation entre Z et la variable réponse Y sans contraindre *a priori* la forme de cette relation. À l'opposé, quand on considère Z comme une variable numérique, on est obligé de poser *a priori* une certaine forme de relation entre Y et Z (une relation linéaire, ou une relation polynômiale, ou une relation exponentielle, ou une relation puissance...). L'inconvénient de découper Z en classes et de considérer cette co-variable comme qualitative

est que le découpage introduit une part d'arbitraire. De plus, le modèle de covariance utilisant les classes de Z (co-variables qualitative) aura généralement davantage de paramètres à estimer que le modèle considérant Z comme une co-variable numérique.

Il est habituel en modélisation de jouer sur la dualité d'interprétation des variables numériques. Lorsqu'une co-variable Z est numérique (comme l'âge des arbres), nous recommandons dans ce cas de procéder en deux temps (comme expliqué dans le § 5.1.1):

1. considérer Z comme une variable qualitative (quitte à la découper en classes) et ajuster un modèle de covariance, ce qui permettra de visualiser la forme de la relation entre Z et les coefficients du modèle;
2. modéliser cette relation par une expression appropriée, et revenir à l'ajustement d'un modèle linéaire ou non-linéaire, en considérant Z comme une variable numérique.

Pour reprendre l'exemple de l'âge des arbres en plantation: supposons que l'âge Z a été découpé en S classes d'âge. La première étape consisterait typiquement en une analyse de co-variance (en supposant que le modèle ait pu être linéarisé):

$$Y_s = a_{0s} + a_{1s}X_{1s} + a_{2s}X_{2s} + \ldots + a_{ps}X_{ps} + \varepsilon$$

avec $s = 1, \ldots, S$. Soit Z_s l'âge médian de la classe d'âge s. On tracerait ensuite le nuage de points de a_{0s} en fonction de Z_s, le nuage de points de a_{1s} en fonction Z_s, ..., le nuage de points de a_{ps} en fonction de Z_s et, pour chaque nuage de points, on rechercherait la forme de la relation s'ajustant à ce nuage de points. Imaginons que a_{0s} varie de façon linéaire en fonction de Z_s, que a_{1s} varie de façon exponentielle en fonction de Z_s, que a_{2s} varie de façon puissance en fonction de Z_s, et que les coefficients a_{3s} à a_{ps} ne varient pas en fonction de Z_s (ce que l'on peut du reste tester formellement). On serait alors amené dans ce cas de figure à ajuster dans un deuxième temps le modèle non-linéaire suivant:

$$Y = \underbrace{b_0 + b_1 Z}_{a_{0s}} + \underbrace{b_2 \exp(-b_3 Z)}_{a_{1s}} X_1 + \underbrace{b_4 Z^{b_5}}_{a_{2s}} X_2 + a_3 X_3 + \ldots + a_p X_p + \varepsilon$$

où l'âge Z est désormais considéré comme une variable numérique. Un tel tarif où intervient une co-variable explicative numérique est appelé tarif *paramétré* (ici par l'âge).

Les co-variables *ordinales* méritent une remarque particulière. Une variable ordinale est une variable qualitative qui définit un ordre. Le mois de l'année, par exemple, est une variable qualitative qui définit un ordre chronologique. Le type de sol le long d'un gradient de fertilité de sols est également une variable ordinale. Les variables ordinales sont généralement traitées comme des variables qualitatives à part entière, mais on perd du coup l'information d'ordre qu'elles apportent. Une alternative consiste à numéroter les modalités ordonnées de la variable ordinale par des valeurs entières et à considérer ensuite la variable ordinale comme une variable numérique. Par exemple, dans le cas des mois de l'année, on pourra poser janvier = 1, février = 2, etc. Cette approche n'a de sens que si les écarts entre les entiers reflètent bien les écarts entre les modalités de la variable ordinale. Par exemple, si on a posé 1 = janvier 2011 jusqu'à 12 = décembre 2011, on posera 1 = janvier 2012 si la réponse est saisonnière cyclique, tandis qu'on posera 13 = janvier 2012 si la réponse présente une tendance continue. Dans le cas de trois types de sol le long d'un gradient de fertilité, on posera 1 = le sol le plus pauvre, 2 = le sol de fertilité intermédiaire et 3 = le sol le plus riche si on pense que la différence de fertilité entre deux sols induit une réponse proportionnelle à cette différence, mais on posera 1 = le sol le plus pauvre, 4 = le sol de fertilité intermédiaire et 9 = le sol le plus riche si on pense que la réponse est proportionnelle au carré de la différence de fertilité.

Cas particulier de l'espèce

Dans le cas de jeux de données plurispécifiques, l'espèce est une co-variable de stratification qui mérite une attention particulière. Si le jeu de données comporte peu d'espèces (moins d'une dizaine approximativement) et qu'il y a suffisamment d'observations par espèce (cf. § 2.2.1), l'espèce pourra être considérée comme une co-variable de stratification comme une autre. On sera donc amené à décliner le modèle en S tarifs spécifiques ou à regrouper des tarifs selon la ressemblance allométrique des espèces.

Lorsque le jeu de données comporte de nombreuses espèces ou que certaines espèces comportent peu d'observations, il est malaisé de traiter l'espèce comme une co-variable de stratification. Une solution dans ce cas consiste à passer par des *traits fonctionnels* d'espèce. Les traits fonctionnels sont ici définis, de manière un peu abusive, comme des variables numériques qui caractérisent l'espèce (Díaz et Cabido, 1997; Rösch *et al.*, 1997; Lavorel et Garnier, 2002; voir Violle *et al.*, 2007 pour une définition plus rigoureuse). Le trait le plus utilisé dans le cas des tarifs de biomasse est la densité du bois. Si on décide d'utiliser des traits fonctionnels pour représenter les espèces, ceux-ci interviennent comme des variables explicatives du modèle au même titre que les variables explicatives caractérisant l'arbre, comme son diamètre ou sa hauteur. Le tarif de biomasse monospécifique à une entrée (par rapport au diamètre) de type puissance, qui sous sa forme linéarisée s'écrit:

$$\ln(B) = a_0 + a_1 \ln(D) + \varepsilon$$

deviendra ainsi dans le cas plurispécifique un tarif de biomasse à deux entrées:

$$\ln(B) = a_0 + a_1 \ln(D) + a_2 \ln(\rho) + \varepsilon$$

si on décide d'utiliser la densité du bois ρ pour représenter l'effet spécifique.

31

Tarif de biomasse dépendant de la densité spécifique du bois

Dans le fil rouge n° 30, l'information sur l'espèce a été prise en compte dans le tarif $\ln(B) = a + b \ln(D^2 H)$ par l'intermédiaire d'une covariable qualitative. On peut à présent chercher à capter cette information à travers la densité spécifique du bois ρ. Le modèle ajusté est donc:

$$\ln(B) = a_0 + a_1 \ln(D^2 H) + a_2 \ln(\rho) + \varepsilon \tag{6.29}$$

avec

$$\mathrm{Var}(\varepsilon) = \sigma^2$$

La densité du bois ayant été mesurée dans le jeu de données pour chaque individu, il faut commencer par calculer la densité du bois moyenne de chaque espèce:

```
dm <- tapply(dat$dens,dat$espèce,mean)
dat <- cbind(dat,dmoy=dm[as.character(dat$espèce)])
```

Le jeu de données `dat` comporte à présent une variable supplémentaire `dmoy` qui donne la densité spécifique du bois. Le modèle est ajusté par la commande:

```
m <- lm(log(Btot)~I(log(dbh^2*haut))+I(log(dmoy)),data=dat[dat$Btot>0,])
summary(m)
```

ce qui donne:

```
                    Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)         -8.38900     0.26452  -31.714   < 2e-16  ***
I(log(dbh^2*haut))   0.85715     0.02031   42.205   < 2e-16  ***
I(log(dmoy))         0.72864     0.17720    4.112  0.000202  ***
```

avec un écart-type résiduel de 0,3442 et $R^2 = 0{,}9806$. Le modèle s'écrit: $\ln(B) = -8{,}38900 + 0{,}85715\ln(D^2H) + 0{,}72864\ln(\rho)$. Vaut-il mieux prendre en compte l'espèce *via* la densité du bois comme on vient de le faire ou en construisant des tarifs spécifiques comme on l'a fait dans le fil rouge n° 30 ? Pour répondre à cette question, on peut comparer le modèle (6.28) au modèle (6.29) en utilisant l'AIC:

```
AIC(m)
```

ce qui donne AIC = 34,17859 pour le tarif spécifique (6.28) et AIC = 33,78733 pour le tarif (6.29) utilisant la densité du bois. Ce dernier est donc légèrement préférable. La différence d'AIC reste cependant faible.

Afin de mieux prendre en compte les variations de densité de bois au sein d'un arbre, il est possible d'analyser les variations inter et intra-spécifique plutôt que d'utiliser une densité moyenne basée sur l'hypothèse que la densité de bois est la même de la moelle vers l'écorce ou du bas vers le haut des arbres (voir chapitre 1). La densité de bois peut être modélisée en prenant en compte des facteurs comme l'espèce, le groupe fonctionnel, la dimension de l'arbre, la position radiale et verticale dans l'arbre. Un premier test de comparaison peut être fait en utilisant le test d'analyse de variance de Friedman, puis le test de différence franchement significative (HSD) de Tuckey. Ces tests permettent de différencier les variables qui influencent le plus la densité de bois. Cette dernière peut ensuite être modélisée en utilisant ces variables (Henry *et al.*, 2010).

32

Tarif de biomasse dépendant de la densité individuelle du bois

Dans le fil rouge n° 31, la densité du bois ρ a été définie au niveau de l'espèce en calculant la moyenne des densités individuelles pour les arbres d'une même espèce. Ajustons à présent un tarif de biomasse basé sur la mesure individuelle de la densité du bois, afin de tenir compte de la variabilité inter-individuelle de densité au sein de l'espèce. Le modèle ajusté est:

$$\ln(B) = a_0 + a_1\ln(D) + a_2\ln(\rho) + \varepsilon$$

avec

$$\mathrm{Var}(\varepsilon) = \sigma^2$$

où ρ est ici, contrairement au fil rouge n° 31, la mesure *individuelle* de la densité du bois. Le modèle est ajusté par la commande:

```
m <- lm(log(Btot)~I(log(dbh))+I(log(dens)),data=dat[dat$Btot>0,])
summary(m)
```

ce qui donne:

```
              Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)   -7.76644     0.20618  -37.668   < 2e-16  ***
I(log(dbh))    2.35272     0.04812   48.889   < 2e-16  ***
I(log(dens))   1.00717     0.14053    7.167  1.46e-08  ***
```

avec un écart-type résiduel de 0,3052 et $R^2 = 0{,}9848$. Le modèle s'écrit: $\ln(B) = -7{,}76644 + 2{,}35272\ln(D) + 1{,}00717\ln(\rho)$. Selon ce modèle, la biomasse dépend de la densité individuelle *via* le terme $\rho^{1{,}00717}$, c'est-à-dire pratiquement ρ. À titre de comparaison, le modèle (6.29) dépendait de la densité spécifique du bois *via* le terme $\rho^{0{,}72864}$. D'un point de vue biologique, l'exposant 1,00717 est plus satisfaisant que l'exposant 0,72864 puisqu'il signifie que la biomasse est le produit d'un volume (qui dépend uniquement des dimensions de l'arbre) et d'une densité. L'écart entre ces deux valeurs d'exposant peut être attribué aux variations inter-individuelles de densité de bois au sein de l'espèce. Cependant, le modèle basé sur la densité individuelle du bois n'a guère d'utilité pratique puisqu'il implique qu'il faudrait mesurer la densité du bois de tout arbre dont on voudrait prédire la biomasse.

6.4.2 Compartiments de l'arbre

La biomasse des arbres est pesée séparément pour chaque compartiment de l'arbre (souche, tronc, grosses branches, petites branches, feuillage, etc.). La biomasse épigée est la somme de ces compartiments. La démarche que l'on a présentée pour l'ajustement d'un tarif pourra être suivie pour chaque compartiment séparément. On construira ainsi un tarif pour la biomasse foliaire, un tarif pour la biomasse des grosses branches, etc. Cette démarche intègre la stratification du jeu de données. Ainsi, on ajustera dans un premier temps un tarif pour chaque compartiment et chaque strate; dans un second temps, en fonction des différences trouvées entre strates, on pourra agréger les strates et/ou paramétrer le tarif de manière à construire un tarif par compartiment pour toutes les strates. La démarche ne s'arrête cependant pas là. On peut poursuivre l'intégration des données pour aller vers un plus petit nombre de modèles plus intégrateurs.

Additivité des compartiments

La biomasse épigée totale étant la somme des biomasses des compartiments, on pourrait penser que le meilleur tarif pour prédire la biomasse épigée est la somme des tarifs prédisant chaque compartiment. En fait, à cause des corrélations qui existent entre les biomasses des différents compartiments, ce n'est pas le cas (Cunia et Briggs, 1984, 1985*a*; Parresol, 1999). Qui plus est, certaines familles de modèles ne sont pas stables par l'addition. C'est le cas en particulier des modèles puissance: la somme de deux fonctions puissance n'est pas une fonction puissance. Si on a ajusté un modèle puissance pour chaque compartiment:

$$\begin{aligned} B^{\text{souche}} &= a_1 D^{b_1} \\ B^{\text{tronc}} &= a_2 D^{b_2} \\ B^{\text{grosses branches}} &= a_3 D^{b_3} \\ B^{\text{petites branches}} &= a_4 D^{b_4} \\ B^{\text{feuillage}} &= a_5 D^{b_5} \end{aligned}$$

la somme $B^{\text{épigée}} = B^{\text{souche}} + B^{\text{tronc}} + B^{\text{grosses branches}} + B^{\text{petites branches}} + B^{\text{feuillage}} = \sum_{m=1}^{5} a_m D^{b_m}$ n'est pas une fonction puissance du diamètre. Les modèles polynômiaux sont en revanche stables par l'addition.

Ajustement d'un modèle multivarié

Pour tenir compte des corrélations qui existent entre les biomasses des compartiments, on peut ajuster les tarifs relatifs aux différents compartiments de manière simultanée plutôt

que séparément. Cette dernière étape dans l'intégration du modèle nécessite une redéfinition de la variable réponse. Comme on veut prédire simultanément les biomasses des différents compartiments, on n'a plus affaire à une variable réponse mais à un *vecteur* réponse $\mathbf{Y}$. La longueur de ce vecteur est égal au nombre M de compartiments. Par exemple, si la variable réponse est la biomasse,

$$\mathbf{Y} = \begin{bmatrix} B^{\text{épigée}} \\ B^{\text{souche}} \\ B^{\text{tronc}} \\ B^{\text{grosses branches}} \\ B^{\text{petites branches}} \\ B^{\text{feuillage}} \end{bmatrix}$$

Si la variable réponse est le logarithme de la biomasse,

$$\mathbf{Y} = \begin{bmatrix} \ln(B^{\text{épigée}}) \\ \ln(B^{\text{souche}}) \\ \ln(B^{\text{tronc}}) \\ \ln(B^{\text{grosses branches}}) \\ \ln(B^{\text{petites branches}}) \\ \ln(B^{\text{feuillage}}) \end{bmatrix}$$

Soit Y_m la variable réponse du m^{e} compartiment (avec $m = 1, \ldots, M$). Sans perte de généralité, on peut considérer que tous les compartiments ont le même jeu $X_1, X_2, \ldots, X_p$ de variables explicatives (si une variable n'intervient pas dans la prédiction d'un compartiment, il suffira de mettre le coefficient correspondant à zéro). Un modèle qui prédit un vecteur réponse plutôt qu'une variable réponse est un modèle multivarié. Une observation pour l'ajustement d'un modèle multivarié consiste en un vecteur $(Y_1, \ldots, Y_M, X_1, \ldots, X_p)$ de longueur $M + p$. Le résidu d'un modèle multivarié est un vecteur $\boldsymbol{\varepsilon}$ de longueur M, égal à la différence entre le vecteur réponse observé et le vecteur réponse prédit.

L'expression d'un modèle M-varié ne diffère des M modèles univariés correspondant à chaque compartiment que par la structure de l'erreur résiduelle; la structure du modèle pour la moyenne ne change pas. Plaçons-nous dans le cas général d'un modèle non-linéaire. Si les M modèles univariés sont:

$$Y_m = f_m(X_1, \ldots, X_p; \theta_m) + \varepsilon_m \tag{6.30}$$

pour $m = 1, \ldots, M$, alors le modèle multivarié s'écrit:

$$\mathbf{Y} = \mathbf{F}(X_1, \ldots, X_p; \boldsymbol{\theta}) + \boldsymbol{\varepsilon}$$

où $\mathbf{Y} = {}^{\text{t}}[Y_1, \ldots, Y_M]$, $\boldsymbol{\theta} = {}^{\text{t}}[\theta_1, \ldots, \theta_M]$, et

$$\mathbf{F}(X_1, \ldots, X_p; \boldsymbol{\theta}) = \begin{bmatrix} f_1(X_1, \ldots, X_p; \theta_1) \\ \vdots \\ f_m(X_1, \ldots, X_p; \theta_m) \\ \vdots \\ f_M(X_1, \ldots, X_p; \theta_M) \end{bmatrix} \tag{6.31}$$

Le vecteur résiduel $\boldsymbol{\varepsilon}$ suit à présent une loi multinormale centrée, de matrice de variance-covariance:

$$\text{Var}(\boldsymbol{\varepsilon}) \equiv \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \zeta_{12} & \cdots & \zeta_{1M} \\ \zeta_{21} & \sigma_2^2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \zeta_{M-1,M} \\ \zeta_{M1} & \cdots & \zeta_{M,M-1} & \sigma_M^2 \end{bmatrix}$$

La matrice $\mathbf{\Sigma}$ est une matrice symétrique à M lignes et M colonnes, telle que $\sigma_m^2 = \text{Var}(\varepsilon_m)$ est la variance résiduelle de la biomasse du m^{e} compartiment et $\zeta_{ml} = \zeta_{lm}$ est la covariance résiduelle entre la biomasse du m^{e} compartiment et celle du l^{e} compartiment. Comme dans le cas univarié, deux résidus correspondant à deux observations différentes sont supposés indépendants: $\boldsymbol{\varepsilon}_i$ est indépendant de $\boldsymbol{\varepsilon}_j$ pour $i \neq j$. La différence vient du fait que les différents compartiments ne sont plus supposés être indépendants les uns des autres.

L'ajustement d'un modèle multivarié tel que (6.31) se fait selon les mêmes principes que les modèles univariés (6.30). Si la matrice de variance-covariance $\mathbf{\Sigma}$ était diagonale (c'est-à-dire $\zeta_{ml} = 0$, $\forall m$, l), alors l'ajustement du modèle multivarié (6.31) serait équivalent à l'ajustement séparé des M modèles univariés (6.30). Dans le cas d'un modèle linéaire, les valeurs estimées des coefficients θ_1, θ_2, …, θ_M résultant de l'ajustement du modèle linéaire M-varié sont identiques aux valeurs obtenues par les ajustements séparés des M modèles linéaires univariés (à condition de toujours garder les mêmes variables explicatives X_1, …, X_p dans tous les cas) (Muller et Stewart, 2006, chapitre 3). Cependant, les tests de significativité associés aux coefficients ne donnent pas les mêmes résultats dans les deux cas. Pour peu que les différents compartiments soient suffisamment corrélés entre eux, l'ajustement simultané des tous les compartiments par le modèle multivarié (6.31) aboutira à une estimation plus précise des coefficients du tarif, donc à des prédictions de la biomasse plus précises.

Harmonisation d'un tarif

Dans certains cas, en particulier dans le contexte du bois énergie, on souhaite prédire la biomasse sèche du tronc à différents diamètres de découpe. Par exemple, on souhaite prédire simultanément la biomasse totale B du tronc, la biomasse B_7 du tronc jusqu'à un diamètre fin bout de 7 cm, et la biomasse B_{10} du tronc jusqu'à un diamètre fin bout de 10 cm. On pourrait considérer le tronc entier, le tronc jusqu'à découpe 7 cm et le tronc jusqu'à découpe 10 cm comme trois compartiments différents et appliquer les mêmes principes d'ajustement que ceux présentés dans le précédent paragraphe. En fait le problème est plus compliqué car, contrairement aux compartiments tronc et feuillage qui sont distincts, les compartiments définis par différents diamètres de découpe sont emboîtés les uns dans les autres: $B = B_7+$ biomasse du tronçon allant du diamètre 7 cm au fin bout, et $B_7 = B_{10}+$ biomasse du tronçon allant du diamètre de découpe 10 cm à 7 cm. Ainsi, le modèle multivarié qui prédit le vecteur (B, B_7, B_{10}) doit faire en sorte que $B > B_7 > B_{10}$ sur tout le domaine de validité du tarif. Le processus consistant à contraindre le modèle multivarié pour qu'il prédise les biomasses des différents compartiments tout en vérifiant la logique de leur emboîtement s'appelle l'*harmonisation* d'un tarif (Parresol, 1999). Jacobs et Cunia (1980) et Cunia et Briggs (1985*b*) ont proposé des solutions à ce problème sous la forme d'équations reliant les coefficients des modèles des différents compartiments. Il faut alors ajuster un modèle M-varié (s'il y a M diamètres de découpe) tout en s'assurant que les coefficients θ_1, …, θ_M correspondant aux M diamètres de découpe vérifient un certain nombre d'équations les reliant. Lorsque les coefficients du modèle multivarié sont estimés par maximum de vraisemblance, leur estimation numérique se ramène à un problème d'optimisation sous contrainte.

Dans le cas de la prédiction du volume ou de la biomasse d'une tige, une alternative au tarif de cubage ou de biomasse est l'intégration du profil de tige (Parresol et Thomas, 1989; Parresol, 1999). Soit $P(h)$ un profil de tige, c'est-à-dire une courbe donnant la surface de la section transversale du tronc en fonction de la hauteur h à partir du sol (h représente aussi la longueur parcourue lorsque l'on suit la tige de son gros bout à son fin bout) (Maguire

et Batista, 1996; Dean et Roxburgh, 2006; Metcalf *et al.*, 2009). Si la section de la tige a une forme approximativement circulaire, le diamètre de l'arbre à la hauteur h peut être calculé comme: $D(h) = \sqrt{4P(h)/\pi}$. La biomasse du tronc jusqu'à un diamètre de découpe D se calcule en intégrant le profil de tige du sol ($h = 0$) jusqu'à la hauteur $P^{-1}(\frac{\pi}{4}D^2)$ qui correspond à ce diamètre:

$$B_D = \int_0^{P^{-1}(\frac{\pi}{4}D^2)} \rho(h)\ P(h)\,\mathrm{d}h$$

où $\rho(h)$ est la densité du bois à la hauteur h. Le volume de la tige jusqu'au diamètre de découpe D se calcule de la même manière, à cela près que ρ est remplacé par 1. L'approche par profil de tige a l'avantage que l'harmonisation du tarif est automatique. C'est une approche qui est cependant conceptuellement différente des tarifs de cubage et de biomasse, avec des problèmes d'ajustement spécifiques (Fang et Bailey, 1999; Parresol, 1999), et qui sort du cadre de ce manuel. À noter pour les très gros arbres pour lesquels la mesure directe de la biomasse est quasiment impossible, l'approche par profil de tiges présente une alternative pertinente (Van Pelt, 2001; Dean *et al.*, 2003; Dean, 2003; Dean et Roxburgh, 2006; Sillett *et al.*, 2010).

7

Utilisation et prédiction

Une fois le tarif de cubage ou de biomasse ajusté, plusieurs utilisations peuvent être faites de ses prédictions. Le plus souvent, il s'agira de prédire le volume ou la biomasse d'arbres dont le volume ou la biomasse n'ont pas été mesurés. Il s'agit là de la *prédiction* proprement dite (§ 7.2–7.4). Parfois, le volume ou la biomasse des arbres auront également été mesurés en plus des variables d'entrée du tarif. Quand on a ainsi à disposition un jeu de données *indépendant* de celui utilisé pour l'ajustement du modèle, et qui contient à la fois la variable réponse et les variables explicatives du modèle, il est possible de faire une *validation* du modèle (§ 7.1). Quand les critères de validation sont appliqués au jeu de données même qui a servi à la calibration du modèle, on parle de *vérification* du modèle. Nous n'insisterons pas sur la vérification d'un modèle, puisque celle-ci est déjà implicite dans l'analyse des résidus du modèle ajusté. Enfin, quand on dispose de tarifs qui existaient préalablement à un tarif ajusté, on peut vouloir également comparer les modèles ou les combiner (§ 7.5).

Domaine de validité du modèle

Avant toute utilisation d'un tarif, il faut s'assurer que les caractéristiques de l'arbre dont on veut prédire le volume ou la biomasse sont dans le *domaine de validité* du tarif (Rykiel, 1996). Si un tarif de cubage ou de biomasse a été ajusté pour des arbres de diamètre compris entre $D_{\min}$ et $D_{\max}$, il n'est en principe pas possible d'utiliser ce tarif pour prédire le volume ou la biomasse d'un arbre de diamètre inférieur à $D_{\min}$ ou supérieur à $D_{\max}$. Il en va de même pour toutes les entrées du tarif. Tous les modèles ne sont cependant pas sujets aux mêmes erreurs lorsqu'on les extrapole en dehors de leur domaine de validité. Les modèles de type puissance restent généralement extrapolables avec une bonne fiabilité en dehors de leur domaine de validité, car ces relations puissances reposent sur un modèle allométrique fractal qui est invariant à toutes les échelles (Zianis et Mencuccini, 2004). Au contraire, les modèles de type polynômial présentent fréquemment des comportements anormaux en dehors de leur domaine de validité (valeurs prédites négatives, par exemple), et ce d'autant plus que le degré du polynôme est élevé.

7.1 Validation d'un tarif

La validation d'un modèle consiste à confronter ses prédictions à des observations indépendantes de celles utilisées pour l'ajustement du modèle (Rykiel, 1996). Soit $(Y'_i, X'_{i1}, \ldots, X'_{ip})$ avec $i = 1, \ldots, n'$ un jeu de données de n' observations indépendant de celui utilisé pour l'ajustement d'un modèle f, où $X'_{i1}, \ldots, X'_{ip}$ sont les variables explicatives et Y'_i est la variable réponse, c'est-à-dire le volume, ou la biomasse, ou une transformée de l'une de ces quantités. Soit

$$\hat{Y}'_i = f(X'_{i1}, \ldots, X'_{ip}; \hat{\theta})$$

la valeur prédite de la variable réponse pour la i^{e} observation, où $\hat{\theta}$ sont les valeur estimées des paramètres du modèle. La validation consiste à comparer les valeurs prédites $\hat{Y}'_i$ aux valeurs observées Y'_i.

7.1.1 Critères de validation

Plusieurs critères, qui sont les pendants des critères utilisés pour évaluer la qualité d'ajustement d'un modèle, peuvent être utilisés pour comparer les prédictions aux observations (Schlaegel, 1982; Parresol, 1999; Tedeschi, 2006), notamment:

- le biais: $\sum_{i=1}^{n'} |Y'_i - \hat{Y}'_i|$
- la somme des carrés des écarts résiduels: $\text{SCE} = \sum_{i=1}^{n'} (Y'_i - \hat{Y}'_i)^2$
- la variance résiduelle: $s^2 = \text{SCE}/(n' - p)$
- l'erreur résiduelle ajustée: $\text{SCE}/(n' - 2p)$
- le R^2 de régression: $R^2 = 1 - s^2/\text{Var}(Y')$
- le critère d'information d'Akaike: $\text{AIC} = n' \ln(s^2) + n' \ln(1 - p/n') + 2p$

où $\text{Var}(Y')$ est la variance empirique de Y' et p est le nombre de paramètres librement estimés du tarif. Les deux premiers critères correspondent à deux normes distinctes de la différence entre le vecteur $(Y'_1, \ldots, Y'_{n'})$ des observations et le vecteur $(\hat{Y}'_1, \ldots, \hat{Y}'_{n'})$ des prédictions: norme L^1 pour le biais et norme L^2 pour la somme des carrés des écarts. Toute autre norme serait également valable. Les trois derniers critères font intervenir le nombre de paramètres utilisés dans le tarif, et sont donc plus appropriés quand il s'agit de comparer différents tarifs.

7.1.2 Validation croisée

Quand on ne dispose pas d'un jeu de données indépendant, il est tentant de partager le jeu de données de calibration en deux sous-jeux de données: un pour l'ajustement du modèle, et un autre pour la validation du modèle. Étant donné que les jeux de données de volume ou de biomasse sont coûteux et souvent de taille limitée, nous ne recommandons pas cette pratique dans le cas des tarifs de cubage ou de biomasse. En revanche, nous recommandons dans ce cas la mise en œuvre d'une *validation croisée* (Efron et Tibshirani, 1993, chapitre 17).

La validation croisée « K fois » consiste à diviser le jeu de données en K parts à peu près égales et à utiliser tour à tour chaque part comme jeu de données de validation, le modèle étant ajusté sur les $K - 1$ parts restantes. Le pseudo-algorithme de la validation croisée « K fois » est le suivant:

1. Partager le jeu de données $\mathcal{S}_n \equiv \{(Y_i, X_{i1}, \ldots, X_{ip}): i = 1, \ldots, n\}$ en K sous-jeux de données $\mathcal{S}_n^{(1)}, \ldots, \mathcal{S}_n^{(K)}$ de tailles à peu près égales (c'est-à-dire avec environ n/K observations dans chaque sous-jeu de données, le total faisant n).
2. Pour k allant de 1 à K:

(a) ajuster le modèle sur le jeu de données privé de sa k^{e} part, *i.e.* sur $\mathcal{S}_n \backslash \mathcal{S}_n^{(k)} = \mathcal{S}_n^{(1)} \cup \ldots \cup \mathcal{S}_n^{(k-1)} \cup \mathcal{S}_n^{(k+1)} \cup \ldots \cup \mathcal{S}_n^{(K)}$;

(b) calculer un critère de validation (cf. § 7.1.1) de ce modèle ajusté en prenant la part restante $\mathcal{S}_n^{(k)}$ comme jeu de données de validation; soit C_k la valeur de ce critère calculé pour $\mathcal{S}_n^{(k)}$.

3. Calculer la moyenne $(\sum_{k=1}^{K} C_k)/K$ des K critères de validation ainsi calculés.

L'absence de chevauchement entre les jeux données utilisés pour l'ajustement du modèle et ceux utilisés pour calculer le critère de validation assure la validité de la démarche. La validation croisée nécessite davantage de calculs qu'une validation simple, mais a l'avantage de tirer profit de toutes les observations disponibles pour l'ajustement du modèle.

Un cas particulier de validation croisée « K fois » est lorsque K est égal au nombre n d'observations disponibles dans le jeu de données. Cette méthode est aussi appelée validation croisée « leave-one-out » et est proche conceptuellement de la technique du Jackknife (Efron et Tibshirani, 1993). Le principe consiste à ajuster le modèle sur $n-1$ observations et à calculer l'erreur résiduelle pour l'observation mise de côté. Il est du reste utilisé en analyse des résidus pour quantifier l'influence des observations (c'est, en particulier, la base du calcul des distances de Cook, cf. Saporta, 1990).

7.2 Prédiction du volume ou de la biomasse d'un arbre

La prédiction à l'aide d'un modèle f consiste à calculer, pour des valeurs données des variables explicatives $X_1, \ldots, X_p$, la valeur prédite $\hat{Y}$ par le modèle de la variable réponse. Une prédiction ne s'arrête pas au calcul de

$$\hat{Y} = f(X_1, \ldots, X_p; \hat{\theta})$$

En effet, l'estimateur $\hat{\theta}$ des paramètres du modèle est un vecteur aléatoire dont la loi découle de la loi des observations utilisées pour ajuster le modèle. Toute prédiction $\hat{Y}$ du modèle est donc elle-même une variable aléatoire dont la loi de distribution découle de la loi des observations utilisées pour ajuster le modèle. Pour traduire cette variabilité intrinsèque de la prédiction, on l'assortira d'un indicateur d'incertitude tel que l'écart-type de la prédiction ou son intervalle de confiance à 95 %.

Il existe plusieurs intervalles de confiance selon que l'on prédit le volume ou la biomasse d'un arbre pris au hasard dans le peuplement, ou celui d'un arbre moyen du peuplement. Nous détaillerons les expressions analytiques de ces intervalles de confiance dans le cas du modèle linéaire (§ 7.2.1), puis dans le cas du modèle non-linéaire (§ 7.2.2). Des expression approchées mais plus simples à calculer de ces intervalles de confiance seront ensuite présentées (§ 7.2.3), avant de s'intéresser au cas des variables transformées (§ 7.2.4).

7.2.1 Prédiction: cas du modèle linéaire

Prédiction par une régression linéaire simple

Soit $\hat{a}$ l'ordonnée à l'origine estimée d'une régression linéaire, et $\hat{b}$ sa pente estimée. La prédiction $\hat{Y}$ de la variable réponse peut s'écrire de deux façons différentes:

$$\hat{Y} = \hat{a} + \hat{b}X \tag{7.1}$$

$$\hat{Y} = \hat{a} + \hat{b}X + \varepsilon \tag{7.2}$$

Dans les deux cas, l'espérance de $\hat{Y}$ est la même puisque $\mathrm{E}(\varepsilon) = 0$. En revanche, la variance de $\hat{Y}$ n'est pas la même dans les deux cas: elle plus élevée dans la deuxième écriture que dans la première. L'interprétation liée à ces deux écritures est la suivante. Mettons que la variable explicative X est le diamètre et la variable réponse Y la biomasse. Le nombre d'arbres dans la forêt entière ayant un diamètre X donné (à quelque chose près qui représente la précision de la mesure) est incommensurable. Si on pouvait mesurer la biomasse de l'intégralité de ces arbres ayant le même diamètre, on trouverait des valeurs variables, oscillant autour d'une certaine valeur moyenne. Quand on cherche à prédire cette biomasse moyenne (sous-entendu: moyenne sur l'ensemble des arbres existants ayant le diamètre X), c'est l'écriture (7.1) de la prédiction qui est valide. En revanche, si on cherche à prédire la biomasse d'un arbre pris au hasard parmi l'ensemble des arbres ayant le diamètre X, c'est l'écriture (7.2) de la prédiction qui est valide. La variabilité de la prédiction est plus forte pour (7.2) que pour (7.1) puisque, en plus de la variabilité de la prédiction de la biomasse moyenne, s'ajoute les différences de biomasse entre arbres dans le second cas.

Cela signifie qu'il y a deux façons de calculer un intervalle de confiance pour une prédiction. Il y a un intervalle de confiance pour la prédiction de la moyenne de Y, et un intervalle de confiance pour la prédiction d'un individu pris au hasard parmi la population sur laquelle la moyenne de Y est calculée. Le deuxième intervalle de confiance est plus large que le premier.

Dans le cas d'une régression linéaire simple, on peut montrer (Saporta, 1990, p.373–374) que l'intervalle de confiance au seuil α pour la prédiction (7.1) de la moyenne est:

$$\hat{a} + \hat{b}X \pm t_{n-2}\,\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(X-\bar{X})^2}{nS_X^2}} \tag{7.3}$$

tandis que l'intervalle de confiance au seuil α pour la prédiction (7.2) d'un arbre pris au hasard est:

$$\hat{a} + \hat{b}X \pm t_{n-2}\,\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(X-\bar{X})^2}{nS_X^2}} \tag{7.4}$$

où t_{n-2} est le quantile $1-\alpha/2$ d'une loi de Student à $n-2$ degrés de liberté, $\bar{X} = (\sum_{i=1}^{n} X_i)/n$ est la moyenne des valeurs observées de X dans le jeu de données ayant servi à l'ajustement du modèle, et $S_X^2 = [\sum_{i=1}^{n}(X_i - \bar{X})^2]/n$ est la variance empirique des valeurs observées de X dans le jeu de données ayant servi à l'ajustement du modèle.

Ces expressions appellent plusieurs remarques. La première, c'est que la différence entre les bornes de l'intervalle de confiance (7.4) pour un arbre pris au hasard et les bornes de l'intervalle de confiance (7.3) pour l'arbre moyen est de l'ordre de $t_{n-2}\hat{\sigma}$. Cette différence fait écho à la différence entre l'écriture (7.2) et (7.1), qui tient au terme résiduel ε dont l'écart-type estimé est $\hat{\sigma}$.

La seconde, c'est que la largeur de l'intervalle de confiance n'est pas constante, mais varie avec X. L'intervalle de confiance est le plus étroit lorsque $X = \bar{X}$ et s'élargit lorsque X s'éloigne de $\bar{X}$.

La troisième remarque, c'est que pour calculer l'intervalle de confiance d'une prédiction selon une régression linéaire, il faut disposer, sinon des données originales ayant servi à ajuster le modèle, du moins de la moyenne $\bar{X}$ de la variable explicative et de son écart-type empirique S_X. Pour peu que les données originales ayant servi à ajuster le modèle ne soient plus disponibles, et que les valeurs de $\bar{X}$ et S_X n'aient pas été documentées, on ne pourra pas calculer l'intervalle de confiance de façon exacte.

33

Intervalle de confiance de $\ln(B)$ **prédit par** $\ln(D)$

Reprenons la régression linéaire simple entre $\ln(B)$ et $\ln(D)$ qui a été ajustée dans le fil rouge n° 7. Soit `m` l'objet contenant le modèle ajusté (cf. fil rouge n° 7). Les intervalles de confiance peuvent être calculés avec la commande `predict`. Par exemple pour un arbre de diamètre 20 cm, l'intervalle de confiance au seuil 95 % pour l'arbre moyen s'obtient par la commande:

```
predict(m,newdata=data.frame(dbh=20),interval="confidence",level=0.95)
```

ce qui donne:

```
          fit        lwr        upr
1  -1.354183  -1.533487  -1.174879
```

Ainsi le modèle prédit $\ln(B) = -1{,}354183$ avec un intervalle de confiance à 95 % allant de $-1{,}533487$ à $-1{,}174879$. Pour un arbre de 20 cm pris au hasard, l'intervalle de confiance s'obtient par la commande:

```
predict(m,newdata=data.frame(dbh=20),interval="prediction",level=0.95)
```

ce qui donne:

```
          fit        lwr        upr
1  -1.354183  -2.305672  -0.4026948
```

La figure 7.1 montre les intervalles de confiance sur l'ensemble de la plage de données.

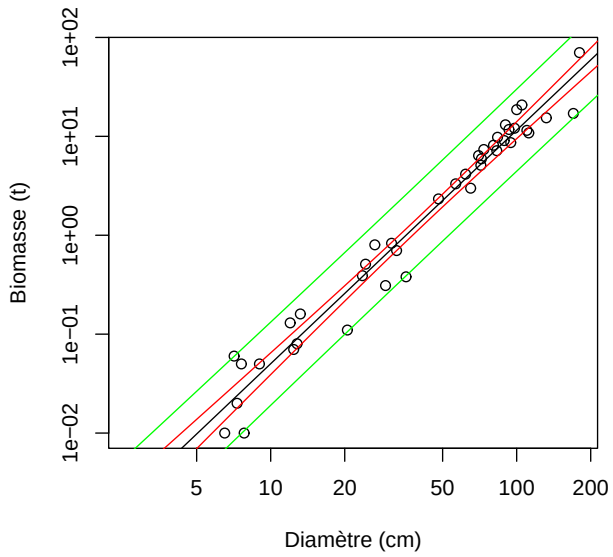


Figure 7.1 – *Données de biomasse en fonction du diamètre (en coordonnées logarithmiques) pour 42 arbres mesurés au Ghana par Henry et al. (2010) (points), prédiction (trait noir) de la régression linéaire simple de* $\ln(B)$ *par rapport à* $\ln(D)$*, et intervalles de confiance de cette prédiction, pour un arbre pris au hasard (trait vert) et pour l'arbre moyen (trait rouge).*

Prédiction par une régression multiple

Les principes de la prédiction exposés dans le cas de la régression linéaire s'étendent immédiatement à la régression multiple. Il y a deux expressions de l'intervalle de confiance: l'une pour la prédiction de l'arbre moyen, l'autre pour la prédiction d'un arbre pris au hasard.

Dans le cas d'une régression multiple de coefficients estimés $\hat{\mathbf{a}} = {}^{\mathrm{t}}[\hat{a}_0, \hat{a}_1, \hat{a}_2, \ldots, \hat{a}_p]$, la valeur prédite $\hat{Y}$ de la variable réponse pour un arbre dont les variables explicatives sont $\mathbf{x} = {}^{\mathrm{t}}[1, X_1, X_2, \ldots, X_p]$, est:

$$\hat{Y} = {}^{\mathrm{t}}\mathbf{x}\,\hat{\mathbf{a}}$$

et l'intervalle de confiance au seuil α de cette prédiction est (Saporta, 1990, p.387):

– pour la prédiction de l'arbre moyen:

$$ {}^{\mathrm{t}}\mathbf{x}\,\hat{\mathbf{a}} \pm t_{n-p-1}\,\hat{\sigma}\sqrt{{}^{\mathrm{t}}\mathbf{x}({}^{\mathrm{t}}\mathbf{X}\mathbf{X})^{-1}\mathbf{x}} \tag{7.5}$$

– pour la prédiction d'un arbre pris au hasard:

$$ {}^{\mathrm{t}}\mathbf{x}\,\hat{\mathbf{a}} \pm t_{n-p-1}\,\hat{\sigma}\sqrt{1 + {}^{\mathrm{t}}\mathbf{x}({}^{\mathrm{t}}\mathbf{X}\mathbf{X})^{-1}\mathbf{x}} \tag{7.6}$$

où $\mathbf{X}$ est la matrice du plan construite à partir des données ayant servi à l'ajustement de la régression multiple. Pour calculer l'intervalle de confiance des prédictions il faut connaître, sinon les données originales ayant servi à l'ajustement du modèle, du moins la matrice $({}^{\mathrm{t}}\mathbf{X}\mathbf{X})^{-1}$. On notera que la variance des prédictions dans le cas (7.6) d'un arbre pris au hasard se compose de deux termes: un terme $\hat{\sigma}^2$ qui représente l'erreur résiduelle et un terme $\hat{\sigma}^2\ {}^{\mathrm{t}}\mathbf{x}({}^{\mathrm{t}}\mathbf{X}\mathbf{X})^{-1}\mathbf{x}$ qui représente la variabilité induite par l'estimation des coefficients du modèle. Dans le cas de l'estimation de l'arbre moyen, le premier terme disparaît et seul reste le second terme.

34

Intervalle de confiance de $\ln(B)$ prédit par $\ln(D)$ et $\ln(H)$

Reprenons la régression linéaire multiple entre $\ln(B)$, $\ln(D)$ et $\ln(H)$ qui a été ajustée dans le fil rouge n° 10. Soit `m` l'objet contenant le modèle ajusté (cf. fil rouge n° 10). Les intervalles de confiance peuvent être calculés avec la commande `predict`. Par exemple pour un arbre de diamètre 20 cm et de hauteur 20 m, l'intervalle de confiance au seuil 95 % pour l'arbre moyen s'obtient par la commande:

```
predict(m,newdata=data.frame(dbh=20,haut=20),interval="confidence",level=0.95)
```

ce qui donne:

```
          fit        lwr        upr
1  -1.195004  -1.380798  -1.009211
```

Ainsi le modèle prédit $\ln(B) = -1{,}195004$ avec un intervalle de confiance à 95 % allant de $-1{,}380798$ à $-1{,}009211$. Pour un arbre de 20 cm de diamètre et de 20 m de haut pris au hasard, l'intervalle de confiance s'obtient par la commande:

```
predict(m,newdata=data.frame(dbh=20,haut=20),interval="prediction",level=0.95)
```

ce qui donne:

```
          fit        lwr         upr
1  -1.195004  -2.046408  -0.3436006
```

7.2.2 Prédiction: cas d'un modèle non-linéaire

Dans le cas général d'un modèle non-linéaire tel que défini par

$$Y = f(X_1, \ldots, X_p; \theta) + \varepsilon$$

avec

$$\varepsilon \sim \mathcal{N}(0,\ kX_1^c)$$

il n'y a pas d'expression explicite exacte des intervalles de confiance des prédictions, comme c'est le cas pour le modèle linéaire. Néanmoins, la δ-méthode permet d'obtenir une expression approchée (et asymptotiquement exacte) des intervalles de confiance (Serfling, 1980). Comme précédemment, il y a deux intervalles de confiance:

– intervalle de confiance pour la prédiction de l'arbre moyen:

$$f(X_1, \ldots, X_p; \hat{\theta}) \pm t_{n-q} \sqrt{{}^{\mathrm{t}}[\mathrm{d}_\theta f(\hat{\theta})]\ \hat{\mathbf{\Sigma}}_{\hat{\theta}}\ [\mathrm{d}_\theta f(\hat{\theta})]} \tag{7.7}$$

– intervalle de confiance pour la prédiction d'un arbre pris au hasard:

$$f(X_1, \ldots, X_p; \hat{\theta}) \pm t_{n-q} \sqrt{\hat{k}^2 X_1^{2\hat{c}} + {}^{\mathrm{t}}[\mathrm{d}_\theta f(\hat{\theta})]\ \hat{\mathbf{\Sigma}}_{\hat{\theta}}\ [\mathrm{d}_\theta f(\hat{\theta})]} \tag{7.8}$$

où q est le nombre de coefficients du modèle (c'est-à-dire la longueur du vecteur θ), $\mathrm{d}_\theta f(\hat{\theta})$ est la valeur en $\theta = \hat{\theta}$ de la différentielle de f par rapport aux coefficients du modèle, et $\hat{\mathbf{\Sigma}}_{\hat{\theta}}$ est une estimation en $\theta = \hat{\theta}$ de la matrice de variance-covariance $\mathbf{\Sigma}_\theta$ de l'estimateur de θ. La différentielle de f par rapport aux coefficients du modèle est le vecteur de longueur q:

$$\mathrm{d}_\theta f(\theta) = {}^{\mathrm{t}}\left[\left(\frac{\partial f(X_1, \ldots, X_p; \theta)}{\partial \theta_1}\right), \ \ldots, \ \left(\frac{\partial f(X_1, \ldots, X_p; \theta)}{\partial \theta_q}\right)\right]$$

où θ_i est le i^{e} élément du vecteur θ, Dans le cas de l'estimateur du maximum de vraisemblance de θ, on peut montrer qu'asymptotiquement lorsque $n \to \infty$ (Saporta, 1990, p.301):

$$\mathbf{\Sigma}_\theta \underset{n\to\infty}{\sim} \mathbf{I}_n(\theta)^{-1} = \frac{1}{n}\, \mathbf{I}_1(\theta)^{-1}$$

où $\mathbf{I}_n(\theta)$ est la matrice de l'information de Fisher apportée par un échantillon de taille n sur le vecteur de paramètres θ. Cette matrice d'information de Fisher a q lignes et q colonnes et se calcule à partir de la dérivée seconde de la log-vraisemblance de l'échantillon:

$$\mathbf{I}_n(\theta) = -\mathrm{E}\left[\frac{\partial^2 \mathcal{L}(\theta)}{\partial \theta^2}\right]$$

Une estimation approchée de la matrice de variance-covariance des paramètres est donc:

$$\hat{\mathbf{\Sigma}}_{\hat{\theta}} = -\left[\left(\frac{\partial^2 \mathcal{L}(\theta)}{\partial \theta^2}\right)\Big|_{\theta=\hat{\theta}}\right]^{-1}$$

En pratique, l'algorithme qui optimise numériquement la log-vraisemblance de l'échantillon donne en même temps une estimation numérique de la dérivée seconde $(\partial^2 \mathcal{L}/\partial \theta^2)$. On obtient donc immédiatement une estimation numérique de $\hat{\mathbf{\Sigma}}_{\hat{\theta}}$.

Comme précédemment, la variance des prédictions dans le cas (7.8) d'un arbre pris au hasard se compose de deux termes: un terme $(\hat{k}X_1^{\hat{c}})^2$ qui représente l'erreur résiduelle et un terme ${}^{\mathrm{t}}[\mathrm{d}_\theta f(\hat{\theta})]\ \hat{\mathbf{\Sigma}}_{\hat{\theta}}\ [\mathrm{d}_\theta f(\hat{\theta})]$ qui représente la variabilité induite par l'estimation des coefficients du modèle. Dans le cas de l'estimation de l'arbre moyen, le premier terme disparaît et seul reste le second terme.

7.2.3 Intervalles de confiance approchés

Le calcul exact des intervalles de confiance des prédictions demande des informations (matrice du plan $\mathbf{X}$ dans le cas du modèle linéaire, matrice de variance-covariance $\hat{\mathbf{\Sigma}}_{\hat{\theta}}$ dans le cas non-linéaire) qui ne sont que très rarement renseignées dans les publications sur les tarifs. Le plus souvent, les publications ne renseignent que sur le nombre n d'observations utilisées pour ajuster le modèle et sur l'écart-type résiduel $\hat{\sigma}$ (cas linéaire) ou $\hat{k}$ et $\hat{c}$ (cas non-linéaire). Parfois, ces informations de base sur l'ajustement ne sont pas même données. Dès lors que $\mathbf{X}$ (cas du modèle linéaire) ou $\hat{\mathbf{\Sigma}}_{\hat{\theta}}$ (cas non-linéaire) ne sont pas données, il n'est pas possible d'utiliser les formules précédemment données pour le calcul des intervalles de confiance. Dans ce cas, on utilisera une méthode approchée.

Erreur résiduelle seule

Bien souvent, seul l'écart-type résiduel $\hat{\sigma}$ (cas linéaire) ou $\hat{k}$ et $\hat{c}$ (cas non-linéaire) est donné. Dans ce cas, un intervalle de confiance approché au seuil α pourra être construit:

– dans le cas d'une régression linéaire:

$$(a_0 + a_1 X_1 + \ldots + a_p X_p) \pm q_{1-\alpha/2}\, \hat{\sigma} \tag{7.9}$$

– dans le cas d'un modèle non-linéaire:

$$f(X_1, \ldots, X_p; \theta) \pm q_{1-\alpha/2}\, \hat{k} X_1^{\hat{c}} \tag{7.10}$$

où $q_{1-\alpha/2}$ est le quantile $1-\alpha/2$ de la loi normale centrée réduite. Cet intervalle de confiance est une retranscription directe de la relation $Y = a_0 + a_1 X_1 + \ldots + a_p X_p + \varepsilon$ avec $\varepsilon \sim \mathcal{N}(0, \hat{\sigma})$ (cas linéaire) ou $Y = f(X_1, \ldots, X_p; \theta) + \varepsilon$ avec $\varepsilon \sim \mathcal{N}(0, \hat{k} X_1^{\hat{c}})$ (cas non-linéaire), où on a volontairement écrit les coefficients du modèle sans accent circonflexe pour souligner qu'il s'agit ici de grandeurs fixes. Ces relations supposent donc implicitement que les coefficients du modèle sont connus de manière exacte et que la seule source de variabilité est l'erreur résiduelle. Autrement dit, l'interprétation de ces intervalles de confiance approchés est la suivante: les intervalles de confiance (7.9) (cas linéaire) et (7.10) (cas non-linéaire) sont ceux que l'on obtiendrait pour la prédiction d'un arbre *pris au hasard* si la taille d'échantillon était *infinie*. On vérifiera en effet que lorsque $n \to \infty$, t_{n-p-1} tend vers $q_{1-\alpha/2}$ et la matrice $({}^{\mathrm{t}}\mathbf{X}\mathbf{X})^{-1}$ dans (7.6) tend vers la matrice nulle (dont tous les coefficients valent zéro). Ainsi l'intervalle de confiance (7.9) est bien la limite de l'intervalle de confiance (7.6) lorsque $n \to \infty$. Il en va de même pour (7.8) et (7.10).

Intervalle de confiance pour l'arbre moyen

Lorsqu'une estimation $\hat{\mathbf{\Sigma}}$ de la matrice de variance-covariance des paramètres est donnée, un intervalle de confiance au seuil α de la prédiction pour l'arbre moyen est:

– dans le cas du modèle linéaire:

$$(\hat{a}_0 + \hat{a}_1 X_1 + \ldots + \hat{a}_p X_p) \pm q_{1-\alpha/2} \sqrt{{}^{\mathrm{t}}\mathbf{x}\hat{\mathbf{\Sigma}}\mathbf{x}} \tag{7.11}$$

où $\mathbf{x}$ est le vecteur ${}^{\mathrm{t}}[X_1, \ldots, X_p]$,

– dans le cas du modèle non-linéaire:

$$f(X_1, \ldots, X_p; \hat{\theta}) \pm q_{1-\alpha/2} \sqrt{{}^{\mathrm{t}}[\mathrm{d}_\theta f(\hat{\theta})]\ \hat{\mathbf{\Sigma}}\ [\mathrm{d}_\theta f(\hat{\theta})]} \tag{7.12}$$

Ces intervalles de confiance considèrent que toute la variabilité de la prédiction provient de l'estimation des coefficients du modèle. De plus, ces intervalles de confiance sont une retranscription directe du fait que les coefficients du modèle suivent une loi multinormale de moyenne leur valeur vraie et de matrice variance-covariance $\hat{\mathbf{\Sigma}}$. En effet, dans le cas linéaire, si $\hat{\mathbf{a}} = {}^{\mathrm{t}}[\hat{a}_1, \ldots, \hat{a}_p]$ suit une loi multinormale de moyenne ${}^{\mathrm{t}}[a_1, \ldots, a_p]$ et de matrice de variance-covariance $\hat{\mathbf{\Sigma}}$, alors la combinaison linéaire ${}^{\mathrm{t}}\mathbf{x}\,\hat{\mathbf{a}}$ suit une loi normale de moyenne ${}^{\mathrm{t}}\mathbf{x}\,\mathbf{a}$ et de variance ${}^{\mathrm{t}}\mathbf{x}\hat{\mathbf{\Sigma}}\mathbf{x}$ (Saporta, 1990, p.85).

Dans le cas du modèle linéaire, on peut montrer que la matrice de variance-covariance de l'estimateur des coefficients du modèle est (Saporta, 1990, p.380): $\mathbf{\Sigma} = \sigma^2({}^{\mathrm{t}}\mathbf{X}\mathbf{X})^{-1}$. Ainsi, une estimation de cette matrice de variance-covariance est: $\hat{\mathbf{\Sigma}} = \hat{\sigma}^2({}^{\mathrm{t}}\mathbf{X}\mathbf{X})^{-1}$. En reportant cette expression dans (7.11), on retrouve bien une expression semblable à (7.5). On vérifie de même, dans le cas non-linéaire, que l'intervalle de confiance (7.12) est une approximation de (7.7).

Dans le cas non-linéaire (7.12), si on veut éviter d'avoir à calculer les dérivées partielles de f, on pourra utiliser une méthode de Monte Carlo. C'est une méthode fondée sur la simulation, consistant à faire Q tirages des coefficients θ selon une loi multinormale de moyenne $\hat{\theta}$ et de matrice de variance-covariance $\hat{\mathbf{\Sigma}}$, à calculer la prédiction pour chacune de ces valeurs simulées, puis à calculer l'intervalle de confiance empirique de ces Q prédictions. Cette méthode est désignée dans la littérature en anglais sous le nom de « population prediction intervals » (Bolker, 2008; Paine *et al.*, 2012). Le pseudo-algorithme est le suivant:

1. Pour k allant de 1 à Q:
   (a) tirer un vecteur $\hat{\theta}^{(k)}$ selon une loi multinormale de moyenne $\hat{\theta}$ et de matrice de variance-covariance $\hat{\mathbf{\Sigma}}$;
   (b) calculer la prédiction $\hat{Y}^{(k)} = f(X_1, \ldots, X_p; \hat{\theta}^{(k)})$.
2. L'intervalle de confiance de la prédiction est l'intervalle de confiance empirique des Q valeurs $\hat{Y}^{(1)}, \ldots, \hat{Y}^{(Q)}$.

Bien souvent, on ne connaît pas la matrice de variance-covariance $\hat{\mathbf{\Sigma}}$, mais on a du moins une estimation des écarts-type des coefficients. Soit $\mathrm{Var}(\hat{a}_i) = \Sigma_i$ (cas linéaire) ou $\mathrm{Var}(\hat{\theta}_i) = \Sigma_i$ (cas non-linéaire) la variance du i^{e} coefficient du modèle. Dans ce cas on négligera la corrélation entre les coefficients du modèle et on approchera la matrice de variance-covariance des coefficients par une matrice diagonale:

$$\hat{\mathbf{\Sigma}} \approx \begin{bmatrix} \hat{\Sigma}_1 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \hat{\Sigma}_p \end{bmatrix}$$

Intervalle de confiance pour un arbre pris au hasard

L'erreur résultant de l'estimation des coefficients du modèle telle que décrite dans le précédent paragraphe peut être cumulée avec l'erreur résiduelle décrite dans l'avant-dernier paragraphe pour construire un intervalle de confiance de la prédiction pour un arbre pris au hasard. Ce sont les variances des prédictions qui s'ajoutent. L'intervalle de confiance au seuil α sera ainsi:

– dans le cas du modèle linéaire:

$$(\hat{a}_0 + \hat{a}_1 X_1 + \ldots + \hat{a}_p X_p) \pm q_{1-\alpha/2}\sqrt{\hat{\sigma}^2 + {}^{\mathrm{t}}\mathbf{x}\hat{\mathbf{\Sigma}}\mathbf{x}}$$

qui est une approximation de (7.6),

– dans le cas non-linéaire:

$$f(X_1, \ldots, X_p; \hat{\theta}) \pm q_{1-\alpha/2} \sqrt{\hat{k}^2 X_1^{2\hat{c}} + {}^{\mathrm{t}}[\mathrm{d}_\theta f(\hat{\theta})]\ \hat{\mathbf{\Sigma}}\ [\mathrm{d}_\theta f(\hat{\theta})]}$$

qui est une approximation de (7.8).

Comme précédemment, si on veut éviter de faire des calculs, on pourra utiliser une méthode de Monte Carlo selon le pseudo-algorithme suivant:

1. Pour k allant de 1 à Q:
   (a) tirer un vecteur $\hat{\theta}^{(k)}$ selon une loi multinormale de moyenne $\hat{\theta}$ et de matrice de variance-covariance $\hat{\mathbf{\Sigma}}$;
   (b) tirer un résidu $\hat{\varepsilon}^{(k)}$ selon une loi normale centrée d'écart-type $\hat{\sigma}$ (cas linéaire) ou $\hat{k} X_1^{\hat{c}}$ (cas non-linéaire);
   (c) calculer la prédiction $\hat{Y}^{(k)} = f(X_1, \ldots, X_p; \hat{\theta}^{(k)}) + \hat{\varepsilon}^{(k)}$.
2. L'intervalle de confiance de la prédiction est l'intervalle de confiance empirique des Q valeurs $\hat{Y}^{(1)}, \ldots, \hat{Y}^{(Q)}$.

Intervalle de confiance avec incertitudes sur les mesures

L'ajustement des tarifs de cubage et de biomasse suppose que les variables explicatives $X_1, \ldots, X_p$ sont connues de manière exacte. Cette hypothèse n'est en réalité qu'une approximation puisque ces grandeurs sont mesurées et sont donc sujettes à une erreur de mesure. Il ne faut pas confondre l'erreur de mesure avec l'erreur résiduelle de la variable réponse: la première est liée à l'instrument de mesure et peut être en principe rendue aussi petite que l'on souhaite en utilisant des instruments de mesure de plus en plus précis; la seconde reflète une variabilité biologique intrinsèque entre les individus. On peut rendre compte de l'impact de l'erreur de mesure sur la prédiction en l'incorporant dans l'intervalle de confiance de la prédiction. Ainsi, les variables explicatives $X_1, \ldots, X_p$ ne sont plus considérées comme fixes mais distribuées selon une certaine loi. Typiquement, pour prédire le volume ou la biomasse d'un arbre de caractéristiques $X_1, \ldots, X_p$, on considérera que la i^{e} caractéristique est distribuée selon une loi normale de moyenne X_i et d'écart-type τ_i. Typiquement, si X_i est un diamètre, on prendra τ_i de l'ordre de 3–5 mm; si X_i est une hauteur, on prendra τ_i de l'ordre de 3 % de X_i pour $X_i \leq 15$ m et de l'ordre de 1 m pour $X_i > 15$ m.

Il est difficile de calculer une expression explicite de l'intervalle de confiance de la prédiction lorsque les variables explicatives sont considérées comme aléatoires, puisque cela implique de calculer des variances de produits de variables aléatoires dont certaines sont corrélées entre elles. La δ-méthode offre une solution analytique approchée (Serfling, 1980). Ou alors, plus simplement, on peut à nouveau utiliser une méthode de Monte Carlo. Le pseudo-algorithme devient:

1. Pour k allant de 1 à Q:
   (a) pour i allant de 1 à p, tirer $\hat{X}_i^{(k)}$ selon une loi normale de moyenne X_i et d'écart-type τ_i;
   (b) tirer un vecteur $\hat{\theta}^{(k)}$ selon une loi multinormale de moyenne $\hat{\theta}$ et de matrice de variance-covariance $\hat{\mathbf{\Sigma}}$;
   (c) tirer un résidu $\hat{\varepsilon}^{(k)}$ selon une loi normale centrée d'écart-type $\hat{\sigma}$ (cas linéaire) ou $\hat{k} X_1^{\hat{c}}$ (cas non-linéaire);
   (d) calculer la prédiction $\hat{Y}^{(k)} = f(\hat{X}_1^{(k)}, \ldots, \hat{X}_p^{(k)}; \hat{\theta}^{(k)}) + \hat{\varepsilon}^{(k)}$.

2. L'intervalle de confiance de la prédiction est l'intervalle de confiance empirique des Q valeurs $\hat{Y}^{(1)}, \ldots, \hat{Y}^{(Q)}$.

Cet intervalle de confiance correspond en l'occurrence à la prédiction d'un arbre pris au hasard. Pour obtenir l'intervalle de confiance pour l'arbre moyen, il suffit d'appliquer le même pseudo-algorithme en remplaçant l'étape (c) par:

(...)

(c) poser $\hat{\varepsilon}^{(k)} = 0$;

(...)

7.2.4 Transformation inverse des variables

On a vu dans la section 6.1.5 comment une transformation de variable pouvait linéariser un modèle qui initialement ne se conformait pas aux hypothèses du modèle linéaire. La transformation de variable agit à la fois sur la moyenne et sur l'erreur résiduelle. Il en sera de même pour la transformation inverse, avec des implications sur le calcul de l'espérance des prédictions. La transformation logarithmique est la plus répandue. Cependant, d'autres types de transformation existent également.

Transformation logarithmique

Considérons d'abord le cas de la transformation logarithmique sur le volume ou la biomasse, qui est de loin le cas le plus fréquent pour les tarifs de cubage et de biomasse. Supposons qu'une transformation logarithmique a été appliquée à la biomasse B pour ajuster un modèle linéaire par rapport à des variables explicatives $X_1, \ldots, X_p$:

$$Y = \ln(B) = a_0 + a_1 X_1 + \ldots + a_p X_p + \varepsilon \tag{7.13}$$

avec

$$\varepsilon \underset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma)$$

Cela équivaut à dire que $\ln(B)$ suit une loi normale de moyenne $a_0 + a_1 X_1 + \ldots + a_p X_p$ et d'écart-type σ ou encore, par définition, que B suit une loi log-normale de paramètres $a_0 + a_1 X_1 + \ldots + a_p X_p$ et σ. L'espérance de cette loi log-normale est:

$$\mathrm{E}(B) = \exp\left(a_0 + a_1 X_1 + \ldots + a_p X_p + \frac{\sigma^2}{2}\right)$$

Par rapport au modèle inverse de (7.13) qui est $B = \exp(a_0 + a_1 X_1 + \ldots + a_p X_p)$, la transformation inverse de l'erreur résiduelle induit un biais de prédiction qui peut être corrigé en multipliant la prédiction $\exp(a_0 + a_1 X_1 + \ldots + a_p X_p)$ par le coefficient correcteur (Parresol, 1999):

$$\mathrm{CF} = \exp\left(\frac{\hat{\sigma}^2}{2}\right) \tag{7.14}$$

On prendra garde que dans les tarifs de biomasse reportés dans la littérature et qui ont été ajustés après transformation logarithmique de la biomasse, ce coefficient correcteur est parfois inclus dans l'expression du tarif, et parfois pas.

Lorsque le logarithme décimal $\log_{10}$ a été utilisé pour la transformation de variable plutôt que le logarithme népérien, le coefficient correcteur est:

$$\mathrm{CF} = \exp\left[\frac{(\hat{\sigma}\ln 10)^2}{2}\right] \approx \exp\left(\frac{\hat{\sigma}^2}{0{,}3772}\right)$$

35

Coefficient de correction de la biomasse prédite

Reprenons l'exemple du tarif de biomasse ajusté dans le fil rouge n° 31 par régression multiple sur les données log-transformées:

$$\ln(B) = -8{,}38900 + 0{,}85715\ln(D^2H) + 0{,}72864\ln(\rho)$$

Si on revient aux données de départ en utilisant la fonction exponentielle (sans tenir compte du facteur de correction), on obtient une prédiction sous-estimée: $B = \exp(-8{,}38900) \times (D^H)^{0{,}85715}\rho^{0{,}72864} = 2{,}274 \times 10^{-4}(D^2H)^{0{,}85715}\rho^{0{,}72864}$. Soit `m` l'objet contenant le modèle ajusté (cf. fil rouge n° 31). Le facteur de correction $\mathrm{CF} = \exp(\hat{\sigma}^2/2)$ s'obtient par la commande:

```
exp(summary(m)$sigma^2/2)
```

et vaut ici 1,061035. Le modèle correct est donc: $B = 2{,}412 \times 10^{-4}(D^2H)^{0{,}85715}\rho^{0{,}72864}$.

Transformation quelconque

Dans le cas général, soit ψ une transformation de variable de la biomasse (ou du volume) telle que la variable réponse $Y = \psi(B)$ puisse être prédite par une régression linéaire par rapport à des variables explicatives $X_1, \ldots, X_p$. On supposera la fonction ψ dérivable et inversible. Comme $\psi(B)$ suit une loi normale de moyenne $a_0 + a_1X_1 + \ldots + a_pX_p$ et d'écart-type σ, $B = \psi^{-1}[\psi(B)]$ a pour espérance (Saporta, 1990, p.26):

$$\mathrm{E}(B) = \int \psi^{-1}(x)\ \phi(x)\ \mathrm{d}x \tag{7.15}$$

où ϕ est la densité de probabilité de la loi normale de moyenne $a_0 + a_1X_1 + \ldots + a_pX_p$ et d'écart-type σ. Cette espérance est généralement différente de $\psi^{-1}(a_0 + a_1X_1 + \ldots + a_pX_p)$: la transformation de variable induit un biais de prédiction quand on revient à la variable de départ par la transformation inverse. L'inconvénient de la formule (7.15) est qu'elle nécessite le calcul d'une intégrale.

Lorsque l'écart-type résiduel σ est petit, la δ-méthode (Serfling, 1980) fournit une expression approchée de ce biais de prédiction:

$$\begin{aligned}\mathrm{E(B)} &\simeq \psi^{-1}[\mathrm{E}(Y)] + \frac{1}{2}\mathrm{Var}(Y)\ (\psi^{-1})''[\mathrm{E}(Y)] \\ &\simeq \psi^{-1}(a_0 + a_1X_1 + \ldots + a_pX_p) + \frac{\sigma^2}{2}\ (\psi^{-1})''(a_0 + a_1X_1 + \ldots + a_pX_p)\end{aligned}$$

Estimation « smearing »

La méthode d'estimation « smearing » (que l'on pourrait traduire par « étalante ») est une méthode non-paramétrique de correction du biais de prédiction lorsque l'on applique une transformation inverse à la variable réponse d'un modèle linéaire (Duan, 1983; Taylor, 1986; Manning et Mullahy, 2001). Étant donné que l'on peut réécrire l'équation (7.15) de l'espérance de la biomasse (ou du volume) sous la forme:

$$\begin{aligned}\mathrm{E}(B) &= \int \psi^{-1}(x)\ \phi_0(x - a_0 - a_1X1 - \ldots - a_pX_p)\ \mathrm{d}x \\ &= \int \psi^{-1}(x + a_0 + a_1X1 + \ldots + a_pX_p)\ \mathrm{d}\Phi_0(x)\end{aligned}$$

où ϕ_0 (respectivement Φ_0) est la densité de probabilité (respectivement la fonction de répartition) de la loi normale centrée d'écart-type σ, la méthode smearing consiste à remplacer Φ_0 par la fonction de répartition empirique des résidus de l'ajustement du modèle, soit:

$$\begin{aligned} B_{\text{smearing}} &= \int \psi^{-1}(x + a_0 + a_1 X_1 + \ldots + a_p X_p) \times \frac{1}{n} \sum_{i=1}^{n} \delta(x - \hat{\varepsilon}_i)\, \mathrm{d}x \\ &= \frac{1}{n} \sum_{i=1}^{n} \psi^{-1}(a_0 + a_1 X_1 + \ldots + a_p X_p + \hat{\varepsilon}_i) \end{aligned}$$

où δ est la masse de Dirac en zéro et $\hat{\varepsilon}_i$ est le résidu du modèle ajusté pour la i^{e} observation. Cette méthode de correction du biais de prédiction a l'avantage d'être à la fois très générale et facile à calculer. Elle a l'inconvénient qu'il faut connaître les résidus $\hat{\varepsilon}_i$ de l'ajustement du modèle. Ce n'est pas un problème quand on ajuste soi-même un modèle à des données, mais c'en est un quand on utilise un tarif publié pour lequel les résidus ne sont pas donnés.

Dans le cas particulier de la transformation logarithme, ψ^{-1} est la fonction exponentielle, et donc l'estimation smearing de la biomasse est: $\exp(a_0 + a_1 X_1 + \ldots + a_p X_p) \times \text{CF}_{\text{smearing}}$, où le coefficient correcteur smearing est:

$$\text{CF}_{\text{smearing}} = \frac{1}{n} \sum_{i=1}^{n} \exp(\hat{\varepsilon}_i)$$

Étant donné que $\hat{\sigma}^2 = (\sum_{i=1}^{n} \hat{\varepsilon}_i^2)/(n - p - 1)$, le coefficient correcteur smearing est différent du coefficient correcteur (7.14). Néanmoins, dans la limite où $\hat{\sigma} \to 0$, les deux coefficients correcteurs sont équivalents.

36

Estimation « smearing » de la biomasse

Reprenons encore l'exemple du tarif de biomasse ajusté dans le fil rouge n° 31 par régression multiple sur les données log-transformées:

$$\ln(B) = -8{,}38900 + 0{,}85715 \ln(D^2 H) + 0{,}72864 \ln(\rho)$$

Le coefficient correcteur smearing s'obtient par la commande:

```
mean(exp(residuals(m)))
```

où `m` est l'objet contenant le modèle ajusté, et vaut dans cet exemple 1,059859. À titre de comparaison, le coefficient correcteur calculé précédemment (fil rouge n° 35) valait 1,061035.

7.3 Prédiction du volume ou de la biomasse d'un peuplement

Pour prédire le volume ou la biomasse d'un peuplement à l'aide d'un tarif de biomasse, il n'est pas possible de mesurer les entrées du tarif sur tous les arbres du peuplement. Les entrées du tarif ne vont être mesurées que sur un échantillon d'arbres du peuplement. Le volume ou la biomasse des arbres de cet échantillon sera calculé à l'aide du tarif, puis extrapolé à l'ensemble du peuplement. La prédiction du volume ou de la biomasse d'un peuplement comporte donc deux sources de variabilité: l'une liée à la prédiction individuelle

par le tarif, et l'autre liée à l'échantillonnage des arbres au sein du peuplement. Tenir compte rigoureusement de ces deux sources de variabilité dans la prédiction à l'échelle du peuplement amène à des problèmes complexes de double échantillonnage, que nous avons évoqués aux paragraphes 2.1.2 et 2.3 (Parresol, 1999).

Le problème est légèrement simplifié quand l'échantillon des arbres utilisé pour construire le tarif est indépendant de l'échantillon des arbres dont les entrées ont été mesurées. Dans ce cas, on peut considérer que l'erreur de prédiction liée au tarif est indépendante de l'erreur d'échantillonnage. Supposons que n placettes d'échantillonnage de surface unitaire A aient été mises en place dans le peuplement, dont la superficie totale est $\mathcal{A}$. Soit N_i le nombre d'arbres trouvés dans la i^{e} placette ($i = 1, \ldots, n$) et soient $X_{ij1}, \ldots, X_{ijp}$ les p variables explicatives mesurées sur le j^{e} arbre de la i^{e} placettes ($j = 1, \ldots, N_i$). Cunia (1965, 1987*b*) a considéré le cas particulier où la biomasse est prédite par régression multiple à partir des p variables explicatives. L'estimation de la biomasse du peuplement est alors:

$$\begin{aligned}\hat{B} &= \frac{\mathcal{A}}{n}\sum_{i=1}^{n}\frac{1}{A}\sum_{j=1}^{N_i}(\hat{a}_0 + \hat{a}_1 X_{ij1} + \ldots + \hat{a}_p X_{ijp}) \\ &= \hat{a}_0\Big(\frac{\mathcal{A}}{nA}\sum_{i=1}^{n} N_i\Big) + \hat{a}_1\Big(\frac{\mathcal{A}}{nA}\sum_{i=1}^{n}\sum_{j=1}^{N_i} X_{ij1}\Big) + \ldots + \hat{a}_p\Big(\frac{\mathcal{A}}{nA}\sum_{i=1}^{n}\sum_{j=1}^{N_i} X_{ijp}\Big)\end{aligned}$$

où $\hat{a}_0, \ldots, \hat{a}_p$ sont les coefficients estimés de la régression. Posons $X_0^\star = (\mathcal{A}/nA)\sum_{i=1}^{n} N_i$ et pour tout $k = 1, \ldots, p$,

$$X_k^\star = \frac{\mathcal{A}}{nA}\sum_{i=1}^{n}\sum_{j=1}^{N_i} X_{ijk}$$

La biomasse estimée du peuplement s'écrit alors:

$$\hat{B} = \hat{a}_0 X_0^\star + \hat{a}_1 X_1^\star + \ldots + \hat{a}_p X_p^\star$$

Ce qui est intéressant, c'est que la variabilité de $\hat{\mathbf{a}} = {}^{\text{t}}[\hat{a}_0, \ldots, \hat{a}_p]$ dépend entièrement de l'ajustement au tarif et pas de l'échantillonnage du peuplement, tandis que la variabilité de $\mathbf{x} = {}^{\text{t}}[X_0^\star, \ldots, X_p^\star]$ dépend au contraire entièrement de l'échantillonnage et pas du tarif. À partir du moment où ces deux erreurs sont indépendantes,

$$\mathrm{E}(\hat{B}) = \mathrm{E}({}^{\text{t}}\hat{\mathbf{a}}\mathbf{x}) = {}^{\text{t}}\mathrm{E}(\hat{\mathbf{a}})\,\mathrm{E}(\mathbf{x})$$

et

$$\mathrm{Var}(\hat{B}) = {}^{\text{t}}\mathbf{a}\boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{a} + {}^{\text{t}}\mathbf{x}\boldsymbol{\Sigma}_{\mathbf{a}}\mathbf{x}$$

où $\boldsymbol{\Sigma}_{\mathbf{a}}$ est la matrice $(p+1)\times(p+1)$ de variance-covariance des coefficients du modèle tandis que $\boldsymbol{\Sigma}_{\mathbf{x}}$ est la matrice $(p+1)\times(p+1)$ de variance-covariance de l'échantillon de $\mathbf{x}$. La première matrice découle de l'ajustement du tarif tandis que la seconde découle de l'échantillonnage du peuplement. Ainsi, l'erreur pour la prédiction de la biomasse du peuplement se décompose en la somme de deux termes, dont un est lié à l'erreur de prédiction du tarif et l'autre à l'erreur d'échantillonnage du peuplement.

Plus généralement, le principe est exactement le même que lorsque l'on a considéré page 184 une incertitude liée à la mesure des variables explicatives $X_1, \ldots, X_p$. Une erreur de mesure n'est pas de même nature qu'une erreur d'échantillonnage. Mais d'un point de vue mathématique, les calculs sont les mêmes: cela revient dans les deux cas à considérer que les variables explicatives $X_1, \ldots, X_p$ sont aléatoires et non plus fixes. On pourra ainsi, dans le cas général, utiliser une méthode de Monte Carlo pour estimer la biomasse du peuplement. Le pseudo-algorithme de cette méthode de Monte Carlo est comme précédemment (cf. p.184):

1. Pour k allant de 1 à Q, où Q est le nombre d'itérations de Monte Carlo:
   (a) pour i allant de 1 à p, tirer $\hat{X}_i^{(k)}$ selon une loi correspondant à la variabilité d'échantillonnage du peuplement (cette loi dépend du type d'échantillonnage mené, de la taille et du nombre de placettes inventoriées, etc.);
   (b) tirer un vecteur $\hat{\theta}^{(k)}$ selon une loi multinormale de moyenne $\hat{\theta}$ et de matrice de variance-covariance $\hat{\mathbf{\Sigma}}$;
   (c) calculer la prédiction $\hat{Y}^{(k)} = f(\hat{X}_1^{(k)}, \ldots, \hat{X}_p^{(k)}; \hat{\theta}^{(k)})$.
2. L'intervalle de confiance de la prédiction est l'intervalle de confiance empirique des Q valeurs $\hat{Y}^{(1)}, \ldots, \hat{Y}^{(Q)}$.

7.4 Expansion et conversion des tarifs

Il arrive que l'on dispose d'un tarif qui prédise une grandeur qui n'est pas exactement celle dont on a besoin, tout en étant fortement reliée à celle-ci. Par exemple, on dispose d'un tarif qui prédit la biomasse sèche du tronc, alors que l'on souhaite connaître la biomasse épigée totale de l'arbre. Ou alors, on dispose d'un tarif qui prédit le volume du tronc alors que l'on souhaite connaître sa biomasse sèche. Plutôt que de renoncer à utiliser un tarif qui ne prédit pas exactement ce que l'on souhaite, il est préférable d'utiliser ce tarif par défaut tout en le corrigeant par un facteur. On pourra utiliser des facteurs de *conversion* pour convertir un volume en biomasse (et *vice versa*), des facteurs d'*expansion* pour extrapoler une partie au tout, ou des combinaisons des deux. Ainsi Henry *et al.* (2011) proposent trois méthodes d'obtention de la biomasse totale:

- la biomasse du tronc est le produit du volume du tronc et de la densité spécifique du bois ρ;
- la biomasse épigée est le produit de la biomasse du tronc et d'un facteur d'expansion de la biomasse (FEB);
- la biomasse épigée est le produit du volume du tronc et d'un facteur de conversion et d'expansion de la biomasse (FCEB = FEB $\times\rho$).

Il existe des valeurs tabulées de ces différents facteurs de conversion et d'expansion. Ces valeurs sont souvent très variables, car elles intègrent implicitement différentes sources de variabilité. Aussi précis le tarif par défaut soit-il, on perd le plus souvent le bénéfice de cette précision dès lors qu'on utilise un facteur d'expansion ou de conversion, puisque l'erreur de la prédiction cumule toutes les sources d'erreur intervenant dans son calcul.

Pour les tarifs qui utilisent la hauteur comme entrée alors qu'on ne dispose pas de cette information, on pourra utiliser un modèle corollaire qui prédit la hauteur en fonction des entrées disponibles (typiquement un modèle de la relation hauteur–diamètre). Comme pour les facteurs de conversion et d'expansion, cela introduit une source d'erreur additionnelle.

7.5 Arbitrage entre différents tarifs

Quand on veut prédire le volume ou la biomasse d'arbres donnés, il arrive fréquemment que l'on ait plusieurs tarifs à disposition. Par exemple, pour une espèce donnée, plusieurs tarifs ont été ajustés à différents endroits. Ou bien on dispose d'un tarif local et d'un tarif pan-tropical. Arbitrer entre les différents tarifs disponibles n'est pas toujours chose aisée (Henry *et al.*, 2011). Vaut-il mieux par exemple choisir un tarif spécifique local ajusté à peu de données (donc *a priori* non biaisé mais avec une forte variabilité de prédiction) ou un tarif multispécifique pan-tropical ajusté à de nombreuses données (donc potentiellement

biaisé mais avec une faible variabilité de prédiction) ? On voit ainsi que de nombreux critères de choix entrent potentiellement en ligne de compte: la qualité du modèle (la taille de son domaine de validité, sa capacité à extrapoler des prédictions, etc.), sa spécificité (avec à un extrême les modèles monospécifiques locaux et à l'autre extrême les modèles pluri-spécifiques pan-tropicaux), la taille du jeu de données utilisé pour ajuster le modèle (donc, implicitement, la variabilité de ses prédictions). L'arbitrage entre différents tarifs existants ne doit pas être confondu avec la sélection de modèles évoquée dans la section 6.3.2. Dans la sélection de modèles, les coefficients des modèles ne sont pas encore connus, et on cherche le modèle qui s'ajuste le mieux aux données quand on estime ses coefficients. Dans l'arbitrage de modèles, on a affaire à des modèles déjà ajustés, dont les coefficients sont connus.

Souvent, l'arbitrage entre différents tarifs doit être fait sans donnée de biomasse ou de volume. Mais le cas qui nous intéresse désormais est lorsque l'on dispose d'un jeu de données de référence $\mathcal{S}_n$, avec n observations de la variable réponse (volume ou biomasse) et des variables explicatives.

7.5.1 Comparaison sur la base de critères de validation

Lorsque l'on dispose d'un jeu de données de référence $\mathcal{S}_n$, on peut comparer les différents tarifs disponibles sur la base des critères de validation définis au paragraphe 7.1.1, en utilisant $\mathcal{S}_n$ comme jeu de données de validation. Dans la mesure où les modèles n'ont pas forcément le même nombre p de paramètres, et selon le principe de parcimonie, on favorisera les critères de validation qui dépendent de p de façon à pénaliser les modèles ayant beaucoup de paramètres.

Quand il s'agit de comparer un tarif candidat bien précis, supposé être le « meilleur », à différents tarifs concurrents, on pourra comparer les prédictions du tarif candidat aux prédictions des concurrents. Pour cela, on regardera si les prédictions des tarifs concurrents entrent ou non dans l'intervalle de confiance au seuil α des prédictions du tarif candidat.

7.5.2 Choix d'un modèle

Le choix d'un tarif peut être fait par rapport à un « vrai » tarif f que l'on ne connaît pas mais que l'on suppose exister. Soit M le nombre de tarifs dont on dispose. On notera en abrégé $\hat{f}_m$ la fonction des p variables explicatives qui prédit le volume ou le biomasse selon le m^{e} tarif. Cette fonction est aléatoire puisqu'elle dépend de coefficients estimés, donc ayant leur propre distribution. La loi de distribution de $\hat{f}_m$ décrit ainsi la variabilité des prédictions selon le m^{e} tarif, telle que décrite dans le paragraphe 7.2. Les M tarifs peuvent avoir des formes très différentes: peut-être que le tarif $\hat{f}_1$ correspondra à une fonction puissance, le tarif $\hat{f}_2$ à une fonction polynômiale, etc. On suppose par ailleurs qu'il existe une fonction f des p variables explicatives qui décrit la « vraie » relation entre la variable réponse (volume ou biomasse) et ces variables explicatives. On ne connaît pas cette « vraie » relation. On ne sait pas quelle forme elle a. Mais chacun des M tarifs estimés peut être vu comme une approximation de la « vraie » relation f.

Dans la théorie de la sélection de modèles (Massart, 2007), l'écart entre la vraie relation f et un tarif $\hat{f}_m$ est quantifiée par une fonction γ qu'on appelle la fonction de *perte*. Par exemple, la fonction de perte pourra être la norme L^2 de la différence entre f et $\hat{f}_m$:

$$\gamma(f,\, \hat{f}_m) = \int_{x_1} \dots \int_{x_p} [f(x_1,\, \dots,\, x_p) - \hat{f}_m(x_1,\, \dots,\, x_p)]^2 \, \mathrm{d}x_1 \dots \mathrm{d}x_p$$

On appelle *risque* (noté R) l'espérance de la perte par rapport à la loi de distribution de $\hat{f}_m$

(c'est-à-dire en intégrant sur la variabilité des prédictions du tarif):

$$R = \mathrm{E}[\gamma(f,\, \hat{f}_m)]$$

Le meilleur tarif parmi les M disponibles est celui qui minimise le risque. Le problème est que l'on ne connaît pas la vraie relation f, donc ce « meilleur » tarif n'est pas connu non plus. Dans la théorie de la sélection de modèles, on appelle ce meilleur tarif un *oracle*. Le tarif choisi va finalement être celui tel que le risque de l'oracle reste borné pour une vaste famille de fonctions f. De façon intuitive, le tarif choisi est celui tel que l'écart entre ce tarif et la « vraie » relation reste limité, quelle que soit cette « vraie » relation (dans les limites d'une gamme de possibilités réalistes). Nous ne développerons pas plus avant cette théorie, qui sort du cadre de ce manuel.

7.5.3 Moyenne bayésienne de modèles

Plutôt que de choisir un tarif parmi M disponibles, avec le risque de ne pas choisir le « meilleur », une alternative consiste à combiner les M tarifs concurrents en un nouveau tarif. C'est ce qui s'appelle en anglais le « Bayesian model averaging ». La moyenne bayésienne de modèles a été très utilisée pour les modèles de prédiction climatique (Raftery *et al.*, 2005; Furrer *et al.*, 2007; Berliner et Kim, 2008; Smith *et al.*, 2009) mais reste encore peu utilisée pour les modèles forestiers (Li *et al.*, 2008; Picard *et al.*, 2012). Soit $\mathcal{S}_n = \{(Y_i,\, X_{i1},\, \ldots,\, X_{ip}),\, i = 1, \ldots, p\}$ un jeu de données de référence avec n observations de la variable réponse Y et des p variables explicatives. La moyenne bayésienne de modèles considère que la loi de distribution de la variable réponse Y est un mélange de M lois:

$$g(Y|X_1,\, \ldots,\, X_p) = \sum_{m=1}^{M} w_m\, g_m(Y|X_1,\, \ldots,\, X_p)$$

où g est la densité de distribution de la loi de Y, g_m est la densité de distribution de la loi conditionnelle de Y sachant que le modèle m est le « meilleur », et w_m est le poids du m^{e} modèle dans le mélange, que l'on peut interpréter comme la probabilité *a posteriori* que le m^{e} modèle soit le « meilleur ». Les probabilités *a posteriori* w_m reflètent la qualité de l'ajustement des modèles aux données, et ont une somme égale à un: $\sum_{m=1}^{M} w_m = 1$.

Comme dans la sélection de modèles évoquée au paragraphe précédent, la moyenne bayésienne de modèles suppose qu'il existe une « vraie » relation (mais qui demeure inconnue) entre la variable réponse et les p variables explicatives, et que chaque tarif s'écarte de cette « vraie » relation selon une loi normale d'écart-type σ_m. Autrement dit, la densité g_m est la densité de la loi normale de moyenne $f_m(x_1,\, \ldots,\, x_p)$ et d'écart-type σ_m, où f_m est la fonction de p variables correspondant au m^{e} tarif. Ainsi,

$$g(Y|X_1,\, \ldots,\, X_p) = \sum_{m=1}^{M} w_m\, \phi(Y;\, f_m(x_1,\, \ldots,\, x_p),\, \sigma_m)$$

où $\phi(\cdot;\mu,\, \sigma)$ est la densité de probabilité de la loi normale d'espérance μ et d'écart-type σ. Le tarif f_{moy} résultant de la combinaison des M tarifs concurrents est défini comme l'espérance du modèle de mélange, c'est-à-dire:

$$f_{\mathrm{moy}}(X_1,\, \ldots,\, X_p) = \mathrm{E}(Y|X_1,\, \ldots,\, X_p) = \sum_{m=1}^{M} w_m\, f_m(X_1,\, \ldots,\, X_p)$$

Ainsi le tarif résultant de la combinaison des M tarifs concurrents est une moyenne pondérée de ces M tarifs, le poids du modèle m étant la probabilité *a posteriori* que ce modèle m

soit le meilleur. On peut également calculer la variance des prédictions selon le tarif f_{moy} résultant de la combinaison des M tarifs concurrents:

$$\begin{aligned}\text{Var}(Y|X_1, \ldots, X_p) &= \sum_{m=1}^{M} w_m \Big[f_m(X_1, \ldots, X_p) - \sum_{l=1}^{M} w_l \, f_l(X_1, \ldots, X_p) \Big]^2 \\ &\quad + \sum_{m=1}^{M} w_m \, \sigma_m^2\end{aligned}$$

Le premier terme correspond à la variance inter-modèles et traduit la variabilité des prédictions d'un modèle à l'autre; le second terme correspond à la variance intra-modèle et traduit l'erreur conditionnelle de prédiction sachant que le modèle est le meilleur.

Pour pouvoir utiliser le tarif f_{moy} en lieu et place des M tarifs $f_1, \ldots, f_M$, il reste à estimer les poids $w_1, \ldots, w_M$ et les écart-types intra-modèle $\sigma_1, \ldots, \sigma_M$. Ces $2M$ paramètres sont estimés à partir du jeu de données de référence $\mathcal{S}_n$ en utilisant un algorithme EM (Dempster *et al.*, 1977; McLachlan et Krishnan, 2008). L'algorithme EM introduit des variables latentes z_{im} telles que z_{im} est la probabilité *a posteriori* que le tarif m soit le meilleur modèle pour l'observation i de $\mathcal{S}_n$. Les variables latentes z_{im} prennent des valeurs entre 0 et 1. L'algorithme EM est itératif et alterne entre deux étapes à chaque itération: l'étape E (comme « expectation ») et l'étape M (comme « maximisation »). L'algorithme EM est le suivant:

1. Choisir des valeurs initiales $w_1^{(0)}, \ldots, w_M^{(0)}, \sigma_1^{(0)}, \ldots, \sigma_M^{(0)}$ des $2M$ paramètres à estimer.
2. Alterner les deux étapes:
   (a) étape E: calculer la valeur de z_{im} à l'itération j sachant les valeurs des paramètres à l'itération $j-1$:

$$z_{im}^{(j)} = \frac{w_m^{(j-1)} \, \phi[Y_i; f_m(X_{i1}, \ldots, X_{ip}), \sigma_m^{(j-1)}]}{\sum_{k=1}^{M} w_k^{(j-1)} \, \phi[Y_i; f_k(X_{i1}, \ldots, X_{ip}), \sigma_k^{(j-1)}]}$$

   (b) étape M: estimer les paramètres à l'itération j en utilisant comme poids les valeurs courantes des z_{im}, c'est-à-dire:

$$\begin{aligned} w_m^{(j)} &= \frac{1}{n} \sum_{i=1}^{n} z_{im}^{(j)} \\ {\sigma_m^{(j)}}^2 &= \frac{\sum_{i=1}^{n} z_{im}^{(j)} [Y_i - f_m(X_{i1}, \ldots, X_{ip})]^2}{\sum_{i=1}^{n} z_{im}^{(j)}} \end{aligned}$$

   tant que $\sum_{m=1}^{M} |w_m^{(j)} - w_m^{(j-1)}| + \sum_{m=1}^{M} |\sigma_m^{(j)} - \sigma_m^{(j-1)}|$ est plus grand qu'un seuil infinitésimal fixé (par exemple 10^{-6}).
3. La valeur estimée de w_m est $w_m^{(j)}$ et la valeur estimée de σ_m est $\sigma_m^{(j)}$.

Conclusions et recommandations

Les méthodes d'estimation du volume et de la biomasse des arbres sont en constante évolution. Obtenir des estimations qui soient les plus proches de la réalité est de plus en plus recherché. Les tarifs de cubage et de biomasse n'ont pas suivi la même évolution en fonction des zones écologiques considérées. En zones tropicales sèches, où la problématique de l'approvisionnement en bois de feu est ancienne, les équations allométriques ont principalement été développées pour la quantification du bois de chauffe. En zone tropicale humide, où l'exploitation est principalement pour le bois d'œuvre, c'est pour établir des tarifs de cubage que les équations ont majoritairement été développées. Aujourd'hui, les préoccupations liées aux changements climatiques sont grandissantes et l'intérêt pour les tarifs de biomasse est équivalent dans les forêts sèches et humides.

Les mesures de biomasse devraient augmenter dans les années à venir et satisfaire les besoins d'estimation des stocks de carbone et de compréhension de la contribution des écosystèmes terrestres au cycle du carbone. L'expérience acquise sur les tarifs de cubage a montré qu'il fallait deux à trois mille observations pour estimer le volume du tronc d'*une* essence donnée avec une précision acceptable pour couvrir la variabilité comprise dans son aire géographique de répartition (CTFT, 1989). À titre de comparaison, le tarif de biomasse de Chave *et al.* (2005), qui est un des tarifs de biomasse les plus utilisés aujourd'hui, a été calibré à partir de 2410 observations. Et il s'agit d'un tarif pan-tropical, couvrant toutes les essences et toutes les zones écologiques, des zones sèches aux zones humides ! La similitude entre ces deux tailles d'échantillons, alors que la variabilité diffère de plusieurs ordres de grandeur, montre qu'il y a encore, dans le domaine de la mesure de la biomasse, une marge de progression considérable pour arriver à explorer l'ensemble de la variabilité naturelle. Et cela d'autant plus que la biomasse, qui englobe tous les compartiments de l'arbre, a probablement une variabilité intrinsèque bien plus élevée que le volume du seul tronc.

Accroître la fiabilité des tarifs de biomasse est liée à l'augmentation du nombre d'observations disponibles. Mais mesurer la biomasse aérienne d'un arbre demande un effort de mesure bien plus considérable que mesurer le volume de son tronc. L'effort nécessaire est encore plus conséquent lorsqu'il s'agit de la biomasse racinaire. À l'heure actuelle, il est peu probable que de vastes campagnes de mesure pour la biomasse aérienne et racinaire puissent être financées. À l'instar de Chave *et al.* (2005), la construction de nouvelles équations allométriques devra donc reposer sur des compilations de jeux de données collectés à différents endroits par des équipes indépendantes. Les méthodes standardisées de mesure de biomasse, et statistiques d'ajustement de modèles capables d'intégrer de l'information complémentaire *via* des co-variables explicatives sont donc cruciales pour permettre des progrès dans le domaine de l'estimation de la biomasse des arbres dans les prochaines années. L'apport des expérimentations sur des peuplements réguliers (effets de l'ontogénie, de la densité de plantation, de la fertilité des sols ou de la fertilisation, de la silviculture plus généralement) permettront d'éclairer la construction de ces modèles génériques.

Contrairement aux manuels existants, nous avons voulu que ce manuel couvre l'ensemble de la démarche de construction d'une équation allométrique, du terrain à la prédiction, en passant par l'ajustement du modèle. Nous ne prétendons cependant pas avoir couvert toutes

les situations possibles. Nombreux sont les cas qui nécessitent le développement de méthodes spécifiques. Les grands arbres à contreforts, par exemple, posent un défi pour la prédiction de leur biomasse — à commencer par le fait que leur diamètre à hauteur de poitrine, qui est la première variable d'entrée de la plupart des tarifs, n'est pas mesurable. Les arbres creux, les figuiers étrangleurs, les baobabs ou les grandes épiphytes, sont autant d'espèces et de particularités qui ne vont pas permettre le suivi des méthodes proposées dans ce manuel sans poser de problèmes. Des méthodes dendrométriques nouvelles devront vraisemblablement être développées pour traiter ces cas spécifiques. L'utilisation de la modélisation en trois dimensions, la photogrammétrie, le radar et le laser, que ce soit au sol ou aéroporté, seront des outils qui faciliteront ou révolutionneront les méthodes d'estimation de la biomasse et, peut-être, remplaceront à terme la tronçonneuse et le peson.

Le domaine des statistiques est également en chantier. Une confrontation du rapport de Whraton et Cunia (1987) aux méthodes d'ajustement aujourd'hui utilisées montre les progrès, dans la discipline forestière, de l'emploi de méthodes statistiques de plus en plus pointues, que nous avons tentées de présenter didactiquement dans ce manuel. La prise en compte de la variabilité intra-tige pourrait devenir commune pour ajuster les tarifs de biomasse dans le futur.

L'amélioration des méthodes de mesure et d'ajustement des modèles, l'accroissement des mesures de terrain, ne contribueront à l'amélioration des processus de recherches scientifiques et d'estimation de la biomasse des arbres que si les modèles et les méthodes produits sont mis à disposition de manière transparente. Nombreuses sont les données qui restent dans les bibliothèques et qui ne sont jamais publiées dans des journaux scientifiques et sur internet. De plus, pour un pays ne disposant pas de données de biomasse pour certaines de ses zones écologiques, les données présentes dans les pays voisins ou dans des zones écologiques identiques ne sont pas facilement accessibles. Aussi, nous encourageons les acteurs forestiers à identifier les données déjà disponibles pour les zones écologiques ou pays d'intérêts. Les données peuvent être intégrées dans une base de données et servir pour l'identification des lacunes. Une fois les lacunes identifiées, des mesures de terrain peuvent être effectuées en utilisant les conseils et le fil rouge proposés dans ce manuel.

Afin de permettre la continuité de l'effort d'amélioration des estimations, il est nécessaire de mettre en place un système d'archivage des données. Le système d'archivage est le point de départ de l'amélioration des estimations futures. Un système d'archivage robuste doit permettre de réduire les efforts des équipes futures pour comprendre et recalculer les estimations existantes. Par ailleurs, il est important de mettre en place des méthodes qui soient cohérentes dans le temps. Le manuel propose différentes méthodes de mesure. Il est préférable d'adopter une méthode qui puisse être reproduite et qui soit le moins dépendante de facteurs financiers, technologiques ou humains. Dans le cas où une méthode alternative est développée pour des raisons pratiques, celle-ci doit être signalée et rendue accessible pour permettre au prochain manuel de mieux prendre en compte la diversité des méthodologies possibles. Enfin, il est préférable d'adopter des méthodes simples et reproductibles.

Bibliographie

AFNOR, 1985. Bois – détermination de la masse volumique. NF B51-005, AFNOR. 64

AGO, 2002. Field measurement procedures for carbon accounting. Bush for Greenhouse Report 2, Australian Greenhouse Office, Canberra, Australia. 30

Akaike, H., 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6): 716–723. 156

Alder, D., 1980. *Estimation des volumes et accroissement des peuplements forestiers – Vol. 2. Étude et prévision de la production.* Études FAO : forêts n° 22/2. FAO, Rome. 26

Andrews, J.A. et Siccama, T.G., 1995. Retranslocation of calcium and magnesium at the heartwood-sapwood boundary of Atlantic white cedar. *Ecology*, 76(2): 659–663. 24

Araújo, T.M., Higuchi, N. et de Carvalho, J.A., 1999. Comparison of formulae for biomass content determination in a tropical rain forest site in the state of Pará, Brazil. *Forest Ecology and Management*, 117(1-3): 43–52. 106

Archibald, S. et Bond, W.J., 2003. Growing tall vs growing wide: tree architecture and allometry of *Acacia karroo* in forest, savanna, and arid environments. *Oikos*, 102(1): 3–14. 23

Assmann, E., 1970. *The Principles of Forest Yield Study.* Pergamon Press, Oxford, UK. 24, 25

Augusto, L., Meredieu, C., Bert, D., Trichet, P., Porté, A., Bosc, A., Lagane, F., Loustau, D., Pellerin, S., Danjon, F., Ranger, J. et Gelpe, J., 2008. Improving models of forest nutrient export with equations that predict the nutrient concentration of tree compartments. *Annals of Forest Science*, 65(8): 808. 24

Basuki, T.M., van Laake, P.E., Skidmore, A.K. et Hussin, Y.A., 2009. Allometric equations for estimating the above-ground biomass in tropical lowland *Dipterocarp* forests. *Forest Ecology and Management*, 257(8): 1684–1694. 106

Batho, A. et García, O., 2006. De Perthuis and the origins of site index: a historical note. *Forest Biometry, Modelling and Information Science*, 1: 1–10. 24

Becking, J.H., 1953. Einige gesichtspunkte für die durchführung von vergleichenden durchforstungsversuchen in gleichälterigen beständen. *In 11e Congrès de l'Union Internationale des Instituts de Recherches Forestiers, Rome, 1953 : comptes rendus.* IUFRO, pp. 580–582. 216

Bellefontaine, R., Petit, S., Pain-Orcet, M., Deleporte, P. et Bertault, J.G., 2001.

Les arbres hors forêt : vers une meilleure prise en compte. Cahier FAO Conservation n° 35. FAO, Rome. 33

Bergès, L., Nepveu, G. et Franc, A., 2008. Effects of ecological factors on radial growth and wood density components of sessile oak (*Quercus petraea* Liebl.) in Northern France. *Forest Ecology and Management*, 255(3-4): 567–579. 24, 27

Berliner, L.M. et Kim, Y., 2008. Bayesian design and analysis for superensemble-based climate forecasting. *Journal of Climate*, 21(9): 1891–1910. 191

Bloom, A.J., Chapin, F.S. et Mooney, H.A., 1985. Resource fimitation in plants–an economic analogy. *Annual Review of Ecology and Systematics*, 16: 363–392. 28

Bohlman, S. et O'Brien, S., 2006. Allometry, adult stature and regeneration requirement of 65 tree species on Barro Colorado Island, Panama. *Journal of Tropical Ecology*, 22(2): 123–136. 23

Bolker, B., 2008. *Ecological Models and Data in R*. Princeton University Press, Princeton, New Jersey, États-Unis. 183

Bontemps, J.D., Hervé, J.C. et Dhôte, J.F., 2009. Long-term changes in forest productivity: a consistent assessment in even-aged stands. *Forest Science*, 55(6): 549–564. 26

Bontemps, J.D., Hervé, J.C., Leban, J.M. et Dhôte, J.F., 2011. Nitrogen footprint in a long-term observation of forest growth over the twentieth century. *Trees–Structure and Function*, 25(2): 237–251. 26

Bormann, F.H., 1953. The statistical efficiency of sample plot size and shape in forest ecology. *Ecology*, 34(3): 474–487. 48, 49

Bouchon, J., 1974. Les tarifs de cubage. Rapport technique, ENGREF, Nancy, France. 31

Bouriaud, O., Leban, J.M., Bert, D. et Deleuze, C., 2005. Intra-annual variations in climate influence growth and wood density of Norway spruce. *Tree Physiology*, 25(6): 651–660. 27

Box, G.E.P. et Draper, N.R., 1987. *Empirical Model Building and Response Surfaces*. Wiley series in probability and mathematical statistics. Wiley, New York, New York, États-Unis. 42

Bozdogan, H., 1987. Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, 52(3): 345–370. 156

Bradley, P.N., 1988. Survey of woody biomass on farms in western Kenya. *Ambio*, 17(1): 40–48. 30

Brown, I.F., Martinelli, L.A., Thomas, W.W., Moreira, M.Z., Victoria, R.A. et Ferreira, C.A.C., 1995. Uncertainty in the biomass of Amazonian forests: An example from Rondônia, Brazil. *Forest Ecology and Management*, 75(1-3): 175–189. 41

Brown, S., 1997. *Estimating Biomass and Biomass Change of Tropical Forests: a Primer*.

FAO Forestry Paper n° 134. FAO, Rome. 45, 106

Brown, S., Gillespie, A.J.R. et Lugo, A.E., 1989. Biomass estimation methods for tropical forests with applications to forest inventory data. *Forest Science*, 35(4): 881–902. 106, 125

Burdon, R.D., Kibblewhite, R.P., Walker, J.C.F., Megraw, E.R. et Cown, D.J., 2004. Juvenile versus mature wood: a new concept, orthogonal to corewood versus outerwood, with special reference to *Pinus radiata* and *P. taeda*. *Forest Science*, 50(4): 399–415. 27

Burnham, K.P. et Anderson, D.R., 2004. Multimodel inference: understanding AIC and BIC in model selection. *Sociological Methods & Research*, 33(2): 261–304. 156

Burnham, K.P. et Anderson, D.R., 2002. *Model Selection and Multimodel Inference. A Practical Information-Theoretic Approach*. Springer Science+Business Media, Inc., New York, New York, États-Unis, 2e éd. 156

Cailliez, F., 1980. *Forest volume estimation and yield prediction. Volume estimation*. Études FAO forêts, vol. 1. FAO, Rome. 31

Cairns, M.A., Brown, S., Helmer, E.H. et Baumgardner, G.A., 1997. Root biomass allocation in the world's upland forests. *Oecologia*, 111(1): 1–11. 28

Calama, R., Barbeito, I., Pardos, M., del Río, M. et Montero, G., 2008. Adapting a model for even-aged *Pinus pinea* L. stands to complex multi-aged structures. *Forest Ecology and Management*, 256(6): 1390–1399. 28

Cavaignac, S., Nguyen Thé, N., Melun, F. et Bouvet, A., 2012. élaboration d'un modèle de croissance pour l'Eucalyptus gundal. *FCBA INFO*, p. 16. 27

Charru, M., Seynave, I., Morneau, F. et Bontemps, J.D., 2010. Recent changes in forest productivity: An analysis of national forest inventory data for common beech (*Fagus sylvatica* L.) in north-eastern France. *Forest Ecology and Management*, 260(5): 864–874. 26

Chave, J., Andalo, C., Brown, S., Cairns, M.A., Chambers, J.Q., Eamus, D., Fölster, H., Fromard, F., Higuchi, N., Kira, T., Lescure, J.P., Nelson, B.W., Ogawa, H., Puig, H., Riéra, B. et Yamakura, T., 2005. Tree allometry and improved estimation of carbon stocks and balance in tropical forests. *Oecologia*, 145(1): 87–99. 106, 193

Chave, J., Coomes, D., Jansen, S., Lewis, S.L., Swenson, N.G. et Zanne, A.E., 2009. Towards a worldwide wood economics spectrum. *Ecology Letters*, 12(4): 351–366. 24

Chave, J., Riéra, B. et Dubois, M.A., 2001. Estimation of biomass in a neotropical forest of French Guiana: spatial and temporal variability. *Journal of Tropical Ecology*, 17(1): 79–96. 106

Chave, J., Condit, R., Aguilar, S., Hernandez, A., Lao, S. et Perez, R., 2004. Error propagation and scaling for tropical forest biomass estimates. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 359(1443): 409–420. 40, 46, 51

Chave, J., Condit, R., Lao, S., Caspersen, J.P., Foster, R.B. et Hubbell, S.P., 2003. Spatial and temporal variation of biomass in a tropical forest: results from a large census plot in Panama. *Journal of Ecology*, 91(2): 240–252. 48, 49, 51

Cochran, W.G., 1977. *Sampling Techniques*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, New York, New York, États-Unis, 3e éd. 33, 38, 39, 42

Colin-Belgrand, M., Ranger, J. et Bouchon, J., 1996. Internal nutrient translocation in chesnut tree stemwood: III. Dynamics across an age series of *Castanea sativa* (Miller). *Annals of Botany*, 78(6): 729–740. 24

Cotta, H., 1804. *Principes fondamentaux de la science forestière*. Bouchard-Huzard, Paris. 31

Courbaud, B., Goreaud, F., Dreyfus, P. et Bonnet, F.R., 2001. Evaluating thinning strategies using a tree distance dependent growth model: some examples based on the CAPSIS software "uneven-aged spruce forests" module. *Forest Ecology and Management*, 145(1): 15–28. 28

Cressie, N., 1993. *Statistics for Spatial Data*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, New York, New York, États-Unis, 2e éd. 48

CTFT, 1989. *Mémento du forestier*. Ministère de la Coopération et du Développement, Paris, 3e éd. 33, 40, 42, 43, 48, 193

Cunia, T., 1964. Weighted least squares method and construction of volume tables. *Forest Science*, 10(2): 180–191. 31, 125

Cunia, T., 1965. Some theory on reliability of volume estimates in a forest inventory sample. *Forest Science*, 11(1): 115–128. 188

Cunia, T., 1987*a*. Construction of tree biomass tables by linear regression techniques. *In* E.H. Whraton et T. Cunia (eds.), *Estimating tree biomass regressions and their error. Proceedings of the workshop on tree biomass regression functions and their contribution to the error of forest inventory estimates, May 26–30, 1986, Syracuse, N.Y. – Part E.* USDA Forest Service, Northeastern Forest Experiment Station, Broomall, Pennsylvania, États-Unis, General Technical Report no NE-117, pp. 27–36. 125

Cunia, T., 1987*b*. Error of forest inventory estimates: its main components. *In* E.H. Whraton et T. Cunia (eds.), *Estimating tree biomass regressions and their error. Proceedings of the workshop on tree biomass regression functions and their contribution to the error of forest inventory estimates, May 26–30, 1986, Syracuse, N.Y. – Part E.* USDA Forest Service, Northeastern Forest Experiment Station, Broomall, Pennsylvania, États-Unis, General Technical Report no NE-117, pp. 1–14. 34, 39, 46, 188

Cunia, T., 1987*c*. An optimization model for subsampling trees for biomass measurement. *In* E.H. Whraton et T. Cunia (eds.), *Estimating tree biomass regressions and their error. Proceedings of the workshop on tree biomass regression functions and their contribution to the error of forest inventory estimates, May 26–30, 1986, Syracuse, N.Y. – Part E.* USDA Forest Service, Northeastern Forest Experiment Station, Broomall, Pennsylvania, États-Unis, General Technical Report no NE-117, pp. 109–118. 34, 39, 46, 49

Cunia, T., 1987*d*. An optimization model to calculate the number of sample trees and plots. *In* E.H. Whraton et T. Cunia (eds.), *Estimating tree biomass regressions and their error. Proceedings of the workshop on tree biomass regression functions and their contribution to the error of forest inventory estimates, May 26–30, 1986, Syracuse, N.Y. – Part E.* USDA Forest Service, Northeastern Forest Experiment Station, Broomall, Pennsylvania, États-Unis, General Technical Report n° NE-117, pp. 15–24. 34, 39, 46, 49

Cunia, T. et Briggs, R.D., 1984. Forcing additivity of biomass tables: some empirical results. *Canadian Journal of Forest Research*, 14: 376–384. 171

Cunia, T. et Briggs, R.D., 1985*a*. Forcing additivity of biomass tables: use of the generalized least squares method. *Canadian Journal of Forest Research*, 15: 23–28. 171

Cunia, T. et Briggs, R.D., 1985*b*. Harmonizing biomass tables by generalized least squares. *Canadian Journal of Forest Research*, 15: 331–340. 173

de Vries, P.G., 1986. *Sampling Theory for Forest Inventory – A Teach-Yourself Course.* Springer-Verlag, Berlin. 38, 47

Dean, C., 2003. Calculation of wood volume and stem taper using terrestrial single-image close-range photogrammetry and contemporary software tools. *Silva Fennica*, 37(3): 359–380. 174

Dean, C. et Roxburgh, S., 2006. Improving visualisation of mature, high-carbon sequestering forests. *Forest, Biometry, Modelling and Information Science*, 1: 48–69. 174

Dean, C., Roxburgh, S. et Mackey, B., 2003. Growth modelling of *Eucalyptus regnans* for carbon accounting at the landscape scale. *In* A. Amaro, D. Reed et P. Soares (eds.), *Modelling Forest Systems.* CAB International Publishing, Wallingford, Royaume-Uni, pp. 27–39. 174

Deans, J.D., Moran, J. et Grace, J., 1996. Biomass relationships for tree species in regenerating semi-deciduous tropical moist forest in Cameroon. *Forest Ecology and Management*, 88(3): 215–225. 41

Decourt, N., 1973. Production primaire, production utile : méthodes d'évaluation, indices de productivité. *Annales des Sciences Forestières*, 30(3): 219–238. 25

Deleuze, C., Blaudez, D. et Hervé, J.C., 1996. Fitting a hyperbolic model for height versus girth relationship in spruce stands. Spacing effects. *Annales des Sciences Forestières*, 53(1): 93–111. 27

Dempster, A.P., Laird, N.M. et Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1): 1–38. 192

Dhôte, J.F., 1990. *Modèles de la dynamique des peuplements forestiers : articulation entre les niveaux de l'arbre et du peuplement. Applications à la sylviculture des hêtraies.* Thèse de doctorat, Université Claude Bernard-Lyon I, Lyon, France. 27

Dhôte, J.F., 1991. Modélisation de la croissance des peuplements réguliers de hêtre : dynamique des hiérarchies sociales et facteurs de production. *Annales des Sciences Forestières*,

48(4): 389–416. 24

Dhôte, J.F., 1996. A model of even-aged beech stands productivity with process-based interpretations. *Annales des Sciences Forestières*, 53(1): 1–20. 26

Díaz, S. et Cabido, M., 1997. Plant functional types and ecosystem function in relation to global change. *Journal of Vegetation Science*, 8: 463–474. 169

Dietz, J. et Kuyah, S., 2011. Guidelines for establishing regional allometric equations for biomass estimation through destructive sampling. Report of the carbon benefits project: Modelling, measurement and monitoring, World Agroforestry Centre (ICRAF), Nairobi, Kenya. 30

Dietze, M.C., Wolosin, M.S. et Clark, J.S., 2008. Capturing diversity and interspecific variability in allometries: A hierarchical approach. *Forest Ecology and Management*, 256(11): 1939–1948. 23

Djomo, A.N., Ibrahima, A., Saborowski, J. et Gravenhorst, G., 2010. Allometric equations for biomass estimations in Cameroon and pan moist tropical equations including biomass data from Africa. *Forest Ecology and Management*, 260(10): 1873–1885. 46, 106

Dong, J., Kaufmann, R.K., Myneni, R.B., Tucker, C.J., Kauppi, P.E., Liski, J., Buermann, W., Alexeyev, V. et Hughes, M.K., 2003. Remote sensing estimates of boreal and temperate forest woody biomass: carbon pools, sources, and sinks. *Remote Sensing of Environment*, 84: 393–410. 30

Dreyfus, P., 2012. Joint simulation of stand dynamics and landscape evolution using a tree-level model for mixed uneven-aged forests. *Annals of Forest Science*, 69(2): 283–303. 28

Duan, N., 1983. Smearing estimate: a nonparametric retransformation method. *Journal of the American Statistical Association*, 78(383): 605–610. 31, 186

Durbin, J. et Watson, G.S., 1971. Testing for serial correlation in least squares regression. III. *Biometrika*, 58(1): 1–19. 115

Ebuy Alipade, J., Lokombé Dimandja, J.P., Ponette, Q., Sonwa, D. et Picard, N., 2011. Biomass equation for predicting tree aboveground biomass at Yangambi, DRC. *Journal of Tropical Forest Science*, 23(2): 125–132. 41

Efron, B. et Tibshirani, R.J., 1993. *An Introduction to the Bootstrap*. Monographs on Statistics and Applied Probability n° 57. Chapman & Hall, New York, New York, États-Unis. 176, 177

Eichhorn, F., 1904. Beziehungen zwischen Bestandshöhe und Bestandsmasse. *Allgemeine Forst- und Jagdzeitung*, 80: 45–49. 25, 216

Enquist, B.J., Brown, J.H. et West, G.B., 1998. Allometric scaling of plant energetics and population density. *Nature*, 395(6698): 163–165. 23, 105

Enquist, B.J., West, G.B., Charnov, E.L. et Brown, J.H., 1999. Allometric scaling of production and life-history variation in vascular plants. *Nature*, 401(6756): 907–911. 23,

105

Enquist, B.J., 2002. Universal scaling in tree and vascular plant allometry: toward a general quantitative theory linking plant form and function from cells to ecosystems. *Tree Physiology*, 22(15-16): 1045–1064. 24

Eyre, F.H. et Zillgitt, W.M., 1950. Size-class distribution in old-growth northern hardwoods twenty years after cutting. Station Paper 21, U.S. Department of Agriculture, Forest Service, Lake States Forest Experiment Station, Saint Paul, Minnessota, États-Unis. 28

Fairfield Smith, H., 1938. An empirical law describing heterogeneity in the yields of agricultural crops. *Journal of Agricultural Science*, 28: 1–23. 48, 50

Fang, Z. et Bailey, R.L., 1999. Compatible volume and taper models with coefficients for tropical species on Hainan island in southern China. *Forest Science*, 45(1): 85–100. 174

FAO, 2006. *Global Forest Resources Assessment 2005. Progress towards sustainable forest management*. FAO Forestry Paper, vol. 147. Food and Agriculture Organization of the United Nations, Rome. 29

Favrichon, V., 1998. Modeling the dynamics and species composition of tropical mixed-species uneven-aged natural forest: effects of alternative cutting regimes. *Forest Science*, 44(1): 113–124. 28

Fonweban, J.N. et Houllier, F., 1997. Tarif de peuplement et modèle de production pour *Eucalyptus saligna* au Cameroun. *Bois et Forêts des Tropiques*, 253: 21–36. 96

Fournier-Djimbi, M., 1998. Le matériau bois : structure, propriétés, technologie. Cours, ENGREF, Département de foresterie rurale et tropicale, Montpellier, France. 64

Franc, A., Gourlet-Fleury, S. et Picard, N., 2000. *Introduction à la modélisation des forêts hétérogènes*. ENGREF, Nancy, France. 28, 105

Furnival, G.M., 1961. An index for comparing equations used in constructing volume tables. *Forest Science*, 7(4): 337–341. 128, 161

Furrer, R., Knutti, R., Sain, S.R., Nychka, D.W. et Meehl, G.A., 2007. Spatial patterns of probabilistic temperature change projections from a multivariate Bayesian analysis. *Geophysical Research Letters*, 34(L06711): 1–4. 191

Gambill, C.W., Wiant, H. V., J. et Yandle, D.O., 1985. Optimum plot size and BAF. *Forest Science*, 31(3): 587–594. 49, 50

García, O., 2003. Dimensionality reduction in growth models: an example. *Forest Biometry, Modelling and Information Science*, 1: 1–15. 26

García, O., 2011. Dynamical implications of the variability representation in site-index modelling. *European Journal of Forest Research*, 130(4): 671–675. 24, 26

Gayon, J., 2000. History of the concept of allometry. *American Zoologist*, 40(5): 748–758. 23

Gehring, C., Park, S. et Denich, M., 2004. Liana allometric biomass equations for Amazonian primary and secondary forest. *Forest Ecology and Management*, 195: 69–83. 32

Genet, A., Wernsdörfer, H., Jonard, M., Pretzsch, H., Rauch, M., Ponette, Q., Nys, C., Legout, A., Ranger, J., Vallet, P. et Saint-André, L., 2011. Ontogeny partly explains the apparent heterogeneity of published biomass equations for *Fagus sylvatica* in central Europe. *Forest Ecology and Management*, 261(7): 1188–1202. 28, 55

Gerwing, J.J., Schnitzer, S.A., Burnham, R.J., Bongers, F., Chave, J., DeWalt, S.J., Ewango, C.E.N., Foster, R., Kenfack, D., Martínez-Ramos, M., Parren, M., Parthasarathy, N., Pérez-Salicrup, D.R., Putz, F.E. et Thomas, D.W., 2006. A standard protocol for liana censuses. *Biotropica*, 38(2): 256–261. 32

Gerwing, J.J. et Farias, D.L., 2000. Integrating liana abundance and forest stature into an estimate of total aboveground biomass for an eastern Amazonian forest. *Journal of Tropical Ecology*, 16(3): 327–335. 32

Gibbs, H.K., Brown, S., Niles, J.O. et Foley, J.A., 2007. Monitoring and estimating tropical forest carbon stocks: making REDD a reality. *Environmental Research Letters*, 2(4): 1–13. Doi:10.1088/1748-9326/2/4/045023. 29

Gomat, H.Y., Deleporte, P., Moukini, R., Mialounguila, G., Ognouabi, N., Saya, R.A., Vigneron, P. et Saint-André, L., 2011. What factors influence the stem taper of *Eucalyptus*: growth, environmental conditions, or genetics? *Annals of Forest Science*, 68(1): 109–120. 27

Gonzalez, P., Asner, G.P., Battles, J.J., Lefsky, M.A., Waring, K.M. et Palace, M., 2010. Forestcarbondensities and uncertainties from Lidar, QuickBird, and field measurements in California. *Remote Sensing of Environment*, 114(7): 1561–1575. 30

Gould, S.J., 1979. An allometric interpretation of species-area curves. The meaning of the coefficient. *American Naturalist*, 114(3): 335–343. 105

Gould, S.J., 1966. Allometry and size in ontogeny and phylogeny. *Biological Reviews*, 41(4): 587–638. 23

Gould, S.J., 1971. Geometric similarity in allometric growth: a contribution to the problem of scaling in the evolution of size. *American Naturalist*, 105(942): 113–136. 23

Goupy, J., 1999. *Plans d'expériences pour surfaces de réponse*. Dunod, Paris. 42

Gourlet-Fleury, S. et Houllier, F., 2000. Modelling diameter increment in a lowland evergreen rain forest in French Guiana. *Forest Ecology and Management*, 131(1-3): 269–289. 28

Gourlet-Fleury, S., Rossi, V., Rejou-Mechain, M., Freycon, V., Fayolle, A., Saint-André, L., Cornu, G., Gérard, J., Sarrailh, J.M., Flores, O., Baya, F., Billand, A., Fauvet, N., Gally, M., Henry, M., Hubert, D., Pasquier, A. et Picard, N., 2011. Environmental filtering of dense-wooded species controls above-ground biomass stored in African moist forests. *Journal of Ecology*, 99(4): 981–990. 28, 64

Gregoire, T.G. et Dyer, M.E., 1989. Model fitting under patterned heterogeneity of

variance. *Forest Science*, 35(1): 105–125. 31

Guilley, E., Hervé, J.C. et Nepveu, G., 2004. The influence of site quality, silviculture and region on wood density mixed model in *Quercus petraea* Liebl. *Forest Ecology and Management*, 189(1-3): 111–121. 24, 27

Hairiah, K., Sitompul, S.M., van Noordwijk, M. et Palm, C.A., 2001. *Methods for sampling carbon stocks above and below ground*. ASB Lecture Note n° 4B. International Centre for Research in Agroforestry (ICRAF), Bogor, Indonesia. 30

Härdle, W. et Simar, L., 2003. *Applied Multivariate Statistical Analysis*. Springer-Verlag, Berlin. 101

Hart, H.M.J., 1928. *Stamtal en dunning: een orienteerend onderzoek naar de beste plantwijdte en dunningswijze voor den djati*. Ph.D. thesis, Wageningen University, Wageningen, The Netherlands. 216

Hawthorne, W., 1995. *Ecological Profiles of Ghanaian Forest Trees*. Tropical Forestry Paper n° 29. Oxford Forestry Institute, Department of Plant Sciences, University of Oxford, Oxford, UK. 74

Hebert, J., Rondeux, J. et Laurent, C., 1988. Comparaison par simulation de 3 types d'unités d'échantillonnage en futaies feuillues de hêtre (*Fagus silvatica* l.). *Annales des Sciences Forestières*, 45(3): 209–221. 49

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. 7, 8, 9, 13, 24, 32, 71, 88, 90, 91, 96, 97, 102, 103, 105, 106, 117, 118, 123, 124, 130, 131, 132, 139, 140, 141, 142, 158, 159, 160, 161, 163, 170, 179

Henry, M., Picard, N., Trotta, C., Manlay, R., Valentini, R., Bernoux, M. et Saint-André, L., 2011. Estimating tree biomass of sub-Saharan African forests: a review of available allometric equations. *Silva Fennica*, 45(3B): 477–569. 40, 105, 189

Hitchcock, H.C.I. et McDonnell, J.P., 1979. Biomass measurement: a synthesis of the literature. *In Proceedings of IUFRO workshop on forest resource inventories, July 23-26, 1979*. SAF-IUFRO, Fort Collins, Colorado, États-Unis, pp. 544–595. 31

Hoeting, J.A., Madigan, D., Raftery, A.E. et Volinsky, C.T., 1999. Bayesian model averaging: a tutorial. *Statistical Science*, 14(4): 382–417. 137

Hofstad, O., 2005. Review of biomass and volume functions for individual trees and shrubs in Southeast Africa. *Journal of Tropical Forest Science*, 17(1): 151–162. 105

Holmgren, P., Masakha, E.J. et Sjöholm, H., 1994. Not all African land is being degraded: a recent survey of trees on farms in Kenya reveals rapidly increasing forest resources. *Ambio*, 23(7): 390–395. 30

Huxley, J.S., 1924. Constant differential growth-ratios and their significance. *Nature*, 114: 895–896. 23

Ikonen, V.P., Kellomäki, S., Väisänenet, H. et Peltola, H., 2006. Modelling the distribution of diameter growth along the stem in Scots pine. *Trees–Structure and Function*, 20(3): 391–402. 27

Jackson, R.B., Canadell, J., Ehleringer, J.R., Mooney, H.A., Sala, O.E. et Schulze, E.D., 1996. A global analysis of root distributions for terrestrial biomes. *Oecologia*, 108(3): 389–411. 28

Jacobs, M.W. et Cunia, T., 1980. Use of dummy variables to harmonize tree biomass tables. *Canadian Journal of Forest Research*, 10: 483–490. 173

Johnson, F.A. et Hixon, H.J., 1952. The most efficient size and shape of plot to use for cruising in old growth Douglas-fir timber. *Journal of Forestry*, 50: 17–20. 48

Keller, M., Palace, M. et Hurtt, G., 2001. Biomass estimation in the Tapajos National Forest, Brazil. Examination of sampling and allometric uncertainties. *Forest Ecology and Management*, 154(3): 371–382. 48

Kelly, J.F. et Beltz, R.C., 1987. A comparison of tree volume estimation models for forest inventory. Research Paper SO-233, U.S. Department of Agriculture, Forest Service, Southern Forest Experiment Station, New Orleans, Louisiane, États-Unis. 31

Ketterings, Q.M., Coe, R., van Noordwijk, M., Ambagau, Y. et Palm, C.A., 2001. Reducing uncertainty in the use of allometric biomass equations for predicting above-ground tree biomass in mixed secondary forests. *Forest Ecology and Management*, 146(1-3): 199–209. 45, 106

King, D.A., 1996. Allometry and life history of tropical trees. *Journal of Tropical Ecology*, 12: 25–44. 23

Knapic, S., Louzada, J.L. et Pereira, H., 2011. Variation in wood density components within and between *Quercus faginea* trees. *Canadian Journal of Forest Research*, 41(6): 1212–1219. 24

Kozak, A., 1970. Methods for ensuring additivity of biomass components by regression analysis. *Forestry Chronicle*, 46(5): 402–405. 32

Lahti, T. et Ranta, E., 1985. The SLOSS principle and conservation practice: an example. *Oikos*, 44(2): 369–370. 49

Lanly, J.P., 1981. *Manuel d'inventaire forestier, avec références particulières aux forêts tropicales hétérogènes*. Études FAO : forêts n° 27. FAO, Rome. 47

Larson, P.R., 1963. Stem form development of forest trees. *Forest Science Monograph*, 5: 1–42. 27

Lavorel, S. et Garnier, E., 2002. Predicting changes in community composition and ecosystem functioning from plant traits: revisiting the Holy Grail. *Functional Ecology*, 16: 545–556. 169

Lefsky, M.A., Cohen, W.B., Harding, D.J., Parker, G.G., Acker, S.A. et Gower, S.T., 2002. Lidar remote sensing of above-ground biomass in three biomes. *Global Ecology*

and Biogeography, 11(5): 393–399. 30

Levillain, J., Thongo M'Bou, A., Deleporte, P., Saint-André, L. et Jourdan, C., 2011. Is the simple auger coring method reliable for below-ground standing biomass estimation in *Eucalyptus* forest plantations? *Annals of Botany*, 108(1): 221–230. 74, 75, 77

Li, Y., Andersen, H.E. et McGaughey, R., 2008. A comparison of statistical methods for estimating forest biomass from light detection and ranging data. *Western Journal of Applied Forestry*, 23(4): 223–231. 191

Loetsch, F. et Haller, K.E., 1973. *Forest Inventory. Statistics of Forest Inventory and Information from Aerial Photographs*. BLV Verlagsgesellschaft mbH, Munchen, Allemagne. 47

Louppe, D., Koua, M. et Coulibaly, A., 1994. Tarifs de cubage pour *Afzelia africana* Smith en forêt de Badénou (nord Côte d'Ivoire). Rapport technique, Institut des Forêts (IDEFOR), département foresterie, Côte d'Ivoire. 96

MacDicken, K.G., 1997. A guide to monitoring carbon storage in forestry and agroforestry projects. Report of the forest carbon monitoring program, Winrock Internationl Institute for Agricultural Development, Arlington, Virginie, États-Unis. 30

Magnus, J.R. et Neudecker, H., 2007. *Matrix Differential Calculus with Applications in Statistics and Econometrics*. Wiley series in probability and statistics. John Wiley and Sons, Chichester, Royaume-Uni, 3e éd. 120, 126

Magnussen, S., Kleinn, C. et Picard, N., 2008*a*. Two new density estimators for distance sampling. *European Journal of Forest Research*, 127(3): 213–224. 51

Magnussen, S., Picard, N. et Kleinn, C., 2008*b*. A gamma-poisson distribution of the point to the k nearest event distance. *Forest Science*, 54(4): 429–441. 51

Maguire, D.A. et Batista, J.L.F., 1996. Sapwood taper models and implied sapwood volume and foliage profiles for coastal Douglas-fir. *Canadian Journal of Forest Research*, 26: 849–863. 173

Maniatis, D., Saint-André, L., Temmerman, M., Malhi, Y. et Beeckman, H., 2011. The potential of using xylarium wood samples for wood density calculations: a comparison of approaches for volume measurements. *iForest – Biogeosciences and Forestry*, 4: 150–159. 67

Manning, W.G. et Mullahy, J., 2001. Estimating log models: to transform or not to transform? *Journal of Health Economics*, 20: 461–494. 186

Martinez-Yrizar, A., Sarukhan, J., Perez-Jimenez, A., Rincon, E., Maass, J.M., Solis-Magallanes, A. et Cervantes, L., 1992. Above-ground phytomass of a tropical deciduous forest on the coast of Jalisco, México. *Journal of Tropical Ecology*, 8: 87–96. 106

Massart, P., 2007. *Concentration Inequalities and Model Selection. École d'Été de Probabilités de Saint-Flour XXXIII – 2003*. Lecture Notes in Mathematics no 1896. Springer-Verlag, Berlin Heidelberg. 190

McCarthy, M.C. et Enquist, B.J., 2007. Consistency between an allometric approach and optimal partitioning theory in global patterns of plant biomass allocation. *Functional Ecology*, 21(4): 713–720. 28

McLachlan, G.J. et Krishnan, T., 2008. *The EM Algorithm and Extensions*. Wiley Series in Probability and Statistics. John Wiley & Sons, Hoboken, New Jersey, États-Unis, 2e éd. 192

Meredieu, C., Perret, S. et Dreyfus, P., 2003. Modelling dominant height growth: effect of stand density. *In* A. Amaro, D. Reed et P. Soares (eds.), *Modelling Forest Systems. Proceedings of the IUFRO 4.01 and 4.11 Conference, Instituto Superior de Gestão and Instituto Superior de Agronomia, Sesimbra, Portugal, 2-5 June 2002*. CAB International Publishing, Wallingford, Royaume-Uni, pp. 111–121. 26

Metcalf, C.J.E., Clark, J.S. et Clark, D.A., 2009. Tree growth inference and prediction when the point of measurement changes: modelling around buttresses in tropical forests. *Journal of Tropical Ecology*, 25(1): 1–12. 174

Mokany, K., Raison, R.J. et Prokushkin, A.S., 2006. Critical analysis of root : shoot ratios in terrestrial biomes. *Global Change Biology*, 12(1): 84–96. 28

Monreal, C.M., Etchevers, J.D., Acosta, M., Hidalgo, C., Padilla, J., López, R.M., Jiménez, L. et Velázquez, A., 2005. A method for measuring above- and below-ground C stocks in hillside landscapes. *Canadian Journal of Soil Science*, 85(Special Issue): 523–530. 30

Muller, K.E. et Stewart, P.W., 2006. *Linear Model Theory. Univariate, Multivariate and Mixed Models*. Wiley series in probability and statistics. John Wiley & Sons, Hoboken, New Jersey, États-Unis. 173

Muller-Landau, H.C., Condit, R.S., Chave, J., Thomas, S.C., Bohlman, S.A., Bunyavejchewin, S., Davies, S., Foster, R., Gunatilleke, S., Gunatilleke, N., Harms, K.E., Hart, T., Hubbell, S.P., Itoh, A., Kassim, A.R., Lafrankie, J.V., Lee, H.S., Losos, E., Makana, J.R., Ohkubo, T., Sukumar, R., Sun, I.f., Nur Supardi, M.N., Tan, S., Thompson, J., Valencia, R., Villa Muñoz, G., Wills, C., Yamakura, T., Chuyong, G., Dattaraja, H.S., Esufali, S., Hall, P., Hernandez, C., Kenfack, D., Kiratiprayoon, S., Suresh, H.S., Thomas, D., Vallejo, M.I. et Ashton, P., 2006. Testing metabolic ecology theory for allometric scaling of tree size, growth and mortality in tropical forests. *Ecology Letters*, 9(5): 575–588. 24, 105

Myers, R.H. et Montgomery, D.C., 2002. *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*. Wiley series in probability and statistics. Wiley, New York, New York, États-Unis. 42

Namaalwa, J., Eid, T. et Sankhayan, P., 2005. A multi-species density-dependent matrix growth model for the dry woodlands of Uganda. *Forest Ecology and Management*, 213(1-3): 312–327. 28

Návar, J., 2009. Allometric equations for tree species and carbon stocks for forests of northwestern Mexico. *Forest Ecology and Management*, 257(2): 427–434. 106

Návar, J., Méndez, E. et Dale, V., 2002. Estimating stand biomass in the Tamaulipan thornscrub of northeastern Mexico. *Annals of Forest Science*, 59(8): 813–821. 31, 32

Navarro, M.N.V., Jourdan, C., Sileye, T., Braconnier, S., Mialet-Serra, I., Saint-André, L., Dauzat, J., Nouvellon, Y., Epron, D., Bonnefond, J.M., Berbigier, P., Rouzière, A., Bouillet, J.P. et Roupsard, O., 2008. Fruit development, not GPP, drives seasonal variation in NPP in a tropical palm plantation. *Tree Physiology*, 28(11): 1661–1674. 75

Nelson, B.W., Mesquita, R., Pereira, L.G., Garcia Aquino de Souza, J.S., Teixeira Batista, G. et Bovino Couto, L., 1999. Allometric regressions for improved estimate of secondary forest biomass in the central Amazon. *Forest Ecology and Management*, 117(1-3): 149 –167. 106

Ngomanda, A., Moundounga Mavouroulou, Q., Engone Obiang, N.L., Midoko Iponga, D., Mavoungou, J.F., Lépengué, N., Picard, N. et Mbatchi, B., 2012. Derivation of diameter measurements for buttressed trees, an example from Gabon. *Journal of Tropical Ecology*, 28(3): 299–302. 137

Nicolini, É., Chanson, B. et Bonne, F., 2001. Stem growth and epicormic branch formation in understorey beech trees (*Fagus sylvatica* L.). *Annals of Botany*, 87(6): 737–750. 27

Nogueira, E.M., Fearnside, P.M., Nelson, B.W., Barbosa, R.I. et Keizer, E.W.H., 2008. Estimates of forest biomass in the Brazilian Amazon: New allometric equations and adjustments to biomass from wood-volume inventories. *Forest Ecology and Management*, 256(11): 1853–1867. 106

Nogueira, E.M., Nelson, B.W. et Fearnside, P.M., 2006. Volume and biomass of trees in central Amazonia: influence of irregularly shaped and hollow trunks. *Forest Ecology and Management*, 227(1-2): 14–21. 32

Paine, C.E.T., Marthews, T.R., Vogt, D.R., Purves, D., Rees, M., Hector, A. et Turnbull, L.A., 2012. How to fit nonlinear plant growth models and calculate growth rates: an update for ecologists. *Methods in Ecology and Evolution*, 3(2): 245–256. 183

Pardé, J., 1980. Forest biomass. *Forestry Abstracts*, 41(8): 343–362. 31

Pardé, J. et Bouchon, J., 1988. *Dendrométrie*. ENGREF, Nancy, France, 2e éd. 26, 31, 35, 40, 41, 42, 43

Parresol, B.R., 1993. Modeling multiplicative error variance: an example predicting tree diameter from stump dimensions in baldcypress. *Forest Science*, 39(4): 670–679. 31

Parresol, B.R., 1999. Assessing tree and stand biomass: a review with examples and critical comparisons. *Forest Science*, 45(4): 573–593. 31, 32, 46, 125, 161, 171, 173, 174, 176, 185, 188

Parresol, B.R., 2001. Additivity of nonlinear biomass equations. *Canadian Journal of Forest Research*, 31(5): 865–878. 31

Parresol, B.R. et Thomas, C.E., 1989. A density-integral approach to estimating stem

biomass. *Forest Ecology and Management*, 26: 285–297. 173

Patenaude, G., Hill, R.A., Milne, R., Gaveau, D.L.A., Briggs, B.B.J. et Dawson, T.P., 2004. Quantifying forest above ground carbon content using LiDAR remote sensing. *Remote Sensing of Environment*, 93(3): 368–380. 30

Pearson, T. et Brown, S., 2005. Guide de mesure et de suivi du carbone dans les forêts et prairies herbeuses. Report, Winrock International, Arlington, Virginie, États-Unis. 30

Peng, C., 2000. Growth and yield models for uneven-aged stands: past, present and future. *Forest Ecology and Management*, 132(2-3): 259–279. 28

Perot, T., Goreaud, F., Ginisty, C. et Dhôte, J.F., 2010. A model bridging distance-dependent and distance-independent tree models to simulate the growth of mixed forests. *Annals of Forest Science*, 67(5): 502. 28

Philippeau, G., 1986. Comment interpréter les résultats d'une analyse en composantes principales ? Manuel de Stat-ITCF, Institut Technique des Céréales et des Fourrages, Paris. 101

Picard, N. et Bar-Hen, A., 2007. Estimation of the density of a clustered point pattern using a distance method. *Environmental and Ecological Statistics*, 14(4): 341–353. 48, 51

Picard, N. et Favier, C., 2011. A point-process model for variance-occupancy-abundance relationships. *American Naturalist*, 178(3): 383–396. 48

Picard, N. et Franc, A., 2001. Aggregation of an individual-based space-dependent model of forest dynamics into distribution-based and space-independent models. *Ecological Modelling*, 145(1): 69–84. 28

Picard, N., Kouyaté, A.M. et Dessard, H., 2005. Tree density estimations using a distance method in mali savanna. *Forest Science*, 51(1): 7–18. 51

Picard, N., Sylla, M.L. et Nouvellet, Y., 2004. Relationship between plot size and the variance of the density estimator in West African savannas. *Canadian Journal of Forest Research*, 34(10): 2018–2026. 48

Picard, N., Henry, M., Mortier, F., Trotta, C. et Saint-André, L., 2012. Using Bayesian model averaging to predict tree aboveground biomass. *Forest Science*, 58(1): 15–23. 191

Picard, N., Yalibanda, Y., Namkosserena, S. et Baya, F., 2008. Estimating the stock recovery rate using matrix models. *Forest Ecology and Management*, 255(10): 3597–3605. 28

Ponce-Hernandez, R., Koohafkan, P. et Antoine, J., 2004. *Assessing carbon stocks and modelling win-win scenarios of carbon sequestration through land-use changes*. FAO, Rome. 30

Porté, A. et Bartelink, H.H., 2002. Modelling mixed forest growth: a review of models for forest management. *Ecological Modelling*, 150(1-2): 141–188. 28, 29

Press, W.H., Teukolsky, S.A., Vetterling, W.T. et Flannery, B.P., 2007. *Numerical Recipes: The Art of Scientific Computing*. Cambridge University Press, Cambridge, Royaume-Uni, 3e éd. 149

Preßler, M.R., 1864. *Das Gesetz der Stammbildung*. Arnoldische Buchhandlung, Leipzig, Allemagne. 216

Pretzsch, H., 2009. *Forest Dynamics, Growth and Yield: From Measurement to Model*. Springer-Verlag, Berlin. 24

Pukkala, T., Lähde, E. et Laiho, O., 2009. Growth and yieldmodels for unevensizedforest stands in Finland. *Forest Ecology and Management*, 258(3): 207–216. 28

Putz, F.E., 1983. Liana biomass and leaf area of a “tierra firme" forest in the Rio Negro Basin, Venezuela. *Biotropica*, 15(3): 185–189. 32

R Development Core Team, 2005. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienne, Autriche. 21

Raftery, A.E., Gneiting, T., Balabdaoui, F. et Polakowski, M., 2005. Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, 133(5): 1155–1174. 191

Reed, D.D. et Green, E.J., 1985. A method of forcing additivity of biomass tables when using nonlinear models. *Canadian Journal of Forest Research*, 15(6): 1184–1187. 32

Reinecke, L.H., 1933. Perfecting a stand-density index for even-aged forests. *Journal of Agricultural Research*, 46(7): 627–638. 216

Reyes, G., Brown, S., Chapman, J. et Lugo, A.E., 1992. Wood densities of tropical tree species. General Technical Report SO-88, USDA Forest Service, Southern Forest Experiment Station, New Orleans, Louisiane, États-Unis. 64

Rivoire, M., Genet, A., Didier, S., Nys, C., Legout, A., Longuetaud, F., Cornu, E., Freyburger, C., Motz, A., Bouxiero, N. et Saint-André, L., 2009. Protocole d’acquisition de données volume-biomasse-minéralomasse, Bure. Rapport technique, INRA, Nancy, France. 55

Rösch, H., Van Rooyen, M.W. et Theron, G.K., 1997. Predicting competitive interactions between pioneer plant species by using plant traits. *Journal of Vegetation Science*, 8: 489–494. 169

Russell, C., 1983. *Nutrient cycling and productivity of native and plantation forests at Jari Florestal, Para, Brazil*. Ph.D. thesis, University of Georgia, Athens, Georgia, États-Unis. 41

Rutishauser, E., Wagner, F., Herault, B., Nicolini, E.A. et Blanc, L., 2010. Contrasting above-ground biomass balance in a Neotropical rain forest. *Journal of Vegetation Science*, 21: 672–682. 51

Rykiel, E.J.J., 1996. Testing ecological models: the meaning of validation. *Ecological Modelling*, 90: 229–244. 175, 176

Saatchi, S.S., Houghton, R.A., Dos Santos Alvalá, R.C., Soares, J.V. et Yu, Y., 2007. Distribution of aboveground live biomass in the Amazon basin. *Global Change Biology*, 13(4): 816–837. 30

Saint-André, L., Laclau, J.P., Bouillet, J.P., Deleporte, P., Miabala, A., Ognouabi, N., Baillères, H., Nouvellon, Y. et Moukini, R., 2002*a*. Integrative modelling approach to assess the sustainability of the *Eucalyptus* plantations in Congo. *In* G. Nepveu (ed.), *Connection between Forest Resources and Wood Quality: Modelling Approaches and Simulation Software. Proceedings of the Fourth workshop IUFRO S5.01.04, Harrison Hot Springs, British Columbia, Canada, September 8-15, 2002*. IUFRO, pp. 611–621. 26, 27

Saint-André, L., Laclau, J.P., Deleporte, P., Ranger, J., Gouma, R., Saya, A. et Joffre, R., 2002*b*. A generic model to describe the dynamics of nutrient concentrations within stemwood across an age series of a eucalyptus hybrid. *Annals of Botany*, 90(1): 65–76. 24, 60

Saint-André, L., Laclau, J.P., P., D., Gava, J.L., Gonçalves, J.L.M., Mendham, D., Nzila, J.D., Smith, C., du Toit, B., Xu, D.P., Sankaran, K.V., Marien, J.N., Nouvellon, Y., Bouillet, J.P. et R, 2008. Slash and litter management effects on *Eucalyptus* productivity: a synthesis using a growth and yield modelling approach. *In* E.K.S. Nambiar (ed.), *Site Management and Productivity in Tropical Plantation Forests. Proceedings of Workshops in Piracicaba (Brazil) 22-26 November 2004 and Bogor (Indonesia) 6-9 November 2006*. CIFOR, Bogor, Indonesia, pp. 173–189. 26

Saint-André, L., Leban, J.M., Houllier, F. et Daquitaine, R., 1999. Comparaison de deux modèles de profil de tige et validation sur un échantillon indépendant. Application à l'épicéa commun dans le nord-est de la France. *Annals of Forest Science*, 56(2): 121–132. 27

Saint-André, L., Thongo M'Bou, A., Mabiala, A., Mouvondy, W., Jourdan, C., Roupsard, O., Deleporte, P., Hamel, O. et Nouvellon, Y., 2005. Age-related equations for above- and below-ground biomass of a *Eucalyptus* hybrid in Congo. *Forest Ecology and Management*, 205(1-3): 199–214. 31, 45, 55, 152, 167

Saporta, G., 1990. *Probabilités, analyse des données et statistique*. Technip, Paris. 36, 38, 41, 47, 177, 178, 180, 181, 183, 186

Savage, V.M., Deeds, E.J. et Fontana, W., 2008. Sizing up allometric scaling theory. *PLoS Computational Biology*, 4(9): e1000171. 28

Schlaegel, B.E., 1982. Testing, reporting, and using biomass estimation models. *In* C.A. Gresham (ed.), *Proceedings of the 3rd Annual Southern Forest Biomass Workshop*. Belle W. Baruch Forest Science Institute, Clemson University, Clemson, South Carolina, États-Unis, pp. 95–112. 176

Schnitzer, S.A., DeWalt, S.J. et Chave, J., 2006. Censusing and measuring lianas: a quantitative comparison of the common methods. *Biotropica*, 38(5): 581–591. 32

Schnitzer, S.A., Rutishauser, S. et Aguilar, S., 2008. Supplemental protocol for liana censuses. *Forest Ecology and Management*, 255: 1044–1049. 32

Schreuder, H.T., Banyard, S.G. et Brink, G.E., 1987. Comparison of three sampling methods in estimating stand parameters for a tropical forest. *Forest Ecology and Management*, 21(1-2): 119–127. 49

Schreuder, H.T., Gregoire, T.G. et Wood, G.B., 1993. *Sampling methods for multiressource forest inventory.* Wiley & Sons, New York, New York, États-Unis. 47, 51

Serfling, R.J., 1980. *Approximation Theorems of Mathematical Statistics.* Wiley Series in Probability and Statistics. John Wiley & Sons, New York, New York, États-Unis. 181, 184, 186

Shaw, J.D., 2006. Reineke's stand density index: Where are we and where do we go from here? *In Driving Changes in Forestry. Proceedings of the Society of American Foresters 2005 National Convention, October 19-23, 2005, Fort Worth, Texas, USA*. Society of American Foresters, Bethesda, Maryland, États-Unis, pp. 1–13. 26

Shinozaki, K., Yoda, K., Hozumi, K. et Kira, T., 1964*a*. A quantitative analysis of plant form - the pipe model theory. I. Basic analyses. *Japanese Journal of Ecology*, 14: 97–104. 23, 27

Shinozaki, K., Yoda, K., Hozumi, K. et Kira, T., 1964*b*. A quantitative analysis of plant form - the pipe model theory. II. Further evidence of the theory and its application on forest ecology. *Japanese Journal of Ecology*, 14: 133–139. 23, 27

Shiver, B.D. et Borders, B.E., 1996. *Sampling techniques for forest ressource inventory.* Wiley & Sons, New York, New York, États-Unis. 38, 39, 47

Sillett, S.C., Van Pelt, R., Koch, G.W., Ambrose, A.R., Carroll, A.L., Antoine, M.E. et Mifsud, B.M., 2010. Increasing wood production through old age in tall trees. *Forest Ecology and Management*, 259(5): 976–994. 174

Skovsgaard, J.P. et Vanclay, J.K., 2008. Forest site productivity: a review of the evolution of dendrometric concepts for even-aged stands. *Forestry*, 81(1): 13–31. 24, 26

Smith, R.L., Tebaldi, C., Nychka, D. et Mearns, L.O., 2009. Bayesian modeling of uncertainty in ensembles of climate models. *Journal of the American Statistical Association*, 104(485): 97–116. 191

Soares, P. et Tomé, M., 2002. Height-diameter equation for first rotation eucalypt plantations in Portugal. *Forest Ecology and Management*, 166(1-3): 99–109. 27

St.-Onge, B., Hu, Y. et Vega, C., 2008. Mapping the height and above-ground biomass of a mixed forest using lidar and stereo Ikonos images. *International Journal of Remote Sensing*, 29(5): 1277–1294. 30

Stoyan, D. et Stoyan, H., 1994. *Fractals, Random Shapes and Point Fields.* John Wiley & Sons, Chichester, Royaume-Uni. 48

Tateno, R., Hishi, T. et Takeda, H., 2004. Above- and belowground biomass and net primary production in a cool-temperate deciduous forest in relation to topographical changes in soil nitrogen. *Forest Ecology and Management*, 193(3): 297–306. 28

Taylor, J.M.G., 1986. The retransformed mean after a fitted power transformation. *Journal of the American Statistical Association*, 81: 114–118. 31, 186

Tedeschi, L.O., 2006. Assessment of the adequacy of mathematical models. *Agricultural Systems*, 89(2-3): 225–247. 176

Thompson, S.K., 1992. *Sampling*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, New York, New York, États-Unis. 38, 39

Thornley, J.H., 1972. A balanced quantitative model for root: shoot ratios in vegetative plants. *Annals of Botany*, 36(2): 431–441. 28

Tomé, M., Barreiro, S., Paulo, J.A. et Tomé, J., 2006. Age-independent difference equations for modelling tree and stand growth. *Canadian Journal of Forest Research*, 36(7): 1621–1630. 28

Valinger, E., 1992. Effects of thinning and nitrogen fertilization on stem growth and stem form of *Pinus sylvestris* trees. *Scandinavian Journal of Forest Research*, 7(1-4): 219–228. 27

Vallet, P., Dhôte, J.F., Le Moguédec, G., Ravart, M. et Pignard, G., 2006. Development of total aboveground volume equations for seven important forest tree species in France. *Forest Ecology and Management*, 229(1-3): 98–110. 27

Vallet, P. et Pérot, T., 2011. Silver fir stand productivity is enhanced when mixed with Norway spruce: evidence based on large-scale inventory data and a generic modelling approach. *Journal of Vegetation Science*, 22(5): 932–942. 28

van Breugel, M., Ransijn, J., Craven, D., Bongers, F. et Hall, J.S., 2011. Estimating carbon stock in secondary forests: Decisions and uncertainties associated with allometric biomass models. *Forest Ecology and Management*, 262(8): 1648–1657. 40, 45, 46, 51

Van Pelt, R., 2001. *Forest Giants of the Pacific Coast*. Global Forest Society, Vancouver, Canada. 174

Vanclay, J.K., 1994. *Modelling Forest Growth and Yield – Applications to Mixed Tropical Forests*. CAB International Publishing, Wallingford, Royaume-Uni. 28

Vanclay, J.K., 2009. Tree diameter, height and stocking in even-aged forests. *Annals of Forest Science*, 66(7): 702. 26

Verzelen, N., Picard, N. et Gourlet-Fleury, S., 2006. Approximating spatial interactions in a model of forest dynamics as a means of understanding spatial patterns. *Ecological Complexity*, 3(3): 209–218. 28

Violle, C., Navas, M.L., Vile, D., Kazakou, E., Fortunel, C., Hummel, I. et Garnier, E., 2007. Let the concept of trait be functional! *Oikos*, 116: 882–892. 169

Wagner, F., Rutishauser, E., Blanc, L. et Herault, B., 2010. Effects of plot size and census interval on descriptors of forest structure and dynamics. *Biotropica*, 42(6): 664–671. 48, 49, 51

Weiskittel, A.R., Hann, D.W., Hibbs, D.E., Lam, T.Y. et Bluhm, A.A., 2009.

Modeling top height growth of red alder plantations. *Forest Ecology and Management*, 258(3): 323–331. 26

West, G.B., Brown, J.H. et Enquist, B.J., 1997. A general model for the origin of allometric scaling laws in biology. *Science*, 276: 122–126. 23, 105

West, G.B., Brown, J.H. et Enquist, B.J., 1999. A general model for the structure and allometry of plant vascular systems. *Nature*, 400(6745): 664–667. 23, 28, 105

West, P.W., 2009. *Tree and Forest Measurement*. Springer-Verlag, Berlin, 2e éd. 47, 51

White, J.F. et Gould, S.J., 1965. Interpretation of the coefficient in the allometric equation. *American Naturalist*, 99(904): 5–18. 23

Whraton, E.H. et Cunia, T. (eds.), 1987. *Estimating tree biomass regressions and their error. Proceedings of the workshop on tree biomass regression functions and their contribution to the error of forest inventory estimates, May 26–30, 1986, Syracuse, N.Y. – Part E*, General Technical Report no NE-117. USDA Forest Service, Northeastern Forest Experiment Station, Broomall, Pennsylvania, États-Unis. 51, 125, 194

Yamakura, T., Hagihara, A., Sukardjo, S. et Ogawa, H., 1986. Tree size in a mature dipterocarp forest stand in Sebulu, East Kalimantan, Indonesia. *Southeast Asian Studies*, 23(4): 452–478. 106

Zeide, B., 1980. Plot size optimization. *Forest Science*, 26(2): 251–257. 49, 50

Zianis, D. et Mencuccini, M., 2004. On simplifying allometric analyses of forest biomass. *Forest Ecology and Management*, 187(2-3): 311–332. 24, 175

Zianis, D., Muukkonen, P., Mäkipää, R. et Mencuccini, M., 2005. *Biomass and Stem Volume Equations for Tree Species in Europe*. Silva Fennica Monographs no 4. The Finnish Society of Forest Science and The Finnish Forest Research Institute, Vantaa, Finland. 24, 40, 105

Glossaire

Nous précisons dans ce glossaire la définition de certains termes techniques non usuels, ou de termes qui sont utilisés dans ce manuel avec une acception différente de leur acception courante.

Additivité. Propriété d'un système d'équations allométriques ajustées aux différents compartiments de l'arbre et à l'arbre dans son intégralité, telle que la somme des prédictions pour chacun des compartiments redonne bien la prédiction pour l'intégralité de l'arbre.

Aliquote. Partie extraite d'un compartiment de l'arbre, dont la mesure permet de mesurer l'ensemble du compartiment par règle de trois.

Allométrie. Relation statistique, à l'échelle d'une population, entre deux caractéristiques de taille des individus de cette population. Le plus souvent cette relation a une forme puissance. Exemple: il y a une allométrie chez les vertébrés entre la masse du corps adulte et la taille du cerveau.

Biomasse. Masse de la matière organique vivante ou morte d'un organisme, exprimée en masse de matière sèche. Pour un arbre, l'unité de mesure est le kg ou ses multiples. Par extension, la biomasse d'une zone est la somme des biomasses des organismes qui se trouvent dans cette zone. L'unité de mesure est alors le kg (ou ses multiples) par unité de surface.

Bruit blanc. En probabilités, processus aléatoire générant des variables aléatoires qui sont toutes indépendantes entre elles.

Compartiment. Une partie d'un arbre, généralement déterminée de sorte que les organes au sein d'un compartiment aient des densités (rapport biomasse sèche sur volume frais) semblables (le feuillage, le tronc, les grosses branches, etc., sont des compartiments).

Covariance. Quantité qui mesure la variation simultanée de deux variables aléatoires, c'est-à-dire que la covariance devient plus positive pour chaque couple de valeurs qui diffèrent de leur moyenne dans le même sens, et plus négative pour chaque couple de valeurs qui diffèrent de leur moyenne dans le sens opposé. La covariance d'une variable aléatoire et de cette même variable aléatoire donne la variance.

Doublet. Collection de deux valeurs numériques.

Fractal. Objet dont la structure est invariante par changement d'échelle.

Hart-Becking (indice de). En sciences forestières, indice établi par Hart (1928) et Becking (1953) qui mesure le stockage d'un peuplement à partir de la distance moyenne entre arbres et la hauteur dominante du peuplement. Cet indice est calculé comme le ratio de la distance moyenne entre arbres sur la hauteur dominante, multiplié par 100.

Hétéroscédasticité. Contraire de l'homoscédasticité, c'est-à-dire lorsque la variance de l'erreur résiduelle d'un modèle n'est pas constante (et varie typiquement avec l'une des variables explicatives du modèle).

Homoscédasticité. Propriété vérifiée par l'erreur résiduelle d'un modèle lorsque la variance de cette erreur est constante. L'homoscédasticité est l'une des conditions requises pour l'ajustement d'un modèle linéaire.

Loi de Eichhorn. En sciences forestières, loi empirique énoncée par Eichhorn (1904) qui stipule: le volume spécifique d'un peuplement équienne, monospécifique et de couvert fermé, n'est fonction que de sa hauteur dominante. Il s'agit là de la seconde loi de Eichhorn; la première loi de Eichhorn stipule: la hauteur dominante d'un peuplement équienne, monospécifique et de couvert fermé, n'est fonction que de l'âge, de l'essence, et des conditions de station.

Loi de Pressler. En sciences forestières, loi empirique énoncée par Preßler (1864) qui stipule: la surface du cerne de croissance d'un arbre augmente linéairement du haut de l'arbre jusqu'à la base fonctionnelle de son houppier, et reste constante de la base du houppier jusqu'au bas de l'arbre.

Minéralomasse. Quantité d'éléments minéraux dans la biomasse.

Masse de Dirac. Distribution (au sens statistique du terme) concentrée en une valeur x_0 d'une variable aléatoire continue (c'est-à-dire que la probabilité que la variable aléatoire soit $< x$ vaut 0 pour $x < x_0$ et 1 pour $x > x_0$).

Monte Carlo. Se dit d'une méthode visant à calculer une valeur numérique à l'aide de simulation d'un processus aléatoire.

Position sociale. Pour un arbre, position de sa couronne dans la canopée, qui détermine sa hiérarchie vis-à-vis de la compétition pour la lumière (on parle aussi de statut social). On distingue souvent les arbres dominants, co-dominants et dominés.

Reinecke Density Index (RDI). En sciences forestières, indice établi par Reinecke (1933) qui mesure le stockage d'un peuplement à partir de son nombre de tiges par hectare (= densité du peuplement) et du diamètre de l'arbre de surface terrière moyenne (= diamètre quadratique moyen). Cet indice est calculé comme le ratio de la densité du peuplement sur sa densité maximale, telle que déterminée à partir du diamètre quadratique moyen par la droite d'auto-éclaircie.

Variable ordinale. Variable qui prend des valeurs discrètes telles que les différentes modalités de ces valeurs discrètes peuvent être ordonnées. Par exemple le mois de l'année est une variable ordinale (les mois peuvent être rangés par ordre chronologique).

Variance. Quantité qui mesure la dispersion d'une variable aléatoire autour de sa valeur moyenne. Elle est calculée comme la moyenne des carrés des écarts à la moyenne.

Lexique des symboles mathématiques

Symboles latins

a valeur estimée d'un coefficient d'un modèle prédictif
A surface d'une placette d'échantillonnage
$\mathcal{A}$ surface du peuplement
b valeur estimée d'un coefficient d'un modèle prédictif
B biomasse d'une aliquote, d'un compartiment (tronc, branches, feuillage...), d'un arbre ou d'un peuplement
CV_X coefficient de variation d'une grandeur X
c exposant d'une loi puissance
C définition 1: circonférence d'un arbre; définition 2: coût d'échantillonnage; définition 3: un critère de validation d'un modèle
D diamètre d'un arbre
D_0 diamètre dominant du peuplement
E précision d'estimation d'une grandeur estimée
f une fonction reliant une variable réponse à une ou plusieurs variables explicatives
F indice de Furnival
g une fonction
G surface terrière d'un arbre ou d'un peuplement
h une hauteur comprise entre zéro (le sol) et la hauteur H de l'arbre
H hauteur d'un arbre
H_0 hauteur dominante du peuplement
$\mathbf{I}_n$ matrice d'information de Fisher pour un échantillon de taille n
k coefficient multiplicateur d'une loi puissance
K nombre de parts pour une validation croisée
ℓ vraisemblance d'un échantillon
$\mathcal{L}$ log-vraisemblance d'un échantillon
L longueur d'un billon de bois
M définition 1: nombre de compartiments de biomasse au sein d'un arbre; définition 2: nombre de modèles concurrents prédisant une même variable réponse
n taille d'un échantillon
N définition 1: nombre total d'unités d'échantillonnage (arbre ou parcelle) dans le peuplement; définition 2: densité d'un peuplement (nombre de tiges par hectare)
$\mathcal{N}$ la loi normale (également appelée loi gaussienne, ou loi de Laplace-Gauss)
p nombre de variables explicatives d'un modèle (ordonnée à l'origine non comprise)
P profil de tige (courbe donnant la surface de la section du tronc en fonction de la hauteur)
q définition 1: nombre de paramètres estimés d'un modèle; définition 2: quantile de la loi normale centrée réduite
Q nombre d'itérations Monte Carlo
R définition 1: coefficient de détermination d'un modèle; définition 2 (dans la théorie de

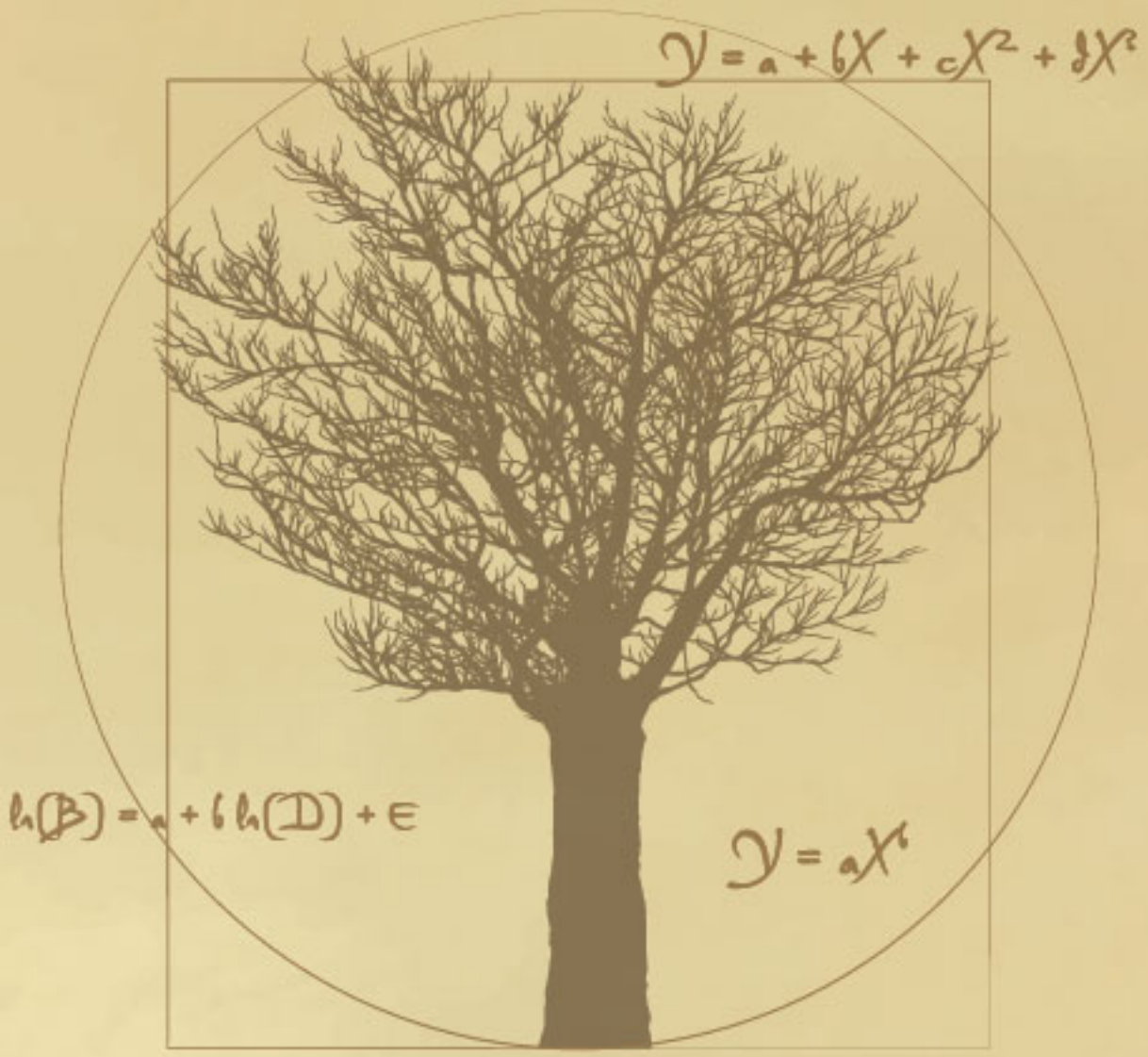
sélection de modèle): un risque; définition 3: rayon d'un billon de bois
S nombre de strates d'une stratification
S_X écart-type empirique d'une variable X
$\mathcal{S}_n$ un jeu de données comportant n observations
t_n quantile d'une loi de Student à n degré de liberté
T âge d'une plantation
V volume d'un billon, d'un arbre ou d'un peuplement
w définition 1: poids d'une observation dans la régression pondérée; définition 2: poids d'un modèle dans un mélange de modèles
X une variable (en général variable explicative d'un modèle)
$\mathbf{x}$ un vecteur de variables explicatives
$\mathbf{X}$ matrice du plan pour un modèle linéaire
Y une variable (en général variable réponse d'un modèle)
$\mathbf{Y}$ vecteur réponse d'un modèle multivarié
z une variable latente pour l'algorithme EM
Z une variable (en général une co-variable définissant une stratification du jeu de données)

Symboles grecs

α définition 1: valeur « vraie » (inconnue) d'un coefficient d'un modèle prédictif; définition 2: seuil de confiance d'un intervalle de confiance (généralement 5 %)
β valeur « vraie » (inconnue) d'un coefficient d'un modèle prédictif
γ fonction de perte (dans la théorie de la sélection de modèle)
δ masse de Dirac
Δ un écart de valeur pour une grandeur donnée
ε erreur résiduelle d'un modèle prédictif
$\boldsymbol{\varepsilon}$ vecteur des erreurs résiduelles d'un modèle multivarié
ζ covariance résiduelle entre deux compartiments
η coefficient de retrait volumique
θ un ensemble de paramètres d'un modèle
$\boldsymbol{\theta}$ un vecteur de paramètres d'un modèle multivarié
ϑ un ensemble de paramètres
μ espérance d'une variable aléatoire = moyenne « vraie » (inconnue) d'une grandeur à estimer
ξ paramètre de la transformation de Box-Cox
ρ densité du bois
σ écart-type de l'erreur résiduelle d'un modèle prédictif
$\boldsymbol{\Sigma}$ matrice de variance-covariance d'une loi multinormale
τ écart-type « vrai » (inconnu) d'une grandeur à estimer
ϕ densité de probabilité de la loi normale
ψ fonction définissant une transformation de variable
χ taux d'humidité
χ_0 point de saturation des fibres
ω une proportion (par exemple, la proportion en biomasse fraîche du bois dans un billon)

Symboles non alphabétiques

$\varnothing$ diamètre d'un arbre, d'un billon, d'une branche ou d'une racine



E-ISBN 978-92-5-207347-5