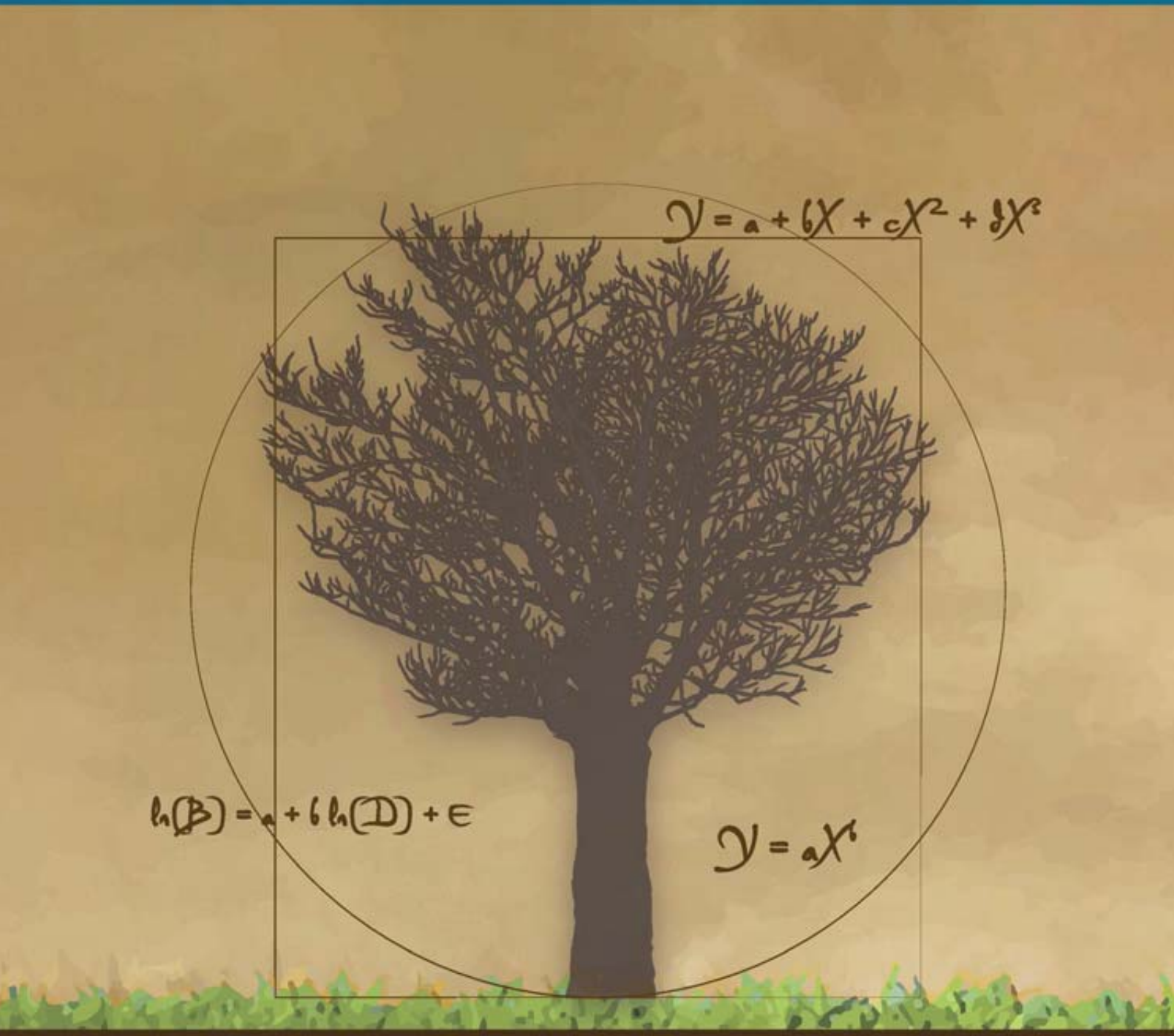
Manuel de construction d'équations allométriques pour l'estimation du volume et la biomasse des arbres. De la mesure de terrain à la prédiction

Nicolas Picard, Laurent Saint-André, Matieu Henry

Manual for building tree volume and biomass allometric equations

From field measurement to prediction



cirad

FAO

Manual for building tree volume and biomass allometric equations

from field measurement to prediction

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
Forestry Department
Food and Agriculture Organization of the United Nations

August 2012

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

Food and Agriculture Organization of the United Nations (FA0)
Viale delle Terme di Caracalla
00153 Rome, Italie

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

Photo credit: 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).

Recommended citation: Picard N., Saint-André L., Henry M. 2012. Manual for building tree volume and biomass allometric equations: from field measurement to prediction. Food and Agricultural Organization of the United Nations, Rome, and *Centre de Coopération Internationale en Recherche Agronomique pour le Développement*, Montpellier, 215 pp.

Contents

List of Figures

List of Photos

List of Tables

List of Red lines

Foreword

In the United Nations Framework Convention on Climate Change, the potential benefits for non-Annex I parties will be based on results that must be measured, reported and verified. The precision of these results therefore has a major impact on potential financial compensation, and the capacity to measure forest carbon stocks is thus of increasing importance for countries who plan to contribute to mitigating climate change through their forest activities. The measurement of these stocks currently calls on techniques that are applied at different scales ranging from field inventories on a local scale to airborne lasers or radar, and ultimately to satellite remote sensing on a national or subregional scale. Indirect measurements of forest carbon stocks, as derived from satellite, Lidar or radar data, are based on calibrated relationships with field measurements. The same applies to inventories. Therefore, and ultimately, any forest carbon stock measurement at some point needs trees to be weighed in the field. This constitutes the keystone on which rests the entire edifice of forest carbon stock estimations, whatever the scale considered.

Also, the allometric equations used to predict the biomass of a tree from easier-to-measure dendrometric characteristics such as tree diameter or height, are key factors in estimating the contribution made by forest ecosystems to the carbon cycle. This manual covers all the steps in the construction of these equations, starting with the measurement of tree biomass in the field. It should therefore prove particularly useful in countries that are not yet in possession of measurements and equations that match their forests.

This *Manual for building tree volume and biomass allometric equations* therefore takes the form of a guide intended for all students, technicians or researchers working to assess forest resources such as volume, biomass and carbon stocks for commercial, bioenergy or climate change mitigation purposes. The methods described herein apply to most forests and ecological areas, but special emphasis has been placed on tropical forest that, more than the others, today require efforts from the international community to measure carbon stocks. The manual includes a red line to guide the reader. This describes concrete cases illustrating the different questions that arise when constructing allometric equations, taking samples, making field measurements, entering data, graphically exploring the data, fitting models and using them for predictions. The data employed in these illustrations are derived from three sites which are very different in terms of forest structure and resources available. The practical advice included in these illustrations should allow readers to tackle most of the problems customarily encountered. This manual may also interest forest biometricians to the extent that it contains not only exhaustive reminders on the mathematical theory of regressions and recent developments therein, but also a great deal of advice on selecting

and handling linear regression models.

215 pages. Numerous illustrations. Bibliography comprising 255 publications

Francis Cailliez

August 2012

Acknowledgements

The authors wish to thank the Food and Agriculture Organization of the United Nations and the UN-REDD programme for funding the publication and translation of this guide.

The authors also wish to thank all those cited below for their contributions to the field campaigns and the data used for the red lines, and for enriching with their personal experience the contents of this guide and accepting to proof-read and translate this guide: 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 and 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 and Prof. Riccardo Valentini.

The authors are also grateful to all those who despite short timelines provided their comments and suggestions for corrections, and their encouragements. This guide benifited greatly from their precious contribution, but the responsibility for its contents is incumbent solely upon its authors.

The methods presented in this guide were developed during measurement campaigns funded by the following projects: ATP Carbone (CIRAD), MODELFOR (ONF), BIOMASSE OPE (ANDRA), SOERE F-ORE-T (GIP ECOFOR), EMERGE (ANR), WAFT (EU), ULCOS (EU), CarboAfrica (EU, contract No. INCO-CT-2004-003729).

Preamble

This guide is intended for students, researchers and engineers who wish to acquire the methodology necessary to construct tree volume, biomass or nutrient content tables. The various models used for this purposes have been grouped together here as they all call upon the same principle: they estimate a variable that is difficult to measure in trees (e.g. volume) using simpler characteristics such as tree diameter at breast height (1.30 m), or tree height or age.

This guide is based on a large number of reference papers, and although it does not claim to present all the possible cases, it does describe techniques that can be used to construct the equations required. The references given in the text are specified when possible as author, year and page such that the reader can easily pursue threads of particular interest. A concrete example (dubbed the "red line") guides the reader and helps acquire knowledge through practical application.

This guide does not require a great many prerequisites. The red lines use Microsoft Excel to prepare files and R (R Development Core Team, 2005) for model fitting. The R command lines used are given in each red line.

1

The foundations of biomass estimations

If we consider trees on a population scale, we see that the different dimensions of an individual are statistically related one with another (Gould, 1966). This relation stems from the ontogenic development of individuals which is the same for all to within life history-related variability. For instance, the proportions between height and diameter, between crown height and diameter, between biomass and diameter follow rules that are the same for all trees, big or small, as long as they are growing under the same conditions (King, 1996; Archibald & Bond, 2003; Bohlman & O'Brien, 2006; Dietze *et al.*, 2008). This is the basic principle of allometry and can be used to predict a tree variable (typically its biomass) from another dimension (e.g. its diameter). An allometric equation is a formula that quantitatively formalizes this relationship. We will speak in this guide of volume, biomass and nutrient content models, respectively. Allometry also has another more restrictive definition: the proportionality relationship between relative increases in dimensions (Huxley, 1924; Gayon, 2000). If we call biomass B and diameter D, this second definition means that there is a coefficient a such that:

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

which integrates to a power relationship: $B = b \times D^a$. Using this restricted definition, an allometric equation is therefore synonymous with a power equation (White & Gould, 1965). Parameter a gives the allometry coefficient (proportionality between relative increases), whereas parameter b indicates the proportionality between cumulated varaibles. It may be necessary to add a y-intercept to this relation that becomes $B = c + bD^a$, where c is the biomass of an individual before it reaches the height at which diameter is measured (e.g. 1.30 m if D is measured at 1.30 m). The power relationship reflects the self-similarity idea of individuals during development (Gould, 1971). It was this principle, and the "pipe model" (Shinozaki *et al.*, 1964a,b), that were used to develop an allometric scaling theory (West *et al.*, 1997, 1999; Enquist *et al.*, 1998, 1999). If we apply certain hypotheses covering biomechanical constraints, tree stability, and hydraulic resistance in the network of water-conducting cells, this theory predicts a power relationship with an exponent of $a = 8/3 \approx 2.67$ between tree biomass and diameter. This relationship is of considerable interest as it is founded on physical principles and a mathematical representation of the cell networks within trees. It has nevertheless been the subject of much discussion, and has on occasion been accused of being too general in nature (Zianis & Mencuccini, 2004; Zianis *et al.*, 2005;

Muller-Landau *et al.*, 2006), even though it is possible to use other allometric coefficients depending on the biomechanical and hydraulic hypotheses selected (Enquist, 2002).

Here in this guide we will adopt the broadest possible definition of allometry, i.e. a linear or non-linear correlation between increases in tree dimensions. The power relationship will therefore simply be considered as one allometric relationship amongst others. But whatever definition is adopted, allometry describes the ontogenic development of individuals, i.e. the growth of trees.

1.1 "Biology": Eichhorn's rule, site index, etc.

Tree growth is a complex biological phenomenon (Pretzsch, 2009) resulting from the activity of buds (primary growth or axis elongation) and the cambium (secondary growth, or axis thickening). Tree growth is eminently variable as dependent upon the individual's genetic heritage, its environment (soil, atmosphere), its developmental stage (tissue aging) and the actions of man (changes in the environment or in the tree itself through trimming or pruning).

Conventionally in biomass studies, trees are divided into homogeneous compartments: trunk wood, bark, live branches, dead branches, leaves, large and medium-size roots, and finally small roots. Biomass is a volume multiplied by a density, while nutrient content is a biomass multiplied by a concentration of nutrient. Volume, density and nutrient concentration change over time, not only in relation to the above-mentioned factors (see for example the review by Chave *et al.*, 2009 on wood density) but also within trees: between compartments but also in relation to radial position (near the pith or the bark), and longitudinal position (near the soil or the crown), see for example regarding the concentrations of nutrient: Andrews & Siccama (1995); Colin-Belgrand *et al.* (1996); Saint-André *et al.* (2002b); Augusto *et al.* (2008); or for wood density: Guilley *et al.* (2004); Bergès *et al.* (2008); Henry *et al.* (2010); Knapic *et al.* (2011). All this has implications for biomass and nutrient content equations, and this section therefore aims to underline a few important notions in forestry which will then be included in the process used to devise this guide's models in "biological" terms (what factors can potentially cause variations?), rather than in purely mathematical terms (which is the best possible equation regardless of its likelihood of reflecting biological processes?). A combination of these two aims is ultimately this guide's very purpose.

1.1.1 Even-aged, monospecific stands

This type of stand is characterized by a relatively homogeneous tree population: all are of the same age and most are of the same species. Tree growth in such stands has been the subject of much study (de Perthuis, 1788 *in* Batho & García, 2006) and the principles described below can be fairly universally applied (Assmann, 1970; Dhôte, 1991; Skovsgaard & Vanclay, 2008; Pretzsch, 2009; García, 2011). It is common practice to distinguish between the stand as a whole then the tree within the stand. This distinction dissociates the different factors that impact on trees growth: site fertility, overall pressure within the stand, and social status. Site fertility, in its broad sense, includes the soil's capacity to nourish the trees (nutrients and water) and the area's general climate (mean lighting, temperature and rainfall, usual recurrence of frost or drought, etc.). The pressure between trees in the stand can be assessed using a number of stand density indices. And the social status of each individual describes its capacity to mobilize resources in its immediate environment.

Photo 1.1 – *Eucalyptus plantation in the Republic of the Congo. Top is the Kissoko area, an illustration of savannah-plantation mosaics. Bottom is the Kondi area in operation, an illustration of the main destinations for eucalyptus wood (logs for paper pulp and charcoal production for the city of Pointe-Noire (photos: L. Saint-André).*

Stand growth

The notion of production in forestry includes standing tree volume (or biomass) and all that has been subtracted from the stand over its lifecycle (by death or thinning). In general, this notion of production, as given in production tables and in most growth and yield models, does not include forest litter (fallen leaves, branches, bark) or root turnover. But in process-based models, and in studies aiming to estimate the carbon or mineral elements content of stands, the notion of production also includes the renewal of these organs. In the rest of this section we will consider production in its reduced definition.

The production of an even-aged, monospecific stand, for a given species in a given region and in a broad range of silvicultures (as long as the canopy is closed), is entirely determined by its mean height. This assertion is known as Eichhorn (1904), or Eichhorn's extended rule when it considers dominant height instead of mean height (Decourt, 1973). It means that the fertility of different sites in a given region modifies only the speed with which this relation is followed. Although the rule has been questioned (see Assmann, 1970), the height growth of dominant trees (H_0) is still the main basic principle in most growth and models (e.g. Dhôte, 1996; García, 2003; Saint-André *et al.*, 2008; Skovsgaard & Vanclay, 2008; Weiskittel *et al.*, 2009; García, 2011). The principle is summarized by Alder (1980 *in* Pardé & Bouchon, 1988) in the following sentence: "The height / age / fertility index relationship is key to predicting the growth of homogeneous stands. It is normally expressed in the form of a cluster of fertility plots". Dominant height growth, as a first approximation, depends only on site fertility (in the broad sense, equivalent to the site index) and on stand age in most even-aged, monospecific, temperate or tropical ecosystems. This is the result of two main factors: because of their status, dominant trees are less sensitive that dominated trees to competition, and height growth is also less sensitive than diameter growth to silviculture (except a specific thinning regime). This means that dominant tree height growth is a

more accurate indicator of site fertility than mean height or diameter growth. To achieve Eichhorn's rule, it is then necessary to couple basal area (or volume) growth and dominant tree height growth. This relationship is also stable for a given species and a given region in a broad range of silvicultures (as a first approximation provided that the canopy is sufficiently dense). An example for beech in France is given by Dhôte (1996).

But there are a few examples where the strict relationship of $H_0 = f(\text{age and fertility})$ is invalid: the Laricio pine in central France (Meredieu *et al.*, 2003) and eucalyptus in the Republic of Congo (Saint-André *et al.*, 2002a). In both cases dominant tree height growth also depends on stand density. The underlying hypothesis is based on poor soil fertility that means strong competition for access to water and mineral resources, even for dominant trees. Over the last few years, a "date" effect has also been clearly shown, both for this relationship and that linking basal area growth and dominant tree height growth, due to global changes (see for example Bontemps *et al.*, 2009, 2011; Charru *et al.*, 2010).

In short, even though it is in dispute and is not necessarily as invariable as was hoped, this first "rule" is nevertheless important as it can be used, in so-called parameterized biomass tables, to introduce the notion of fertility through dominant tree age and height in inventoried stands (and incidentally density) and thereby increase the generic nature of the equations constructed.

Tree growth in a stand

When volume or biomass growth is obtained for an entire stand, efforts must be made to distribute this between the different trees. The relationships employed for individual diameter growth are often of the potential × reducer type, where the potential is given by basal area growth and / or dominant tree growth, and the reducers depend on (*i*) a density index and (*ii*) the tree's social status. A density index may simply be stand density, but researchers have developed other indices such as the Hart-Becking spacing index, based on the growth of trees outside the stand (free growth), or the Reinecke density index (RDI), based on the self-thinning rule (tree growth in a highly dense stand). Both of these have the advantage of being less dependent upon stand age than on density itsel (see Shaw, 2006 or more generally the literature review by Vanclay, 2009). The social status of trees is generally expressed by ratios such as H/H_0 or D/D_0 (where D is tree diameter at breast height, H its height and D_0 the diameter at breast height (dbh) of the dominant tree in the stand), but other relationships may also be used. For example, Dhôte (1990), Saint-André *et al.* (2002a) and more recently Cavaignac *et al.* (2012) used a piecewise linear model to model tree diameter growth: below a certain circumference threshold, trees are overtopped and no longer grow; above the threshold their growth shows a linear relationship with tree circumference. This relationship clearly expresses the fact that dominant trees grow more than dominated trees. The threshold and slope of the relationship change with stand age and silviculture treatments (thinning). Although height growth may also be estimated using potential × reducer relationships, modelers generally prefer height-circumference relationships (Soares & Tomé, 2002). These relationships are saturated (the asymptote is the height of the dominant tree in the stand) and curvilinear. This relationship's parameters also change with age and silviculture treatment (Deleuze *et al.*, 1996).

These two other relationships give the dimension of each tree in stands, and in particular it should be underlined that the density and competition (social status) indices that largely determine the individual growth of trees in stands are factors that can also be integrated into biomass tables. Interesting variables from this standpoint are stand density, the Hart-Becking spacing index, the RDI, then — on an individual scale — the slenderness ratio

(H/D), tree robustness ($D^{1/2}/H$ — Vallet *et al.*, 2006; Gomat *et al.*, 2011), or its social status (H/H_0 ou D/D_0).

Biomass distribution in the tree

Finally, now that stand biomass has been distributed between the trees, we need to attribute it to each compartment in each tree and distribute it along the axes. Regarding the trunk, Pessler's rule is usually employed (or for ecophysiologists its equivalent given by the pipe model described by Shinozaki *et al.*, 1964a,b): (*i*) ring cross-sectionnal area increases in a linear manner from the top of the tree to the functional base of the crown; (*ii*) it is then constant from the base of the crown to the base of the tree. This means that as the tree grows its trunk becomes increasingly cylindrical as the growth rings are wider apart near the crown than at the base. Pressler's rule, however, does not express the mean distribution of wood along the tree (Saint-André *et al.*, 1999) since ring mark area in dominant trees may continue to increase beneath the crown, and in dominated / overtopped trees it may greatly decrease. In extreme cases the growth ring may not be complete at the base of the tree, or may even be entirely missing as for example in the beech (Nicolini *et al.*, 2001). Also, any action on the crown (high or low densities, thinning, trimming) will have an impact on the distribution of annual growth rings and therefore on the shape of the trunk (see the review by Larson, 1963, or the examples given by Valinger, 1992; Ikonen *et al.*, 2006). Wood density is also different at the top and base of the tree (young wood near the crown, high proportion of adult wood toward the base — Burdon *et al.*, 2004) but it may also vary according to the condition of the tree's growth (changes in proportion between early and late wood, or changes in cell structure and properties, see Guilley *et al.*, 2004; Bouriaud *et al.*, 2005; Bergès *et al.*, 2008 for some recent papers). Therefore, the biomass of different trunks of the same dimensions (height, dbh, age) may be the same or different depending on the conditions of tree growth. An increase in trunk volume may be accompanied by a decrease in its wood density (as typically in softwoods), and does not therefore cause a major difference in trunk biomass. Regarding branches and leaves, their biomass is greatly dependent upon tree architecture and therefore on stand density: considering trees of equal dimensions (height, dbh and age), those growing in open stands will have more branches and leaves than those growing in dense stands. The major issue, therefore, in current research on biomass consists in identifying the part of the biomass related to the tree's intrinsic development (ontogenics) and that related to environmental factors (Thornley, 1972; Bloom *et al.*, 1985; West *et al.*, 1999; McCarthy & Enquist, 2007; Savage *et al.*, 2008; Genet *et al.*, 2011; Gourlet-Fleury *et al.*, 2011). And the biomass of the roots depends on the biome, on above-ground biomass, development stage and growing conditions (see for example Jackson *et al.*, 1996; Cairns *et al.*, 1997; Tateno *et al.*, 2004; Mokany *et al.*, 2006).

What emerges from these notions is that growing conditions have an impact not only on the overall amount of biomass produced, but also on its distribution within the tree (above-ground / below-ground proportion, ring stacking, etc.). These potential variations must therefore be taken into account when sampling (particularly for crosscutting trunks and taking different aliquots), but also when constructing volume/biomass tables such that they reflect correctly the different above-ground / below-ground; trunk / branch; leaves / small roots biomass ratios resulting from different growing conditions.

1.1.2 Uneven-aged and/or multispecific stands

The notions described above are also valid for multispecific and uneven-aged stands, but in most cases their translation into equations is impossible in the previous form (Peng, 2000). For example, the notion of dominant height is difficult to quantify for irregular and / or multispecific stands (should we determine a dominant height considering all the trees? Or species by species?). Likewise, what does basal area growth mean for a very irregular stand as can be found in wet tropical forests? And finally, how can we deal with the fact that tree age is often inaccessible (Tomé *et al.*, 2006)? The growth models developed for these stands therefore more finely break down the different scales (biomass produced on the stand scale, distribution between trees and allocation within trees) than do those developed for regular stands. These models may be divided into three types: (1) matrix population models; (2) individual-based models that in general depend on the distances between trees; (3) gap models (see the different reviews by Vanclay, 1994; Franc *et al.*, 2000; Porté & Bartelink, 2002). Matrix population models divide trees into functional groups (with a common growth strategy) and into classes of homogeneous dimensions (generally diameter) then apply a system of matrices that include recruitment, mortality and the transition of individuals from one group to another (see for example Eyre & Zillgitt, 1950; Favrichon, 1998; Namaalwa *et al.*, 2005; Picard *et al.*, 2008). In individual-based models, the tree population is generally mapped and the growth of a given tree depends on its neighbors (see for example Gourlet-Fleury & Houllier, 2000 for a model in tropical forest, or Courbaud *et al.*, 2001 for a model in temperate forest). But, like for the models developed for regular stands, some individual-based models are independent of distances (for example Calama *et al.*, 2008; Pukkala *et al.*, 2009; Vallet & Pérot, 2011; Dreyfus, 2012) and even intermediate models (see Picard & Franc, 2001; Verzelen *et al.*, 2006; Perot *et al.*, 2010). Finally, in gap models, the forest is represented by a set of cells at different stages of the silvigenetic cycle. Here, tree recruitment and mortality are simulated stochastically while tree growth follows identical rules to those of distance-independent, individual-based models (see a review in Porté & Bartelink, 2002).

The fact that these stands are more complicated to translate into equations does not in any way detract from the principles mentioned above for the construction of volume, biomass or nutrient content tables: (*i*) fertility is introduced to broaden the range over which the biomass tables are valid; (*ii*) density indices are used to take account of the degree of competition between trees; and (*iii*) social status is taken into account in addition to tree basic characteristics (height, diameter).

The development of volume/biomass tables for multispecific forests faces a number of constraints in addition to those facing tables for monospecific forests: conception of suitable sampling (which species? how can species be divided into so-called functional groups?), access to the field (above all in tropical forests where stands are often located in protected area where felling is highly regulated or even prohibited for certain species).

1.2 Selecting a method

1.2.1 Estimating the biomass of a biome

There is no single method for estimating biomass stocks, but a number of methods depending on the scale considered (Gibbs *et al.*, 2007). On a national or larger scale, mean values per biome are usually employed (FAO, 2006): the amount of biomass is estimated by multiplying the surface area of each biome by the mean amount of biomass per unit surface area of this biome. And mean amounts per biome are estimated from measurements made on a smaller

Photo 1.2 – *Heterogeneous stands. On the left, multispecific stands on Mont Saint-Anne in Quebec; on the right, multispecific, uneven-aged stands in Costa-Rica (photo: B. Locatelli).*

scale. Biomass on national to landscape scales can be estimated by remote sensing. Regardless of whether this uses satellite-borne optical sensors (Landsat, MODIS), high resolution satellite images (Ikonos, QuickBird), low resolution aerial photographs, satellite-borne radar or microwave sensors (ERS, JERS, Envisat, PALSAR), or laser sensors (Lidar), all these methods assume that field measurements are available to adjust biomass-predicting models to remote sensing observations. When dealing with satellite-borne optical sensors, field data are necessary to calibrate the relationship between biomass and satellite vegetation indices (NDVI, NDFI, AVI, GVI, etc.) (Dong *et al.*, 2003; Saatchi *et al.*, 2007). High-resolution images and aerial photographs provide information on tree crown size and height, and field data are then necessary to relate this information to biomass (for example Bradley, 1988; Holmgren *et al.*, 1994; St.-Onge *et al.*, 2008; Gonzalez *et al.*, 2010). The same applies to Lidar information on the vertical structure of forests, and to radar and microwave information on the vertical distribution of plant water (for example Lefsky *et al.*, 2002; Patenaude *et al.*, 2004). But remote sensing-based methods have their limitations in terms of providing precise biomass measurements (particularly surface areas) and differentiating forest types due to the technical, financial and human resources available. They are also hindered by cloud cover and are susceptible to saturated signals for certain vegetation types.

Therefore, biomass estimation methods on a landscape or greater scale rely on field measurements taken between the landscape and plot scales. Biomass estimations on this scale are based on forest inventory data: inventory of a sample of trees if the area is large, or otherwise a full inventory (particularly in permanent plots of a few hectares). On a smaller scale, individual biomass measurements may be taken by weighing trees and understorey vegetation.

1.2.2 Estimating the biomass of a forest or a set of forests

Estimations of forest biomass or nutrient content based on forest inventories require:

1. an exhaustive or statistical inventory of the trees growing;

2. models to evaluate carbon stocks from the dimensions of the individuals measured;

3. an evaluation of the biomass contained in the necromass (standing dead wood) and in understorey vegetation.

In this guide we focus on the second aspect even though an inventory and quantitative evaluation of the understorey is not necessarily easy to obtain, particularly in highly mixed forests.

Two main methods based on inventory data may be employed to estimate tree carbon or mineral element stocks (MacDicken, 1997; Hairiah *et al.*, 2001; AGO, 2002; Ponce-Hernandez *et al.*, 2004; Monreal *et al.*, 2005; Pearson & Brown, 2005; Dietz & Kuyah, 2011): (1) use of biomass/nutrient content tables: this solution is often adopted as it rapidly yields the carbon or nutrient content of a plot at a given time point. In general, all the different (above-ground, below-ground, ground litter, etc.) compartments of the ecosystem are considered. Trees are specifically felled for these operations. The compartments (cross cuts) may vary depending on application and the field of interest (see chapter 3). (2) Use of models to estimate successively tree volume, wood density and nutrient content. This method has the advantage of dissociating the different components, making it possible to analyze the effect of age and growing conditions independently on one or another of the components. In general, only the trunk can be modeled in detail (between and within rings). The biomass of the other compartments is estimated from volume expansion factors, mean wood density measurements and nutrient concentration. In all cases these models call upon one and the same model type that indifferently groups together "volume tables, biomass tables, nutrient content tables, etc." and which is described here in this guide.

Biomass and nutrient content tables are closely related to volume tables that have been the subject of much study for nearly two centuries. The first tables were published by Cotta, 1804 (*in* Bouchon, 1974) for beech (*Fagus sylvatica*). The principle employed is to link a variable that is difficult to measure (e.g. the volume of a tree, its mass, or its mineral elements content) to variables that are easier to determine, e.g. tree diameter at breast height (DBH) or height. If both characteristics are used, we speak of a double-entry table; if only DBH is used, we speak of a single-entry table. In general, close correlations are obtained and the polynomial, logarithmic and power functions are those most widely used. For more details, readers may see the reviews by Bouchon (1974); Hitchcock & McDonnell (1979); Pardé (1980); Cailliez (1980); Pardé & Bouchon (1988), and more recently by Parresol (1999, 2001).

These functions are relatively simple but have three major snags. First, they are not particularly generic: if we change species or stray outside the calibration range, the equations must be used with caution. The chapter on sampling gives a few pointers on how to mitigate this problem. The key principle is to cover as best as possible the variability of the quantities studied.

The second snag with these functions lies in the very nature of the data processed (volumes, masses, nutrient content). In particular, problems of heteroscedasticity may arise (i.e. non homogeneous variance in biomasses in relation to the regressor). This has little impact on the value of the estimated parameters: the greater the number of trees sampled, the more rapid the convergence toward real parameter values (Kelly & Beltz, 1987). But everything concerning the confidence interval of the estimations is affected:

1. the variance of the estimated parameters is not minimal;
2. this variance is biased; and
3. residual variance is poorly estimated (Cunia, 1964; Parresol, 1993; Gregoire & Dyer, 1989).

If no efforts are made to correct these heteroscedasticity problems, this has little impact on the mean biomass or volume values estimated. By contrast, these corrections are indispensable if we are to obtain correct confidence intervals around these predictions. Two methods are often put forward to correct these heteroscedasticity problems: the first consists in weighting (for instance by the inverse of the diameter or the squared diameter), but here everything depends on the weighting function and in particular on the power applied. The second consists in taking the logarithm of the terms in the equation, but here the simulated values need to be corrected to ensure that the estimated values return to a normal distribution (Duan, 1983; Taylor, 1986). Also, it may well happen that the log transformation does not result in a linear model (Návar *et al.*, 2002; Saint-André *et al.*, 2005).

The third snag is due to the additivity of the equations. Biomass measurements, then the fitting of functions, are often made compartment by compartment. The relations are not immediately additive, and a desired property of the equations system is that the sum of the biomass predictions made compartment by compartment equals the total biomass predicted for the tree (see Kozak, 1970; Reed & Green, 1985; Návar *et al.*, 2002). Three solutions are in general suggested (Parresol, 1999):

1. total biomass is calculated by summing the compartment by compartment biomasses, and the variance of this estimation uses the variances calculated for each compartment and the covariances calculated two by two;

2. additivity is ensured by using the same regressors and the same weights for all the functions, with the parameters of the total biomass function being the sum of the parameters obtained for each compartment;

3. the models are different compartment by compartment but are fitted jointly, and additivity is obtained by constraints on the parameters.

Each method has its advantages and disadvantages. Here in this guide we will fit a model for each compartment and a model for total biomass, while checking that additivity is respected. Concrete examples (dubbed "red lines") will be used throughout this guide as illustrations. These are based on datasets obtained during studies conducted in a natural wet tropical forest in Ghana (Henry *et al.*, 2010).

1.2.3 Measuring the biomass of a tree

Biomass tables create a link between actual individual biomass measurements and biomass estimations in the field based on inventory data. Weighing trees to measure their biomass is therefore an integral part of the approach used to construct allometric equations, and a large part of this guide is therefore devoted to describing this operation. Although the general principles described in chapter 3 (tree segmentation into compartments of homogenous dry weight density, measurement of dry matter / fresh volume ratios in aliquots, and application of a simple rule of three) should yield a biomass estimation for any type of woody plant, this guide will not address all the special cases. Plants that are not trees but potentially have the stature of trees (bamboo, rattan, palms, arborescent ferns, Musaceae, *Pandanus spp.*, etc.) are considered as exceptions.

Plants that use trees as supports for their growth (epiphytes, parasitic plants, climbers, etc.) are another special case (Putz, 1983; Gerwing & Farias, 2000; Gerwing *et al.*, 2006; Gehring *et al.*, 2004; Schnitzer *et al.*, 2006, 2008). Their biomass should be dissociated from that of their host.

Finally, hollow trees, and trees whose trunks are shaped very differently from a cylinder (e.g. *Swartzia polyphylla* DC.), strangler fig, etc., are all exceptions to which biomass tables cannot be applied without specific adjustments (Nogueira *et al.*, 2006).

2

Sampling and stratification

Sampling consists in predicting the characteristics of a set from one part (a sample) of this set. Typically, we want to estimate the volume of wood in a forest, but it is impossible to measure the volume of all the trees one by one. We therefore measure the volume of a sample of trees in the forest, then extrapolate to obtain an estimate for all the trees in the entire forest (CTFT, 1989, p.252). As volume is measured only in a sample of trees, the estimated total volume we obtain suffers from a *sampling error*[1]. Sampling in the strict sense consists in:

1. selecting appropriate trees for the measured set (we tend to speak of a *sampling plan*),
2. selecting a method used to calculate total volume from the measurements (we tend to speak of an *estimator*),

in such a manner to minimize the sampling error.

In conventional sampling theory, the volumes of N trees in the stand are fixed data: the only source of variation in the estimations is the sampling, such that exhaustive sampling would always give the same estimation. Here in this guide we will adopt the so-called super-population approach that was developed in the 1970s (Cochran, 1977). This consists in considering that the volumes of the N trees that make up the stand are random variables, such that the observed stand is only one among others drawn from a super-population. This approach means that we can do away with certain approximations and draw up an optimal sampling plan (something that is often impossible in the conventional approach). It does, however, have the disadvantage of leading to incorrect solutions if the super-population model adopted does not comply with reality.

The sampling method depends on the goal. Therefore, in principle, the question must first be asked as to the purpose of the volume or biomass tables we are seeking to construct. Are they to serve to predict the characteristics of a particular tree that has known entry variables? Are they to serve to predict the characteristics of an average tree from data on entry variables? Are they to serve to predict the total volume of the stand that yielded the trees used to construct the tables, or the total volume of another stand? In both these

[1]The terms in italics are those used in sampling theory; a definition of these terms may be found, for example, in appendix 2 of Bellefontaine *et al.* (2001).

latter cases, are the tables' entry variables measured in all the trees of the stand, or again on a sample of trees? Etc. We can therefore construct a chain that runs from the studied stand to the variable we are looking to predict (Figure 2.1).

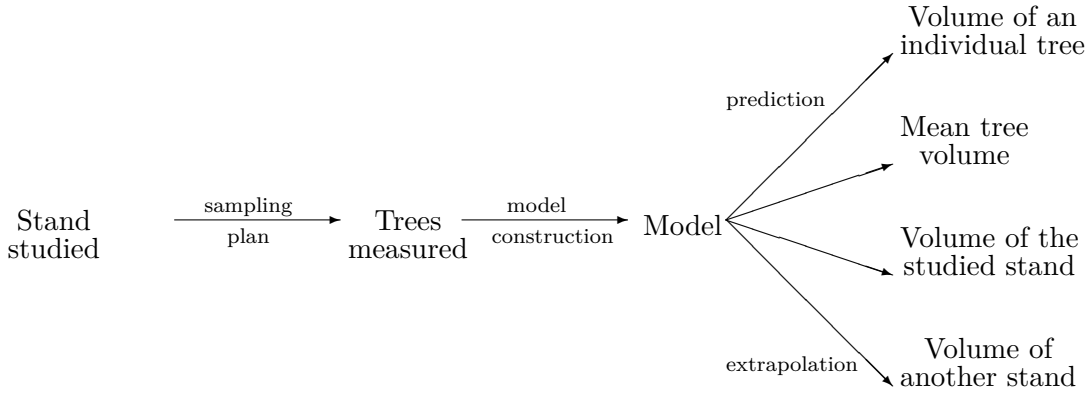


Figure 2.1 – *Chain running from the studied stand to the variables we are looking to predict.*

By following this chain backward, it can be seen that the precision of the predicted variable depends on the precision of the tables' parameters, which itself depends on the sampling plan (number and selection of trees measured) and the variability within the studied stand (Cunia, 1987b). We can therefore set a target for the precision of our predictions which, by a retroactive effect for a given table and a given sampling type, determines the minimum number of trees to be measured. We can also apply an optimization procedure to determine, for a given target precision and a given table, the sampling method that minimizes measurement time or cost (Cunia, 1987c,d). In certain cases the cost of the measurements is the limiting factor. This in particular is the case for biomass measurements of root systems. Here in this case we are not looking so much to reach a given target precision in our prediction but to remain within reasonable cost limits. We can therefore, for a given measurement cost and a given type of table, determine the sampling plan that maximizes the precision of our predictions.

This reasoning, which is often too complex to be followed in a rigorous manner, should be applied on a case-by-case basis as it depends (*i*) on what we are looking to predict, (*ii*) on the type of tables used and (*iii*) on the type of sampling adopted. The fact that we are using a table is in itself a constraint on the sampling method: the total volume of a stand can be estimated from the volume of a sample of trees, without using volume tables. By using volume tables to estimate the total volume of a stand, we are already restricting ourselves to one type of total volume *estimator*.

What is more, the reasoning used to determine the sampling plan that complies with the target precision for our predictions assumes that we know the relation between the precision of our predictions and the precision of the parameters used to construct the tables, and the relation between the precision of the parameters used to construct the tables and sample size, etc. In certain simple cases these relations are explicitly known. But in most cases, as soon as the tables take a more complex form, these relations are not explicit. In this case, the above reasoning is no longer tractable.

This reasoning is delivered a knock-out blow when we realize that: (*i*) volume/biomass tables generally have multiple purposes or unspecified purposes and (*ii*) the form of these tables is generally unknown in advance. This is due to the simple fact that in most cases we wish to use these tables for different ends: to estimate the volume of a particular tree, an average tree, an entire stand, etc. Constructing the tables is ultimately an end in itself,

without any reference to a quantity we wish to predict. Also, the form of the table is most often selected based on the results of an exploratory data analysis, and is therefore not known in advance. Obviously, some relationships such as the power function or second-order polynomials often recur, but no rule can be established *a priori*. It is therefore illusory to look to optimize a sampling plan.

Ultimately, the sampling used to construct volume/biomass tables is generally based on empirical considerations relating to the sampling plan. The choice of the *estimator*, which in fact determines which tables are chosen is made *a posteriori*, depending on the data collected, and independently of the sampling plan.

2.1 Sampling for a simple linear regression

Let us begin with a simple example that clearly illustrates the notions developed above. Let us assume that the trees in a given stand are described by their dbh D, their height H and their volume V. We use volume tables that predict volume V from the variable D^2H. The super-population model we adopt to describe the stand assumes that the relation between V and D^2H is linear, with white noise ε of variance σ^2:

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

where ε follows a normal distribution of zero expectation and standard deviation σ. Let us also assume that the quantity D^2H follows a normal distribution with a mean of μ and a standard deviation of τ. White noise ε incorporates all the factors that cause two trees of the same dbh and the same height not to have necessarily the same volume. Parameters α and β are unknown. To estimate them, we will measure n trees; and will thus obtain a sample of n doublets $(D_1^2H_1, V_1)$, ..., $(D_n^2H_n, V_n)$, then construct the following linear regression:

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

In sampling theory jargon, the entry variables for the volume/biomass tables (diameter, height...) are called *auxiliary* variables. It is important to distinguish these tree-related variables from those such as age which are stand-related. These latter variables are considered as parameters (Pardé & Bouchon, 1988, p.106). Also, the sampling *unit* is the tree. Let us now see how to draw up a sampling plan on the basis of the goal set.

2.1.1 Predicting the volume of a particular tree

Let us assume that the goal is to predict the volume of a tree in a stand of dbh D^* and height H^*. Its predicted volume is of course:

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

The super-population model stipulates that because of white noise ε, two trees selected at random and with the same dbh D and the same height H do not necessarily have the same volume. This means we are faced with intrinsic variability when we measure a particular tree, equal to σ^2. When attempting to predict volumes, this intrinsic variability is supplemented by the variability due to the imprecision of the α and β parameter estimations. We will return to these notions later (in chapter 7). Thus, for a linear regression, the semi-amplitude of the confidence interval at the threshold α (typically 5 %) of V^* is equal to (Saporta, 1990, p.374):

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

where t_{n-2} is the quantile $1-\alpha/2$ of a Student's distribution with $n-2$ degrees of freedom, $\overline{D^2H}_e$ is the empirical mean of the D^2H values measured in the sample:

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

and $\hat{\sigma}$ is an estimate of the standard deviation of the residuals:

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

The lowest value of this semi-amplitude (when $n \to \infty$) is $1.96\,\sigma$. We will now set our precision target for the estimation as a deviation of $E\,\%$ from this incompressible minimum, i.e. in an approximate manner we are looking for sample size n such that:

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

Random sampling

Let us first examine the case where we do not look to optimize the sampling plan, for example by selecting sample trees at random. The empirical mean of D^2H in the sample is therefore an estimation of μ, while the empirical variance of D^2H in the sample is an estimation of τ^2. Thus:

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

By way of a numerical example, let us take $\mu = 5$ m^3 as the mean value of D^2H in the entire stand and $\tau = 1$ m^3 as its standard deviation. If we want to predict the volume of a tree that has a D^2H of 2 m^3 with a precision deviation of $E = 5\,\%$, we will have to measure approximately $n = 98$ trees. We can see here that the expression of n in relation to $D^{*2}H^*$ is symmetrical around μ and passes through a minimum for $D^{*2}H^* = \mu$. As $\mu - 2 = 3$ m^3 and $\mu + 3 = 8$ m^3, we will also need $n = 98$ trees to predict the volume of a tree whose D^2H is 8 m^3 with a precision deviation of 5 %. The $n = 98$ sample size may also be interpreted as that which ensures a precision deviation of at most 5 % (at $\alpha = 5\,\%$) for all predictions in the 2–8 m^3 range.

Optimized sampling

Let us now examine the case where we are looking to optimize the sampling plan in relation to the value of $D^{*2}H^*$. Equation (2.3) shows that precision deviation E is smallest when $\overline{D^2H}_e = D^{*2}H^*$. It is therefore advantageous to select sample trees in such a manner that their empirical mean D^2H value is $D^{*2}H^*$. In practice, as the empirical mean of the sample's D^2H values will never be exactly $D^{*2}H^*$, it is advantageous to maximize the denominator $\sum_i (D_i^2 H_i - \overline{D^2H}_e)^2$, i.e. to maximize the empirical variance of the sample's D^2H values. Ultimately, the sampling plan that maximizes the precision of the volume prediction of trees with D^2H values of $D^{*2}H^*$ consists in selecting $n/2$ trees with D^2H values of $D^{*2}H^* - \Delta$ and $n/2$ trees with D^2H values of $D^{*2}H^* + \Delta$, with Δ being as large as possible (Figure 2.2). This sampling plan means that we can ignore the term dependent upon $D^{*2}H^*$ in

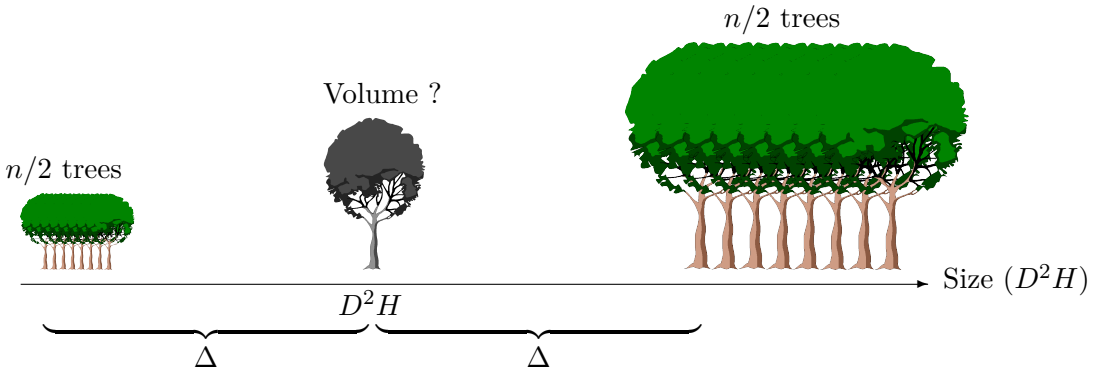


Figure 2.2 – *Sampling plan optimizing volume prediction precision for a particular tree. The size deviation* Δ *must be as large as possible.*

(2.3), thus simplifying the relationship to:

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

For $E = 5\,\%$, this gives us $n = 10$ trees. Optimizing the sampling plan has therefore "saved" us measuring 88 trees compared to the plan that consisted in selecting trees at random. However, the optimized plan is restricted to estimating the volume of a $D^{*2}H^*$sized tree. It is not optimized for estimating the volume of other sized trees. This reasoning therefore clearly has its limitations as volume tables are not generally (or even never) constructed to predict the volume of only one size of tree.

Even more seriously, the optimized sampling plan is also reliant upon the super-population model and may yield erroneous estimations if it does not correspond to reality. This is illustrated in Figure 2.3. The optimized sampling plan for a given $D^{*2}H^*$ size leads to the selection of extreme points (in black in Figure 2.3) for the sample. This situation is critical for a linear regression as two groups of points far removed one from the other will give an elevated R^2 value without us really knowing what is happening between the two. If the linear relation assumed by the super-population model is correct (Figure 2.3 left), then there is no problem: the volume predicted by the tables (shown by a star) will indeed be close to the actual value (gray point). By contrast, if the super-population model is wrong, the predicted volume will be wrong: this is illustrated in Figure 2.3 on the right (where the sample points, in black, are exactly the same as those in Figure 2.3 on the left), where the size-volume relationship is in fact parabolic, not linear. In practice, the shape of the size-volume relationship (and therefore the volume tables) is not known in advance, and it is therefore best to sample trees across the entire size variation interval so as to visualize the nature of the size-volume relationship.

2.1.2 Predicting the volume of a stand

Let us now assume that the goal is to predict the volume of an entire stand. Here, we must first assume that we measure dbh D and height H in *all* the trees in the stand. Let N be the number of trees in the stand (including the n trees in the sample). Modulo renumbering of the trees therefore gives us a measurement of tree volume V for $i = 1, \ldots, n$ and a measurement of tree size D^2H for $i = 1, \ldots, N$. The estimator for the total volume of the

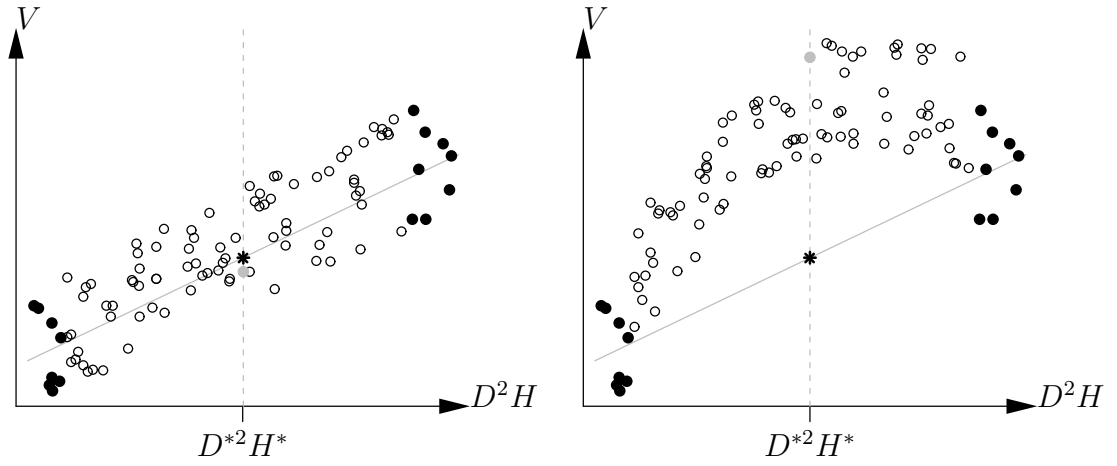


Figure 2.3 – *Volume prediction based on a linear regression constructed using extreme points (in black) when the size-volume relationship is indeed linear (left) and when it is not (right). The black points are the same in both cases. The star indicates the volume predicted by the linear regression constructed using the black points while the gray points indicate the real volume corresponding to* $D^{*2}H^*$.

stand deduced from the volume tables is therefore:

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

which may also be written: $V_{\text{tot}} = N\bar{V}$, where $\bar{V} = a + b\overline{D^2H}$ is the mean volume of the trees in the stand, and $\overline{D^2H} = (\sum_{i=1}^{N} D_i^2 H_i)/N$ is the mean dbh of the trees in the stand. To the extent that the volume tables are obtained by linear regression of $(V_1, \ldots, V_n)$ against $(D_1^2H_1, \ldots, D_n^2H_n)$, the numerical values of the coefficients a and b varify (Saporta, 1990, p.363): $\bar{V}_e = a + b\overline{D^2H}_e$, where $\bar{V}_e = (\sum_{i=1}^{n} V_i)/n$ is the mean volume of the trees in the sample and $\overline{D^2H}_e = (\sum_{i=1}^{n} D_i^2 H_i)/n$ is the mean size of the trees in the sample. By subtracting, we arrive at the following result:

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

In this equation it is important to note that $\bar{V}$ and $\overline{D^2H}$ are mean values for the entire stand while, $\bar{V}_e$ and $\overline{D^2H}_e$ are mean values for the sample. In addition $\overline{D^2H}$, $\overline{D^2H}_e$ and $\bar{V}_e$ are derived from actual measurement, while $\bar{V}$ is the quantity we are looking to estimate.

If we look at equation (2.4), we recognize it as being a type of estimator that is widely used in sampling theory: the *regression estimators*. The theoretical basis of regression estimators is described in detail in Cochran (1977, chapitre 7) and in Thompson (1992, chapitre 8). A description of regression estimators used in forestry is given in de Vries (1986) and in Shiver & Borders (1996, chapitre 6) (the first is more theoretical, the second more applied). The theory of regression estimators applies in the case of a linear relationship between the quantity to be predicted ($\bar{V}$ in the example above) and an auxiliary variable ($\overline{D^2H}$ in the example above). By contrast, this theory is less well developed in the case of non-linear relations or multiple regressions, even though such cases often arise in connection with volume tables.

The semi-amplitude of the confidence interval at α (typically 5 %) of $\bar{V}$ is (Cochran,

1977, p.199; Thompson, 1992, p.83):

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

It can be seen that the minimum of this amplitude is zero, that is reached when entire stand is included in the sample ($n = N$, that means that $\overline{D^2H} = \overline{D^2H}_e$). As before, the optimal sampling plan is such that $\overline{D^2H}_e$ is the closest possible to $\overline{D^2H}$, with a maximal empirical variance of D^2H in the sample.

When deriving the regression estimator we assumed that size D^2H is measured for *all* the trees in the stand in order to arrive at an estimation of total volume V_{tot}. In practice, a more realistic measurements protocol is as follows: the size of the trees in a sample of size $n' < N$ is measured; both size and volume are measured in a subsample of $n < n'$ of this sample. The volume vs dbh regression (i.e. the volume tables) is constructed using subsample values; this is then used to estimate the volume of the sample then, by extrapolation, that of the entire stand. This sampling strategy is called *double sampling*. Its theory is described in Cochran (1977, section 12.6) and, more pragmatically, in Shiver & Borders (1996, chapitre 7). Its use for estimating stand biomass was developed by Cunia (1987b,c,d).

Finally, the properties of the regression estimator (2.4) are well known in conventional sampling theory which does not require the hypothesis of a linear model (2.1) but considers that the only source of variability is the sampling. In the case of a simple random sampling plan and a sufficiently large sample size n, the variance of $\bar{V}$ in conventional theory is approximately (Cochran, 1977, p.195; Shiver & Borders, 1996, p.181):

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

and the semi-amplitude confidence interval at α (typiquement 5 %) of $\bar{V}$ is approximately (Thompson, 1992, p.80; Shiver & Borders, 1996, p.185):

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

This last expression is considered as being more robust than expression (2.5) when the super-population model (2.1) strays from reality (Thompson, 1992, p.84).

In conclusion, this simple example shows both the advantages and limitations of sampling for the construction of volume tables: advantages because sampling theory provides us with the minimum number of trees that need to be measured to reach a given precision in our predictions, and thereby optimize the sampling plan; limitations because this reasoning assumes that the form of the volume tables (and that the underlying super-population model) is known in advance and that the tables will be used for a given application. Neither of these prerequisites is met in practice. Also, the calculations that are relatively simple in the case of the linear model we have just described rapidly become hopelessly complex for more realistic models.

2.2 Sampling to construct volume tables

As a first step, let us consider the problem of predicting the volume or biomass of a particular tree using volume or biomass tables. How many trees need to be measured to construct

the tables (§ 2.2.1)? Which trees in the stand should be selected for measurement? This second question begs another: how should we sort the trees in the sample in relation to volume table entry variables, starting with the tree size (§ 2.2.2)? How, if necessary, should we stratify the sample (§ 2.2.3)? Is it best to select individuals far apart in the forest, or on the contrary inventory all the trees in a given plot (§ 2.2.4)?

2.2.1 Number of trees

Given the limitations of the sampling theory, the number of measured trees for volume or biomass (in other terms, the sample size) is generally selected empirically, based on rules established by experience. A general principle is that, for any given precision, the more variable the material, the larger the sample size: smaller sample sizes are required for a plantation of clones than for a natural tropical forest, for a given species than for a group of species, for a 10 ha plot than for a natural area. In certain cases, such as for root biomass, it is the cost of the measurements that determines the sample size rather than the desired precision of the predictions: the number of trees selected corresponds to an acceptable amount of work to obtain the measurements. As a rough guide, when constructing volume tables, the Mémento du forestier (CTFT, 1989, p.256) recommends measuring about 100 trees "for one or more stands of recent plantation over a small surface area (e.g. silviculture research plots)". Pardé & Bouchon (1988, p.108) recommended the sample sizes given in Table 2.1, that depend on the surface area of the zone in which the volume tables are to be used. Different volume and biomass tables have been compiled by Zianis *et al.* (2005) for Europe and by Henry *et al.* (2011) for sub-Saharan Africa. The sample sizes reported for the volume and biomass tables listed in these literature reviews give some idea of the sampling effort required. Chave *et al.* (2004) showed that when 300 trees were used to construct a biomass table, the resulting biomass estimation for a wet tropical stand (Barro Colorado Island in Panama) had a coefficient of variation of barely 3.1 %. This coefficient of variation rose above 10 % when the number of trees used to construct the biomass table fell below 50, with the coefficient of variation increasing approximately in proportion to $1/\sqrt{n}$ (Chave *et al.*, 2004, Figure 3). Van Breugel *et al.* (2011) noted the same type of decrease in estimation precision with the sample size used to construct the biomass table, for n between 49 and 195 trees.

Table 2.1 – *Number of trees to be measured to establish a volume table for different surface areas over which the tables are to be used: recommendations by Pardé & Bouchon (1988).*

Area	n
Single, homogeneous stand	30
15-ha plot	100
1000-ha forest	400
Natural area	800
Species area	2000 to 3000

The costlier an observation, in terms of measurement time and effort, the more the sampling plan tends to be determined by the sampling effort that workers are willing to make rather than by the desired precision of the estimation. As the above-ground biomass of a tree is more difficult to measure than the volume of its stem, biomass tables tend to be constructed using fewer observations than volume tables. Certain biomass tables are

constructed from measurements in only a few trees (8 trees for Brown *et al.*, 1995 in Brazil, 12 trees for Ebuy Alipade *et al.*, 2011 in the Democratic Republic of Congo, 14 trees for Deans *et al.*, 1996, 15 trees for Russell, 1983 in Brazil). Root tables, that require an even greater measurement effort, are often based on even smaller sample sizes. Tables constructed using such small samples are generally unreliable, and in any case are valid only very locally. However, these small datasets can subsequently be grouped together into more impressive datasets that have advantages for model fitting (on condition that the variability induced by the merging of these datasets can be controlled by effect covariables: age, wood density, etc. or by stratification factors: species, type of vegetal formation...).

2.2.2 Sorting trees

The sorting of trees in the sample on the basis of their size (and more generally on the basis of table entry variables) can in principle be optimized. For example, in the case of a linear regression, the semi-amplitude of the confidence interval at α of the regression slope is (Saporta, 1990, p.367):

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

where t_{n-2} is the quantile $1-\alpha/2$ of a Student's distribution with $n-2$ degrees of freedom, $\hat{\sigma}$ is the empirical standard deviation of the model's residuals, n is sample size and S_X is the empirical standard deviation of the entry variable X in the sample:

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

Therefore, the higher the value of S_X the more precise the estimation of the slope, which, for a set sample size, brings us back to a trees breakdown similar to that shown in Figure 2.2. We have already seen the limitations of this reasoning: although the sampling plan that consists in selecting trees at the two extremities of the size gradient is optimal when the conditions of a linear relationship are met, it results in erroneous estimations when the relationship is not linear (Figure 2.3). It is therefore preferable, in practice, to sample trees right across the size gradient in order to check the form of the relationship between their volume (or their mass) and their size.

The theory of response surfaces (Box & Draper, 1987; Goupy, 1999; Myers & Montgomery, 2002) can be used to optimize tree sorting on the basis of dbh (and more generally on the basis of table entry variables). We will not enter into the details of this theory here, but will simply sketch a few general principles. The first principle is to spread the tree size gradient of the sample as broadly as possible.

If the volume (or mass) variance is constant for all tree sizes, the rule is to measure the same number of trees in each size category (Pardé & Bouchon, 1988, p.108; CTFT, 1989, p.256). It would be a mistake to form a sample by selecting in each size class a number of trees proportional to the magnitude of that class in the stand (in other words select the trees at random). But volume variance is rarely constant; generally it increases with tree size (heteroscedasticity of the residuals). The rule is therefore to increase sampling intensity in the most variable classes, thereby improving precision. In theory, the ideal approach is to measure a number of trees in a given size class proportional to the standard deviation of the tree volume in this class (CTFT, 1989, p.256). In practice, when the entry variable is dbh, an empirical rule consists in selecting the same number of trees per basal area growth class and thus better represent large diameter trees (CTFT, 1989, p.256–257).

This reasoning may be extended to other effect variables. If the table entry variable is D^2H, the trees may be sorted by D^2H. For multi-specific biomass tables, wood density ρ is often used as the entry variable (with the specificity that it concerns the species not the tree). For multi-specific tables using diameter D and specific wood density ρ as entry variable, adequate sorting of the trees in the sample consists in dividing them evenly by dbh class and wood density class.

2.2.3 Stratification

We have already seen for the sorting of trees into size classes that selecting trees at random, and thereby giving all trees and equal *inclusion probability* is a suboptimal sampling plan. Stratification aims to take account of exogenous information to establish homogeneous sampling *strata* and thus improve the precision of our estimations. The principle, in the same manner as previously, is to increase the sampling intensity of the most variable strata (relative to the other strata). Again using the example given in section 2.1, the variance of the regression estimator $\bar{V}$ in the case of a stratified sampling approach becomes (Cochran, 1977, p.202):

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

where h is the stratum, N_h the number of individuals in the stand that belong to stratum h, n_h is the number of individuals in the sample that belong to stratum h, V_{ih} is the volume of the ith individual in stratum h in the sample, $\bar{V}_{eh}$ is the empirical mean volume in stratum h of the sample, etc. This formula replaces (2.6). Let us now illustrate the precision gained by the stratification using a simple numerical example. Let us suppose, to simplify, that we have two strata, each corresponding to 50 % of the stand (such that $N_1/N = N_2/N = 0.5$), and that the sampling in each stratum is such that the second term between brackets in (2.7) is negligible. Let us also assume that $n_1 \ll N_1$ and $n_2 \ll N_2$. The variance of the regression estimator in this case is approximately proportional to:

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

The terms in brackets represent within-strata variances in volume. Let us suppose that the standard deviation of the volume is 4 m^3 in the first stratum and 2 m^3 in the second. Total sample size is set at $n_1 + n_2 = 60$ individuals. If we take account of the stratification, i.e. if we select the number of trees in each stratum in proportion to the frequency N_h/N of the stratum in the stand, then in this case we have the same number of trees in each stratum of the sample: $n_1 = n_2 = 30$ individuals. The variance of the regression estimator is therefore approximately:

$$\frac{4^2}{30-2} + \frac{2^2}{30-2} = 0.71 \ \mathrm{m}^6$$

By contrast, if we set the number of trees in each stratum in proportion to the standard deviation of the volume in the stratum, then: $n_1 = 2n_2$, giving $n_1 = 40$ individuals and $n_2 = 20$ individuals. The variance of the regression estimator is therefore approximately:

$$\frac{4^2}{40-2} + \frac{2^2}{20-2} = 0.64 \ \mathrm{m}^6$$

It can thus be seen that from the standpoint of estimator variance, $30 + 30$ does not equal $40 + 20$. We can also check that the minimum of the function that at n_1 associates $16/(n_1 - 2) + 4/(58 - n_1)$ is obtained for $n_1 = 39.333$.

From the standpoint of sampling theory, stratification aims to increase the precision of the estimation by fitting the sampling plan to the variability in each stratum. But from the standpoint of constructing volume tables, stratification has a second aim that is just as important as the first: check that the relationship between tree volume (or biomass) and size is the same in each stratum and, if needed, break down the tables into as many relations as necessary. This second point is implicit in equation (2.7) that is based on the fitting of a different slope b (see equation 2.2) in each stratum.

In short, stratification aims to explore the variability of the study area to (*i*) if necessary vary the form of the tables in accordance with the strata, and (*ii*) fit the sampling plan to the variability in the strata. Often when constructing volume tables, point (*i*) is predominant over point (*ii*), whereas the reverse is true in sampling theory. Figure 2.4 illustrated these two aims.

Stratification factors

Any factor that can explain the variability in the study area may be envisaged: stand age (above all for plantations), fertility, station, silvicultural treatments, variety or species, altitude, depth of the water table, etc. (Pardé & Bouchon, 1988, p.106; CTFT, 1989, p.255). Stratification factors may be nested: stratification according to morpho-pedological region, then according to fertility in each region, then according to age in each fertility class, then according to density in each age class. The "finenesse" of the stratification factors must also be adapted to match the context. The stratification factors employed are not the same when reasoning on a global scale like Brown (1997), on a landscape scale like van Breugel *et al.* (2011), or on the scale of a plantation of clones like Saint-André *et al.* (2005). Brown (1997) proposes "all-species" tables for climatic areas (dry forests, wet forests). At the other extreme, on an "eddy-correlation" plot, and with the aim of comparing NEP estimations, trees may be stratified according to plot age, season and flux tower "foot-print".

Species as stratification factor

When dealing with natural formations that contain several species, the species may also be considered as stratification factor. Here, it is routine practice to construct volume tables separately for each species (or at least the most abundant species), then attempt to group gender or group together all the species ("all species" table). By merging the datasets we increase sample size which can be advantageous if this compensates for the increase in variability resulting from the mix of different species. In comparison with a mono-specific model, a multi-specific model introduces a prediction bias that may be viewed as between-species variability. As an illustration, Van Breugel *et al.* (2011) quantified the prediction bias resulting from the aggregation of several species. Therefore, merging the data from several species is advantageous if the gain in within-species variability provided by this merger compensates for the between-species variability introduced. However, a check must be made that (*i*) this merger is meaningful and that (*ii*) the sample sizes for the different species are comparable (Figure 2.4). If from the outset we are looking to construct "all species" volume table (which is often the case for natural stands), care must be taken to ensure that the individuals are selected for the sample independently of their species, such that the tables are not biased in favor of one particular species.

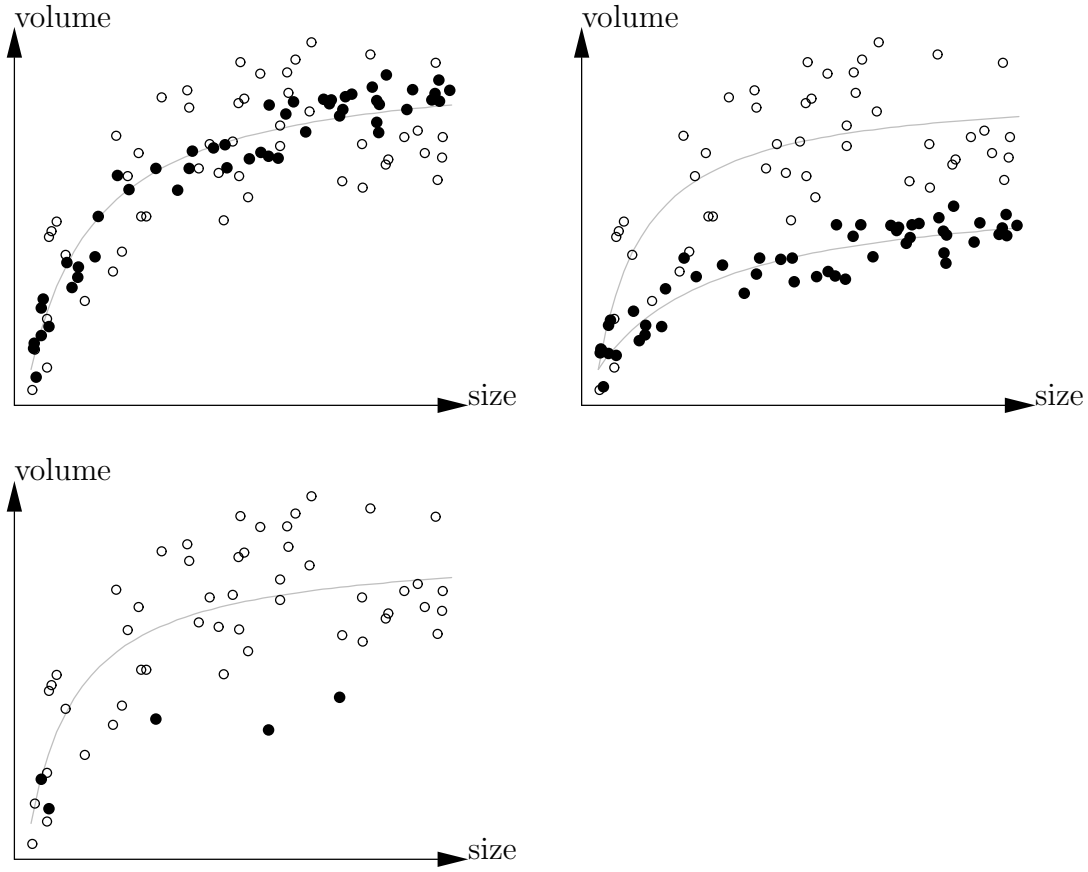


Figure 2.4 – *Predicting volume from tree size in two strata (black and white points): top left: the two strata corresponding to the two variances of the residuals (higher variance for the white than the black points) but the relationship is the same; top right: both the variance and the relation are different in the two strata; bottom: the situation is the same as top right but the second stratum has been subsampled such that we are led to believe that we are dealing with the same relation in the two strata.*

Allocating to strata

Once the strata have been identified, the sampling plan must be adapted to it following empirical rules. If we have an *a priori* estimation of volume variability (for volume tables) or biomass variability (for biomass tables) for each stratum, the empirical rule is to select a sampling intensity in proportion to the standard deviation in each stratum. If no *a priori* estimation of the variability is available, then a constant sampling intensity should be used in each stratum (this does not correspond to random sampling if the strata do not have the same frequency in the stand).

Parameterized models

Strata-related information is then incorporated into the volume model by establishing different models for each stratum. We can then test whether the models for the two strata are significantly different and, if appropriate, merge the two datasets to construct a single model. We could also construct a parameterized model from several models for different strata by following the principle of the mixed model: the model's parameters themselves become functions of the variables defining the strata. These different points will be developed in the sections below, devoted to the actual construction of volume/biomass models. By way of an example, Ketterings *et al.* (2001) developed individual, single-entry biomass models of the power form:

$$B = aD^b$$

where D is the dbh and B biomass of different trees of different species at different sites in the Jambi province of Sumatra in Indonesia. The site factor was then taken into account in parameter b that was written $b = 2 + c$, where c is the parameter in the allometric equation that links height to dbh at each site: $H = kD^c$. The species factor was taken into account in parameter a that was written $a = r\rho$, where ρ is the wood density of the species and r is a constant parameter. The model finally obtained, and valid for all the species at all the sites, is a parameterized model:

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

2.2.4 Selecting trees

Once the composition of the sample has been established, the trees must be identified and measured in the field. Given that these measurements involve a great deal of work and, when involving biomass, are destructive, the trees used must be selected with great care. A strategy adopted by some when constructing biomass tables is to cut all the trees within a given area (for example half a hectare). This has the advantage of "killing two birds with one stone" in that it provides both a biomass estimate for the stand and individual observations for the construction of a model. From a practical standpoint, the space created by the cutting of the first trees facilitates the subsequent cutting of the next. But this strategy has a major drawback: as the tree size distribution of the trees in the stand has only a very small likelihood of coinciding with the desired sorting of the trees in the sample by size class, the tree size distribution in the sample is suboptimal. The same applies for any factor in the structure of the sample (wood density classes, strata, etc.). Also, the scale of this disturbance in the stand may have unexpected consequences. For instance, Djomo *et al.* (2010) reported on a plot that was invaded by ants after the trees had been felled, to such an extent that it was impossible to make any tree biomass measurements. This tree selection strategy should therefore be avoided in areas infested by *Wasmannia* ants as their attacks are highly dangerous.

Rather than selecting all the trees within a given area, preference should be given to selecting stems one by one depending on the requirements of sample constitution. This strategy may be longer to implement as it requires the identification of individual trees. In view of the difficulties inherent to measuring tree biomass (see chapter 3), preference should be given to trees that match the criteria of the sampling plan and are most easily accessible.

2.3 Sampling for stand estimations

Let us now consider the problem of predicting the volume or biomass of a stand. In a rigorously statistical manner, we need to consider the entire error propagation chain as described in Figure 2.1 (Parresol, 1999). This raises questions of double sampling and regression estimator, as described in section 2.1.2. Cunia (1987b,c,d), Chave *et al.* (2004) and van Breugel *et al.* (2011) are rare examples where the entire error propagation chain was indeed taken into account, and where the biomass estimation error for a stand was related to the sample size used to construct the biomass table necessary for this estimation. In practice the problem is generally simplified by considering the table as exact and not suffering from any prediction error. This approximation — which disconnects the sampling of the stand used to predict its volume or biomass, from the sampling of the trees used to construct the model — reduces the sampling to a conventional forest inventory problem.

We will not dwell here on this question of forest inventory, first because it is not central to this guide's purpose, and second because entire books have already been devoted to the subject (Loetsch & Haller, 1973; Lanly, 1981; de Vries, 1986; Schreuder *et al.*, 1993; Shiver & Borders, 1996; West, 2009). We will nevertheless describe a few developments relating to the estimation of stand biomass.

2.3.1 Sampling unit

Whereas it is entirely suitable, when constructing a biomass model, to select trees for the sample in an individual manner, this sampling strategy is unrealistic when estimating the biomass of a stand. We should rather in this case opt for a strategy consisting in measuring all the trees within a given area, and if necessary repeating this approach on another area to increase sample size. This area, or plot, thus becomes the sampling unit. Let n be the number of plots inventoried, N_i be the number of trees found in the ith plot ($i = 1, \ldots, n$), and B_{ij} be the biomass of the jth tree in the ith plot ($j = 1, \ldots, N_i$), calculated using the biomass model and the measured characteristics of the tree. The number N_i is random, but for a given tree the prediction of B_{ij} is considered as deterministic. The biomass of the ith plot is therefore: $B_i = \sum_{j=1}^{N_i} B_{ij}$.

Let A be the area of a sampling plot and $\mathcal{A}$ be the area of the stand. In the super-population model the biomass of the stand is therefore estimated by: $(\mathcal{A}/A)\,\bar{B}$, where $\bar{B} = (\sum_{i=1}^{n} B_i)/n$ is the mean biomass of a plot. It is generally considered that A and $\mathcal{A}$ are accurately known parameters. The stand biomass estimation error therefore stems from that of mean biomass $\bar{B}$.

2.3.2 Relation between coefficient of variation and plot size

According to the central limit theorem, the confidence interval at the threshold α for the expectation of plot biomass is approximately (this expression is accurate when the biomass follows a normal distribution, or when the number of plots approaches infinity) (Saporta,

1990, p.304):

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

where t_{n-1} is the quantile $1-\alpha/2$ of a Student distribution with $n-1$ degrees of freedom, and S_B is the empirical standard deviation of plot biomass:

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

By definition, the precision of the estimation E at the threshold α is the ratio of the semi-amplitude of the confidence interval at the threshold α to the mean biomass:

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

where $\mathrm{CV}_B = S_B/\bar{B}$ is the biomass coefficient of variation. By rounding t_{n-1} to 2, the sample size n required to reach a given estimation precision E is therefore:

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

The biomass coefficient of variation for a plot of area A is therefore the essential factor when constructing the sampling plan. Also, as area A of the plots is not in principle known, the relation between the biomass coefficient of variation and the area A of the plots must in fact be determined.

The exact derivation of this relation between A and CV_B needs a model capable of describing the spatial distribution of the trees. The theory of point processes meets this need (Cressie, 1993; Stoyan & Stoyan, 1994). It is feasible, in a point process, to calculate exactly the relationship between A and CV_B but this is rather complex (Picard *et al.*, 2004; Picard & Bar-Hen, 2007). The exact calculation brings two things to our notice:

1. although plot *shape* has an effect on the coefficient of variation (as already demonstrated empirically, see Johnson & Hixon, 1952; Bormann, 1953), it has a negligible effect compared to plot size;

2. the relation between A and CV_B can be approached by a power relation (Fairfield Smith, 1938; Picard & Favier, 2011):

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

In practice it is this power relation that is most often specified. Intuitively, the value $c = 0.5$ corresponds to a random spatial biomass distribution within the stand; a value of $0 < c < 0.5$ corresponds to an clustered spatial biomass distribution; and a value $c > 0.5$ corresponds to a regular spatial biomass distribution (CTFT, 1989, p.284). Using biomass data for a large plot in Paracou, French Guiana, Wagner *et al.* (2010) found:

$$\mathrm{CV}_B = 557 \times A^{-0.430} \qquad (A \text{ in m}^2,\ \mathrm{CV}_B \text{ in \%})$$

If compared to the above definition, this corresponds to a slightly aggregated spatial biomass distribution. In the Brazilian Amazon, Keller *et al.* (2001) found the following relationship (fitted to the data in their Figure 4 with $R^2 = 0.993$ on log-transformed data):

$$\mathrm{CV}_B = 706 \times A^{-0.350} \qquad (A \text{ in m}^2,\ \mathrm{CV}_B \text{ in \%})$$

Table 2.2 – *Coefficient of variation in relation to plot size: data drawn from Table 5 by Chave et al. (2003) for Barro Colorado Island, 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

The lowest (absolute) value for the exponent is indicative of a more aggregated distribution than in French Guiana. A similar study was conducted by Chave *et al.* (2003) using data from a 50 ha plot on Barro Colorado Island, Panama. Chave *et al.* (2003) in their Table 5 reported 95 % interval amplitude values not for the biomass expectation of a plot, but for biomass expectation by unit area. The amplitude of the 95 % confidence interval for the biomass expectation of a plot therefore corresponds to the amplitude reported by Chave *et al.* (2003) times plot area, i.e.:

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

where Δ is the amplitude of the 95 % confidence interval reported by Chave *et al.* (2003) in their Table 5. We can deduce from this:

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

where μ is mean biomass per unit area, and reported as 274 $Mg\,ha^{-1}$ in the study by Chave *et al.* (2003). Table 2.2 supplements Table 5 by Chave *et al.* (2003) with the calculated value of CV_B. The CV_B values given in Table 2.2 closely fit ($R^2 = 0.998$ on log-transformed data) the following power relation with plot size:

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

Biomass variability (expressed by the value of the multiplier $k = 942$) was greater than at Paracou, but the spatial structure of the biomass (expressed by the exponent $c = 0.45$) fairly similar to that observed in Paracou by Wagner *et al.* (2010). Also, the fact that c approached 0.5 showed that the biomass was only slightly spatially aggregated. Chave *et al.* (2003) underlined that there was no significant spatial autocorrelation of the biomass (which would correspond to $c = 0.5$, or to a constant value of Δ).

2.3.3 Selecting plot size

The size of sampling plots may be selected in such a way to optimize the precision of the estimation for a given sampling effort (Bormann, 1953; Schreuder *et al.*, 1987; Hebert *et al.*, 1988), or in a manner to minimize sampling effort for a given estimation precision (Zeide, 1980; Gambill *et al.*, 1985; Cunia, 1987c,d). These two points of view are dual one to the

other and lead to the same optimum. The sampling effort may be quantified simply by the sampling rate $n \times A/\mathcal{A}$ or, more realistically, by a cost whose expression is more complex. Let us examine these two options.

Fixed sampling rate

At a constant sampling rate, the area A and the number n of sampling plots are linked by an inversely proportional relation: $n \propto 1/A$. Choosing appropriate plot size boils down to the following question: "is it better to select a few large plots or many small plots?", also called the SLOSS compromise (for "single large or several small"; Lahti & Ranta, 1985). If we calculate the relation $n \propto 1/A$ in (2.8) (and considering that $t_{n-1}/\sqrt{n-1}$ is little different from $2/\sqrt{n}$):

$$E \propto 2\ \mathrm{CV}_B\ \sqrt{A}$$

If the spatial distribution of the biomass is random, $\mathrm{CV}_B \propto A^{-0,5}$ and therefore the precision of the estimation E is independent of the area A of the plots. If the spatial distribution of the biomass is aggregated, $\mathrm{CV}_B \propto A^{-c}$ with $c < 0.5$ and therefore $E \propto A^{0,5-c}$ with $0.5 - c > 0$: then the smaller the plot area A, the greater the precision of the estimation (low value of E). In this case, and with a fixed sampling rate, it is therefore better to select several small plots than a few large plots. This is demonstrated by Table 2.2, where we can see that the value of Δ decreases when A decreases (this decrease is only slight as c is close to 0.5). If the spatial distribution of the biomass is regular, $\mathrm{CV}_B \propto A^{-c}$ with $c > 0.5$ and therefore $E \propto A^{0,5-c}$ with $0.5 - c < 0$: the larger the plot area A the better the precision of the estimation (low value of E). In this case, and with a fixed sampling rate, it is therefore better to select a few large plots than many small plots.

The parameters measured in biology in most cases have an aggregated spatial distribution ($c < 0.5$), sometimes random distribution ($c = 0.5$), and only rarely a regular distribution ($c > 0.5$) (Fairfield Smith, 1938). In other words, the SLOSS compromise is in most cases resolved to the benefit of a multitude of small plots. If we push this reasoning to its limit, we see that the precision of the estimation would be optimal (E minimal) for $A = 0$, i.e. by selecting an infinite number of plots of zero size! This illustrates the limits of this reasoning. When we quantify the sampling effort by the sampling rate $nA/\mathcal{A}$, we assume implicitly that the sampling cost, i.e. the time or money necessary for the sampling, is proportional to nA. This means that we consider only an area-based cost, i.e. a sampling cost that is proportional to the inventoried area.

Sampling cost

In actual fact, the area-based cost is only one component of the sampling cost. Strictly speaking, the inventory made of the sampling plots, and whose duration is indeed proportional to the inventoried area, is not the only task that takes time. Staking out the boundaries of the sampling plots also takes time. And this time is proportional to the cumulated length of their boundaries: this is therefore a linear cost. Moving from one plot to another also takes time. It is therefore more realistic to measure sampling effort by a cost that includes all these tasks rather than simply by the sampling rate. If we measure this cost in terms of time, and if the sampling plots are square, the sampling cost would for example be (Zeide, 1980; Gambill *et al.*, 1985):

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

where α is inventory time by unit area, β is staking out time by unit length ($4\sqrt{A}$ corresponds to the boundaries of a square plot of area A), γ is travel speed, and $d(n, A)$ is the length

of the path linking the n sampling plots. This sampling cost expression can be further supplemented to take account of other tasks. The reasoning followed in the previous section consisted of $\beta = \gamma = 0$. If $\beta > 0$ and $\gamma > 0$, the solution to the SLOSS compromise is now no longer $A = 0$ in the case of aggregated spatial biomass distribution ($c < 0.5$).

Other constraints

However, it would be too restrictive to limit the question of sampling the biomass of a stand to a question of estimation precision. Often, the question is not limited to estimating a stand's biomass since a number of goals are pursued simultaneously. For instance, we may also be interesting in determining biomass variations over time. Mortality processes must also be considered in this case, and the inventoried area may therefore be far larger (Chave *et al.*, 2003, 2004; Rutishauser *et al.*, 2010; Wagner *et al.*, 2010). Or we may be looking to estimate plot biomass in order to determine a relation with satellite image indices and thereby extrapolate the biomass estimation to a landscape scale. In this case the area of the sampling plots is also constrained by the resolution of the satellite images and by the smallest area necessary to calculate satellite remote sensing indices.

Additionally, the type of sampling plan implicitly considered here, i.e. a simple random plan using plots of a fixed size, is rarely the most efficient (*i.e.* with the best estimation precision for a given sampling cost). On the scale of a landscape consisting of different forest types, other sampling strategies may turn out to be more efficient (Whraton & Cunia, 1987; van Breugel *et al.*, 2011). Compared to a simple random plan with a single plot size, a stratified sampling plan will be more costly (as stratification comes at a cost) but would provide a more precise estimation. A cluster sampling plan is less costly (less traveling) but provides a less precise estimation. Specific forest inventory techniques such as distance sampling (Magnussen *et al.*, 2008a,b; Picard & Bar-Hen, 2007; Picard *et al.*, 2005) or using Bitterlich's relascope (Schreuder *et al.*, 1993; West, 2009), share the feature of being based on plots of variable sizes, and can also be more efficient alternatives to fixed-size plot approaches.

3
In the field

The field step is the most crucial as it may generate measurement errors that cannot be corrected. This phase must be governed by three key principles: (*i*) it is better to weigh everything in the field rather than calculate a volume and multiply it later by a density measurement (see chapter 1 and variations in stem shape and wood density); (*ii*) if wood aliquots are taken, it is best to weigh the whole then the aliquot to follow moisture loss; finally (*iii*) as a biomass campaign requires a great deal of work and is very costly to undertake, other measurements may be taken at the same time to avoid a subsequent return to the field (e.g. stem profile, sampling for nutrient content).

Selecting trees to be measured in the field (see chapter 2), whether as individuals or exhaustively over a given area, means that the trees must be marked with paint, measured for circumference if possible at breast height (by circling the trunk with paint at this point) and measured for height. This procedure is used to check that the tree selected indeed corresponds to the sampling plan chosen (case of a selection by individual) and provides for control measures once the tree has been felled. It is also very practical to take a photo of the individual selected and draw a schematic diagram of the tree on the field form. This makes it easier to interpret the data and check the results obtained. In general, trees that are too unusual (crown broken, stem knotty or sinuous) should not be selected unless these stems account for a significant proportion of the stand, or if the aim is to quantify an accident (e.g. crown breaking subsequent to frost). Likewise, trees located in an unrepresentative environment should be excluded (forest edge, clearing, degraded forest, etc.) as their architecture often differs from the other trees in the stand. Finally, field constraints (slope, access, stand non compliant with the stratum, etc.) may well call the original sampling into question.

The general basis for measuring biomass, and even more nutrient content, is the rule of three between fresh biomass measured in the field, fresh biomass of the aliquot, and oven-dried biomass of the aliquot. As the different organs of a tree do not have the same moisture content or the same density, it is preferable to proceed by compartmentalization to take account of variations in tree density and moisture content (and nutrient concentrations when calculating nutrient content). The finer the stratification, the more precise the biomass estimation, but this requires a great deal more work. A compromise must therefore be found between measurement precision and work rate in the field. Conventionally, a tree is divided

into the following compartments: the trunk, distinguishing the wood from the bark, and that should be sawed into logs to determine variations in density and moisture content in relation to section diameter; branches that are generally sampled by diameter class, distinguishing or not the wood from the bark; the smallest twigs, generally including buds; leaves; fruits; flowers; and finally roots by diameter class. An example of a such a compartmentalization is given in Figure 3.1 for beech.

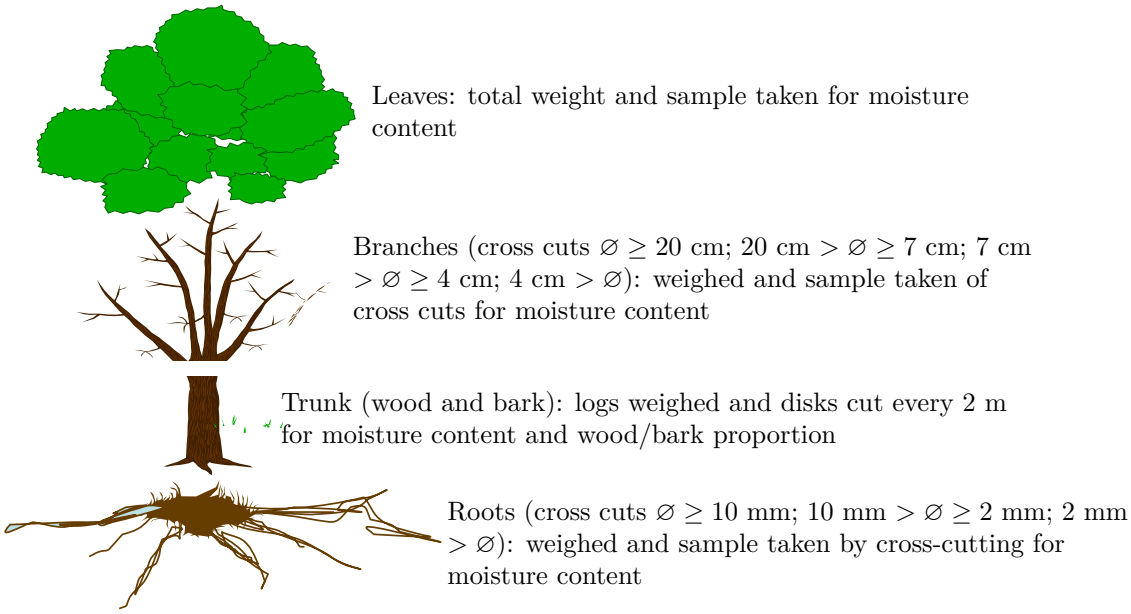


Figure 3.1 – *Example of tree compartmentalization for a biomass and nutrient content campaign on beech in France.*

Full advantage may be made of existing felling operations to sample the trees necessary to establish the table, particularly given that access and the taking of samples is often regulated in forests, and silviculture provides one of the only means to access the trees sought. However, this method may induce bias in tree selection as mainly commercial species are felled. Other species are felled only if they hinder the removal of a tree selected by the timber company, or if they are located on conveyance tracks or in storage areas. Also, trees felled for commercial purposes cannot necessarily be cross-cut into logs of a reasonable size to be weighed in the field. This depends on the capacity of the weighing machines available and log length. These constraints mean that individuals need to be carefully selected and two methods combined: (1) all the non-commercial sections of the tree must be weighed, particularly the branches; (2) the volume and wood density of the trunk must be determined.

Therefore, in the field, there is no standard method as operators must adapt to each situation. But, here in this guide, we will nevertheless describe three typical cases that can be employed subsequently as the basis for conducting any type of field campaign. The first concerns regular forest (derived from regeneration or planted), the second dry forest and the third wet tropical forest. In the first case, all the compartments are weighed directly in the field. In the second, as the trees cannot be felled, the measurements made are semi-destructive. The third case concerns trees that are too large to be weighed in their entirety in the field. Measurements are therefore obtained in three phases: in the field, in the laboratory, then by computation. As the field measurements and computations are specific

to each method, they are presented for each individual case. The laboratory procedures are generally the same in all cases.

3.1 Weighing all compartments directly in the field

The first case we will tackle is the most common and involves directly weighing all the compartments in the field. The operating procedure suggested is the fruit of several field campaigns conducted both in temperate and tropical climates. We will illustrate the case by examples taken from a variety of regular stands: plantations of eucalyptus in Congo (Saint-André *et al.*, 2005), rubber tree in Thailand, and high forests of beech and oak in France (Genet *et al.*, 2011). An example of this methodology, with additional measurements on the tapering of large branches, and sampling for nutrient content, is given by Rivoire *et al.* (2009).

3.1.1 In the field

Logging is a complex business and must be smoothly organized such that all the various teams are always working, with no slack periods (see 3.6 for details on these teams). It must be prepared in advance by the logging manager who preselects the trees and marks them in the forest. This is followed by laboratory work to (*i*) prepare the necessary materials (see 3.5 for details), (*ii*) prepare data record forms (weights of the different compartments, connected measurements), (*iii*) prepare bags to receive the aliquots taken from the trees (see 3.1), (*iv*) explain to the different operators how logging is organized so that all know what to do in the field. Figure 3.2 illustrates an effective organization for a biomass campaign, with seven operations being undertaken at the same time.

Given that limbing time is the longest, it may be useful to start with a large tree (photo 3.3). The site manager accompanies the loggers and places the sample bags at the base of the tree (operation 1). Bag size must be suitable for the size of the sample to be taken. The bags must systematically be labeled with information on compartment, tree and plot. After felling, the first team to intervene is that which measures stem profile (operation 2). Once this team has finished, it moves on to the second tree that the loggers have felled in the meantime while the limbing teams start work on the first tree (operation 3). A 12-tonne tree (about 90–100 cm in diameter) takes about half a day to process in this way. By the time limbing has been completed on the first tree, the loggers have had time to fell sufficient trees in advance to keep the profiles team busy for the rest of the day. The loggers can now return to the first tree for cross-cutting into logs, and to cut disks (operation 4). Once the cross-cutting and disk cutting has been completed on the first tree, the loggers can move on to the second tree that in the meantime has been limbed. The leaves, logs and branches from the first tree can now be weighed (operation 5) while the site manager takes samples of leaves and branches (operation 6). All the samples, including the disks, must be carried to the sample weighing area (operation 7). Once the stem profiling team has finished all the trees for the day, it can come to this area to complete the weighing operations.

This chronological process is valid under temperate climatic conditions. In the tropics, it is impossible to wait for the end of the day before weighing the samples. The samples taken must therefore be weighed at the same time as the logs and brushwood. If the samples cannot be weighed *in situ*, they must be weighed in the laboratory, but in this case they must be shipped in a tightly sealed box to reduce evaporation of the water they contain. This should be considered as a last resort since weighing in the field is more reliable.

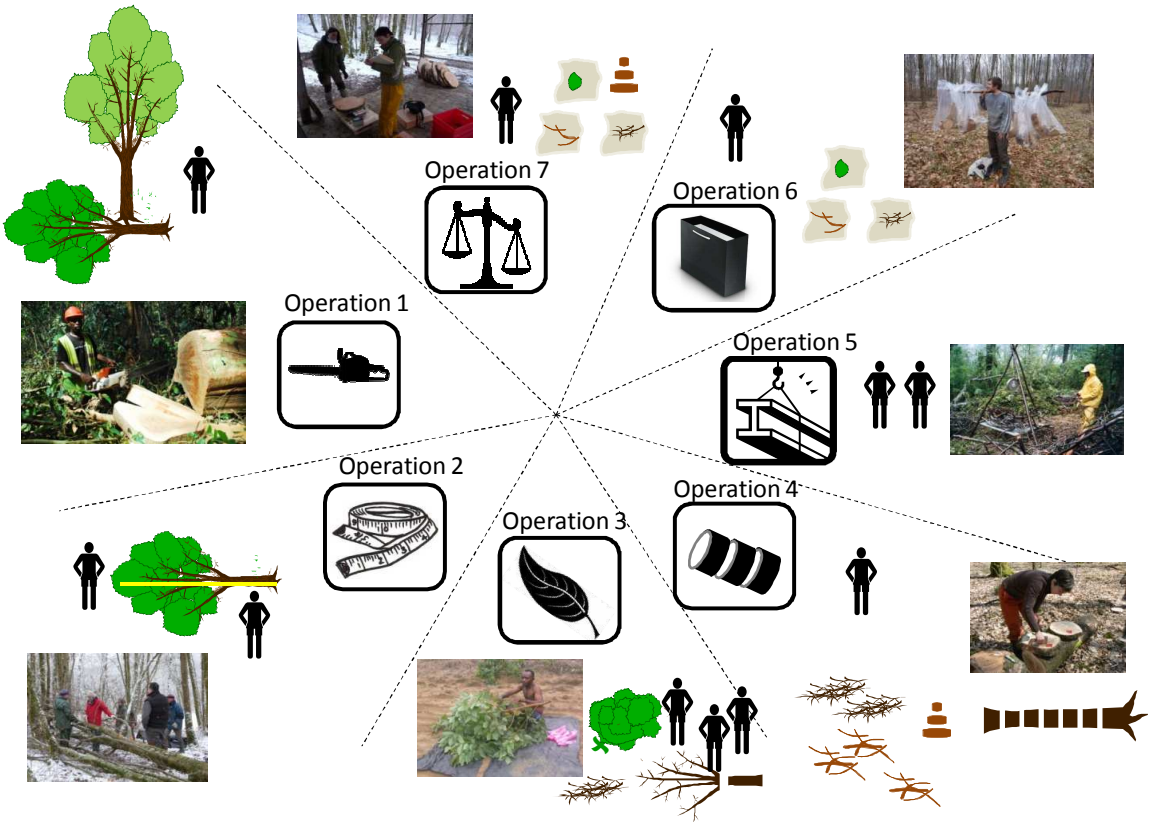


Figure 3.2 – *Organization of a biomass measurement site with 7 different operations. Operation 1, site preparation and felling of the trees (photo: L. Saint-André); operation 2, measurement of felled trees: stem profile, marking for cross-cutting (photo: M. Rivoire); operation 3, stripping of leaves and limbing (photo: R. D'Annunzio and M. Rivoire); operation 4, cross-cutting into logs and disks (photo: C. Nys); operation 5, weighing of logs and brushwood (photo: J.-F. Picard); operation 6, sampling of branches (photo: M. Rivoire); operation 7, sample weighing area (photo: M. Rivoire).*

PHOTO 3.3 – *Measurement campaign in a high forest coppice in France. On the left, arriving at the site and deploying equipment (photo: M. Rivoire); on the right, felling the first tree (photo: L. Saint-André).*

Felling (operation 1)

The logger prepares the tree selected while the technicians cut any small stems that may hinder the tree's fall and clear the area. A tarpaulin may be spread on the ground so as not to lose any leaves when felling (photo 3.4). As the tree's fall may bring down branches from other crowns, the technicians separate these from those of the selected tree.

Photo 3.4 – *Biomass campaign in a eucalyptus plantation in Congo. On the left, stripping leaves onto a tarpaulin (photo: R. D'Annunzio). On the right, operations completed for a tree, with bags in the weighing area containing leaves, logs and brushwood (photo: L. Saint-André).*

Measuring the tree (operation 2)

Stem profiles are then measured (photo 3.5). As the trunk is never cut at ground level, it is vital to mark the tree at breast height (1.30 m) with paint prior to felling, and place the measuring tape's 1.30 m graduation on this mark once the tree has been felled. This avoids any introduction of bias in section location (the shift induced by cutting height). Circumferences are generally measured every meter or, more usefully for constructing stem profile models, by percentage of total height. This latter method, however, is far more difficult to apply in the field. If it is impossible to measure a circumference because the trunk is flat on the ground, use tree calipers to take two diameter measurements, one perpendicular to the other. When moving up the trunk, the cross-cutting points agreed with the forest management organization (or timber buyer) should be marked on the trunk with paint or a logger's mark.

If the tree is straight and has a clearly identified main stem, there is no need to choose a principal axis. But if the stem is very twisted or branched (crown of hardwoods), the principal axis needs to be identified. This should then be marked, for instance with paint. The principal axis differs from the others by having the largest diameter at each fork in the trunk. All branched axes on the main stem are considered to be branches. It is possible, in the case of multistemmed trees, either to include each stem in the principal stem (photo 3.6), or consider each stem as an individual and in this case trace the principal axis on each.

The length of the trunk and the position of the last live branch and major forks must then be determined. Various heights can be measured on the felled tree, for example: height at the small end < 1 cm, height at cross-cut diameter 4 cm, and height at cross-cut diameter

PHOTO 3.5 – *Left, biomass campaign in Ghana in a teak high forest: measurement of branching (photo: S. Adu-Bredu). Right, biomass campaign in France in a high forest coppice: stem profile measurements (photo: M. Rivoire).*

PHOTO 3.6 – *Biomass campaign in rubber tree plantations in Thailand. Left, a multi-stemmed felled tree (3 stems on the same stump): limbing and stripping of the leaves. Right, leaves mixed together prior to aliquoting (photos: L. Saint-André).*

7 cm. The measurements made on the felled tree can then be compared with those made in standing trees during forest inventories. This checks dataset consistency and can be used to correct aberrant data, though differences may of course arise from the imprecision of pre-felling height measurements (generally 1 m), or from stem sinuosity or breakages when measuring post-felling lengths.

Cross-cutting (operations 3 and 4)

Ideally, the tree should be cut into 2 m logs to take account of wood density and moisture content variations in the stem. Once the tree has been prepared, the branches (and if necessary the leaves) must be separated from the trunk. The branches should then be cut to prepare bundles by small end class. When dealing with a stand of temperate hardwoods, the stem is generally cut into diameter classes of > 20 cm, 20–7 cm, 7–4 cm, < 4 cm. For eucalyptus in the DR of Congo, the branches were split into two groups: < 2 cm and > 2 cm. Branch bundles were prepared on an iron frame with two pieces of strong string (see 3.5 and photo 3.14). If the branches bear leaves, these leaves should be stripped off the twigs. Tarpaulins should be used in this case to avoid losing any leaves. If the leaves are difficult to remove from the woody axes (e.g. holm oak or softwoods), a subsampling strategy should be adopted (see the example below for Cameroon). The leaves are placed in large plastic bags for weighing. Limbing and leaf stripping take time, and sufficient human resources (number of teams) should be allocated so as not to slow the loggers in their work. The limbs of the tree are often large in diameter (> 20 cm) and should be processed in the same manner as for the trunk, by cross-cutting into logs and disks.

This cross-cutting should take place once the limbs have been separated from the main trunk. A disk about 3–5 cm thick should be taken from the stump, then every x meters (photo 3.7). Log length x depends on the dimensions of the tree and the agreements established with the body in charge of the forest or the timber company. As this field work is tedious and time-consuming, full advantage should be made of this time spent in the field to take multiple samples (e.g. an extra disk for more detailed measurements of wood density or nutrient content — see for instance Saint-André *et al.*, 2002b, for nutrient concentrations in the stems of eucalyptus). It is important to indicate the position of each disk taken. These disks must be weighed *in situ* on the day the tree was processed to reduce moisture losses (this operation requires work by two people — generally the stem profiling team performs this task by taking a break from its work a little earlier to weigh the disks, see 3.2).

Weighing logs and brushwood bundles (operation 5)

The logs and bundles produced are weighed in the field (photo 3.7) and within a short timeframe to ensure that the measurements for a given tree were taken under the same relative humidity conditions. The most practical approach here is to use hanging scales attached to a digger. The bundles are attached to the scales and fresh weight measured. The string and the tarpaulin can then be recovered for reuse.

Taking aliquots (operations 6 and 7)

Once the bundles have been weighed, aliquots are taken from each to estimate branch moisture content. Samples of different diameters should be taken from different branches to be representative of the architecture of a standard branch: sampling a single branch could lead to bias if it is more wet or dry than the others. The branches should be split into four diameter groups (class 1: $0 < \varnothing \leq 4$ cm, class 2: $4 < \varnothing \leq 7$ cm, class 3: $7 < \varnothing \leq 20$ cm,

PHOTO 3.7 – *Biomass campaign in an oak grove. Left, disks taken and placed in large bags for transfer to the samples weighing area; middle, samples weighing area; right, digger used to weigh logs (photos: C. Nys).*

and class 4: $\varnothing > 20$ cm). Samples about 10 cm long should be taken of the class 1 branches. A similar principle applies for the other branches, but as they are larger in diameter, disks are taken instead of pieces 10 cm long. About 9, 6 and 3 disks should be taken for classes 2, 3 and 4, respectively. These figures are only a rough guide but are nevertheless derived from a number of different campaigns conducted in different ecosystems. These aliquots are then placed in specially prepared paper bags (previously placed at the foot of the tree, see the first step). These paper bags are then placed in a plastic bag for a given tree to ensure that there is no mix up of samples between trees.

To avoid sampling bias, the same person must take all the samples and work in a systematic manner that is representative of the variability in each branch size class. To minimize bias when measuring moisture content, all the samples should be taken to the weighing area (same place as the disks) and should be weighed in their paper bag before being processed in the laboratory. Although it is best to weigh the samples in the field, if this proves impossible then any loss of moisture should be avoided. For instance the use of a cool box is very highly recommended. When sampling leaves, they should first be mixed together and the sample should be taken randomly from the middle of the heap. This mixing-sampling operation should be conducted five or six times for each tree (photo 3.6). The samples from each tree should then be placed in the same bag (the amount taken should be adapted to leaf size and heterogeneity, and particularly the proportion of green to senescent leaves — one standard plastic bag is generally perfectly adequate).

3.1.2 In the laboratory

If the trunk disks cannot be measured immediately they should be stocked in the fresh air on wooden battens so that air can circulate freely between them (to avoid mould). They can now be left to dry provided that fresh biomass was measured in the field. But if they were not weighed in the field, they must be weighed immediately on arrival.

For aliquots that were weighed in a bag in the field, the weight of the empty bag must now be determined (if possible measure each bag, or if very damaged, weigh 10 to 20 bags and calculate mean weight). This value should then be deducted from the field measurements. If a fresh bag is used when drying the aliquots, this must bear all the necessary information.

The oven should be set at 70°C for leaves, flowers and fruits, or at 65°C if the aliquots are subsequently to undergo chemical analysis. A temperature of 105°C should be used for biomass determinations, and only for wood. At least three controls should be weighed every day for all sample categories until a constant weight is reached. This avoids withdrawing actual samples from the oven for daily controls. Leaf weight generally stabilizes after two

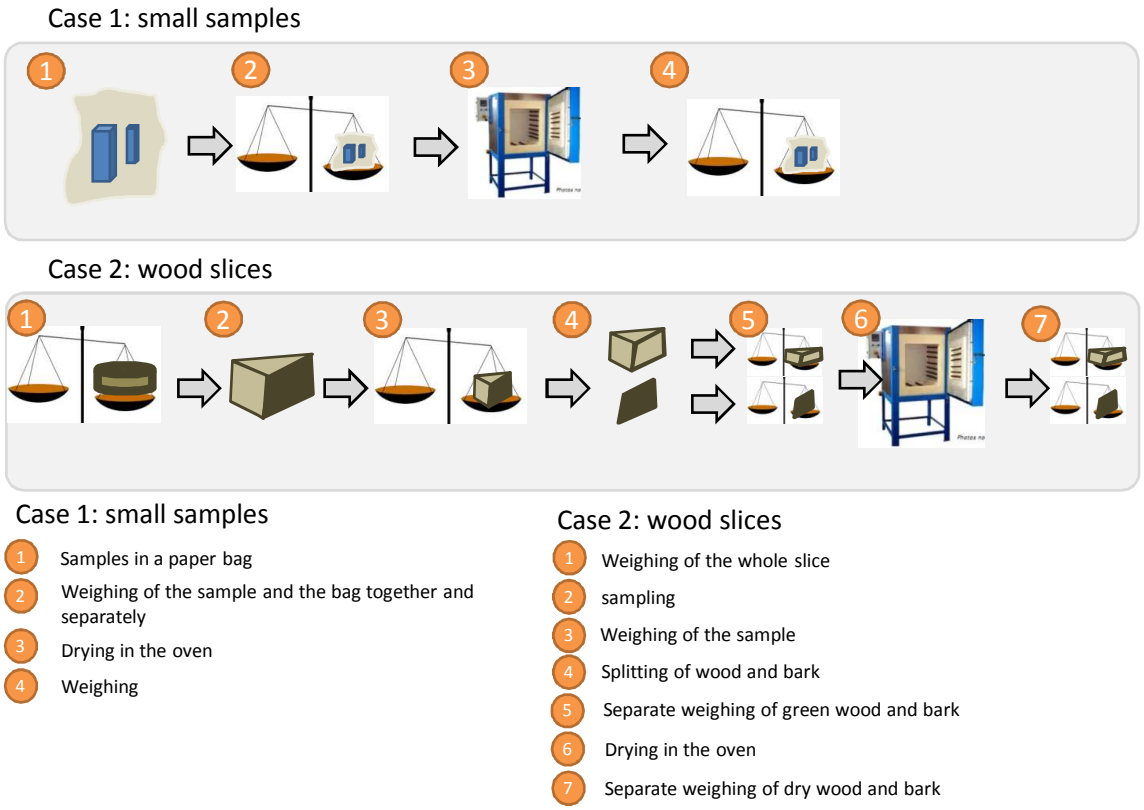


Figure 3.3 – *Weighing procedure for samples on arrival at the laboratory.*

days, and woody structures after about one week depending on the size of the sample.

Figure 3.3 shows the operating procedure to be adopted for measuring the samples taken. Laboratory measurements begin by weighing the wet samples with their bag (measurement checking that made in the field). If the disks are too large, they may be subsampled. In this case it is vital to weight the full disk again, then the sample taken. Moisture lost from the disk between the field and the laboratory should be added to this laboratory measurement after drying the disk sample in the oven. If a great deal of time has elapsed between the field phase and the laboratory phase, and if this step in the protocol is overlooked, this may cause very considerable errors of up to 60 or 70% in dry biomass estimations. The bark is generally removed using a bark knife or a wood chisel (photo 3.8) — placing the disks in a freezer when still wet may facilitate this operation (e.g. for oak). Samples of the bark and wood are then measured and dried in the oven (take care not to place too many bags at once in the oven).

3.1.3 Calculations

Calculating the biomass of the trunk

For each log i, measure the circumference at the two extremities: circumference C_{1i} at the small end is the circumference of the disk taken at the small end and circumference C_{2i} at the large end is the circumference of the disk taken at the large end. This can be employed to calculate the volume of the fresh log using the truncated cone volume formula (or Newton's

PHOTO 3.8 – *Laboratory measurements: (A) debarking, (B) weighing the wood, (C) weighing the bark (photos: L. Saint-André), (D) oven-drying samples, (E) regular weighing to constant weight (photos: M. Henry).*

formula):

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

where L_i is the length of log i, and $R_{1i} = C_{1i}/(2\pi)$ et $R_{2i} = C_{2i}/(2\pi)$ are the radii of log i at its two extremities. This volume may be measured with bark (circumferences being measured in the field) or without bark (circumferences being measured on the disks after debarking in the laboratory). Fresh volume with bark is widely employed when selling timber whereas the second measurement is used to check data consistency for it is used to calculate wood density in the tree.

It should be noted that other formulas are also used to calculate the volume of a log. That most commonly used is Huber's formula (based on circumference measured in the middle of the log) and Smalian's formula (based on the quadratic mean of the circumferences measured at the top and bottom of the log). But if the logs are short (1 or 2 m), their shape is little different from that of a cone, with very little tapering, and the difference between the formulas is very slight.

Also, for each sample taken of log i we calculate:

- the proportion of fresh biomass in the wood (without the bark):

$$\omega_{\text{fresh wood},i} = \frac{B^{\text{aliquot}}_{\text{frresh wood},i}}{B^{\text{aliquot}}_{\text{fresh wood},i} + B^{\text{aliquot}}_{\text{fresh bark},i}}$$

  where $B^{\text{aliquot}}_{\text{fresh wood},i}$ is the fresh biomass of the wood (without the bark) in the sample of log i, and $B^{\text{aliquot}}_{\text{fresh bark},i}$ is the fresh biomass of the bark in the sample of log i;

- the moisture content of the wood (without the bark):

$$\chi_{\text{wood},i} = \frac{B^{\text{aliquot}}_{\text{dry wood},i}}{B^{\text{aliquot}}_{\text{fresh wood},i}} \tag{3.2}$$

  where $B^{\text{aliquot}}_{\text{dry wood},i}$ is the oven-dried biomass of the wood (without the bark) in the sample of log i;

- the proportion of fresh biomass in the bark:

$$\omega_{\text{fresh bark},i} = 1 - \omega_{\text{fresh wood},i}$$

- the moisture content of the bark:

$$\chi_{\text{bark},i} = \frac{B^{\text{aliquot}}_{\text{dry bark},i}}{B^{\text{aliquot}}_{\text{fresh bark},i}}$$

  where $B^{\text{aliquot}}_{\text{dry bark},i}$ is the oven-dried biomass of the bark in the sample of log i.

We then extrapolate the values obtained for the sample of log i to the whole of log i by rules of three:

- the oven-dried biomass of the wood (without the bark) in log i is:

$$B_{\text{dry wood},i} = B_{\text{fresh},i} \times \omega_{\text{fresh wood},i} \times \chi_{\text{wood},i}$$

  where $B_{\text{fresh},i}$ is the fresh biomass (including the bark) of log i;

- the oven-dried biomass of the bark in log i is:

$$B_{\text{dry bark},i} = B_{\text{fresh},i} \times \omega_{\text{fresh bark},i} \times \chi_{\text{bark},i}$$

- the density of the wood in log i is:

$$\rho_i = \frac{B_{\text{dry wood},i}}{V_{\text{fresh},i}}$$

  where $V_{\text{fresh},i}$ is the fresh volume (without the bark) given by equation (3.1).

The weights of all the logs should then be summed to obtain the dry weight of the trunk:

- the dry biomass of the wood (without the bark) in the trunk is:

$$B_{\text{trunk dry wood}} = \sum_i B_{\text{dry wood},i}$$

  where the sum concerns all logs i that make up the trunk;

- the dry biomass of the bark in the trunk is:

$$B_{\text{trunk dry bark}} = \sum_i B_{\text{dry bark},i}$$

The wood density ρ_i used in the calculation of dry biomass must be the oven-dry wood density, i.e. the ratio of dry biomass (dried in the oven to constant dry weight) to the volume of fresh wood. Care should be taken not to confuse this wood density value with that calculated using the same moisture content for the mass and the volume (i.e. dry mass over dry volume or fresh mass over fresh volume). The AFNOR (1985) standard, however, defines wood density differently, as the ratio of the biomass of the wood dried in the fresh air to the volume of the wood containing 12 % water (Fournier-Djimbi, 1998). But the oven-dry wood density can be calculated from the density of wood containing 12 % water using the expression (Gourlet-Fleury *et al.*, 2011):

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

where ρ_χ is the ratio of the biomass of the wood dried in the fresh air to the volume of the wood containing χ % water (in $\mathrm{g\,cm^{-3}}$), ρ is the ratio of the oven-dry wood density to the fresh volume of the wood (in $\mathrm{g\,cm^{-3}}$), η is the volumetric shrinkage coefficient (figure without dimensions) and χ_0 is the fibers saturation point. The coefficients η and χ_0 vary from one species to another and require information on the technological properties of the species' wood. Also, using findings for ρ and $\rho_{12\%}$ in 379 trees, Reyes *et al.* (1992) established an empirical relationship between oven-dry density ρ and density with 12 % water $\rho_{12\%}$: $\rho = 0.0134 + 0.800\rho_{12\%}$ with a determination coefficient of $R^2 = 0.988$.

Calculating the biomass of the leaves

For each sample i of leaves taken, calculate the moisture content:

$$\chi_{\text{leaf},i} = \frac{B^{\text{aliquot}}_{\text{dry leaf},i}}{B^{\text{aliquot}}_{\text{fresh leaf},i}}$$

where $B^{\text{aliquot}}_{\text{dry leaf},i}$ is the oven-dry biomass of the leaves in sample i, and $B^{\text{aliquot}}_{\text{fresh leaf},i}$ is the fresh biomass of the leaves in sample i. We then, by a rule of three, extrapolate sample i to compartment i from which this sample was drawn:

$$B_{\text{dry leaf},i} = B_{\text{fresh leaf},i} \times \chi_{\text{leaf},i}$$

where $B_{\text{dry leaf},i}$ is the (calculated) dry biomass of the leaves in compartment i, and $B_{\text{fresh leaf},i}$ is the (measured) fresh biomass of the leaves in compartment i. Often the entire crown corresponds to a single compartment. But if the crown has been compartmentalized (e.g. by successive tiers), the total dry weight of the leaves is obtained by adding together all compartments i:

$$B_{\text{dry leaf}} = \sum_i B_{\text{dry leaf},i}$$

Calculating the biomass of the branches

If the branches are very large (e.g. > 20 cm in diameter), proceed as for the trunk, and for the brushwood proceed as for the leaves.

Calculating the biomass of fruits, flowers

The method is identical to that used for the leaves.

3.2 Direct weighing for certain compartments and volume and density measurements for others

The second case we consider here is that where felling constraints mean that only semi-destructive measurements are taken. These combine direct measurements for certain parts of the tree, and volume and density measurements for the others. We will illustrate this by developing an allometric equation for the dry northern forests of Cameroon. The biomass of dry forests is particularly difficult to determine given the architectural complexity of the trees they contain. Human intervention is also particularly significant in dry areas because of the rarity of forest resources and the magnitude of the demand for bioenergy. This is reflected by the trimming, pruning and maintenance practices often employed for trees located in deciduous forests, agro-forest parks and hedges (photo 3.9).

Photo 3.9 – *Trimming shea trees (Vitellaria paradoxa) in north Cameroon (photo: R. Peltier).*

The trees in most dry areas are protected as this woody resource regenerates particularly slowly and is endangered by human activities. Biomass measurements are therefore non-destructive and take full advantage of this trimming to measure the biomass of the trimmed compartments. Grazing activities limit regeneration and small trees are often few in number. This part of the guide will therefore consider only mature trees.

3.2.1 In the field: case of semi-destructive measurements

Generally, the trunk and the large branches are not trimmed; only the small branches are affected. The measurement of fresh biomass (in kg) may be divided into two parts: measuring trimmed fresh biomass and measuring untrimmed fresh biomass (Figure 3.4A).

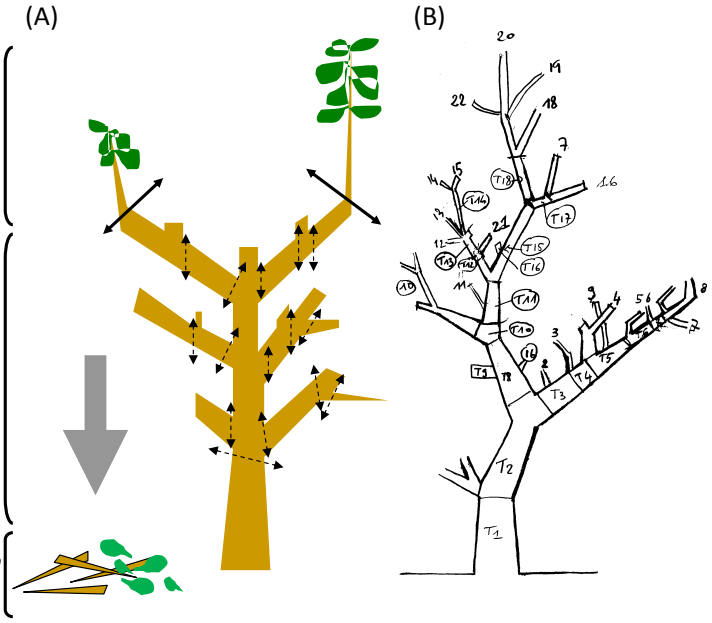


Figure 3.4 – *Determination of total fresh biomass. (A) Separation and measurement of trimmed and untrimmed biomass, (B) numbering of the sections and branches measured on a trimmed tree.*

Trimmed fresh biomass

The branches may be trimmed in compliance with local practices (often using a machete). Use a measuring tape to determine the diameter at the base of each branch. Then separate the leaves from the trimmed branches. Determine (by weighing separately) the fresh biomass of the leaves from the trimmed branches ($B_{\text{trimmed fresh leaf}}$) and the fresh biomass of the wood from the trimmed branches ($B_{\text{trimmed fresh wood}}$). Suitable scales should be used for these weighing operations. If the leaves weigh less than 2 kg, they may be weighed using field-work electronic scales.

Take a random sample of the leaves from the trimmed branches. At least three samples of leaves from three different branches are generally required to constitute the aliquot. Measure its fresh weight ($B^{\text{aliquot}}_{\text{fresh leaf}}$ in g). Take also an aliquot of the wood at random from the trimmed branches, without debarking. Measure its fresh mass ($B^{\text{aliquot}}_{\text{fresh wood}}$ in g) in the field, immediately after cutting. Place these aliquots in numbered plastic bags and send to the laboratory. The fresh volume of the wood aliquot will be measured later in the lab (see § 3.2.2), and the value used to determine mean wood density $\bar{\rho}$.

Untrimmed fresh biomass

Untrimmed biomass is measured indirectly as non-destructive. The different branches in the trimmed tree must first be numbered (Figure 3.4B). The small untrimmed branches must be processed differently from the large branches and the trunk (Figure 3.4A). For the small branches, only basal diameter needs to be measured. The biomass of these small untrimmed branches is estimated from the relationship between their basal diameter and their mass, as explained in section 3.2.3.

The biomass of the trunk and the large branches is estimated from measurements of volumes (V_i in cm^3) and mean wood density ($\bar{\rho}$ in $\text{g}\,\text{cm}^{-3}$). The large branches and trunk should be divided virtually into sections that are then materialized by marking the tree. The

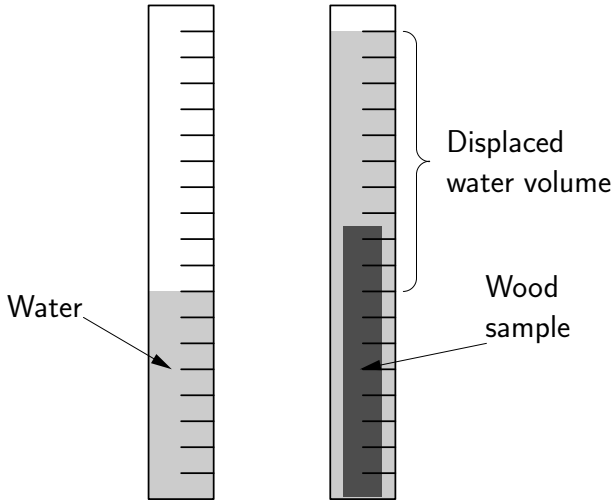


Figure 3.5 – *Measuring sample volume by water displacement.*

volume V_i of each section i is obtained by measuring its diameter (or its circumference) and its length. Sections about one meter in length are preferable in order to consider diameter variations along the length of the trunk and branches.

3.2.2 In the laboratory

First, measure the volume ($V^{\text{aliquot}}_{\text{fresh wood}}$) of the wood aliquot taken from the trimmed compartments. This volume may be measured in different manners (Maniatis *et al.*, 2011). That most commonly employed measures the volume of water displaced when the sample is immersed in water. The volume of water displaced may be measured using a graduated tube of suitable dimensions for the sample (Figure 3.5). Another method consists in cutting the sample into shapes whose volume may be accurately measured. This method requires precision tools and personnel trained in cutting wood.

The wood and leaf aliquots should then be subject to the same laboratory measurements (oven drying, determination of dry weight, etc.) as described in section 3.1.2.

3.2.3 Calculations

The dry biomass of the tree is obtained by the sum of the trimmed dry biomass and the untrimmed dry biomass:

$$B_{\text{dry}} = B_{\text{trimmed dry}} + B_{\text{untrimmed dry}}$$

Calculating trimmed biomass

From the fresh biomass $B^{\text{aliquot}}_{\text{fresh wood}}$ of a wood aliquot and its dry biomass $B^{\text{aliquot}}_{\text{dry wood}}$, calculate as above (see equation 3.2) the moisture content of the wood (including bark):

$$\chi_{\text{wood}} = \frac{B^{\text{aliquot}}_{\text{dry wood}}}{B^{\text{aliquot}}_{\text{fres wood}}}$$

Likewise, calculate the moisture content of the leaves from the fresh biomass $B^{\text{aliquot}}_{\text{fresh leaf}}$ of the leaf aliquot and its dry biomass $B^{\text{aliquot}}_{\text{dry leaf}}$:

$$\chi_{\text{leaf}} = \frac{B^{\text{aliquot}}_{\text{dry leaf}}}{B^{\text{aliquot}}_{\text{fresh leaf}}}$$

Trimmed dry biomass can then be calculated:

$$B_{\text{trimmed dry}} = B_{\text{trimmed fresh wood}} \times \chi_{\text{wood}} + B_{\text{trimmed fresh leaf}} \times \chi_{\text{leaf}}$$

where $B_{\text{trimmed fresh leaf}}$ is the fresh biomass of the leaves stripped from the trimmed branches and $B_{\text{trimmed fresh wood}}$ is the fresh biomass of the wood in the trimmed branches.

Calculating untrimmed biomass

Two calculations are required to calculate the dry biomass of the untrimmed part (i.e. that still standing): one for the small branches, the other for the large branches and the trunk. The untrimmed biomass is the sum of the two results:

$$B_{\text{untrimmed dry}} = B_{\text{untrimmed dry branch}} + B_{\text{dry section}}$$

Each section i of the trunk and the large branches may be considered to be a cylinder of volume (Smalian's formula):

$$V_i = \frac{\pi}{8} L_i \left(D^2_{1i} + D^2_{2i}\right) \tag{3.3}$$

where V_i is the volume of the section i, L_i its lenght, and D_{1i} and D_{2i} are the diameters of the two extremities of section i. The truncated cone volume formula (see equation 3.1) can also be used instead of the cylinder formula (3.3), but the difference between the results will be slight as the tapering over one meter is not very pronounced in trees.

The dry biomass of the large branches and trunk is the product of mean wood density and total volume of the large branches and trunk:

$$B_{\text{dry section}} = \bar{\rho} \times \sum_i V_i \tag{3.4}$$

where the sum corresponds to all the sections in the large branches and the trunk (Figure 3.4B), and where mean wood density is calculated by:

$$\bar{\rho} = \frac{B^{\text{aliquot}}_{\text{dry wood}}}{V^{\text{aliquot}}_{\text{fresh wood}}}$$

Care must be taken to use consistent measurement units. For example, if mean wood density $\bar{\rho}$ in (3.4) is expressed in $\text{g}\,\text{cm}^{-3}$, then volume V_i must be expressed in cm^3, meaning that both length L_i and diameters D_{1i} and D_{2i} must also be expressed in cm. Biomass in this case is therefore expressed in g.

The dry biomass of the untrimmed small branches should then be calculated using a model between dry biomass and basal diameter. This model is established by following the same procedure as for the development of an allometric model (see chapters 4 to 7 herein). Power type equations are often used:

$$B_{\text{dry branch}} = a + bD^c$$

where a, b and c are model parameters and D branch basal diameter, but other regressions may be tested (see 5.1). Using a model of this type, the dry biomass of the untrimmed branches is:

$$B_{\text{untrimmed dry branch}} = \sum_j (a + bD_j^c)$$

where the sum is all the untrimmed small branches and D_j is the basal diameter of the branch j.

3.3 Partial weighings in the field

The third case we envisage is where the trees are too large to be weighed in their entirety by hand. We will illustrate this by developing an allometric equation to estimate the above-ground biomass of trees in a wet tropical forest based on destructive measurements. The method proposed must suit local circumstances and the resources available. And commercial value and the demand for timber are two other factors that must be taken into account when making measurements in forest concessions.

The trees selected must be felled using suitable practices. Variables such as total height, and height of any buttresses can be determined using a measuring tape. The tree's architecture can then be analyzed (Figure 3.6). The proposed approach separates trees that can be weighed manually in the field (e.g. trees of diameter ≤ 20 cm) from those requiring more substantial technical means (trees of diameter > 20 cm).

3.3.1 Trees less than 20 cm in diameter

The approach for trees of diameter ≤ 20 cm, is similar to that described in the first example (§ 3.1). The branches and the trunk are first separated. The fresh biomass of the trunk ($B_{\text{fresh trunk}}$) and branches ($B_{\text{fresh branch}}$, branch, wood and leaves together) is measured using appropriate scales. When measuring the biomass of the leaves, a limited number of branches should be selected at random for each individual. The leaves and the wood in this sample of branches should then be separated. The fresh biomass of the leaves ($B^{\text{sample}}_{\text{fresh leaf}}$) and the fresh biomass of the wood ($B^{\text{sample}}_{\text{fresh wood}}$) in this sample of branches should then be measured separately using appropriate scales. The foliage proportion of the branches must then be calculated:

$$\omega_{\text{leaf}} = \frac{B^{\text{sample}}_{\text{fresh leaf}}}{B^{\text{sample}}_{\text{fresh leaf}} + B^{\text{sample}}_{\text{fresh wood}}}$$

The fresh biomass of the leaves ($B_{\text{fresh leaf}}$) and the wood ($B_{\text{fresh wood}}$) in the branches is then calculated based on this mean proportion of leaves:

$$\begin{aligned} B_{\text{fresh leaf}} &= \omega_{\text{leaf}} \times B_{\text{fresh branch}} \\ B_{\text{fresh wood}} &= (1 - \omega_{\text{leaf}}) \times B_{\text{fresh branch}} \end{aligned}$$

Aliquots of leaves and wood are then taken at different points along the branches and along the trunk. The fresh biomass ($B^{\text{aliquot}}_{\text{fresh leaf}}$ and $B^{\text{aliquot}}_{\text{fresh wood}}$) of the aliquots must then be measured using electronic scales in the field. The aliquots must then be taken to the laboratory for drying and weighing in accordance with the same protocol as that described in the first example (§ 3.1.2). The oven-dry biomass ($B^{\text{aliquot}}_{\text{dry leaf}}$ and $B^{\text{aliquot}}_{\text{oven-dry}}$) of the aliquots can then be used to calculate the water contents of the leaves and wood:

$$\chi_{\text{leaf}} = \frac{B^{\text{aliquot}}_{\text{dry leaf}}}{B^{\text{aliquot}}_{\text{fresh leaf}}}, \qquad \chi_{\text{wood}} = \frac{B^{\text{aliquot}}_{\text{dry wood}}}{B^{\text{aliquot}}_{\text{fresh wood}}}$$

Finally, the dry biomass of the leaves and wood is obtained from their fresh biomass and water contents calculated in the aliquots. The biomass of the wood is obtained by adding its fresh biomass to that of the branches and the trunk:

$$\begin{aligned} B_{\text{dry leaf}} &= \chi_{\text{leaf}} \times B_{\text{fresh leaf}} \\ B_{\text{dry wood}} &= \chi_{\text{wood}} \times (B_{\text{fresh wood}} + B_{\text{fresh trunk}}) \end{aligned}$$

Total dry mass is finally obtained by adding together the dry biomass of the leaves and the dry biomass of the wood:

$$B_{\text{dry}} = B_{\text{dry leaf}} + B_{\text{dry wood}}$$

3.3.2 Trees more than 20 cm in diameter

If a tree is very large, it possesses a very large quantity of branches and leaves, and it is impractical to separate the branches from the trunk. The alternative method proposed here therefore consists in processing the trunk and large branches (basal diameter more than 10 cm) differently from the small branches (basal diameter less than 10 cm). Whereas the large branches of basal diameter > 10 cm are made up only of wood, the small branches of diameter ≤ 10 cm may also bear leaves. The large branches of basal diameter > 10 cm should be processed in the same manner as the trunk. The first step therefore consists in dividing the branches into sections. Whereas the biomass of the sections of diameter > 10 cm may be deduced from their measured volume ($V_{\text{billon},i}$) and mean wood density ($\bar{\rho}$), that of the branches of basal diameter ≤ 10 cm may be estimated using a regression between their basal diameter and the biomass they carry.

Measuring the volume of sections of diameter > 10 cm (trunk or branch)

Once the trunk and the branches of basal diameter > 10 cm have been divided into sections, the volume of each section can be calculated from their length and diameters (or circumferences) at the two extremities (D_{1i} and D_{2i}). A fixed length (e.g. two meters) may be used as the length of each section (photo 3.10A). A shorter section length than the fixed length must be used in some places where a fork prevents the log from assuming a cylindrical shape. The technician must record the length and diameters of each section before drawing a diagram of the tree's architecture (Figure 3.6). This diagram is particularly useful for analyzing and interpreting the results.

Trees of diameter > 20 cm often have buttresses. The volume of the buttresses may be estimated by assuming that they correspond in shape to a pyramid, the upper ridge of which is a quarter ellipse (insert to Figure 3.6; Henry *et al.*, 2010). For each buttress j, we need to measure its height H_j, its width l_j and its length L_j (insert to Figure 3.6).

Aliquots of wood must then be taken from the different sections of diameter > 10 cm (trunk, branches and any buttresses; photo 3.10B). These aliquots of fresh wood should be placed in tightly sealed bags and sent to the laboratory. In the laboratory, their volume ($V^{\text{aliquot}}_{\text{fresh wood}}$) should be measured in accordance with the protocol described in section 3.2.2. They should then be oven-dried and weighed as described in section 3.1.2. Their dry biomass ($B^{\text{aliquot}}_{\text{dry wood}}$) can be calculated from the results.

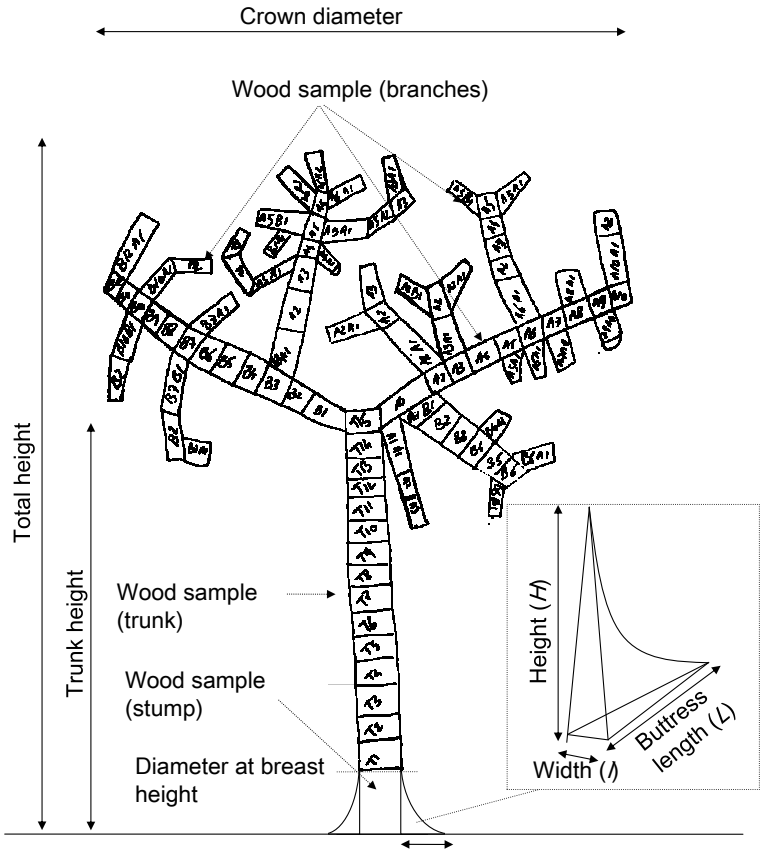


Figure 3.6 – *Diagram showing the different sections of a tree used to calculate its volume.*

Calculating the biomass of sections of diameter > 10 cm (trunk or branch)

As previously described (see equation 3.3), the volume $V_{\log,i}$ of the section i (trunk or branch of basal diameter > 10 cm) is calculated using Smalian's formula:

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

where L_i is the length of the section i, D_{1i} is the diameter at one of the extremities and D_{2i} is the diameter at the other. Given that its shape is that of a pyramid, a different formula is used to calculate the volume $V_{\text{buttress},j}$ of the buttress j:

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

where l_j is the length of the buttress j, L_j its length and H_j its height.

Mean wood density should then be calculated from the ovendry biomass and fresh volume of the wood aliquots:

$$\bar{\rho} = \frac{B_{\text{dry wood}}^{\text{aliquot}}}{V_{\text{fresh wood}}^{\text{aliquot}}}$$

The cumulated dry biomass of the sections (trunk and branches of basal diameter > 10 cm) is therefore:

$$B_{\text{dry section}} = \bar{\rho} \times \sum_i V_{\log,i}$$

PHOTO 3.10 – *Measuring a large tree in the field: (A) measuring the volume of a tree of diameter > 20 cm, (B) taking aliquots of wood from the trunk (photos: M. Henry).*

where the sum corresponds to all the sections, whereas the dry biomass of the buttresses is:

$$B_{\text{dry buttress}} = \bar{\rho} \times \sum_{j} V_{\text{buttress},j}$$

where the sum corresponds to all the buttresses. Alternatively, instead of using mean wood density, a specific density for each compartment (trunk, branches, buttresses) may also be employed. In this case mean wood density $\bar{\rho}$ in the expressions above must be replaced by the specific density of the relevant compartment.

Measuring branches of diameter < 10 cm

The basal diameter of all branches of diameter ≤ 10 cm, should now be measured and their dry biomass can be estimated from the regression between their basal diameter and the dry biomass they carry. This regression can be established from a sample of branches selected in the tree such as to represent the different basal diameter classes present. The leaves and the wood in this sample must now be separated. The fresh biomass of the leaves ($B^{\text{sample}}_{\text{fresh leaf},i}$ in the branch i) and the fresh biomass of the wood ($B^{\text{sample}}_{\text{fresh wood},i}$ in the branch i) are weighed separately in the field for each branch in the sample.

Some branches may be distorted and do not form a ramified architecture. In this case, the volume may be measured and the anomaly must be recorded in the field forms.

Aliquots of wood and leaves should then be taken and their fresh biomass ($B^{\text{aliquot}}_{\text{fresh wood}}$ and $B^{\text{aliquot}}_{\text{fresh leaf}}$) determined immediately in the field. These aliquots should be placed in tightly sealed plastic bags and taken to the laboratory where they are oven-dried and weighed in

accordance with the protocol described in section 3.1.2. Thus yields their ovendry biomass ($B^{\text{aliquot}}_{\text{dry wood}}$ and $B^{\text{aliquot}}_{\text{dry leaf}}$)

Calculating the biomass of branches of diameter < 10 cm

The fresh biomass of the aliquots serves to determine the water content of the leaves and wood:

$$\chi_{\text{leaf}} = \frac{B^{\text{aliquot}}_{\text{dry leaf}}}{B^{\text{aliquot}}_{\text{fresh leaf}}}, \qquad \chi_{\text{wood}} = \frac{B^{\text{aliquot}}_{\text{dry wood}}}{B^{\text{aliquot}}_{\text{fresh wood}}}$$

From this, and for each branch i of the sample of branches, may be determined the dry biomass of the leaves, the dry biomass of the wood, then the total dry biomass of branch i:

$$\begin{aligned} B^{\text{sample}}_{\text{dry leaf},i} &= \chi_{\text{leaf}} \times B^{\text{sample}}_{\text{fresh leaf},i} \\ B^{\text{sample}}_{\text{dry wood},i} &= \chi_{\text{wood}} \times B^{\text{sample}}_{\text{fresh wood},i} \\ B^{\text{sample}}_{\text{dry branch},i} &= B^{\text{sample}}_{\text{dry leaf},i} + B^{\text{sample}}_{\text{dry wood},i} \end{aligned}$$

In the same manner as in section 3.2.3, a biomass table for the branches can then be fitted to the data ($B^{\text{sample}}_{\text{dry branch},i}$, D^{sample}_{i}), where D^{sample}_{i} is the basal diameter of the branch i of the sample. The biomass model for the branches may be established by following the same procedure as for the development of an allometric equation (see chapter 4 to 7 herein). To increase sample size, the model can be established based on all the branches measured for all the trees of the same species or by species functional group (Hawthorne, 1995).

The branch biomass model established in this way can then be used to calculate the dry biomass of all branches of basal diameter ≤ 10 cm:

$$B_{\text{dry branch}} = \sum_i f(D_i)$$

where the sum corresponds to all the branches of basal diameter ≤ 10 cm, D_i is the basal diameter of the branch i, and f is the biomass model that predicts the dry biomass of a branch on the basis of its basal diameter.

Calculating the biomass of the tree

The dry biomass of the tree may be obtained by summing the dry biomass of the sections (trunk and branches of basal diameter > 10 cm), dry biomass of the buttresses, and the dry biomass of the branches of diameter ≤ 10 cm:

$$B_{\text{dry}} = B_{\text{dry sections}} + B_{\text{dry buttresses}} + B_{\text{dry branches}}$$

3.4 Measuring roots

It is far more difficult to determine root biomass than above-ground biomass. The methods we propose here are the results of campaigns conducted in different ecosystems and were compared in a study conducted in Congo (Levillain *et al.*, 2011).

The first step, whatever the ecosystem, consists in drawing a Voronoi[1] diagram around the selected tree. Figure 3.7 indicates the procedure employed: *(i)* draw the segments that

[1] A Voronoi diagram (also called a Voronoi decomposition, a Voronoi partition or Voronoi polygons) is a special kind of decomposition of a metric space determined by distances to a specified family of objects, generally a set of points.

link the selected tree to each of its neighbors; *(ii)* draw the bisectors for each segment, *(iii)* connect the bisectors one to another to bound a space around the tree; *(iv)* this space can then be divided into joined triangles, the area of each zone being easy to calculate when the length of all three sides (a, b and c) is known:

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

where $p = a + b + c$ is the triangle perimeter and A its area. Figure 3.8 illustrates this procedure for coconut plantations in Vanuatu (Navarro *et al.*, 2008).

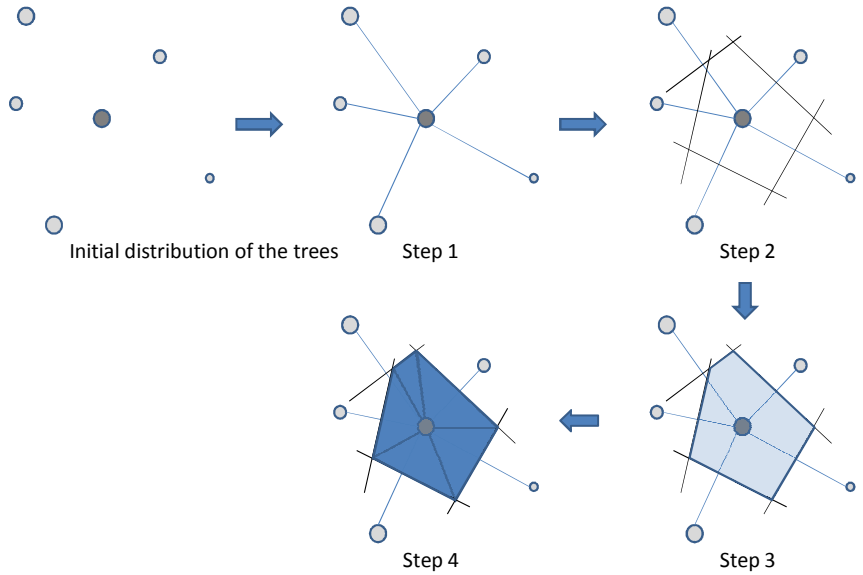


Figure 3.7 – *Method used to draw a Voronoi diagram and its subdivisions around a tree in any neighborhood situation.*

The space established in this manner is not a materialization of the tree's "vital" space. It is simply a manner used to cut space into joined area to facilitate the sampling of below-ground biomass. The major hypothesis here is that the roots of other trees entering this space compensate for those of the selected tree leaving this space.

In the case of multispecific stands or agro-forest, it is sometimes impossible to separate the roots of the different species. In this case its would be very risky to construct individual models (root biomass related to the sampled tree), but root biomass estimations per hectare, without any distinction between species, are perfectly valid.

Sampling methods vary with root size. Levillain *et al.* (2011) conducted a study to compare different methods in the same tree (photo 3.11). They showed that the most profitable, in terms of cost-precision, was to sample the finest roots by coring, whereas medium-size roots require partial excavation and large roots total excavation from the Voronoi space.

The number of cores obtained and the size of the pit dug vary from one ecosystem to another. In Congo, in eucalyptus plantations, the optimum number of cores for 10 % precision is about 300 on the surface (0–10 cm) and about 100 for deeper roots (10–50 and 50–100 cm). Such a level of precision to a depth of 1 m requires 36 man-days. By contrast, if only 30 % precision is sought, this cuts sampling time by 75 %. This example clearly illustrates the usefulness of pre-sampling (see chapitre 2) in order to evaluate the variability of the studied ecosystem then adapt the protocol to goals and desired precision.

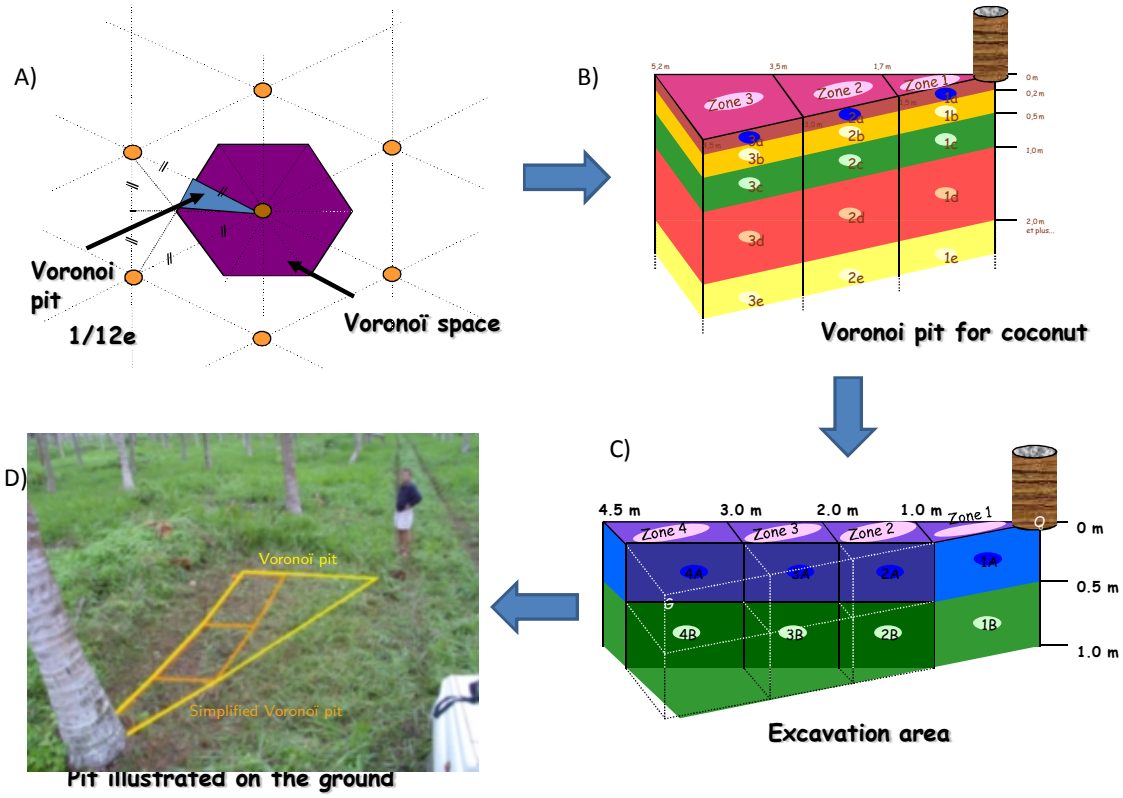


Figure 3.8 – *Typical Voronoi diagram used for root sampling in a coconut plantation in Vanuatu (photo: C. Jourdan). (A) Drawing the Voronoi space and choosing to work on 1/12th of this space; (B) section through the pits dug; (C) protocol simplification given the variability observed in the first sample; (D) tracing the pit on the ground.*

Once the soil samples with the roots have been taken, the cores containing the fine roots can be sorted in the laboratory. On the other hand, given the volume and weight of the earth excavated, the medium-size and large roots must be sorted in the field. In the laboratory, the soil sample should first be washed through a filter and roots recovered either by flotation and/or by sieving. In the field, the samples should be sorted manually on tarpaulins. An air knife can be used on the large and medium-size roots to excavate the root system fully while preserving its architecture. This method, which is particularly advantageous in sandy soils, helps meet biomass and architecture goals but does require the presence of a mobile compressor (photo 3.12).

Once the roots have been sorted and harvested they should be dried in an oven using the same parameters as for above-ground biomass. The fine roots will require roughly the same drying time as for leaves, whereas the large and medium-size roots will need a drying time equivalent to that of branches.

Regarding the stump, a subsample should be taken, preferably vertically to better take account of wood density variations in this part of the tree. Here, the same procedures should be used as for the disks of the trunk.

The calculations then made are the same as for above-ground biomass.

PHOTO 3.11 – *Left, superimposition of different sampling methods (cores, excavations by cubes, partial Voronoi excavation, total Voronoi excavation), from Levillain et al. (2011) (photo: C. Jourdan). Right, manual excavation of large roots in a rubber tree plantation in Thailand (photo: L. Saint-André).*

PHOTO 3.12 – *Air knife being used in Congo to extract the root systems (large and medium-size roots) of eucalyptus. Left, operator with protective equipment (dust, noise); middle, compressor and close up of the air pressure gauge (about 8 bars); right, the result (photos: C. Jourdan).*

3.5 Recommended equipment

3.5.1 Heavy machinery and vehicles

- Cars, trucks, trailer: for transporting personnel, equipment and samples to/from the laboratory.
- Quad bike (if possible): for transporting bulky equipment and samples in the field.
- Digger for weighing brushwood bundles.

3.5.2 General equipment

- Tool box with basic tools.
- Plastic boxes (for storing and transporting materials — about 10).
- Plastic bags (count one or two large bags per tree) to hold all the samples from a given tree and avoid water losses. Paper bags (count one per compartment and per tree) to hold samples immediately after they are taken. Ideally, these bags should be pre-labeled with tree and compartment number (which saves time in the field). But plan also to take a number of unmarked bags and a black, felt-tip pen (indelible ink) to make up for any mistakes made, or so that extra samples may be taken.
- Large tarpaulins for the crown (either cut up for the branch bundles, or spread on the ground to recover leaves torn off the trees).
- Labels (stapled onto the disks), staples and stapler; or fuschine pencil (but, if the samples are to be stored for later measurements of nutrient content, fuchsine should be avoided) (photo 3.13).
- Cutters, machetes, axes, shears, saws (photo 3.13).
- Iron frames for making bundling branches (photo 3.14) or alternatively bins of different sizes.
- Power saws (ideally one power saw suitable for felling trees and another smaller more maneuverable power saw for limbing — photo 3.13).
- Strong string for bundling the branches (given that this will be reused throughout the campaign, the knots should be undoable).
- Very strong large bags (such as grain or sand bags — photo 3.15) for transporting the disks and field samples to vehicles (if these are parked some distance from the felling site).

3.5.3 Computer-entering field data

- Pocket PC (with battery charger and cables) or field record forms on waterproof paper or cardboard, if possible bound in notebooks with plasticized covers front and back.
- Tree identification guide or key for wet tropical forest sites.
- 2B pencils, eraser, pencil sharpener.

PHOTO 3.13 – *Field equipment. Left, equipment for cutting and labeling aliquots; middle, typical gauges for cutting branches; right, power saw and safety equipment (photo: A. Genet).*

PHOTO 3.14 – *Bundling branches. Left, iron frame, tarpaulin and string for bundling branches (photo: A. Genet); middle, bundling branches in the field (photo: M. Rivoire); right, bundle ready for weighing (photo: M. Rivoire).*

PHOTO 3.15 – *Transporting disks and aliquots in a sand or grain bag (photo: J.-F. Picard).*

- Field scales (with 2 batteries and charger) for weighing samples (precise to within 1 g), ideally a full range of scales for different sample weights (a 1 or 2 m log may weigh several hundred kg, whereas the wood disks weigh from a few dozen g to several dozen kg). A digger will facilitate the field weighing of large logs. Straps should therefore be taken into the field to attach two scales to the bucket and self-blocking claws to attach the log.
- Decameter to measure lengths along the trunk (stem profiles).
- Tree calipers and girth tape to measure circumference.
- Tree marking paint gun (for marking standing trees and the main stem in highly developed crowns).
- Logger's mark to indicate where disks should be cut (or marked with the paint gun).

3.5.4 Laboratory equipment

- Ovens.
- Graduated tubes of at least 500 ml.
- Bark knife.
- Secateurs.
- Scales of a 2 to 2000 g capacity (precise to within 0.1 to 1 g).
- Band saw.

3.6 Recommended field team composition

Felling team: one logger, two logging assistants, two persons to clear the area for felling. Count two days to fell 40 trees of 31 to 290 cm girth (mean: 140 cm). All the trees can be felled at the start of the campaign so that this team is freed for tree limbing. If team members are to be kept occupied, with no slack periods, about 10 trees per day (about 20 meters high, i.e. about 10 to 20 disks per tree) need to be ready for cross-cutting at the site.

Stem profiling team: two people (one anchor, one measurer). This team intervenes as soon as the tree has been felled, and therefore follows the felling team. In general, these two teams work in a fairly synchronized manner. The profiling team never has to wait for the felling team, except in very rare cases of felling problems (e.g. large stems, or stems imprisoned by other trees that need to be released).

Limbing team: three persons per processing unit. Each unit includes a power saw handler (with the small, maneuverable power saw) and two bundlers. These teams can be doubled or tripled depending on the size of the crown to be processed. As a rough guide: a tree with a breast-height girth of 200 cm will require three units; between girths of 80 and 200 cm the tree will require two units, and below a girth of 80 cm one unit will suffice. This rough guide includes the fact that the units must not hinder one another in their work. When three teams are working on the same tree, one should be positioned at the base, and should

more upward along the length of the trunk. The two other teams should be positioned on each side of the main axis and should work from approximately the middle of the crown toward the top of the tree.

Logging team: this consists of one logger (cutting stems) and one person to label the disks. In general, the felling team fulfills this function. Once all the trees have been felled, the logger joins the logging team and the logging assistants bolster the limbing units.

Weighing team: three people (digger driver and two other persons to handle the logs and the branch bundles).

Branch sampling team: one or two people.

4

Entering and formatting data

Once the field measurements phase has been completed, and before any analysis, the data collected need to be given structure, which includes data entry, cleansing and formatting.

4.1 Data entry

This consists in transferring the data recorded on the field forms into a computer file. The first step is to choose the software. For a small dataset, a spreadsheet such as Microsoft Excel or OpenOffice Calc may be sufficient. For more substantial campaigns, a database management system such as Microsoft Access or MySQL (www.mysql.com) should be used.

4.1.1 Data entry errors

The data must be entered as carefully as possible to minimize errors. One of the methods used to achieve this is double data entry: one operator enters the data, a second operator (if possible different from the first) repeats the data entry in a manner entirely independent from the first. The two files can then simply be compared to detect any data entry errors. As it is unlikely that the two operators will make the same data entry error, this method guarantees good quality data entry. But it is time consuming and tedious.

Certain details, that have their importance, must also be respected during data entry. First, numbers must be distinguished from character strings. For the statistical software that will subsequently process the data, a number does not have the same role as a character string. This distinction must therefore be made during data entry. A number will be interpreted as the value of a numerical variable, whereas a character string will be interpreted as the category of a qualitative variable. The difference between the two is generally obvious, but not always. Let us, for instance, consider latitudes and longitudes. If we want to calculate the correlation between the latitude or the longitude and another variable (to identify a north-south or east-west gradient), the software must perceive the geographic coordinates as being numbers. These geographic coordinates must not therefore be entered as 7°28'55.1" or 13°41'25.9" as they would be interpreted as qualitative variables, and no calculations would be possible. A possible solution consists in converting geographic coordinates into decimal values. Another solution consists in entering geographic coordinated

in three columns (one for degrees, one for minutes, and the third for seconds).

When entering qualitative variables, care must be taken to avoid entering strings of great length as this enhances the risk of making data entry errors. It is far better to enter a short code and in the meta-information (see below) give the meaning of the code.

Another important detail is the decimal symbol used. Practically all statistical software packages can switch from the comma (symbol used in French) to the period (symbol used in English), therefore one or the other may be used. By contrast, if you choose to use a comma or a period as the decimal symbol, you must stick to this throughout the data entry operation. If you use one, then the other, some of the normally numerical data will be interpreted by the statistical software as character strings.

4.1.2 Meta-information

When entering data, some thought must be given to the meta-information, i.e. the information that accompanies the data, without in itself being measured data. Meta-information will, for instance, give the date on which measurements were taken, and by whom. If codes are used during data entry, the meta-information will specify what these codes mean. For example, it is by no means rare for the names of species to be entered in a shortened form. A species code such as ANO in dry, west African savannah for instance is ambiguous: it could mean *Annona senegalensis* or *Anogeissus leiocarpus*. The meta-information is there to clear up this ambiguity. This meta-information must also specify the nature of the variables measured. For example, if tree girth is measured, then noting "girth" in the data table is in itself insufficient. The meta-information must specify at what height tree girth was measured (at the base, at 20 cm, at 1.30 m...) and, an extremely important point, in which units this girth is expressed (in cm, dm...). It is essential that the units in which each variable was measured must be specified in the meta-information. All too often, data tables fail to specify in which units the data are expressed, and this gives rise to extremely hazardous guessing games.

For the people who designed the measuring system and supervised the actual taking of measurements, these details given in the meta-information are often obvious and they do not see the need to spend time providing this information. But let us imagine the situation facing persons who were not involved in the measurements and who review the data 10 years later. If the meta-information has been correctly prepared, these persons should be able to work on the dataset as if they had constituted it themselves.

4.1.3 Nested levels

The data should be entered in the spreadsheets as one line per individual. If the data include several nested levels, then as many data tables as there are levels should be prepared. Let us imagine that we are looking to construct a table for high forest coppice formations on a regional scale. For multistemmed individuals, the volume of each stem is calculated separately. Individuals should be selected within the land plots, themselves selected within the forests distributed across the study area. This therefore gives rise to four nested levels: the forest, which includes several plots; the plot, which includes several trees; the tree, which includes several stems; and finally the stem. Four data tables should therefore be prepared (Figure 4.1). Each data table will include the data describing the individuals of the corresponding level, with one worksheet line per individual. For example, this first data table will give the area of each forest. The second will give the geographic coordinates of each plot. The third data table will give the species and its coordinates for each tree in the plot. Finally, the fourth data table will give the volume and size of each stem. Therefore,

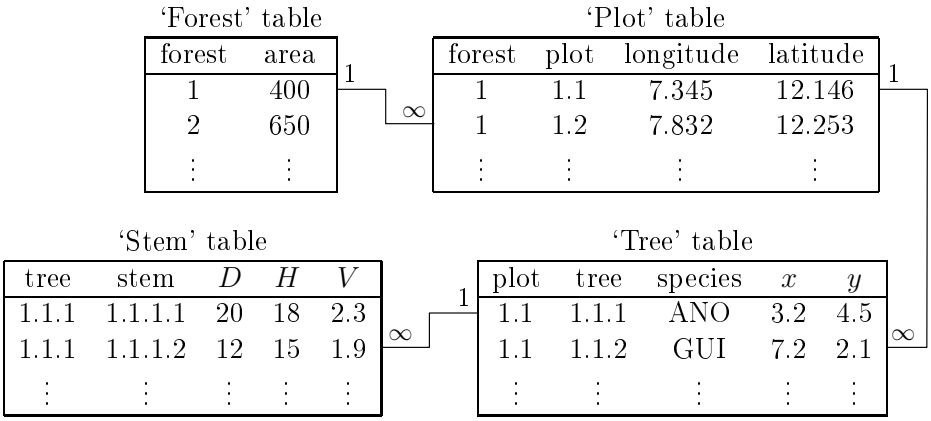


Figure 4.1 – *Example of four data tables for four nested levels.*

Table 4.1 – *Data entry with four nested levels in a single data table.*

forest	area	plot	longitude	latitude	tree	species	x	y	stem	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	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

to each line in a data table will correspond several lines in the lower level data table. A code should be used to establish the correspondence between the lines in the different data tables. For instance, the number of the forest should be repeated in the "forest" and "plot", data tables, the number of the plot should be repeated in the "plot" and "tree", data tables, and the number of the tree should be repeated in the "tree" and "stem" data tables (Figure 4.1).

Giving the data this structure minimizes the amount of information entered, and therefore minimizes data entry errors. An alternative consists in entering all the data in a single data table, as indicated for instance in Table 4.1. This is not recommended as data is unnecessarily repeated and therefore the risk of making data entry errors is greatly increased. For instance, in Table 4.1 we have deliberately introduced a data entry error in the second line of the data table, where the area of forest 1, which should be 400 ha, here is 401 ha. By unnecessarily repeating the information, we find ourselves multiplying this type of inconsistency that must subsequently be corrected.

An effective method that can be used to resolve these nested level problems is to build a relational database. This type of database is specifically intended to manage different inter-related data tables. It removes all inconsistencies, such as that illustrated in Table 4.1 by systematically checking relational integrity between the tables. But building a relational database requires technical knowledge, and a person with skills in this field may need to be called upon.

In brief, when entering data, it is advisable to:

- avoid repeating the same information,
- prefer relational databases,
- provide additional information (meta-information),
- take great care with units,
- distinguish between qualitative and quantitative information,
- check the data,
- reduce or correct missing data.

4.2 Data cleansing

Data cleansing requires toing-and-froing between the data record forms and the statistical software (or use of specially designed data cleansing software). This step aims to remove all data inconsistencies. And, if the measurement system is still in place, it may require certain measurements to be repeated. Data cleansing removes:

- aberrant data. For example, a tree 50 meters in diameter.
- inconsistent data. For example, a tree with a trunk biomass of 755 kg and a total biomass of 440 kg, or a tree that is 5 cm in diameter and 40 m in height.
- false categories for qualitative variables. For example, software that differentiates between upper case and lower case will interpret "yes" and "Yes" as two different categories, whereas in fact they are the same.

The difficulty inherent to detecting aberrant data lies in choosing the threshold between what is a normal measurement and what is an aberrant measurement. It may happen that aberrant data stem from a change in unit during data entry. If the field data record form mentions 1.2 kg then 900 g for leaf biomass measurements, care must be taken to enter 1.2 and 0.9 (en kg), or 1200 and 900 (in g), not 1.2 and 900. Inconsistent data are more difficult to detect as they require the comparison of several variables one with another. In the example given above, it is perfectly normal for a tree to have a trunk biomass of 755 kg, as it is for a tree to have a total biomass of 440 kg, but obviously the two values cannot be simultaneously correct for the same tree. Likewise, there is nothing abnormal about a tree having a diameter of 5 cm, or a tree being 40 m high. It is simply that a tree with a diameter of 5 cm and a height of 40 m is abnormal.

Aberrant data and data inconsistencies can be detected using descriptive statistics and by plotting two variables one against the other: an examination of means, quantiles, and maximum and minimum values, will often detect aberrant data; plots of two variables one against the other can be used to detect inconsistent data. For instance, in the above example, we could plot total biomass against trunk biomass, and check that all the points are located above the line $y = x$. Plots of height against diameter, volume against diameter, etc. can also be used to detect abnormal data. The categories describing qualitative variables could also be inspected by counting the number of observations per category. Two qualitative variables could then be crossed by building the corresponding contingency table. It should be assumed, during this inspection, that the statistical software has correctly interpreted numerical data as numerical data and qualitative data as qualitative data.

False categories often stem from sloppy data entry. Involuntary spelling mistakes are often made when entering the names of species in full, and this generates false categories. False categories may also be very ambiguous and difficult to correct. Let us consider the example of a dataset of trees in Central African wet tropical forest which, among other things, contains two species: alombi (*Julbernardia seretii*) and ilomba (*Picnanthus angolensis*). Let us assume that, by mistake, the false category "alomba" was entered. This false category differs from the true alombi and ilomba categories by only one letter. But which is the right category? Accents are also often the source of false categories, depending on whether text was entered with or without accents. Let us for example consider the entering of a color: for the person entering the data (in French) it may be perfectly clear that "vert foncé" and "vert fonce" are the same thing, but the software will consider these as two different categories. Masculines and feminines (again in French) may also pose a problem. "Feuille vert clair" and "feuille verte clair" will be understood by the software as two different categories. A false category often found is generated by a space. The "green" category (no space) and the "green␣" category (with a space, shown here by "␣") will be understand by the software as two different categories. This false category is particularly difficult to detect as the space is invisible on the screen and the user therefore has the impression that the two categories are the same. All invisible characters (carriage return, tabulation, etc.) and characters that appear in the same fashion on the screen but have different ASCII codes, may generate the same type of confusing error.

False categories can be avoided by using data entry templates that restrict the entering of qualitative variable to a choice among admissible categories. Automatic data cleansing scripts, that remove incorrect spaces, check accents or letter upper or lower case, and check that qualitative variables take their value from a list of admissible categories, are a necessity for large datasets.

4.3 Data formatting

This formatting consists in organizing the data into a format suitable for the calculations needed to construct the model. Typically, this consists of a model with one line per statistical individual (a tree for an individual model, a plot for a stand table) and as many columns as there are descriptive variables (i.e. variables predicting biomass, volume...), and as there are effect variables (dbh, height...). This formatting phase may require advanced data handling techniques. In some cases the data will need to be aggregated from one descriptive level to another. For example, if we are looking to construct an individual model for a coppice, and the measurements were made on stems, the data concerning the stems of a given stump will need to be aggregated: add together the volumes and masses, and calculate the equivalent diameter (i.e. the quadratic mean) of the stump from the diameters of its stems. Another example is the construction of a stand model from measurements in individual trees. In this case the data regarding the trees need to be aggregated into data characterizing the stand (volume per hectare, dominant height, etc.). In the opposite case, datasets need to be split. For example, we have calculated the volume of trees selected at random in a multispecific stand, and we are looking to construct a separate model for the five dominant species: the dataset needs to be split on the basis of tree species.

This data formatting will be greatly facilitated if the data were initially entered in an appropriate format. In this matter, relational databases have the advantage of offering a search language that can be used to construct the appropriate tables easily. In Microsoft Excel, the notion of "pivot table" can also usefully be employed to format data.

Red line dataset

To illustrate specific points in this guide we will be using a dataset collected in Ghana by Henry *et al.* (2010). This dataset gives the dry biomass of 42 trees belonging to 16 different species in a wet tropical forest. The diameter at breast height of each tree was measured, together with its height, the diameter of its crown, the mean density of its wood, its volume, and its dry biomass in five compartments: branches, leaves, trunk, buttresses, and total biomass.

Table 4.2 gives the data Henry *et al.* (2010) as they should be presented in a spreadsheet. The data table in the spreadsheet takes the form of a data rectangle; it should not contain any empty lines or empty columns, and its layout must not deviate from this matrix-like format. Decorative effects, indentations, cells left empty to "lighten" its appearance are to be avoided as the statistical software will be unable to read a dataset that deviates from a matrix format. Column headings should be reduced to short words, or even abbreviations. Information on the significance of these variables, and their units, should be presented separately in the meta-information.

If information needs to be entered about species, this should be entered in a second table as this involves two nested levels: the species level, with several trees per species, and the tree level nested in the species level. Thus, if we are looking to enter the ecological guild and the vernacular name of the species, this will generate a second Table 4.3 specific to the species where the scientific name of the species is used to create the link between Table 4.2 and Table 4.3.

Data reading. Let us suppose that the data formatted in the matrix format have been saved in the Excel file `Henry_et_al2010.xls`, the first worksheet of which, called `biomass`, contains Table 4.2. In R software, these data are read using the commands:

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

The data are then stored in the object `dat`.

Data cleansing. A few checks may be performed to verify data quality. In R, the command `summary` generates basic descriptive statistics for the variables in a data table:

```
summary(dat)
```

In particular for diameter at breast height, the result is:

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

Thus tree dbh varies from 2.6 cm to 180 cm, with a mean of 58.59 cm and a median of 59.25 cm. Basic descriptive statistics for total dry biomass are as follows:

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

The total dry biomass of the largest tree is 70.24 tonnes. The dry biomass of the smallest tree in the dataset is zero. As biomass is expressed in tonnes to two decimal places, this zero value is not aberrant, it means simply that the dry biomass of this tree is less than 0.01 tonnes = 10 kg. This zero value will nevertheless pose a problem in the future when we want to log-transform the data.

Finally, we can check that the total dry biomass is indeed the sum of the biomasses of the four other compartments:

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

The greatest difference in absolute value is 0.01 tonnes, which indeed corresponds to the precision of the data (two significant figures). The dataset therefore does not contain any inconsistencies at this level.

Table 4.2 – *Tree biomass data from Henry et al. (2010) in Ghana.* `dbh` *is diameter at breast height in cm,* `heig` *is height in m,* `houp` *is the diameter of the crown in m,* `dens` *is mean wood density in* $g\,cm^{-3}$, `volume` *is volume in* m^3, `Bbran` *is the dry biomass of the branches in tonnes,* `Bfol` *is the dry biomass of the foliage in tonnes,* `Btronc` *is the dry biomass of the trunk in tonnes,* `Bctf` *is the dry mass of the buttresses in tonnes, and* `Btot` *is total dry biomass in tonnes.*

species	dbh	heig	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 – *Data on the species sampled by Henry et al. (2010) in Ghana.*

guild	species	vernacular
non-pioneer light demanding	Heritiera utilis	Nyankom
non-pioneer light demanding	Tieghemella heckelii	Makore
non-pioneer light demanding	Piptadeniastrum africanum	Dahoma
non-pioneer light demanding	Aubrevillea kerstingii	Dahomanua
non-pioneer light demanding	Afzelia bella	Papao-nua
pioneer	Cecropia peltata	Odwuma
pioneer	Ceiba pentandra	Onyina
pioneer	Nauclea diderrichii	Kusia
pioneer	Daniellia thurifera	Sopi
shade bearer	Guarea cedrata	Kwabohoro
shade bearer	Strombosia glaucescens	Afena
shade bearer	Garcinia epunctata	Nsokonua
shade bearer	Drypetes chevalieri	Katreka
shade bearer	Cola nitida	Bese
shade bearer	Nesogordonia papaverifera	Danta
shade bearer	Dialium aubrevilliei	Dua bankye

5

Graphical exploration of the data

A graphic exploration of the data is the first step in their analysis. It consists in visually assessing relations between variables in order to gain an idea of the type of model to be fitted. In practice, points are plotted on a graph as a cluster, the coordinates of which correspond to two variables: effect variable on the x axis and response variable on the y axis. A graph can be plotted only for at most two variables simultaneously (three-dimensional graphs cannot in practice be assessed visually). To explore graphically the relations between p variables (where $p > 1$), a total of $p(p-1)/2$ graphs should be plotted of one variable against the other and/or efforts should be made to construct combined effect variables from several effect variables (we will return to this point in § 5.1.1).

Let us now suppose that we have one response variable Y (volume, biomass...) and p effect variables $X_1, X_2, \ldots, X_p$ (dbh, height...). The aim of the graphical exploration is not to select from among the p effect variables those that will in fact be used for the model: when we select variables we need to be able to test the significant nature of a variable, which only comes later in the next phase of model fitting. The p effect variables are therefore considered as fixed, and we look to determine the form of model that best links variable Y to variables X_1 to X_p. A model is made up of two terms: the mean and the error (or residual). The graphical exploration aims to specify both the form of the mean relation and the form of the error, without any concern for the value of model parameters (dealt with in the next model-fitting step). The mean relation may be linear or non linear, linearizable or not; the residual error may be additive or multiplicative, of constant variance (homoscedasticity) or not (heteroscedasticity). By way of an illustration, Figure 5.1 shows four examples where the relation is linear or not, and the variance of the residuals is constant or not.

The graphical exploration of the data phase is also necessary to avoid falling into the traps of "blind" fitting: the impression can be gained that the model accurately fits the data, whereas in fact this result is an artifact. This is illustrated in 5.2 for a linear relation. In all four cases shown in this figure the R^2 of the linear regression of Y against X is high, while in actual fact the linear relation $Y = a + bX + \varepsilon$ does not fit the data. In Figure 5.2A, the cluster of points is split into three subsets, each of which shows a linear relation between Y and X with a negative correlation coefficient. But the three subsets are located along a line that has a positive slope, i.e. that reflecting the linear regression. In 5.2B, the cluster of points, except one outlier (very likely an aberrant value), does not show any relation

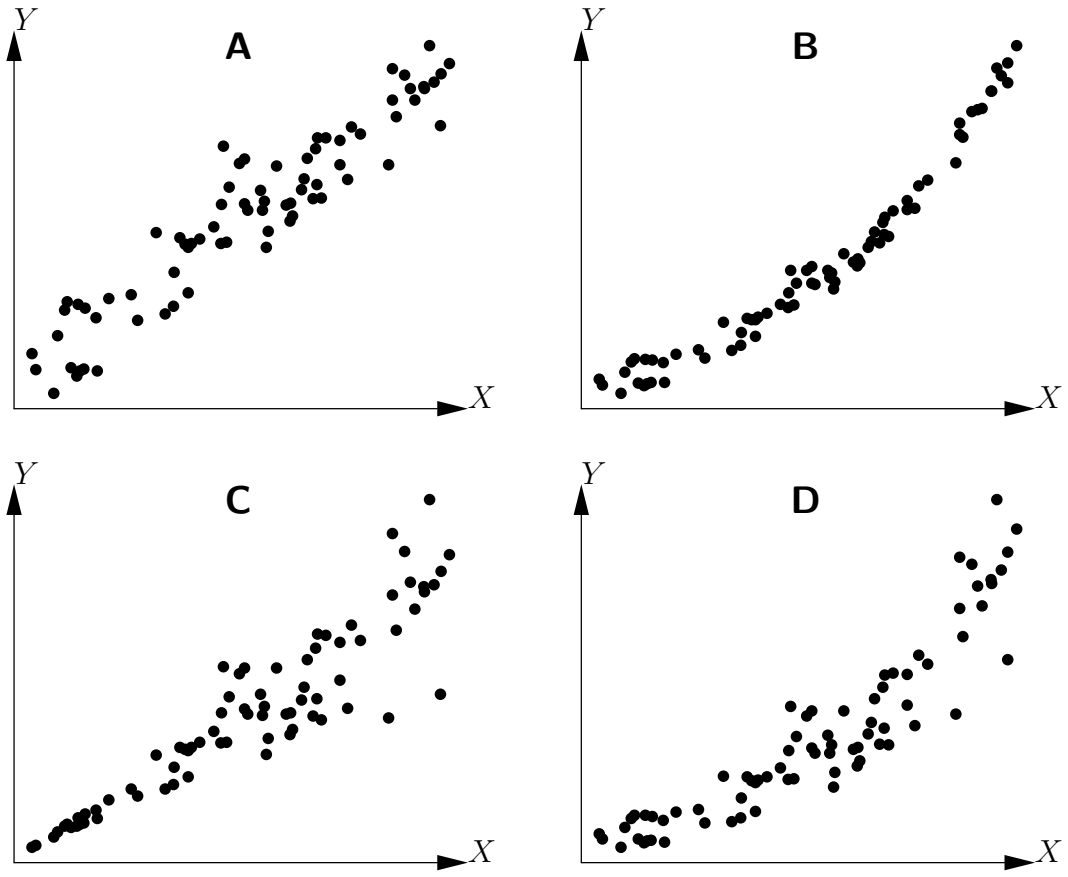


Figure 5.1 – *Example of relations between two variables X and Y: linear relation and constant variance of the residuals, (B) non linear relation and constant variance of the residuals, (C) linear relation and non constant variance of the residuals, (D) non linear relation and non constant variance of the residuals.*

between Y and X. But the outlier is sufficient to make it look as though there is a positive relation between Y and X. In 5.2C, the relation between Y and X is parabolic. Finally, in 5.2D, the cluster of points, except one outlier, is located along a line with a positive slope. In this case, a linear relation between Y and X is well suited to describing the data, once the outlier has been removed. This outlier artificially reduces the value of R^2 (contrary to the situation in Figure 5.2B where the outlier artificially increases R^2).

As indicated by its name, this graphical exploration phase is more an exploration than a systematic method. Although some advice can be given in matters of finding the right model, the process requires both experience and intuition.

5.1 Exploring the mean relation

Here in this section we will focus on the graphical method used to determine the nature of the mean relation between two variables X and Y, i.e. on how to find the form of the function f (if it exists !) such that $\mathrm{E}(Y) = f(X)$. When there is only one effect variable X, the graphical exploration consists in plotting the cluster of points Y against$X$.

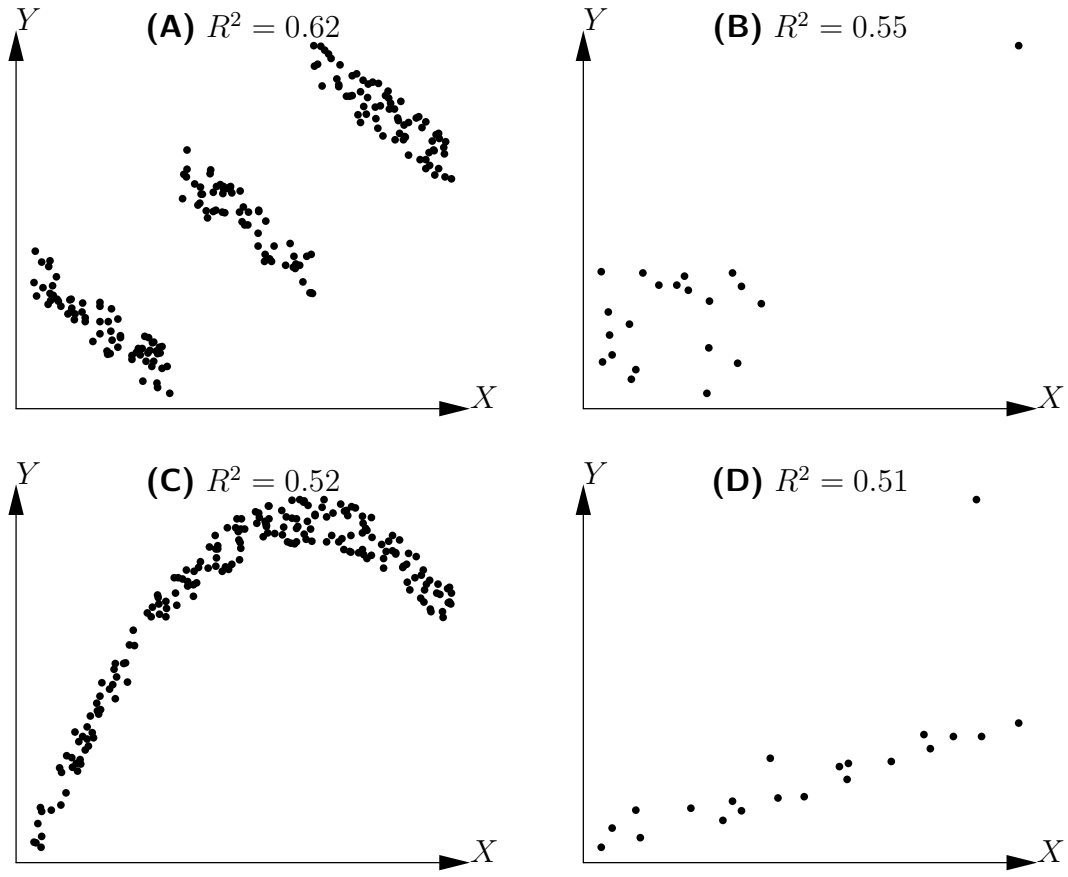


Figure 5.2 – *Determination coefficients (R^2) of linear regressions for clusters of points not showing a linear relation.*

②

Exploring the biomass-dbh relation

To visualize the form of the relation between total dry biomass and dbh, the biomass cluster of points should be plotted against dbh. As the dataset has already been read (see red line 1), the following command can be used to plot the cluster of points:

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

The resulting scatter plot is shown in Figure 5.3. This cluster of points is of the same type as that in Figure 5.1D: the relation between biomass and dbh is not linear, and the variance of the biomass increases with dbh.

As the scatter plot graphical method is applicable only to one effect variable, we should attempt to achieve this when there are several effect variables. Let us first explain this point.

5.1.1 When there is more than one effect variable

The first step is to determine whether it is possible to form a single combined effect variable out of the several effect variables considered. For example, if we are looking to predict trunk

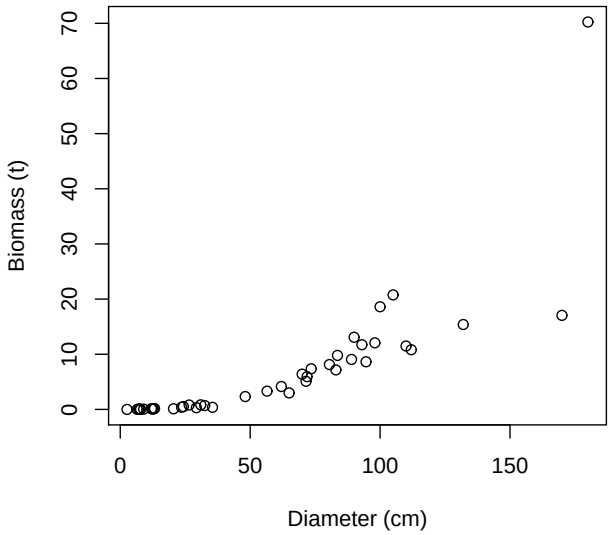


Figure 5.3 – *Scatter plot of total dry biomass (tonnes) against diameter at breast height (cm) for the 42 trees measured in Ghana by Henry et al. (2010).*

volume from its dbh D and its height H, we can be sure that the new variable D^2H will be an effective predictor. We have therefore taken two effect variables, D and H and used them to form a new (single!) effect variable, D^2H. As an illustration of this, Louppe *et al.* (1994) created the following individual volume model for *Afzelia africana* in the Badénou forest reserve in Côte d'Ivoire:

$$V = -0.0019 + 0.04846C^2H$$

where V is total volume in m^3, C is circumference at 1.30 m in m and H is height in m. Although this table is double-entry (girth and height), it in fact contains only one effect variable: C^2H. Another example is the stand volume model established by Fonweban & Houllier (1997) in Cameroon for *Eucalyptus saligna*:

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

where V is stand volume in $m^3\,ha^{-1}$, G is basal area in $m^2\,ha^{-1}$, H_0 is the height of the dominant tree in the stand, N is stand density (number of stems per hectare) and β are constants. Although this table is triple-entry (basal area, dominant height and density), it in fact contains only two effect variables: G and the ratio H_0/N.

③

Exploring the biomass–D^2H relation

Compared with a double-entry biomass model using dbh D and height H, the D^2H data item constitutes an approximation of trunk volume (to within the form factor) and can therefore be used as a combined effect variable. The scatter plot of biomass points against D^2H is obtained by the command:

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

and the result is shown in Figure 5.4. This cluster of points is of the same type as that in Figure 5.1C: the relation between biomass and D^2H is linear, but the variance of the biomass increases with D^2H.

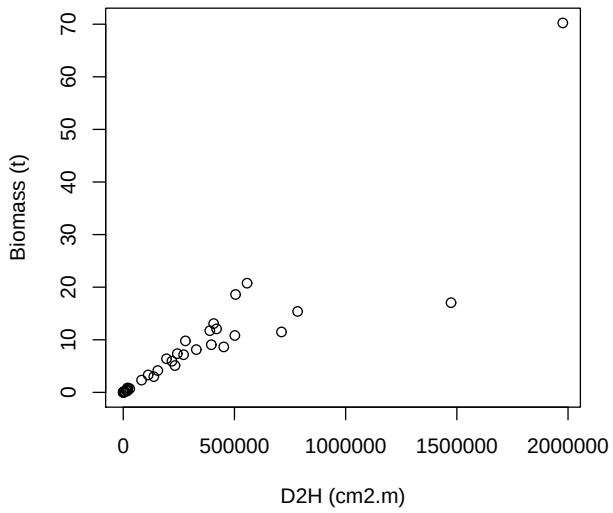


Figure 5.4 – *Scatter plot of total dry biomass (tonnes) against D^2H, where D is diameter at breast height (cm) and H height (m) for the 42 trees measured in Ghana by Henry et al. (2010).*

Assume that, once the effect variables have been aggregated, we are still left with p effect variables $X_1, \ldots, X_p$ (with $p > 1$). We could first explore the p relations between Y and each of the p effect variables. As these take the form of relations between two variables, the graphical methods presented hereafter are perfectly suitable. However, this approach often yields little information as the relation between Y and p variables does not boil down to the p relations between Y and each of the p variables separately. A simple example can clearly illustrate this: let us suppose that variable Y is (to within the errors) the sum of two effect variables:

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

where ε is a zero expectation error, and variables X_1 and X_2 are related such that X_1 varies between 0 and $-X_{\max}$ and, for a given X_1, X_2 varies between $\max(0, -X1)$ and $\min(X_{\max}, 1 - X_1)$. Figure 5.5 shows the two plots of Y against each of the effect variables X_1 and X_2 using simulated data for this model (with $X_{\max} = 5$). The cluster of points does not appear to have any particular structure, and no model is therefore detected that generates $\mathrm{E}(Y) = X_1 + X_2$.

When we have two effect variables, one way to solve the problem is by conditioning. This consists in looking at the relation between the response variable Y and one of the effect variables (let us say X_2) conditionally on the values of the other effect variable (in this case X_1). In practice, we divide the dataset into classes based on the value of X_1, then we explore the relation between Y and X_2 in each data subset. If we consider again the previous example, we divide the values of X_1 into 12 large intervals of 0.5 units: the first ranging from -5 to -4.5, the second from -4.5 to -4, etc., up to the last interval that ranges from 0.5 to 1. The dataset shown in Figure 5.5 has therefore been divided into 12 data subsets based on the 12 classes of X_1 values, and 12 plots can be constructed

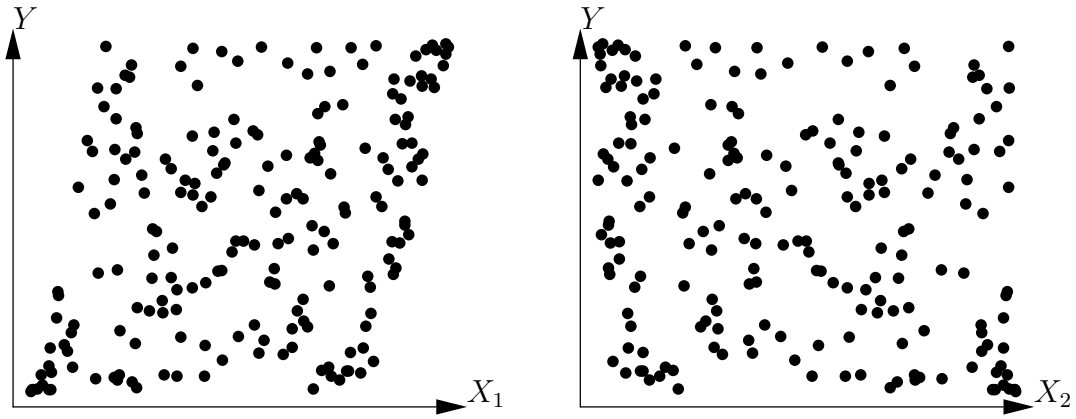


Figure 5.5 – *Plots of variable Y against each of the effect variables X_1 and X_2 such that $E(Y) = X_1 + X_2$.*

of Y against X_2 for each of these 12 data subsets. The result is shown in Figure 5.6. If we superimpose all the plots in Figure 5.6 this will give the plot shown on the right in Figure 5.5. These plots show that, for any given value of X_1, relation between Y and X_2 is indeed linear. Additionally, we can see that the slope of the line of Y plotted against X_2 for any given X_1 has the same constant value for all values of X_1. This graphical exploration therefore shows that the model takes the form:

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

where a is a constant coefficient (in this case it has a value of 1, but the graphical exploration is not concerned with parameter values), and $f(X_1)$ is the y-intercept of the line generated by plotting Y against X_2 for a given value of X_1. This y-intercept potentially varies with X_1, in accordance with function f that remains to be determined.

In order to explore the form of function f, we can plot a linear regression for each of the 12 data subsets of Y and X_2 corresponding to the 12 classes of X_1 values. The y-intercept of each of these 12 lines is then recorded, called y_0, and a graph is plotted of y_0 against the middle of each class of X_1 values. This plot is shown in Figure 5.7 for the same simulated data as previously. This graphical exploration shows that the relation between y_0 and X_1 is linear, i.e.: $f(X_1) = bX_1 + c$. Altogether, the graphical exploration based on the conditioning relative to X_1 showed that an adequate model is:

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

Given that the variables X_1 and X_2 play symmetrical roles in the model (5.1), the conditioning is also symmetrical with regard to these two variables. Here we have studied the relation between Y and X_2 conditionally on X_1, but would have arrived, in the same manner, at the same model had we explored the relation between Y and X_1 conditionally on X_2.

In this example, the relation between Y and X_2 for a given X_1 is a line whose slope is independent of X_2: it may be said that there is no interaction between X_1 and X_2. A model with an interaction would, for example, correspond to: $\mathrm{E}(Y) = X_1 + X_2 + X_1X_2$. In this case, the relation between Y and X_2 at a given X_1 is a line whose slope, which is $1 + X_1$, is indeed dependent on X_1. The conditioning used means that we can without additional difficulty explore the form of models that include interactions between effect variables.

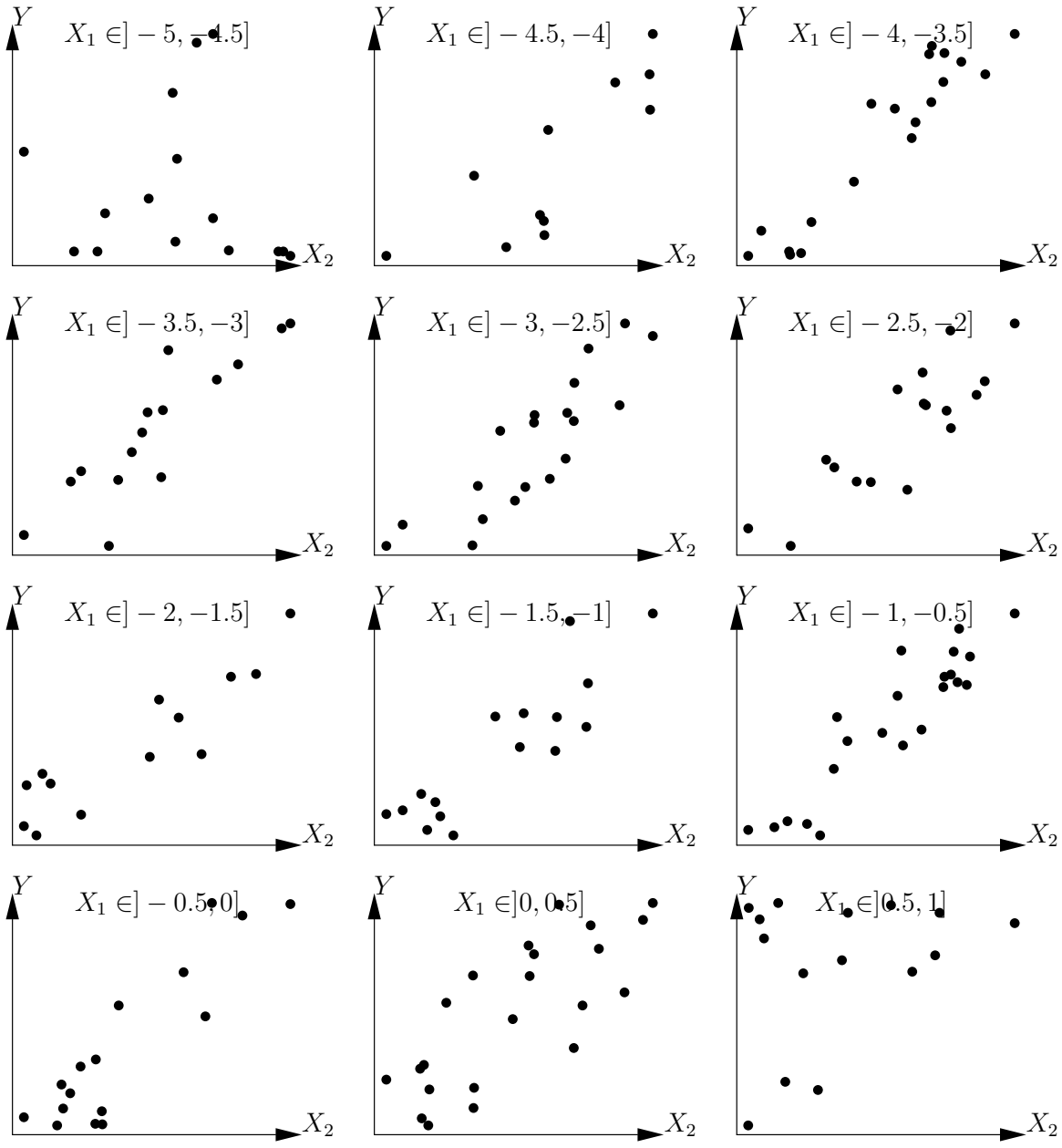


Figure 5.6 – *Plots of a response variable against an effect variable X_2 for each of the data subsets defined by the value classes of another effect variable X_1, with $E(Y) = X_1 + X_2$.*

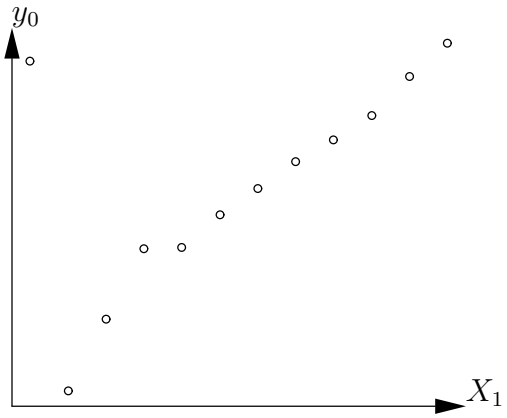


Figure 5.7 – *Plot of the y-intercepts of linear regressions between Y and X_2 for data subsets corresponding to classes of X_1 values against the middle of these classes, for data simulated in accordance with the model $Y = X_1 + X_2 + \varepsilon$.*

This conditioning can, in principle, be extended to any number of effect variables. For three effect variables X_1, X_2, X_3 for example, we could explore the relation between Y and X_3 with X_1 and X_2 fixed and note f the function defining this relation and $\theta(X_1, X_2)$ the parameters of f (that potentially depend on X_1 and X_2):

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

We would then explore the relation between θ and the two variables X_1 and X_2. Again, we would condition by exploring the relation between θ and X_2 with X_1 fixed; and note g the function defining this relation and $\phi(X_1)$ the parameters of g (that potentially depend on X_1):

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

Finally, we would explore the relation between ϕ and X_1; i.e. h the function defining this relation. Ultimately, the model describing the data corresponds to:

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

This reasoning may, in principle, be extended to any number of effect variables, but it is obvious, in practice, that it becomes difficult to implement for $p > 3$. This conditioning requires a great deal of data as each data subset, established by the value classes of the conditional variables, must include sufficient data to allow a valid graphical exploration of the relations between variables. When using three effect variables, the data subsets are established by crossing classes of X_1 and X_2 values (for example). If the full dataset includes n observations, if X_1 and X_2 are divided into 10 value classes, and if the data are evenly distributed between these classes, then each data subset will contain only $n/100$ observations! Therefore, in practice, and unless the dataset is particularly huge, it is difficult to use the conditioning method for more than two effect variables.

The number of entries employed when fitting biomass or volume tables is in most cases limited to at most two or three, meaning that we generally are not faced with the problem of conducting the graphical exploration with a large number of effect variables. Otherwise, multivariate analyses such as the principal components analysis may be usefully employed (Philippeau, 1986; Härdle & Simar, 2003). These analyses aim to project the observations made onto a reduced subspace (usually in two or three dimensions), constructed from linear

combinations of effect variables and in such a manner to maximize the variability of the observations in this subspace. In other words, these multivariate analyses can be used to visualize relations between variables while losing as little information as possible, which is precisely the aim of graphical exploration.

④

Conditioning on wood density

Let us now explore the relation between biomass, D^2H and wood density ρ. Let us define `n` classes of wood density such that each class contains approximately the same number of observations:

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

The object `d` defines the limits of the density classes while object `i` contains the number of the density class to which each observation belongs. The plot of biomass against D^2H on a log scale, with different symbols and colors for the different density classes, may be obtained by the command:

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

and is shown in Figure 5.8 for `n` $= 4$ wood density classes. Pre-empting chapter 6, we will now plot the linear regression between $\ln(B)$ and $\ln(D^2H)$ for each subset of observations corresponding to each wood density class:

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

When we plot the y-intercepts of the regression and its slopes against median wood density for the class:

```
dmid <- (d[-1]+d[-n])/2
plot(dmid,m[1,],xlab="Wood density (g/cm3)",ylab="y-Intercept")
plot(dmid,m[2,],xlab="Wood density (g/cm3)",ylab="Slope")
```

this at first sight does not appear to yield any particular relation (Figure 5.9).

5.1.2 Determining whether or not a relation is adequate

Henceforth, we will assume that we have a single effect variable X and that we are looking to explore the relation between X and the response variable Y. The first step is to plot all the cluster of points on a graph, with X on the x-axis and Y on the y-axis. We then need to guess visually the function that passes through the middle of this cluster of points while following its shape. In fact, the human eye is not very good at discriminating between similar shapes. For example, Figure 5.10 shows three scatter plots corresponding to the three following models, but not in the right order (the error term here has been set at zero):

power model:	$Y = aX^b$
exponential model:	$Y = a\exp(bX)$
polynomial model:	$Y = a + bX + cX^2 + dX^3$

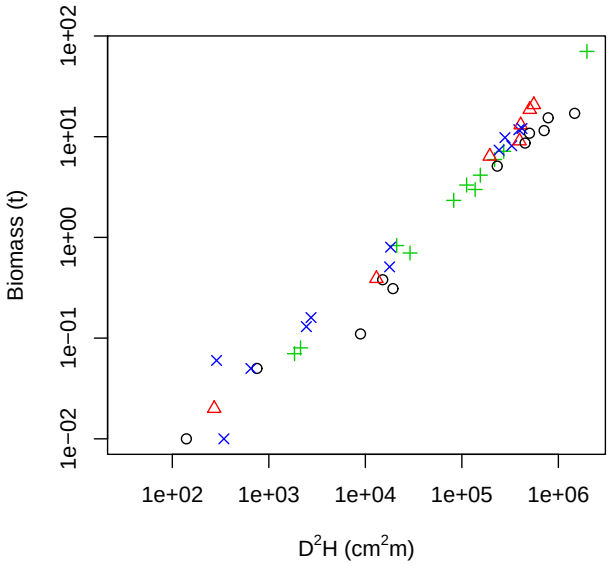


Figure 5.8 – *Scatter plot (log-transformed data) of total dry biomass (tonnes) against D^2H, where D is diameter at breast height (cm) and H is height (m) for the 42 trees measured in Ghana by Henry et al. (2010) with different symbols for the different wood density classes: black circles, $0.170 \leq \rho < 0.545$ g cm^{-3}; red triangles, $0.545 \leq \rho < 0.600$ g cm^{-3}; green plus signs, $0.600 \leq \rho < 0.650$ g cm^{-3}; blue crosses, $0.650 \leq \rho < 0.780$ g cm^{-3}.*

The three scatter plots look fairly similar, and it would be no mean feat to sort out which one corresponds to which model.

By contrast, the human eye is very good at detecting whether a relation is linear or not. To detect visually whether or not the shape of a cluster of points fits a function, it is therefore greatly in our interests when possible to use a variables transformation that renders the relation linear. For instance, in the case of the power model, $Y = aX^b$ yields $\ln Y = \ln a + b \ln X$, and this variables transformation:

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

renders the relation linear. In the case of the exponential model, $Y = a \exp(bX)$ yields $\ln Y = \ln a + bX$, and this variables transformation:

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

renders the relation linear. By contrast, neither of these two transformation is able to render the polynomial relation linear. If we apply these transformations to the data presented in Figure 5.10, we should be able to detect which of the scatter plots corresponds to which model. Figure 5.11 shows the three scatter plots after transforming the variables as in (5.3). The first scatter plot takes the form of a straight line while the two others are still curves. The scatter plot on the left of Figure 5.10 therefore corresponds to the exponential model.

Figure 5.12 shows the three scatter plots after transforming the variables as in (5.2).The second scatter plot takes the form of a straight line while the two others are still curves. The scatter plot in the center of Figure 5.10 therefore corresponds to the power model. By deduction, the scatter plot on the right of Figure 5.10 corresponds to the polynomial model.

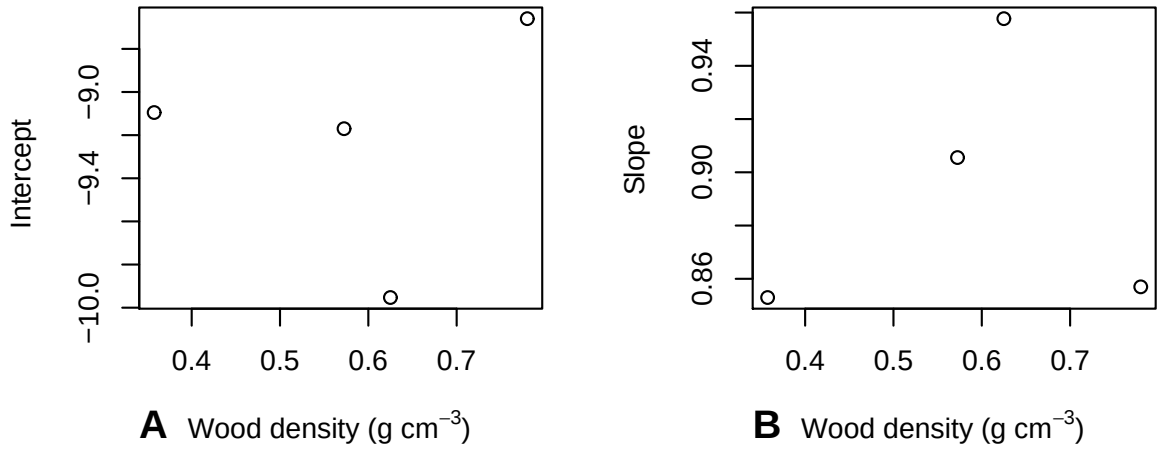


Figure 5.9 – *Y-intercept a and slope b of the linear regression $\ln(B) = a + b\ln(D^2H)$ conditioned on wood density class, against the median wood density of the classes. The regressions were plotted using the data from the 42 trees measured by Henry et al. (2010) au Ghana.*

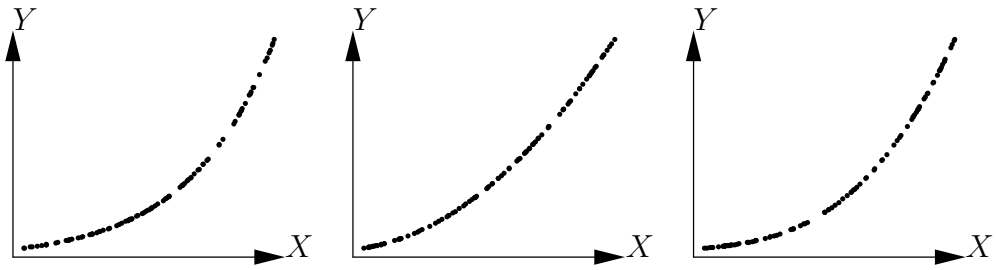


Figure 5.10 – *Three scatter points corresponding to three models: power model, exponential model and polynomial model, but not in the right order.*

It is not always possible to find a variable transformation that renders the relation linear. This, for instance, is precisely the case of the polynomial model $Y = a + bX + cX^2 + dX^3$: no transformation of X into X' and of Y into Y' was found such that the relation between Y' and X' is a straight line, whatever the coefficients a, b, c et d. It must also be clearly understood that the linearity we are talking about here is that of the relation between the response variable Y and the effect variable X. It is not linearity in the sense of the linear model, which describes linearity with regards to model coefficients (thus, the model $Y = a + bX^2$ is linear in the sense of the linear model whereas it describes a non-linear relation between X and Y).

When no transformation of the variable is able to render the relation between X and Y linear, the best approach is to fit the model and determine visually whether the fitted model passes through the middle of the cluster of points and follows its shape. Also, in this case, we have much to gain by examining the plot of the residuals against the fitted values.

⑤

Exploring the biomass–dbh relation: variables transformation

Let us use the log transformation to transform simultaneously both dbh and biomass.

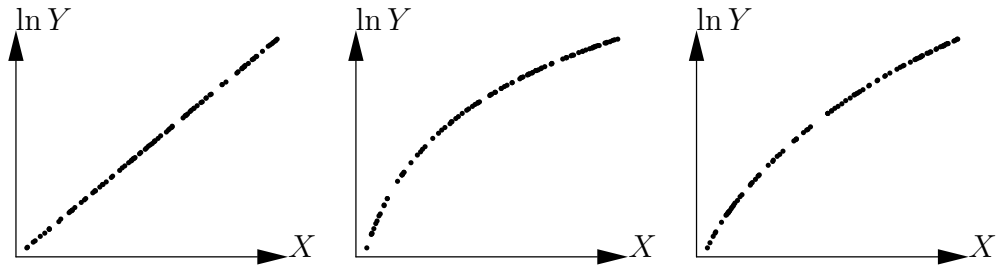


Figure 5.11 – *Application of the variables transformation* $X \to X$, $Y \to \ln Y$ *to the scatter plots given in 5.10.*

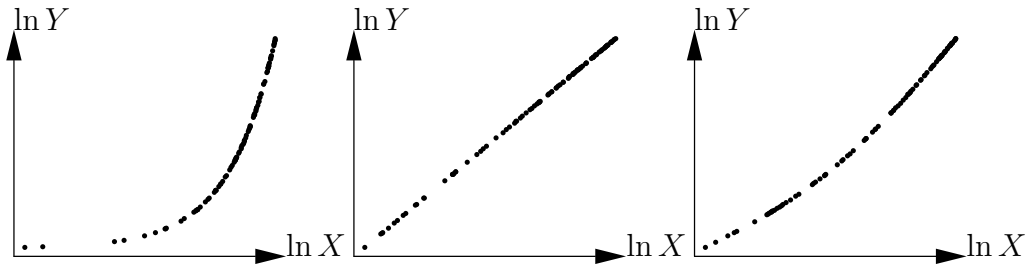


Figure 5.12 – *Application of the variables transformation* $X \to \ln X$, $Y \to \ln Y$ *to the scatter plots shown in 5.10.*

The scatter plot of the log-transformed data is obtained as follows:

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

The resulting scatter plot is shown in Figure 5.13. The log transformation has rendered linear the relation between biomass and dbh: the relation between $\ln(D)$ and $\ln(B)$ is a straight line and the variance of $\ln(B)$ does not vary with diameter (like in Figure 5.1A).

⑥

Exploring the biomass–D^2H relation: variables transformation

Let us use the log transformation to transform simultaneously both D^2H and biomass. The scatter plot of the log-transformed data is obtained as follows:

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

The resulting scatter plot is shown in Figure 5.14. The log transformation has rendered linear the relation between biomass and D^2H: the relation between $\ln(D^2H)$ and $\ln(B)$ is a straight line and the variance of $\ln(B)$ does not vary with D^2H (like in Figure 5.1A).

5.1.3 Catalog of primitives

If we consider the tables calculated by Zianis *et al.* (2005) for Europe, by Henry *et al.* (2011) for sub-Saharan Africa, or more specifically by Hofstad (2005) for southern Africa, this gives

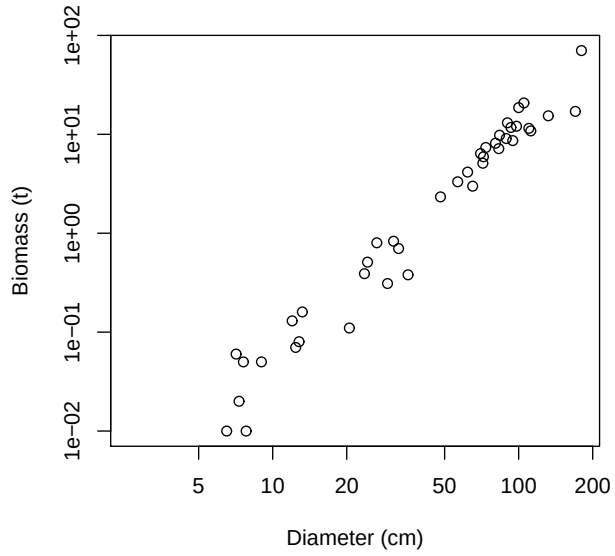


Figure 5.13 – *Scatter plot (log-transformed data) of total dry biomass (tonnes) against diameter at breast height (cm) for the 42 trees measured in Ghana by Henry et al. (2010).*

an idea as to the form of the models most commonly found in the literature on biomass and volume tables. This shows that two models are often used: the power model and the (second, or at most, third order) polynomial model. These two models are therefore useful starting points for the graphical exploration of data when looking to construct a volume or biomass model. The power model $Y = aX^b$ is also known as an allometric relation and the literature contains a number of biological interpretations of this model (Gould, 1979; Franc *et al.*, 2000, § 1.1.5). In particular, the "metabolic scaling theory" (Enquist *et al.*, 1998, 1999; West *et al.*, 1997, 1999) predicts in a theoretical manner, and based on a fractal description of the internal structure of trees, that the biomass of a tree is related to its dbh by a power relation where the exponent is $8/3 \approx 2.67$:

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

where ρ is specific wood density. Although metabolic scaling theory has been widely criticized (Muller-Landau *et al.*, 2006), it does at least have the merit of laying a biological basis for this often observed power relation.

In addition to this power model $B = aD^b$ and the second-order polynomial model $B = a_0 + a_1 D + a_2 D^2$, and without pretending in any way to be exhaustive, the following biomass tables are also often found (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. double-entry model of the power form in relation to the variable D^2H: $B = a(\rho D^2 H)^b$
2. double-entry model: $\ln(B) = a_0 + a_1 \ln(D) + a_2 \ln(H) + a_3 \ln(\rho)$
3. single-entry model: $\ln(B) = a_0 + a_1 \ln(D) + a_2[\ln(D)]^2 + a_3[\ln(D)]^3 + a_4 \ln(\rho)$,

where ρ is specific wood density. To within the form factor, the variable D^2H corresponds to the volume of the trunk, which explains why it is so often used as effect variable. The

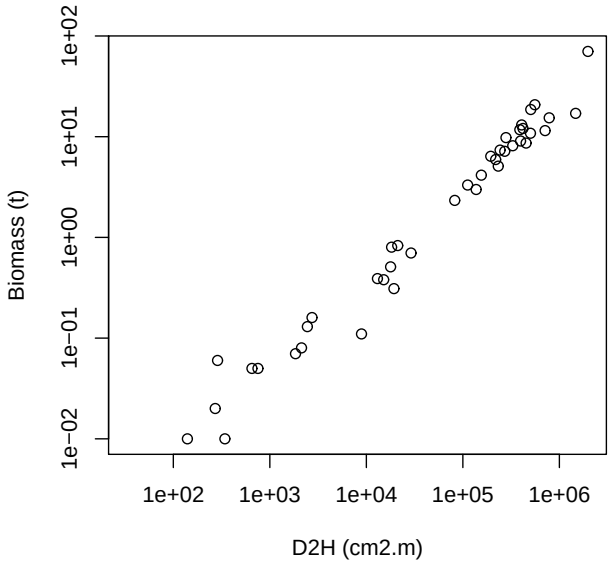


Figure 5.14 – *Scatter plot (log-transformed data) of total dry biomass (tonnes) against D^2H, where D is diameter at breast height (cm) and H is height (m) for the 42 trees measured in Ghana by Henry et al. (2010).*

second equation may be considered to be a generalization of the first for if we apply the log transformation, the first equation is equivalent to: $\ln(B) = \ln(a)+2b\ln(D)+b\ln(H)+b\ln(\rho)$. The first equation is therefore equivalent to the second in the particular case where $a_2 = a_3 = a_1/2$. Finally, the third equation may be considered to be a generalization of the power model $B = aD^b$.

More generally, Table 5.1 gives a number of functions that are able to model the relation between the two variables. It is indicated when variables have been transformed to render linear the relation between X and Y. It should be noted that the modified power model is simply a re-writing of the exponential model, and that the root model is simply a re-writing of the modified exponential model. It should also be noted that a large proportion of these models are simply special cases of more complex models that contain more parameters. For example, the linear model is simply a special case of the polynomial model, and the Gompertz model is simply a special case of the Sloboda model, etc.

The p order polynomial model should be used with caution as polynomials are able to fit any form provided the p order is sufficiently high (the usual functions can all be decomposed in a polynomials database: this is the principle of limited development). In practice, we may have a polynomial that very closely follows the form of the cluster of points across the range of values available, but which assumes a very unlikely form outside this range. In other words, the polynomial model may present a risk when extrapolating, and the higher the p order, the greater the risk. In practice, therefore, fitting greater than 3rd order polynomials should be avoided.

5.2 Exploring variance

Let us now consider the model's error term ε that links the response variable Y to an effect variable X. Exploring the form of the variance corresponds basically to answering the question: is the variance of the residuals constant (homoscedasticity) or not (heteroscedas-

Table 5.1 – *A few models linking two variables.*

Name	Equation	Transformation
Polynomial models		
linear	$Y = a + bX$	same
parabolic ou quadratic	$Y = a + bX + cX^2$	
p-order polynomial	$Y = a_0 + a_1X + a_2X^2 + \ldots + a_pX^p$	
Exponential models		
exponential or Malthus	$Y = a\exp(bX)$	$Y' = \ln Y$, $X' = X$
modified exponential	$Y = a\exp(b/X)$	$Y' = \ln Y$, $X' = 1/X$
logarithm	$Y = a + b\ln X$	$Y' = Y$, $X' = \ln X$
reciprocal log	$Y = 1/(a + b\ln X)$	$Y' = 1/Y$, $X' = \ln X$
Vapor pressure	$Y = \exp(a + b/X + c\ln X)$	
Power law models		
power	$Y = aX^b$	$Y' = \ln Y$, $X' = \ln X$
modified power	$Y = ab^X$	$Y' = \ln Y$, $X' = X$
shifted power	$Y = a(X - b)^c$	
geometric	$Y = aX^{bX}$	$Y' = \ln Y$, $X' = X\ln X$
modified geometric	$Y = aX^{b/X}$	$Y' = \ln Y$, $X' = (\ln X)/X$
root	$Y = ab^{1/X}$	$Y' = \ln Y$, $X' = 1/X$
Hoerl's model	$Y = ab^X X^c$	
modified Hoerl's model	$Y = ab^{1/X}X^c$	
Production-density models		
inverse	$Y = 1/(a + bX)$	$Y' = 1/Y$, $X' = X$
quadratic inverse	$Y = 1/(a + bX + cX^2)$	
Bleasdale's model	$Y = (a + bX)^{-1/c}$	
Harris's model	$Y = 1/(a + bX^c)$	
Growth models		
saturated growth	$Y = aX/(b + X)$	$Y' = X/Y$, $X' = X$
mononuclear or Mitscherlich's model	$Y = a[b - \exp(-cX)]$	
Sigmoidal models		
Gompertz	$Y = a\exp[-b\exp(-cX)]$	
Sloboda	$Y = a\exp[-b\exp(-cX^d)]$	
logistic or Verhulst	$Y = a/[1 + b\exp(-cX)]$	
Nelder	$Y = a/[1 + b\exp(-cX)]^{1/d}$	
von Bertalanffy	$Y = a[1 - b\exp(-cX)]^3$	
Chapman-Richards	$Y = a[1 - b\exp(-cX)]^d$	
Hossfeld	$Y = a/[1 + b(1 + cX)^{-1/d}]$	
Levakovic	$Y = a/[1 + b(1 + cX)^{-1/d}]^{1/e}$	
multiple multiplicative factor	$Y = (ab + cX^d)/(b + X^d)$	
Johnson-Schumacher	$Y = a\exp[-1/(b + cX)]$	
Lundqvist-Matérn ou de Korf	$Y = a\exp[-(b + cX)^d]$	
Weibull	$Y = a - b\exp(-cX^d)$	
Miscellaneous models		
hyperbolic	$Y = a + b/X$	$Y' = Y$, $X' = 1/X$
sinusoidal	$Y = a + b\cos(cX + d)$	
heat capacity	$Y = a + bX + c/X^2$	
Gaussian	$Y = a\exp[-(X - b)^2/(2c^2)]$	
rational fraction	$Y = (a + bX)/(1 + cX + dX^2)$	

ticity)? The reply to this question depends implicitly on the precise form of the relation that will be used to fit the model. For example, for the power relation $Y = aX^b$, we can

- either fit the non-linear model $Y = aX^b + \varepsilon$, which corresponds to estimating parameters a and b directly;
- or fit the linear model $Y' = a' + bX' + \varepsilon$ to the transformed data $Y' = \ln Y$ and $X' = \ln X$, which corresponds to estimating parameters $a' = \ln a$ and b.

These two options are obviously not interchangeable as the error term ε (which we assume follows a normal distribution of constant standard deviation) does not play the same role in the two cases. In the first case the error is additive with respect to the power model. In the second case the error is additive with respect to the linearized model. Therefore, if we return to the power model:

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

which corresponds to a multiplicative error with respect to the power model, where ε' follows a lognormal distribution. The difference between these two options is illustrated in Figure 5.15. The additive error is expressed by a constant variance in the plot (a) of Y against X and by a variance that decreases with X in the plot (c) of the same, but log-transformed, data. The multiplicative error is expressed by a variance that increases with X in the plot (b) of Y against X and by a constant variance in the plot (d) of the same, but log-transformed, data.

Thus, when the model linking Y to X is linearized by a variables transformation, this affects both the form of the mean relation and that of the error term. This property may be used to stabilize residuals that vary with X rendering them constant, but this point will be addressed in the next chapter. For the moment, we are looking to explore the form of the error $Y - \mathrm{E}(Y)$ in relation to X, without transforming the variables X and Y.

As the form of the mean relation $\mathrm{E}(Y) = f(X)$ has already been determined graphically, a simple visual examination of the plot of Y against X will show whether the points are evenly distributed on each side of the curve f for all values of X. Plots (a) and (b) in Figure 5.1, for example, illustrate the case of residuals of constant variance at all values of X, whereas plots (c) and (d) in the same figure show the case of residuals that increase in variance with X. More complex relations, like that illustrated in Figure 5.16, may also be imagined. In the particular case illustrated by Figure 5.16, the variance of the residuals fluctuates periodically with X. However, there is little risk of encountering such situations in practice in the context of biomass or volume tables. In almost all cases we need to choose between two situations: the variance of the residuals is constant or increases with X. In the first case, nothing further needs to be done. In the second case, there is no need to investigate the exact form of the relation between X and the variance of the residuals, but simply from the outset adopt a power model to link the variance of the residuals ε to X:

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

The values of the coefficients α and β can be estimated at the same time as those of the model's other coefficients during the model fitting phase, which will be considered in the next chapter.

5.3 Exploring doesn't mean selecting

By way of a conclusion, it should be underlined that this graphical exploration does not aim to select a single type of model, but rather to sort between models that are able to describe

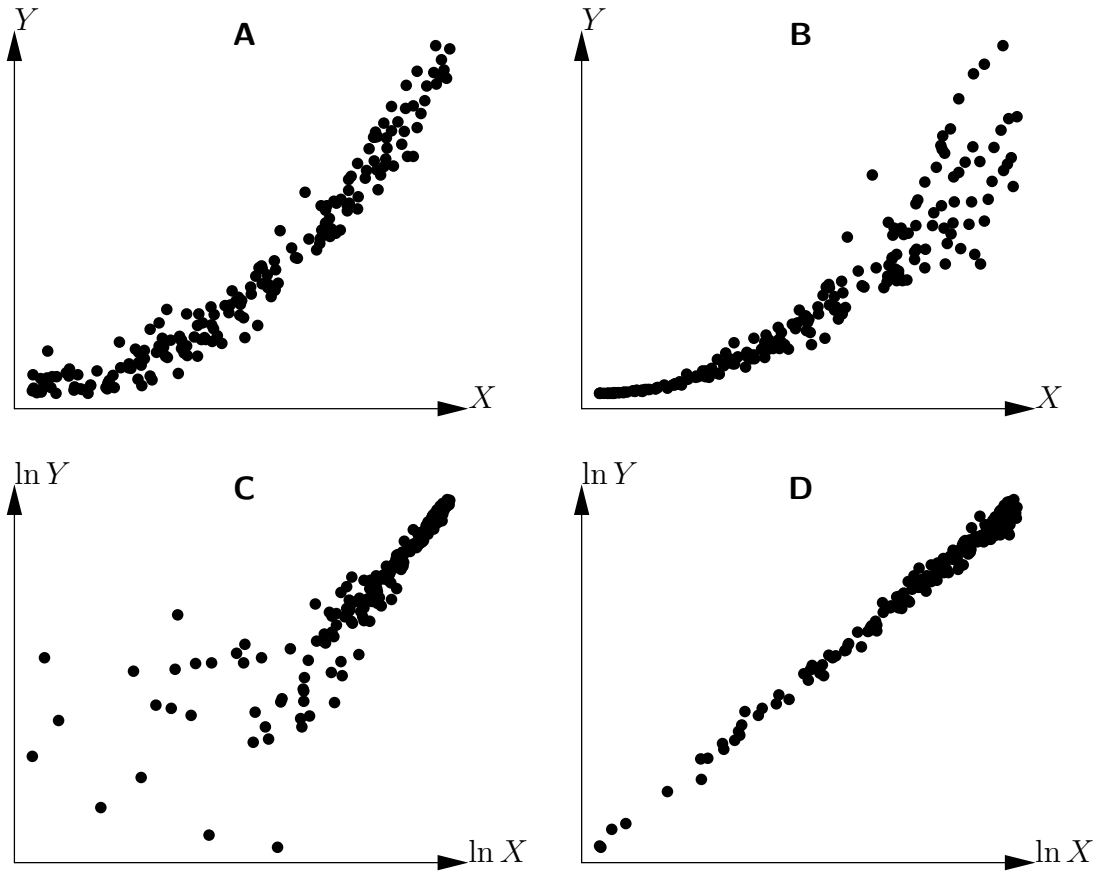


Figure 5.15 – *Power model with additive (A and C) or multiplicative (B and D) error. Plot (C) (respectively D) results from plot (A) (respectively B) by transformation of the variables* $X \to \ln X$ *and* $Y \to \ln Y$.

the dataset and those that are not. Far from seeking to select THE model assumed to be the best at describing the data, efforts should be made to select three or four candidate models likely to be able to describe the data. The final choice between these three or four models selected during the graphical exploration phase can then be made after the data fitting phase that is described in the next chapter.

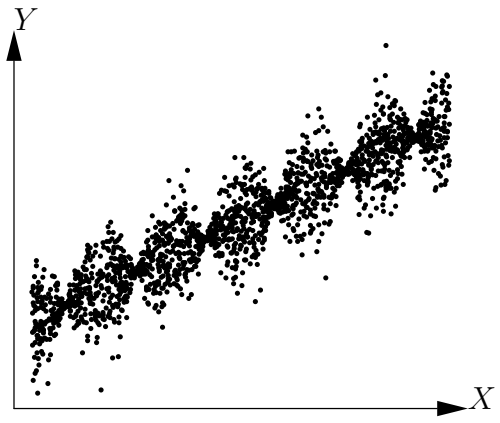


Figure 5.16 – *Scatter plot of points generated by the model* $Y = a + bX + \varepsilon$, *where* ε *follows a normal distribution with a mean of zero and a standard deviation proportional to the cosine of* X.

6
Model fitting

Fitting a model consists in estimating its parameters from data. This assumes, therefore, that data are already available and formatted, and that the mathematical expression of the model to be fitted is known. For example, fitting the power model $B = aD^b$ consists in estimating coefficients a and b from a dataset that gives the values of B_i and D_i from the biomass and dbh of n trees ($i = 1, \ldots, n$). The response variable (also in the literature called the output variable or the dependent variable) is the variable fitted by the model. There is only one. In this guide, the response variable is always a volume or a biomass. Effect variables are the variables used to predict the response variable. There may be several, and their number is denoted by p. Care must be taken not to confuse effect variables and model entry data. The model $B = a(D^2H)^b$ possesses a single effect variable (D^2H) but two entries (dbh D and height H). Conversely, the model $B = a_0 + a_1D + a_2D^2$ possesses two effect variables (D and D^2) but only a single entry (dbh D). Each effect variable is associated with a coefficient to be calculated. To this must be added, if necessary, the y-intercept or a multiplier such that the total number of coefficients to be calculated in a model with p effect variables is p or $p + 1$.

An observation consists of the data forming the response variable (volume or biomass) and the effect variables for a tree. If we again consider the model $B = aD^b$, an observation consists of the doublet (B_i, D_i). The number of observations is therefore n. An observation stems from a measurement in the field. The prediction provided by the model is the value it predicts for the response variable from the data available for the effect variables. A prediction stems from a calculation. For example, the prediction provided by the model $B = aD^b$ for a tree of dbh D_i is $\hat{B}_i = aD_i^b$. There are as many predictions as there are observations. A key concept in model fitting is the residual. The residual, or residual error, is the difference between the observed value of the response variable and its prediction. Again for the same example, the residual of the ith observation is: $\varepsilon_i = B_i - \hat{B}_i = B_i - aD_i^b$. There are as many residuals as there are observations. The lower the residuals, the better the fit. Also, the statistical properties of the model stem from the properties that the residuals are assumed to check *a priori*, in particular the form of their distribution. The type of a model's fitting is therefore directly dependent upon the properties of its residuals.

In all the models we will be considering here, the observations will be assumed to be independent or, which comes to the same thing, the residuals will be assumed to be inde-

pendent: for each $i \neq j$, ε_i is assumed to be independent of ε_j. This independence property is relatively easy to ensure through the sampling protocol. Typically, care must be taken to ensure that the characteristics of a tree measured in a given place do not affect the characteristics of another tree in the sample. Selecting trees that are sufficiently far apart is generally enough to ensure this independence. If the residuals are not independent, the model can be modified to take account of this. For example, a spatial dependency structure could be introduced into the residuals to take account of a spatial auto-correlation between the measurements. We will not be considering these models here as they are far more complex to use.

In all the models we will be looking at, we assume also that the residuals have a normal distribution with zero expectation. The zero mean of the residuals is in fact a property that stems automatically from model fitting, and ensures that the model's predictions are not biased. It is the residuals that are assumed to have a normal distribution, not the observations. This hypothesis in fact causes little constraint for volume or biomass data. In the unlikely case where the distribution of the residuals is far from normal, efforts could be made to fit other model types, e.g. the generalized linear model, but this will not be addressed here in this guide. The hypotheses that the residuals are independent and follow a normal distribution are the first two hypotheses underlying model fitting. We will see a third hypothesis later. It should be checked that these two hypotheses are actually valid. To the extent that these hypotheses concern model residuals, not the observations, they cannot be tested until they have been calculated, i.e. until the model has been fitted. These hypotheses are therefore checked a posteriori, after model fitting. The models we will look at here are also robust with regard to these hypotheses, i.e. the predictive quality of the fitted models is acceptable even when the independence and the normal distribution of the residuals are not perfectly satisfied. For this reason, we will not look to test these two hypotheses very formally. In practice we will simply perform a visual verification of the plots.

6.1 Fitting a linear model

The linear model is the simplest of all models to fit. The word *linear* means here that the model is *linearly* dependent on its coefficients. For example, $Y = a+bX^2$ and $Y = a+b\ln(X)$ are linear models because the response variable Y is linearly dependent upon coefficients a and b, even if Y is not linearly dependent on the effect variable X. Conversely, $Y = aX^b$ is not a linear model because Y is not linearly dependent on coefficient b. Another property of the linear model is that the residual is additive. This is underlined by explicitly including the residual ε in the model's formula. For example, a linear regression of Y against X will be written: $Y = a + bX + \varepsilon$.

6.1.1 Simple linear regression

A simple linear regression is the simplest of the linear models. It assumes (*i*) that there is only one effect variable X, (*ii*) that the relation between the response variable Y and X is a straight line:

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

where a is the y-intercept of the line and b its slope, and (*iii*) that the residuals are of constant variance: $\mathrm{Var}(\varepsilon) = \sigma^2$. For example, the model

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

is a typical simple linear regression, with response variable $Y = \ln(B)$ and effect variable $X = \ln(D)$. It corresponds to a power model for biomass: $B = \exp(a)D^b$. This model is often used to fit a single-entry biomass model. Another example is the double-entry biomass model:

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

The hypothesis whereby the residuals are of constant variance is added to the two independence and normal distribution hypotheses (we also speak of homoscedasticity). These three hypotheses may be summarized by writing:

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

where $\mathcal{N}(\mu,\, \sigma)$ designates a normal distribution of expectation μ and standard deviation σ, the tilde "$\sim$" means "is distributed in accordance with", and "i.i.d." is the abbreviation of "independently and identically distributed".

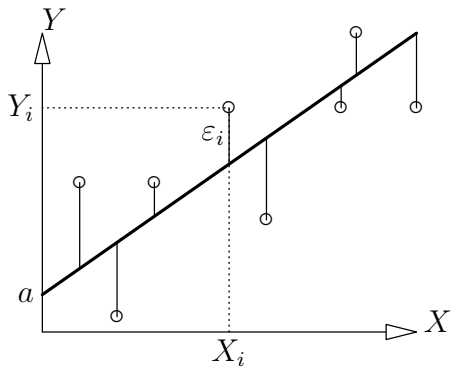


Figure 6.1 – *Actual observations (points) and plot of the regression (thick line) and residuals (thin lines).*

Estimating coefficients

Figure 6.1 shows the observations and the plot of predicted values. The best fit is that which minimizes the residual error. There are several ways of quantifying this residual error. From a mathematical standpoint, this is equivalent with choosing a norm to measure ε and various norms could be used. The norm that is commonly used is the L_2 norm, which is equivalent with quantifying the residual difference between the actual observations and the predictions by summing the squares of the residuals, which is also called the sum of squares (SS):

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

The best fit is therefore that which minimizes SS. In other words, the estimations $\hat{a}$ and $\hat{b}$ of the coefficients a and b are values of a and b that minimize the sum of squares:

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

This minimum may be obtained by calculating the partial derivatives of SS in relation to a and b, and by looking for the values of a and b that cancel these partial derivatives.

Simple calculations yield the following results: $\hat{b} = \widehat{\text{Cov}}(X, Y)/S_X^2$ and $\hat{a} = \bar{Y} - \hat{b}\bar{X}$, where $\bar{X} = (\sum_{i=1}^n X_i)/n$ is the empirical mean of the effect variable, $\bar{Y} = (\sum_{i=1}^n Y_i)/n$ is the empirical mean of the response variable,

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

is the empirical variance of the effect variable, and

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

is the empirical covariance between the effect variable and the response variable. The estimation of the residual variance is:

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

Because this method of estimating the coefficients is based on minimizing the sum of squares, it is called the *least squares* method (sometimes it is called the "ordinary least squares" method to differentiate it from the weighted least squares method we will see in § 6.1.3). This estimation method has the advantage of providing an explicit expression of the estimated coefficients.

Interpreting the results of a regression

When fitting a simple linear regression, several outputs need to be analyzed. The determination coefficient, more commonly called R^2, measures the quality of the fit. R^2 is directly related to the residual variance since:

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

where $S_Y^2 = [\sum_{i=1}^n (Y_i - \bar{Y})^2]/n$ is the empirical variance of Y. The difference $S_Y^2 - \hat{\sigma}^2(n-2)/n$ between the variance of Y and the residual variance corresponds to the variance explained by the model. The determination coefficient R^2 can be interpreted as being the ratio between the variance explained by the model and the total variance. It is between 0 and 1, and the closer it is to 1, the better the quality of the fit. In the case of a simple linear regression, and only in this case, R^2 is also the square of the linear correlation coefficient (also called Pearson's coefficient) between X and Y. We have already seen in chapter 5 (particularly in Figure 5.2) that the interpretation of R^2 has its limitations.

In addition to estimating values for coefficients a and b, model fitting also provides the standard deviations of these estimations (i.e. the standard deviations of estimators $\hat{a}$ and $\hat{b}$), and the results of significance tests on these coefficients. A test is performed on the y-intercept a, which tests the null hypothesis that $a = 0$, and likewise a test is performed on the slope b, which tests the hypothesis that $b = 0$.

Finally, the result given by the overall significance test for the model is also analyzed. This test is based on breaking down the total variance of Y into the sum of the variance explained by the model and the residual variance. Like in an analysis of variance, the test used is Fisher's test which, as test statistic, uses a weighted ratio of the explained variance over the residual variance. In the case of a simple linear regression, and only in this case, the test for the overall significance of the model gives the same result as the test on the null hypothesis $b = 0$. This can be grasped intuitively: a line linking X to Y is significant only if its slope is not zero.

Checking hypotheses

Model fitting may be brought to a conclusion by checking that the hypotheses put forward for the residuals are in fact satisfied. We will not consider here the hypothesis that the residuals are independent for this has already been satisfied thanks to the sampling plan adopted. If there is a natural order in the observations, we could possibly use the Durbin-Watson test to test if the residuals are indeed independent (Durbin & Watson, 1971). The hypothesis that the residuals are normally distributed can be checked visually by inspecting the quantile–quantile graph. This graph plots the empirical quantiles of the residuals against the theoretical quantiles of the standard normal distribution. If the hypothesis that the residuals are normally distributed is satisfied, then the points are approximately aligned along a straight line, as in Figure 6.2 (right plot).

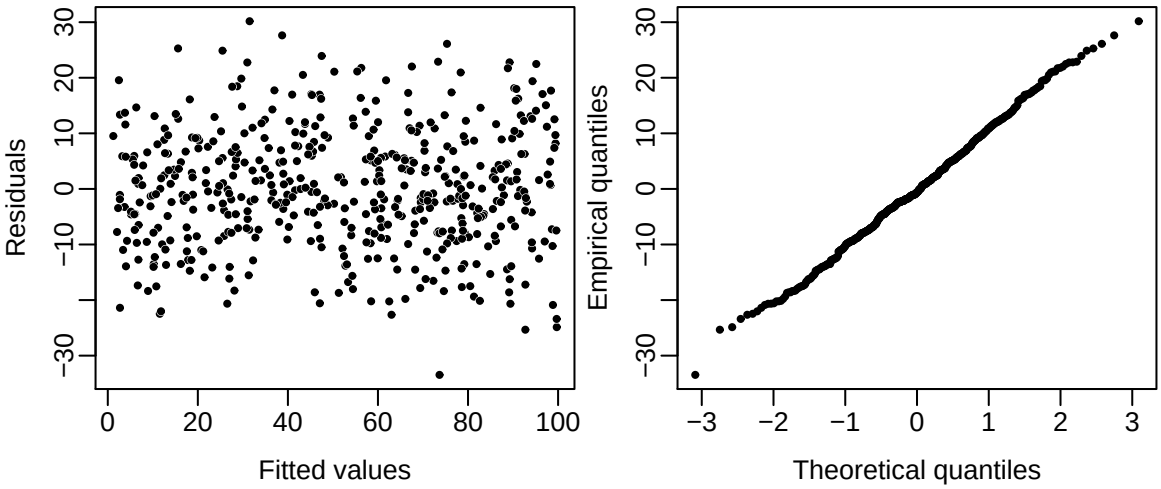


Figure 6.2 – *Residuals plotted against fitted values (left) and quantile–quantile plot (right) when the normal distribution and constant variance of the residuals hypotheses are satisfied.*

In the case of fitting volume or biomass models, the most important hypothesis to satisfy is that of the constant variance of the residuals. This can be checked visually by plotting the cluster of points for the residuals $\varepsilon_i = Y_i - \hat{Y}_i$ in function to predicted values $\hat{Y}_i = \hat{a} + \hat{b}X_i$.

If the variance of the residuals is indeed constant, this cluster of points should not show any particular trend and no particular structure. This for instance is the case for the plot shown on the left in Figure 6.2. By contrast, if the cluster of points shows some form of structure, the hypothesis should be questioned. This for instance is the case in Figure 6.3 where the cluster of points for the residuals plotted against fitted values forms a funnel shape. This shape is typical of a residual variance that increases with the effect variable (which we call heteroscedasticity). If this is the case, a model other than a simple linear regression must be fitted.

In the case of biological data such as tree volume or biomass, heteroscedasticity is the rule and homoscedasticity the exception. This means simply that the greater tree biomass (or volume), the greater the variability of this biomass (or volume). This increase in the variability of biomass with increasing size is a general principle in biology. Thus, when fitting biomass or volume models, simple linear regression using biomass as response variable ($Y = B$) is generally of little use. The log transformation (i.e. $Y = \ln(B)$) resolves this problem and therefore the linear regressions we use for adjusting models nearly always use log-transformed data. We will return at length to this fundamental point later.

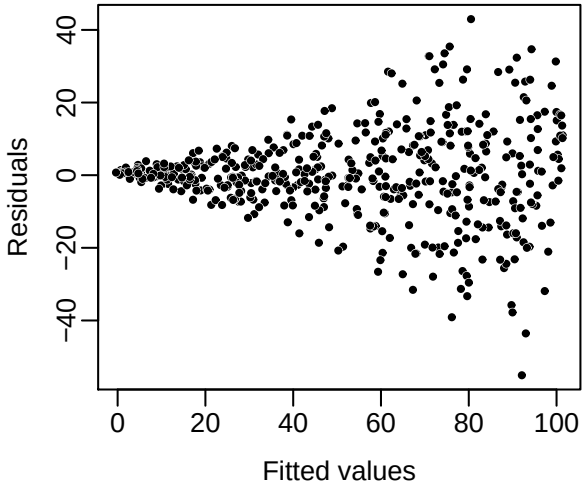


Figure 6.3 – *Plot of residuals against fitted values when the residuals have a non constant variance (heteroscedasticity).*

⑦

Simple linear regression between $\ln(B)$ **and** $\ln(D)$

The exploratory analysis (red line 5) showed that the relation between the log of the biomass and the log of the dbh was linear, with a variance of $\ln(B)$ that was approximately constant. A simple linear regression may therefore be fitted to predict $\ln(B)$from $\ln(D)$:

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

where

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

The regression is fitted using the ordinary least squares method. Given that we cannot apply the log transformation to a value of zero, zero biomass data (see red line 1) must first be withdrawn from the dataset:

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

The residual standard deviation is $\hat{\sigma} = 0.462$, R^2 is 0.9642 and the model is highly significant (Fisher's test: $F_{1,39} = 1051$, p-value $< 2.2 \times 10^{-16}$). The values of the coefficients are given in the table below:

```
             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  ***
```

The first column in the table gives the values of the coefficients. The model is therefore: $\ln(B) = -8.42722 + 2.36104\ln(D)$. The second column gives the standard deviations of the coefficient estimators. The third column gives the result of the test on the null hypothesis that the coefficient is zero. Finally, the fourth column gives the p-value of this test. In the present case, both the slope and the y-intercept are significantly different from zero. We must now check graphically that the hypotheses of the linear regression are satisfied:

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

The result is shown in Figure 6.4. Even though the quantile–quantile plot of the residuals appears to have a slight structure, we will consider that the hypotheses of the simple linear regression are suitably satisfied.

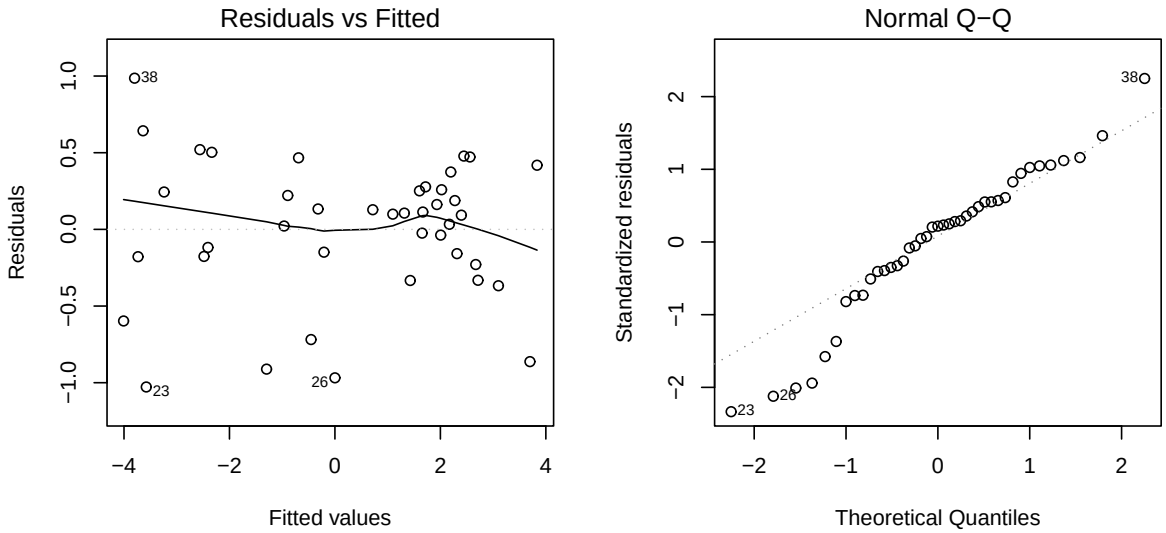


Figure 6.4 – *Residuals plotted against fitted values (left) and quantile–quantile plot (right) of the residuals of the simple linear regression of* $\ln(B)$ *against* $\ln(D)$ *fitted for the 42 trees measured by Henry et al. (2010) in Ghana.*

⑧

Simple linear regression between $\ln(B)$ **and** $\ln(D^2H)$

The exploratory analysis (red line 6) showed that the relation between the log of the biomass and the log of D^2H was linear, with a variance of $\ln(B)$ that was approximately constant. We can therefore fit a simple linear regression to predict $\ln(B)$ from $\ln(D^2H)$:

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

where

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

The regression is fitted by the ordinary least squares method. Given that we cannot apply the log transformation to a value of zero, zero biomass data (see red line 1) must first be withdrawn from the dataset:

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

The residual standard deviation is $\hat{\sigma} = 0.4084$, R^2 is 0.972 and the model is highly significant (Fisher's test: $F_{1,39} = 1356$, p-value $< 2.2 \times 10^{-16}$). The values of the coefficients are as follows:

```
                    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  ***
```

The first column in the table gives the values of the coefficients. The model is therefore: $\ln(B) = -8.99427 + 0.87238 \ln(D^2H)$. The second column gives the standard deviations of the coefficient estimators. The third column gives the result of the test on the null hypothesis that "the coefficient is zero". Finally, the fourth column gives the p-value of this test. In the present case, both the slope and the y-intercept are significantly different from zero.

We must now check graphically that the hypotheses of the linear regression are satisfied:

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

The result is shown in Figure 6.5. Even though the plot of the residuals against the fitted values appears to have a slight structure, we will consider that the hypotheses of the simple linear regression are suitably satisfied.

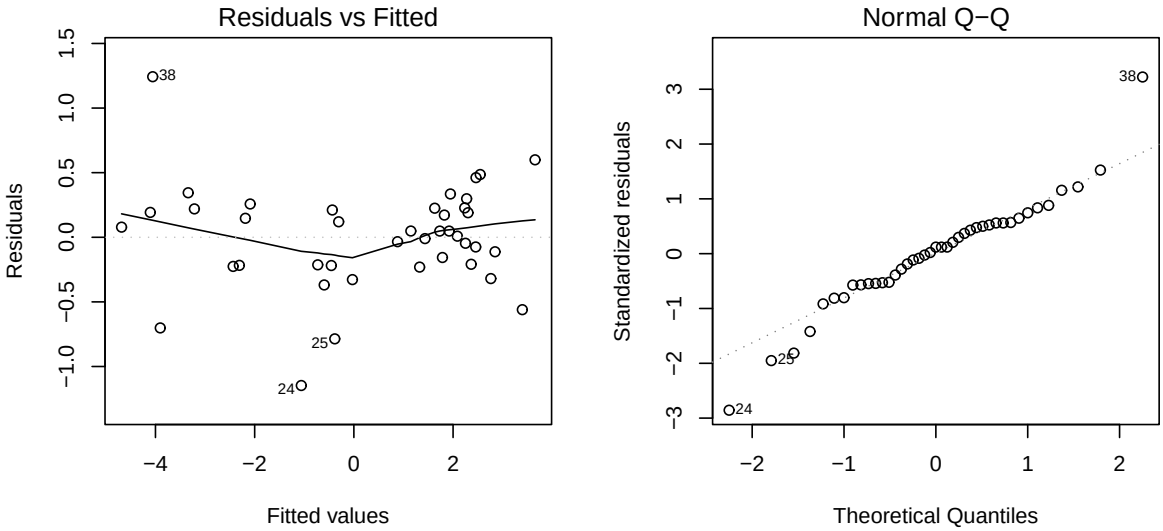


Figure 6.5 – *Residuals plotted against fitted values (left) and quantile–quantile plot (right) of the residuals of the simple linear regression of* $\ln(B)$ *against* $\ln(D^2H)$ *fitted for the 42 trees measured by Henry et al. (2010) in Ghana.*

6.1.2 Multiple regression

Multiple regression is the extension of simple linear regression to the case where there are several effect variables, and is written:

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

where Y is the response variable, $X_1, \ldots, X_p$ the p effect variables, $a_0, \ldots, a_p$ the coefficients to be estimated, and ε the residual error. Counting the y-intercept a_0, there are $p + 1$ coefficients to be estimated. Like for simple linear regression, the variance of the residuals is assumed to be constant and equal to σ^2:

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

The following biomass models are examples of multiple regressions:

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

where ρ is wood density. In all these examples the response variable is the log of the biomass: $Y = \ln(B)$. As model (6.4) generalizes (6.2) by adding dependency on wood specific density: typically (6.4) should be preferred to (6.2) when the dataset is multispecific. Model (6.5) generalizes (6.2) by considering that the exponent associated with height H is not necessarily half the exponent associated with dbh. It therefore introduces a little more flexibility into the form of the relation between biomass and D^2H. Model (6.6) generalizes (6.2) by considering that there are both several species and that biomass is not quite a power of D^2H. Model (6.7) generalizes (6.1) by considering that the relation between $\ln(B)$ and $\ln(D)$ is not exactly linear. It therefore offers a little more flexibility in the form of this relation. Model (6.8) is the extension of (6.7) to take account of the presence of several species in a dataset.

Estimating coefficients

In the same manner as for simple linear regression, the estimation of the coefficients is based on the least squares method. Estimators $\hat{a}_0$, $\hat{a}_1$, ..., $\hat{a}_p$ are the values of coefficients a_0, a_1, ..., a_p that minimize the sum of squares:

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

where X_{ij} is the value of the jth effect variable for the ith observation ($i = 1, \ldots, n$ and $j = 1 \ \ldots, p$). Once again, estimations of the coefficients may be obtained by calculating the partial derivatives of SS in relation to the coefficients, and by looking for the values of the coefficients that cancel these partial derivatives. These computations are barely more complex than for simple linear regression, on condition that they are arranged in a matrix form. Let $\mathbf{X}$ be the matrix with n lines and p columns, called the design matrix, containing the values observed for the effect variables:

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

Let $\mathbf{Y} = {}^{\mathrm{t}}[Y_1, \ldots, Y_n]$ be the vector of the n values observed for the response variable, and $\mathbf{a} = {}^{\mathrm{t}}[a_0, \ldots, a_p]$ be the vector of the $p + 1$ coefficients to be estimated. Thus

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

is none other than the vector $\hat{\mathbf{Y}}$ of the n values fitted by the response variable model. Using these matrix notations, the sum of squares is written:

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

And using matrix differential calculus (Magnus & Neudecker, 2007), we finally obtain:

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

The estimation of the residual variance is:

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

Like for simple linear regression, this estimation method has the advantage of providing an explicit expression of the coefficients estimated. As simple linear regression is a special case of multiple regression (case where $p = 1$), we can check that the matrix expressions for estimating the coefficients and $\hat{\sigma}$ actually — when $p = 1$ — give again the expressions given previously with a simple linear regression.

Interpreting the results of a multiple regression

In the same manner as for simple linear regression, fitting a multiple regression provides a determination coefficient R^2 corresponding to the proportion of the variance explained by the model, values $\hat{\mathbf{a}}$ for model coefficients a_0, a_1, …, a_p, standard deviations for these estimations, the results of significance tests on these coefficients (there are $p+1$ — one for each coefficient — null hypotheses $a_i = 0$ for $i = 0, \ldots, p$), and the result of the test on the overall significance of the model.

As previously, R^2 has a value of between 0 and 1. The higher the value, the better the fit. However, it should be remembered that the value of R^2 increases automatically with the number of effect variables used. For instance, if we are predicting Y using a polynomial with p orders in X,

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

R^2 will automatically increase with the number of orders p. This may give the illusion that the higher the number of orders p in a polynomial, the better its fit. This of course is not the case. If the number p of orders is too high, this will result in model over-parameterization. In other words, R^2 is not a valid criterion on which model selection may be based. We will return to this point in section 6.3.

Checking hypotheses

Like simple linear regression, multiple regression is based on three hypotheses: that the residuals are independent, that they follow a normal distribution and that their variance is constant. These hypotheses may be checked in exactly the same manner as for the simple linear regression. To check that the residuals follow a normal distribution, we can plot a quantile–quantile graph and verify visually that the cluster of points forms a straight line. To check that the variance of the residuals is constant, we can plot the residuals against the fitted values and verify visually that the cluster of points does not show any particular trend. The same restriction as for the simple linear regression applies to volume or biomass data that nearly always (or always) show heteroscedasticity. For this reason, multiple regression is generally applicable for fitting models only on log-transformed data.

⑨ **Polynomial regression between $\ln(B)$ and $\ln(D)$**

The exploratory analysis (red line 5) showed that the relation between the log of the biomass and the log of the dbh was linear. We can ask ourselves the question of whether this relation is truly linear, or has a more complex shape. To do this, we must construct a polynomial regression with p orders, i.e. a multiple regression of $\ln(B)$ against $\ln(D)$, $[\ln(D)]^2$, ..., $[\ln(D)]^p$:

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

where

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

The regression is fitted by the ordinary least squares method. As the log transformation stabilizes the residual variance, the hypotheses on which the multiple regression is based are in principle satisfied. For a second-order polynomial, the polynomial regression is fitted by the following code:

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

This yields:

```
                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
```

with $R^2 = 0.9642$. And as for a third-order polynomial regression:

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

it yields:

```
               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
```

with $R^2 = 0.9648$. Finally, a fourth-order polynomial regression:

```
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))
```

yields:

```
               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
```

with $R^2 = 0.9666$. Adding higher order terms above 1 therefore does not improve the model. The coefficients associated with these terms were not significantly different from zero. But the model's R^2 continued to increase with the number p of orders in the polynomial. R^2 is not therefore a reliable criterion for selecting the order of the polynomial. The plots fitted

by these different polynomials may be superimposed on the biomass-dbh cluster of points: with object `m` indicating the linear regression of $\ln(B)$ against $\ln(D)$ fitted in red line 7,

```
with(dat,plot(dbh,Btot,xlab="Dbh (cm)",ylab="Biomass (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))))
```

The plots are shown in Figure 6.6: the higher the order of the polynomial, the more deformed the plot in order to fit the data, with increasingly unrealistic extrapolations outside the range of the data (typical of model over-parameterization).

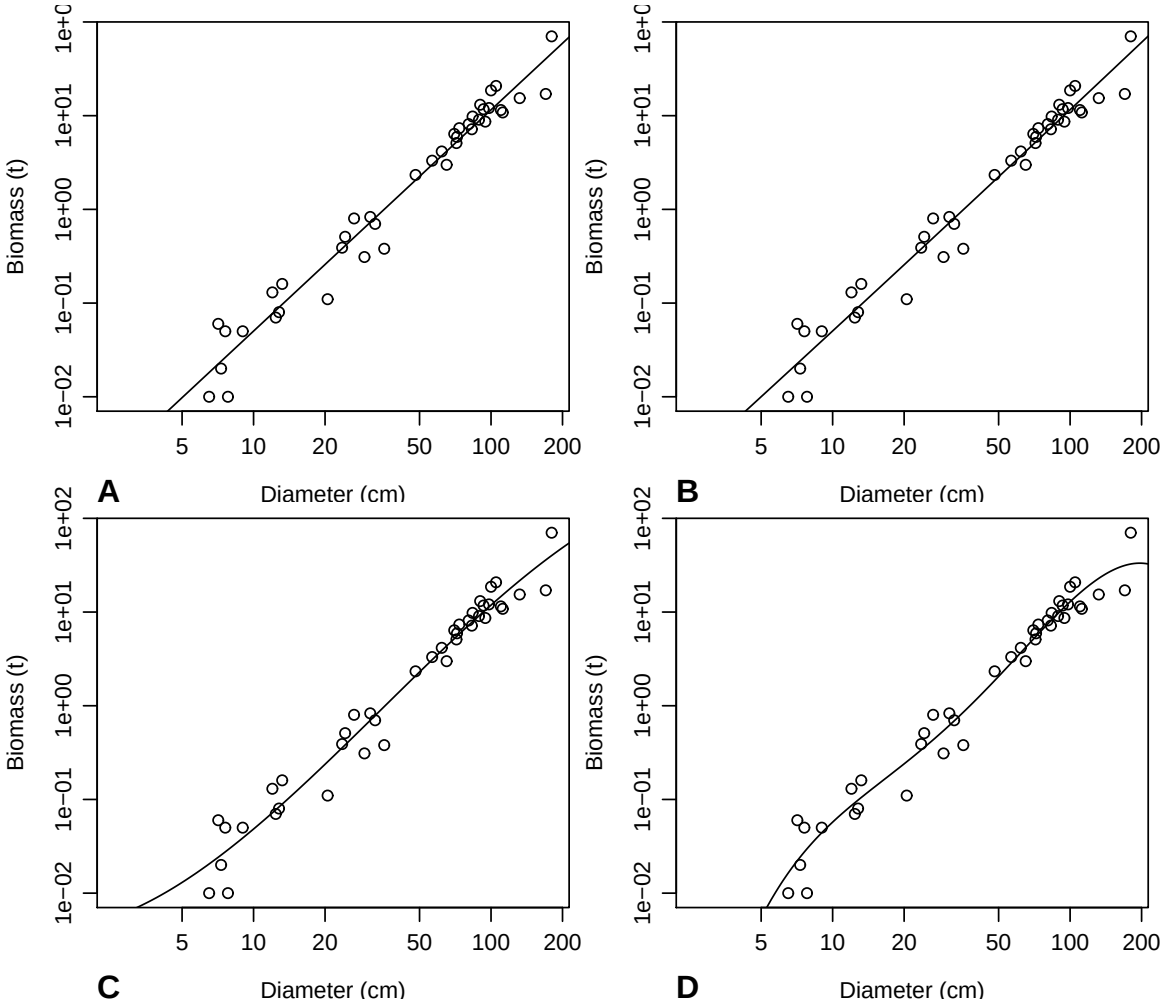


Figure 6.6 – *Biomass against dbh for 42 trees in Ghana measured by Henry et al. (2010) (points), and predictions (plots) by a polynomial regression of* $\ln(B)$ *against* $\ln(D)$*: (A) first-order polynomial (straight line); (B) second-order polynomial (parabolic); (C) third-order polynomial; (D) fourth-order polynomial.*

10

Multiple regression between $\ln(B)$, $\ln(D)$ and $\ln(H)$

The graphical exploration (red lines 3 and 6) showed that the combined variable D^2H was linked to biomass by a power relation (i.e. a linear relation on a log scale): $B = a(D^2H)^b$. We can, however, wonder whether the variables D^2 and H have the same exponent b, or whether they have different exponents: $B = a \times (D^2)^{b_1} H^{b_2}$. Working on the log-transformed data (which in passing stabilizes the residual variance), means fitting a multiple regression of $\ln(B)$ against $\ln(D)$ and $\ln(H)$:

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

where

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

The regression is fitted by the ordinary least squares method. Fitting this multiple regression:

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

yields:

```
              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  **
```

with a residual standard deviation of 0.4104 and $R^2 = 0.9725$. The model is highly significant (Fisher's test: $F_{2,38} = 671.5$, p-value $< 2.2 \times 10^{-16}$). The model, all of whose coefficients are significantly different from zero, is written: $\ln(B) = -8.9050 + 1.8654 \ln(D) + 0.7083 \ln(H)$. By applying the exponential function to return to the starting data, the model becomes: $B = 1.357 \times 10^{-4} D^{1.8654} H^{0.7083}$. The exponent associated with height is a little less than half that associated with dbh, and is a little less than the exponent of 0.87238 found for the combined variable D^2H (see red line 8). An examination of the residuals:

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

shows nothing in particular (Figure 6.7).

6.1.3 Weighted regression

Let us now suppose that we want to fit directly a polynomial model of biomass B against D. For example, for a second-order polynomial:

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

As mentioned earlier, the variability of the biomass nearly always (if not always...) increases with tree dbh D. In other words, the variance of ε increases with D, in contradiction with the homoscedasticity hypothesis required by multiple regression. Therefore, model (6.9) cannot be fitted by multiple regression. Log transformation stabilizes the residual variance (we will return to this in section 6.1.5). If we take $\ln(B)$ as the response variable, the model becomes:

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

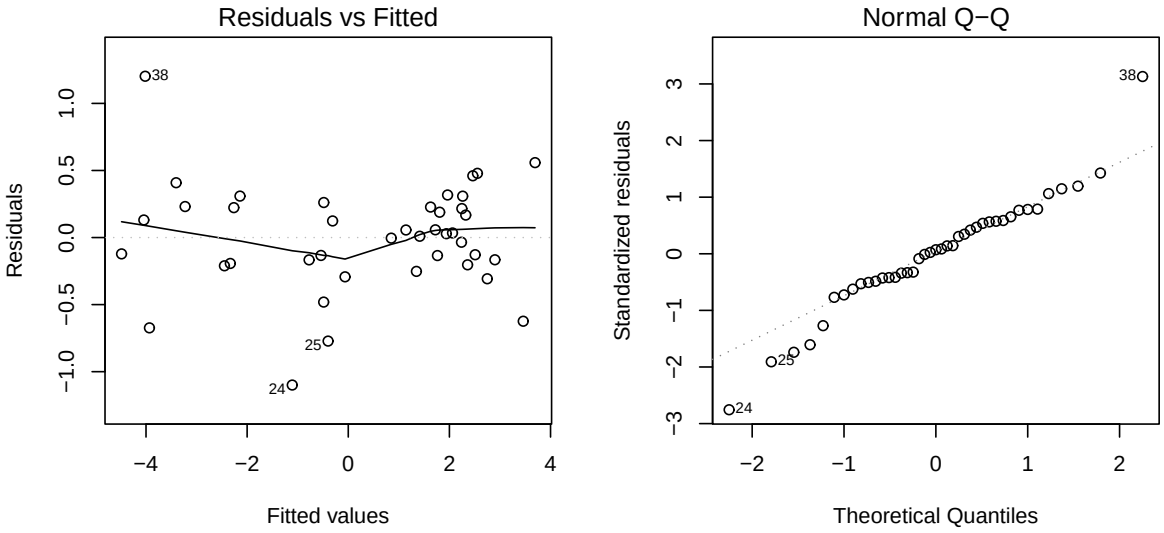


Figure 6.7 – *Residuals plotted against fitted values (left) and quantile–quantile plot (right) for residuals of the multiple regression of* $\ln(B)$ *against* $\ln(D)$ *and* $\ln(H)$ *fitted for the 42 trees measured by Henry et al. (2010) in Ghana.*

It is reasonable to assume that the variance of the residuals in such a model is constant. But unfortunately, this is no longer a linear model as the dependency of the response variables on the coefficients a_0, a_1 and a_2 is not linear. Therefore, model (6.10) cannot be fitted by a linear model. We will see later (§ 6.2) how to fit this non-linear model.

A weighted regression can be used to fit a model such as (6.9) where the variance of the residuals is not constant, while nevertheless using the formalism of the linear model. It may be regarded as extending multiple regression to the case where the variance of the residuals is not constant. Weighted regression was developed in forestry between the 1960s and the 1980s, particularly thanks to the work of Cunia (1964, 1987a). It was widely used to fit linear tables (Whraton & Cunia, 1987; Brown *et al.*, 1989; Parresol, 1999), before being replaced by more efficient fitting methods we will see in section 6.1.4.

A weighted regression is written in an identical fashion to the multiple regression (6.3):

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

except that it is no longer assumed that the variance of the residuals is constant. Each observation now has its own residual variance σ_i^2:

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

Each observation is associated with a positive *weight* w_i (hence the term "weighted" to describe this regression), which is inversely proportional to the residual variance:

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

The proportionality coefficient between w_i and $1/\sigma_i^2$ does not need to be specified as the method is insensitive to any renormalization of the weights (as we shall see in the next section). Associating each observation with a weight inversely proportional to its variance is quite natural. If an observation has a very high residual variance, this means that it has very high intrinsic variability, and it is therefore quite natural that it has less weight in model fitting. As we cannot estimate n weights from n observations, we must model the

weighting. When dealing with biological data such as biomass or volume, heteroscedasticity of the residuals corresponds nearly always to a power relation between the residual variance and the size of the trees. We may therefore assume that among the p effect variables of the weighted regression, there is one (typically tree dbh) such that σ_i is a power relation of this variable. Without loss of generality, we can put forward that this variable is X_1, such that:

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

where $k > 0$ and $c \geq 0$. In consequence:

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

The exponent c cannot be estimated in the same manner as a_0, a_1, …, a_p, but must be fixed in advance. This is the main drawback of this fitting method. We will see later how to select a value for exponent c. By contrast, the multiplier k does not need to be estimated as the weights w_i are defined only to within a multiplier factor. In practice therefore we can put forward $w_i = X_{i1}^{-2c}$.

Estimating coefficients

The least squares method is adjusted to take account of the weighting of the observations. We therefore speak of the weighted least squares method. For a fixed exponent c, estimations of coefficients a_0, …, a_p have values that minimize the weighted sum of squares:

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

or as a matrix:

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

where $\mathbf{W}$ is the $n \times n$ diagonal matrix with w_i on its diagonal:

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

The least SS is obtained for (Magnus & Neudecker, 2007):

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

This minimum does not change when all the weights w_i are multiplied by the same scalar, clearly proving that the method is not sensitive to normalization of weights. We can check that the estimation yielded by the weighted least squares method applied to observations X_{ij} and Y_i gives the same result as that yielded by the ordinary least squares method applied to observations $\sqrt{w_i}\, X_{ij}$ and $\sqrt{w_i}\, Y_i$. Like previously, this fitting method has the advantage that the estimations of the coefficients have an explicit expression.

Interpreting results and checking hypotheses

The results of the weighted regression are interpreted in exactly the same fashion as for the multiple regression. The same may also be said of the residuals-related hypotheses, except that the residuals are replaced by the weighted residuals $\varepsilon'_i = \sqrt{w_i}\,\varepsilon_i = \varepsilon_i / X_i^c$. The graph of the weighted residuals ε'_i against the fitted values must not show any particular trend (like in Figure 6.8B). If the cluster of points given by plotting the residuals against the fitted values takes on a funnel-like shape open toward the right (as in Figure 6.8A), then the value of exponent c is too low (the lowest possible value being zero). If the cluster of points takes on a funnel-like shape closed toward the right (as in Figure 6.8C), then the value of exponent c is too high.

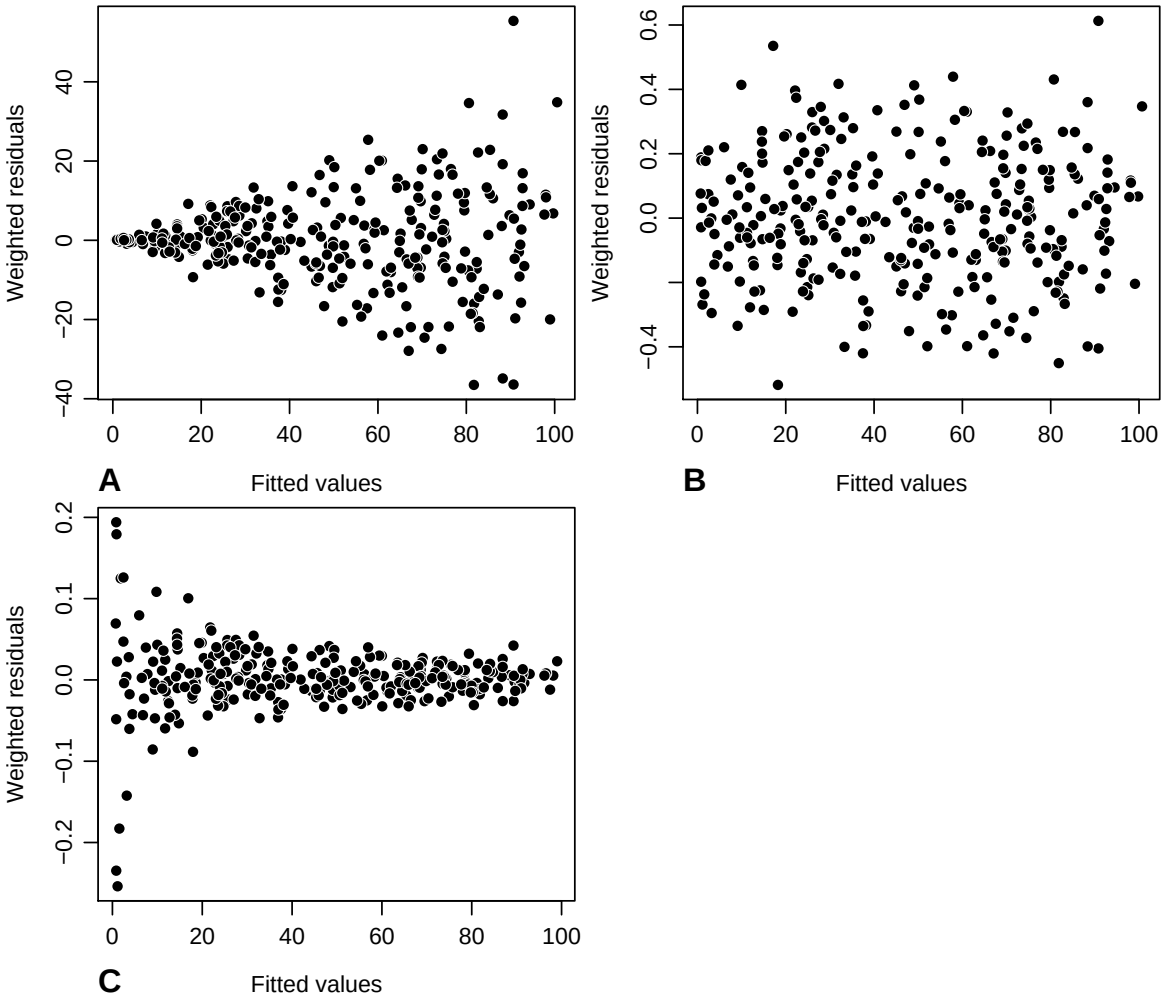


Figure 6.8 – *Plot of weighted residuals against fitted values for a weighted regression: (A) the value of exponent c is too low for the weighting; (B) the value of exponent c is appropriate; (C) the value of exponent c is too high. It should be noted that as the value of c increases, the rank of the values for the weighted values* ε / X^c *decreases.*

Choosing the right weighting

A crucial point in weighted regression is the prior selection of a value for exponent c that defines the weighting. Several methods can be used to determine c. The first consists in proceeding by trial and error based on the appearance of the plot of the residuals against the fitted values. As the appearance of the plot provides information on the pertinence of the value of c (Figure 6.8), several values of c can simply be tested until the cluster of points formed by plotting the weighted residuals against the fitted values no longer shows any particular trend.

As linear regression is robust with regard to the hypothesis that the variance of the residuals is constant, there is no need to determine c with great precision. In most cases it is enough to test integers of c. In practice, the weighted regression may be fitted for c values of 0, 1, 2, 3 or 4 (it is rarely useful to go above 4), and retain the integer value that yields the best appearance for the cluster of points in the plot of the weighted residuals against the fitted values. This simple method is generally amply sufficient.

If we are looking to obtain a more precise value for exponent c, we can calculate approximately the conditional variance of response variable Y as we know X_1:

1. divide X_1 into K classes centered on X_{1k} ($k = 1, \ldots, K$);
2. calculate the empirical variance, σ_k^2, of Y for the observations in class k (avec $k = 1, \ldots, K$);
3. plot the linear regression of $\ln(\sigma_k)$ against $\ln(X_{1k})$.

The slope of this regression is an estimation of c.

The third way of estimating c consists in looking for the value of c that minimizes Furnival (1961) index. This index is defined on page 154.

11

Weighted linear regression between B and D^2H

The exploratory analysis of the relation between biomass and D^2H showed (red line 3) that this relation is linear, but that the variance of the biomass increases with D^2H. We can therefore fit a weighted linear regression of biomass B against D^2H:

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

where

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

The linear regression is fitted by the weighted least squares method, meaning that we must beforehand know the value of exponent c. Let us first estimate coefficient c for the weighting of the observations. To do this we must divide the observations into dbh classes and calculate the standard deviation of the biomass in each dbh class:

```
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))
```

Object `D` contains the bounds of the dbh classes that are calculated such to have 5 classes containing approximately the same number of observations: Object `i` contains the number of the dbh class to which each observation belongs. Figure 6.9, obtained by the command:

```
with(sdB,plot(D,sdB,log="xy",xlab="Diameter (cm)",ylab=
"Biomass standard deviation (t)"))
```

shows a plot of the standard deviation of the biomass against the median dbh of each dbh class, on a log scale. The points are roughly aligned along a straight line, confirming that the power model is appropriate for modeling the residual variance. The linear regression of the log of the standard deviation of the biomass against the log of the median dbh for each class, fitted by the command:

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

yields:

```
              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  **
```

The slope of the regression corresponds to c, and is 2. Thus, the standard deviation σ of the biomass is approximately proportional to D^2, and we will select a weighting of the observations that is inversely proportional to D^4. The weighted regression of biomass B against D^2H with this weighting, and fitted by the command:

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

yields:

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

An examination of the result of this fitting shows that the y-intercept is not significantly different from zero. This leads us to fit a new weighted regression of biomass B against D^2H without an intercept:

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

which yields:

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

The model is therefore: $B = 2.747 \times 10^{-5} D^2H$, with $R^2 = 0.8897$ and a residual standard deviation $k = 0.0003513$ tonnes cm^{-2}. The model is highly significant (Fisher's test: $F_{1,41} = 330.8$, p-value $< 2.2 \times 10^{-16}$). As this model was fitted directly on non-transformed data, it should be noted that it was unnecessary to withdraw the observations of zero biomass (unlike the situation in red line 8). Figure 6.10A, obtained by the command:

```
plot(fitted(m),residuals(m)/dat$dbh^2,xlab="Fitted values",ylab=
"Weighted residuals")
```

gives a plot of the weighted residuals against the fitted values. By way of a comparison, Figure 6.10B shows a plot of the weighted residuals against the fitted values when the weighting is too low (with weights inversely proportional to D^2):

```
m <- lm(Btot~-1+I(dbh^2*heig),data=dat,weights=1/dat$dbh^2)
plot(fitted(m),residuals(m)/dat$dbh,xlab="Fitted values",ylab="Weighted residuals")
```

whereas 6.10C shows a plot of the weighted residuals against the fitted values when the weighting is too high (with weights inversely proportional to D^5):

```
m <- lm(Btot~-1+I(dbh^2*heig),data=dat,weights=1/dat$dbh^5)
plot(fitted(m),residuals(m)/dat$dbh^2.5,xlab="Fitted values",ylab=
"Weighted residuals")
```

Thus, the coefficient $c = 2$ for the weighting indeed proves to be that which is suitable.

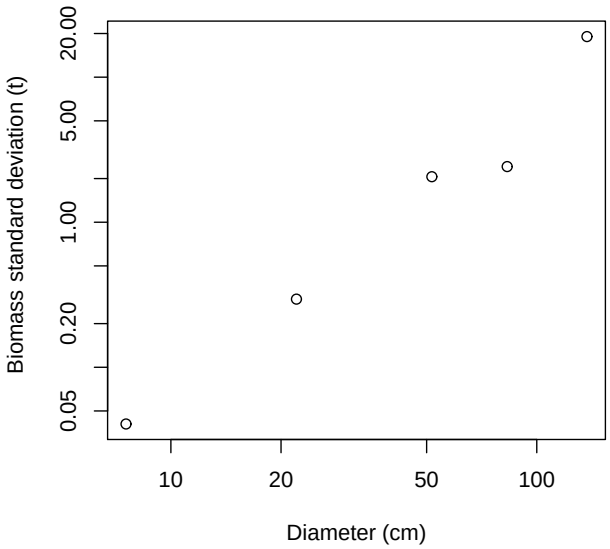


Figure 6.9 – *Plot of standard deviation of biomass calculated in five dbh classes against median dbh of the class (on a log scale), for 42 trees measured in Ghana by Henry et al. (2010).*

12

Weighted polynomial regression between B and D

The exploratory analysis (red line 2) showed that the relation between biomass and dbh was parabolic, with the variance of the biomass increasing with dbh. Log-transformation rendered the relation between biomass and dbh linear, but we can also model the relation between biomass and dbh directly by a parabolic relation:

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

where

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

In red line 11 we saw that the value $c = 2$ of the exponent was suitable for modeling the conditional standard deviation of the biomass when we know the dbh. We will therefore fit the multiple regression using the weighted least squares method with weighting of the observations proportional to $1/D^4$:

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

which yields:

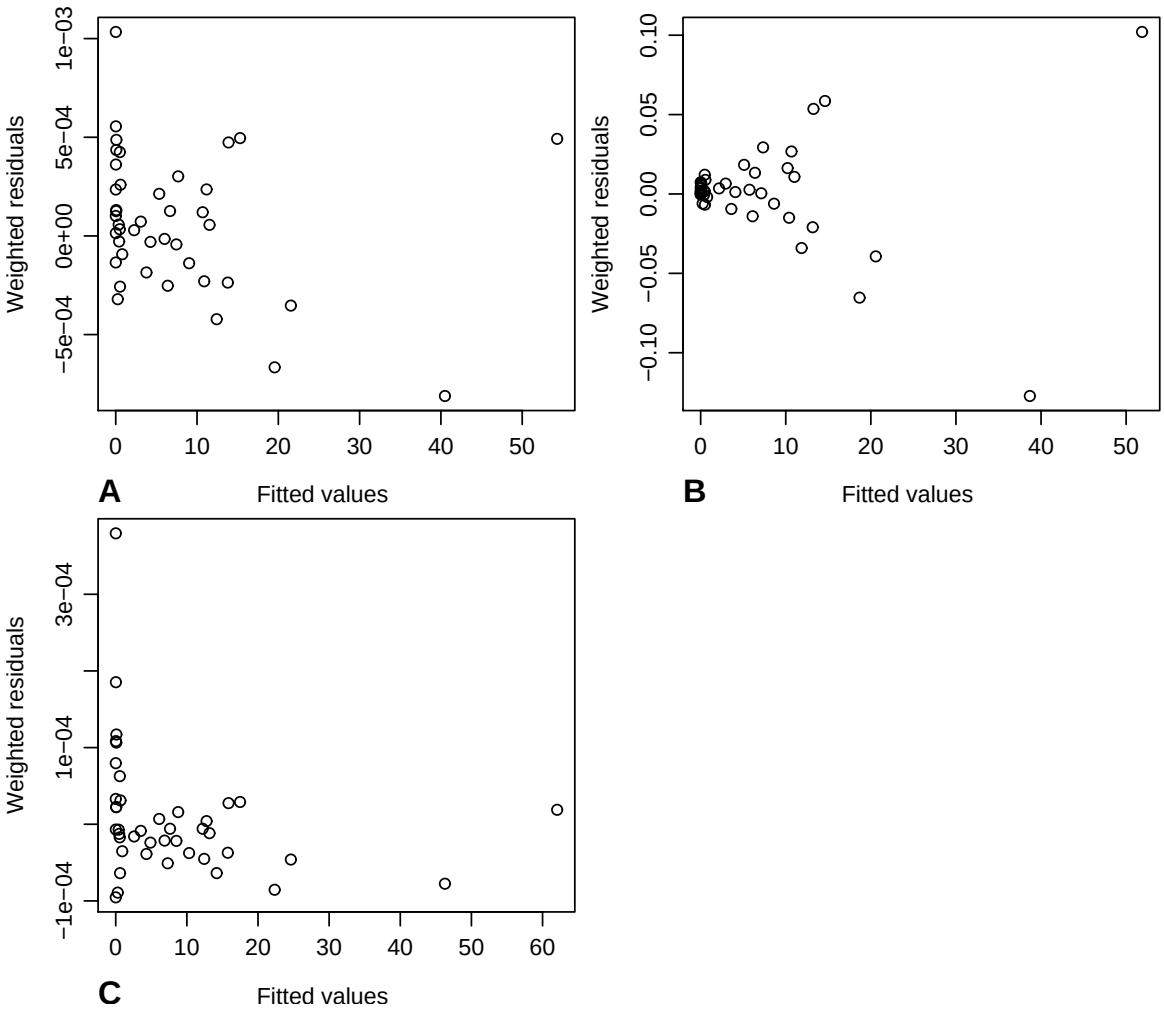


Figure 6.10 – *Plot of weighted residuals against fitted values for a weighted regression of biomass against D^2H for 42 trees measured in Ghana by Henry et al. (2010): (A) the weighting is inversely proportional to D^4; (B) the weighting is inversely proportional to D^2; (C) the weighting is inversely proportional to D^5.*

```
               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  ***
```

with a residual standard deviation $k = 0.0003882$ tonnes cm^{-2} and $R^2 = 0.8709$. The y-intercept is not significantly different from zero. We can therefore again fit a parabolic relation but without the intercept:

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

which yields:

```
             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  ***
```

with a residual standard deviation $k = 0.0003985$ tonnes cm^{-2} and $R^2 = 0.8615$. The model is highly significant (Fisher's test: $F_{2,40} = 124.4$, p-value $= 2.2 \times 10^{-16}$) and is written: $B = -3.840 \times 10^{-3} D + 1.124 \times 10^{-3} D^2$. Figure 6.11 obtained by the command:

```
plot(fitted(m),residuals(m)/dat$dbh^2,xlab="fitted values",ylab=
"Weighted residuals")
```

gives a plot of the weighted residuals against the fitted values.

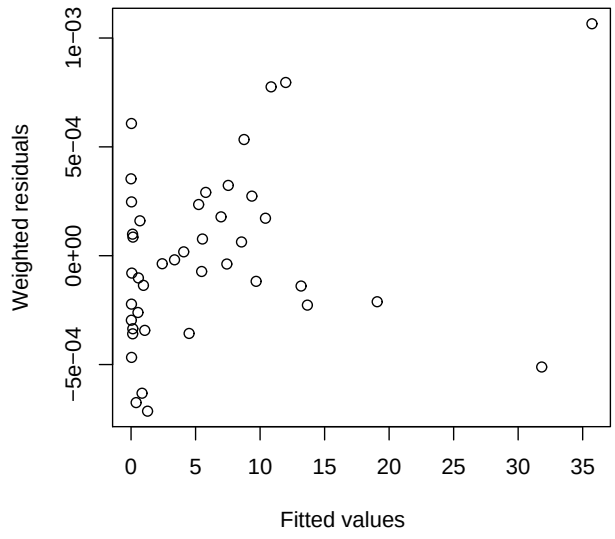


Figure 6.11 – *Plot of weighted residuals against fitted values for a weighted regression of biomass against D and D^2 for 42 trees measured in Ghana by Henry et al. (2010).*

6.1.4 Linear regression with variance model

An alternative to the weighted regression is to use explicitly a model for the variance of the residuals. As previously, it is realistic to assume that there is an effect variable (without loss of generality, the first) such that the residual standard deviation is a power function of this variable:

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

where $k > 0$ and $c \geq 0$. The model is therefore written:

$$Y = a_0 + a_1X_1 + a_2X_2 + \ldots + a_pX_p + \varepsilon \tag{6.12}$$

where:

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

In its form the model is little different from the weighted regression. Regarding its content, it has one fundamental difference: the coefficients k and c are now model parameters that need to be estimated, in the same manner as coefficients a_0, a_1, …, a_p. Because of these k and c parameters that need to be estimated, the least squares method can no longer be used to estimate model coefficients. Another estimation method has to be used, the maximum likelihood method. Strictly speaking, the model defined by (6.11) and (6.12) is not a linear model. It is far closer conceptually to the non-linear model we will see in section 6.2. We will not go any further here in presenting the non-linear model: the model fitting method defined by (6.11) and (6.12) will be presented as a special case of the non-linear model in section 6.2.

13

Linear regression between B and D^2H with variance model

Pre-empting section 6.2, we will fit a linear regression of biomass against D^2H by specifying a power model on the residual variance:

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

where

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

We will see later (§ 6.2) that this model is fitted by maximum likelihood. This regression is in spirit very similar to the previously-constructed weighted regression of biomass against D^2H (red line 11), except that exponent c used to define the weighting of the observations is now a parameter that must be estimated in its own right rather than a coefficient that is fixed in advance. The linear regression with variance model is fitted as follows:

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

and yields (we will return in section 6.2 to the meaning of the start object `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
```

with an estimated value of exponent $c = 1.977736$. Like for the weighted linear regression (red line 11), the y-intercept proves not to be significantly different from zero. We can therefore refit the model without the intercept:

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

which yields:

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

with an estimated value of exponent $c = 1.980263$. This value is very similar to that evaluated for the weighted linear regression ($c = 2$ in red line 11). The fitted model is therefore: $B = 2.740688 \times 10^{-5} D^2 H$, which is very similar to the model fitted by weighted linear regression (red line 11).

14

Polynomial regression between B and D with variance model

Pre-empting section 6.2, we will fit a multiple regression of biomass against D and D^2 by specifying a power model on the residual variance:

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

where

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

We will see later (§ 6.2) that this model is fitted by maximum likelihood. This regression is in spirit very similar to the previously-constructed polynomial regression of biomass against D and D^2 (red line 12), except that exponent c employed to define the weighting of the observations is now a parameter that must be estimated in its own right rather than a coefficient that is fixed in advance. The linear regression with variance model is fitted as follows:

```
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)))
```

and yields (we will return in section 6.2 to the meaning of the start object `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
```

with an estimated value of exponent $c = 2.127509$. Like for the weighted polynomial regression (red line 12), the y-intercept proves not to be significantly different from zero. We can therefore refit the model without the intercept:

```
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)))
```

which yields:

```
            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
```

with an estimated value of exponent $c = 2.139967$. This value is very similar to that evaluated for the weighted polynomial regression ($c = 2$ in red line 12). The fitted model is

therefore: $B = -3.319456 \times 10^{-3} D + 1.067068 \times 10^{-3} D^2$, which is very similar to the model fitted by weighted polynomial regression (red line 12).

6.1.5 Transforming variables

Let us reconsider the example of the biomass model with a single entry (here dbh) of the power type:

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

We have already seen that this is a non-linear model as B is non linearly dependent upon coefficients a and b. But this model can be rendered linear by log transformation. Relation (6.13) is equivalent to: $\ln(B) = \ln(a) + b\ln(D)$, which can be considered to be a linear regression of response variable $Y = \ln(B)$ against effect variable $X = \ln(D)$. We can therefore estimate coefficients a and b (or rather $\ln(a)$ and b) in power model (6.13) by linear regression on log-transformed data. What about the residual error? If the linear regression on log-transformed data is appropriate, this means that $\varepsilon = \ln(B) - \ln(a) - b\ln(D)$ follows a centered normal distribution of constant standard deviation σ. If we return to the starting data and use exponential transformation (which is the inverse transformation to log transformation), the residual error here is a factor:

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

where $\varepsilon' = \exp(\varepsilon)$. Thus, we have moved from an additive error on log-transformed data to a multiplicative error on the starting data. Also, if ε follows a centered normal distribution of standard deviation σ, then, by definition, $\varepsilon' = \exp(\varepsilon)$ follows a log-normal distribution of parameters 0 and σ:

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

Contrary to ε, which has a mean of zero, the mean of ε' is not zero but: $\mathrm{E}(\varepsilon') = \exp(\sigma^2/2)$. The implications of this will be considered in chapter 7.

We can draw two lessons from this example:

1. when we are faced by a non-linear relation between a response variable and one (or several) effect variables, a transformation may render this relation linear;

2. this transformation of the variable affects not only the form of the relation between the effect variable(s) and the response variable, but also the residual error.

Concerning the first point, this variables transformation means that we have two approaches for fitting a non-linear model. The first, when faced with a non-linear relation between a response variable and effect variables, consists in looking for a transformation that renders this relation linear, and thereafter using the approach employed for the linear model. The second consists in fitting the non-linear model directly, as we shall see in section 6.2. Each approach has its advantages and its drawbacks. The linear model has the advantage of providing a relatively simple theoretical framework and, above all, the estimations of its coefficients have explicit expressions. The drawback is that the model linearization step introduces an additional difficulty, and the inverse transformation, if we are not careful, may produce prediction bias (we will return to this in chapter 7). Also, not all models can be rendered linear. For example, no variables transformation can render the following model linear: $Y = a_0 + a_1 X + a_2 \exp(a_3 X)$.

Regarding the second point, we are now, therefore, obliged to distinguish the form of the relation between the response variable and the effect variables (we also speak of the mean model, i.e. the mean of the response variable Y), and the form of the model for the residual error (we also speak of the variance model, i.e. the variance of Y). This transformation of the variable affects both simultaneously. The art of transforming variables therefore lies in tackling the two simultaneously and thereby rendering the model linear with regard to its coefficients and stabilizing the variance of the residuals (i.e. rendering it constant).

Common variable transformations

Although theoretically there is no limit as to the variable transformations we can use, the transformations likely to concern volumes and biomasses are few in number. That most commonly employed to fit models is the log transformation. Given a power model:

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

the log transformation consists in replacing the variable Y by its log: $Y' = \ln(Y)$, and each of the effect variables by their log: $X'_j = \ln(X_j)$. The resulting model corresponds to:

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

where $\varepsilon' = \ln(\varepsilon)$. The inverse transformation is the exponential for all the effect and response variables. In terms of residual error, the log transformation is appropriate if ε' follows a normal distribution, therefore if the error ε is positive and multiplicative. It should be noted that log transformation poses a problem for variables that may take a zero value. In this case, the transformation $X' = \ln(X+1)$ is used rather than $X' = \ln(X)$ (or more generally, $X' = \ln(X + \text{constant})$ if X can take a negative value, e.g. a dbh increment). By way of examples, the following biomass models:

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

may be fitted by linear regression after log transformation of the data.

Given an exponential model:

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

the appropriate transformation consists in replacing variable Y by its log: $Y' = \ln(Y)$, and not transforming the effect variables: $X'_j = X_j$. The resulting model is identical to (6.14). The inverse transformation is the exponential for the response variable and no change for the effect variables. In terms of residual error, this transformation is appropriate if ε' follows a normal distribution, therefore if the error ε is positive and multiplicative. It should be noted that, without loss of generality, we can reparameterize the coefficients of the exponential model (6.15) by applying $b'_j = \exp(b_j)$. Strictly identical writing of exponential model (6.15) therefore yields:

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

For example, the following biomass model:

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

may be fitted by linear regression after this type of variable transformation (with, in this example, $X_j = [\ln(D)]^j$).

The Box-Cox transformation generalizes the log transformation. It is in fact a family of transformations indexed by a parameter ξ. Given a variable X, its Box-Cox transform X'_ξ corresponds to:

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

The Box-Cox transformation can be used to convert the question of choosing a variable transformation into a question of estimating parameter ξ (Hoeting *et al.*, 1999).

A special variable transformation

The usual variable transformations change the form of the relation between the response variable and the effect variable. When the cluster of points (X_i, Y_i) formed by plotting the response variable against the effect variable is a straight line with heteroscedasticity, as shown in Figure 6.12, then variable transformation needs to be employed to stabilize the variance of Y, though without affecting the linear nature of the relation between X and Y. The particular case illustrated by 6.12 occurs fairly often when an allometric equation is fitted between two values that vary in a proportional manner (see for example Ngomanda *et al.*, 2012). The linear nature of the relation between X and Y means that the model has the form:

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

but the heteroscedasticity indicates that the variance ofε is not constant, and this prevents any fitting of a linear regression. A variable transformation in this case consists in replacing Y by $Y' = Y/X$ and X by $X' = 1/X$. By dividing each member of (6.16) by X, the post-variable transformation model becomes:

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

where $\varepsilon' = \varepsilon/X$. The transformed model still corresponds to a linear relation, except that the y-intercept a of the relation between X and Y has become the slope of the relation between X' and Y', and reciprocally, the slope b of the relation between X and Y has become the y-intercept of the relation between X' and Y'. Model (6.17) may be fitted by simple linear regression if the variance of ε' is constant. As $\mathrm{Var}(\varepsilon') = \sigma^2$ means $\mathrm{Var}(\varepsilon) = \sigma^2 X^2$, the variable transformation is appropriate if the standard deviation of ε is proportional to X.

As model (6.17) was fitted by simple linear regression, its sum of squares corresponds to:

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

Here we recognize the expression for the sum of squares for a weighted regression using the weight $w_i = X_i^{-2}$. Thus, the variable transformation $Y' = Y/X$ and $X' = 1/X$ is strictly identical to a weighted regression of weight $w = 1/X^2$.

15

Linear regression between B/D^2 and H

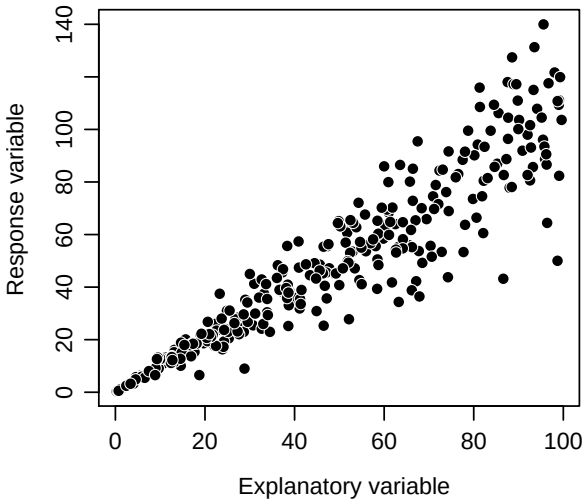


Figure 6.12 – *Linear relation between an effect variable (X) and a response variable (Y), with an increase in the variability of Y with an increase in X (heteroscedasticity).*

We saw in red line 11 that a double-entry biomass model using dbh and height corresponds to: $B = a + bD^2H + \varepsilon$ where $\mathrm{Var}(\varepsilon) \propto D^4$. By dividing each member of the equation by D^2, we obtain:

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

where

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

Thus, the regression of the response variable $Y = B/D^2$ against the two effect variables $X_1 = 1/D^2$ and $X_2 = H$ in principle satisfies the hypotheses of multiple linear regression. This regression is fitted by the ordinary least squares method. Fitting of this multiple regression by the command:

```
summary(lm((Btot/dbh^2)~-1+I(1/dbh^2)+heig,data=dat))
```

yields:

```
             Estimate  Std.  Error  t value  Pr(>|t|)
I(1/dbh^2)  1.181e-03    2.288e-03    0.516     0.608
heig        2.742e-05    1.527e-06   17.957    <2e-16  ***
```

where it can be seen that the coefficient associated with $X_1 = 1/D^2$ is not significantly different from zero. If we now return to the starting data, this means simply that the y-intercept a is not significantly different from zero, something we had already noted in red line 11. Therefore, X_1 may be withdrawn and a simple linear regression may be fitted of $Y = B/D^2$ against $X_2 = H$:

```
with(dat,plot(heig,Btot/dbh^2,xlab="Height (m)",ylab="Biomass/square of dbh
(t/cm2)"))
m <- lm((Btot/dbh^2)~-1+heig,data=dat)
summary(m)
plot(m,which=1:2)
```

The scatter plot of B/D^2 against H is indeed a straight line with a variance of B/D^2 that is approximately constant (Figure 6.13). Fitting the simple linear regression yields:

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

with $R^2 = 0.8897$ and a residual standard deviation of 0.0003513 tonnes cm^{-2}. The model is written: $B/D^2 = 2.747 \times 10^{-5} H$, or if we return to the starting variables: $B = 2.747 \times 10^{-5} D^2 H$. We now need to check that this model is strictly identical to the weighted regression of B against $D^2 H$ shown in red line 11 with weighting proportional to $1/D^4$. The plot of the residuals against the fitted values and the quantile-quantile plot of the residuals are shown in Figure 6.14.

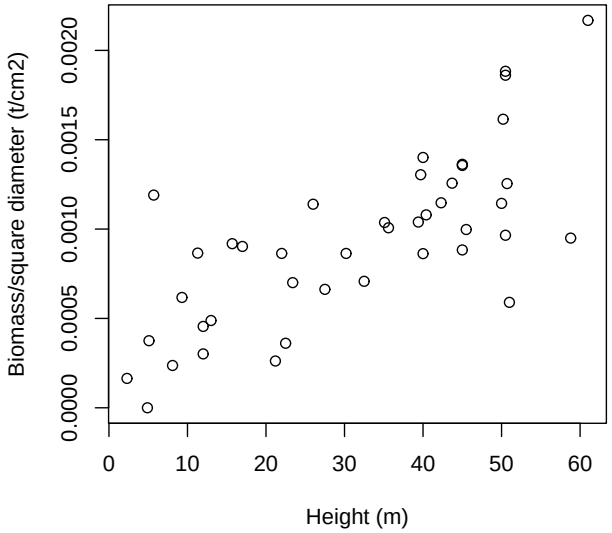


Figure 6.13 – *Scatter plot of biomass divided by the square of the dbh (tonnes cm^{-2}) against height (m) for the 42 trees measured in Ghana by Henry et al. (2010).*

16

Linear regression between B/D^2 and $1/D$

We saw in red line 12 that a polynomial biomass model against dbh corresponded to: $B = a_0 + a_1 D + a_2 D^2 + \varepsilon$ where $\text{Var}(\varepsilon) \propto D^4$. By dividing each member of the equation by D^2, we obtain:

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

where

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

Thus, the regression of the response variable $Y = B/D^2$ against the two effect variables $X_1 = 1/D^2$ and $X_2 = 1/D$ in principle satisfies the hypotheses of multiple linear regression. This regression is fitted by the ordinary least squares method. Fitting this multiple regression by the command:

```
summary(lm((Btot/dbh^2)~I(1/dbh^2)+I(1/dbh),data=dat))
```

yields:

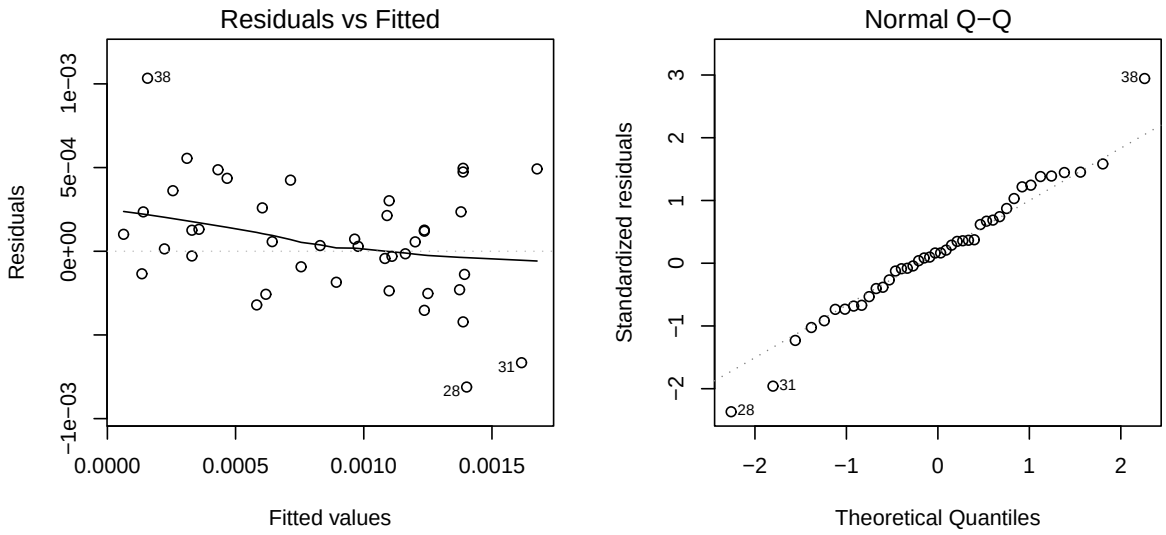


Figure 6.14 – *Residuals plotted against fitted values (left) and quantile–quantile plot (right) of the residuals of the simple linear regression of B/D^2 against H fitted for the 42 trees measured by in Ghana.*

```
              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  **
```

where it can be seen that the coefficient associated with $X_1 = 1/D^2$ is not significantly different from zero. If we now return to the starting data, this means simply that the y-intercept a_0 is not significantly different from zero, something we had already noted in red line 12. Therefore, X_1 may be withdrawn and a simple linear regression may be fitted of $Y = B/D^2$ against $X_2 = 1/D$:

```
with(dat,plot(1/dbh,Btot/dbh^2,xlab="1/Diameter (/cm)",ylab="Biomass/square of
dbh (t/cm2)"))
m <- lm((Btot/dbh^2)~I(1/dbh),data=dat)
summary(m)
plot(m,which=1:2)
```

The scatter plot of B/D^2 against $1/D$ is approximately a straight line with a variance of B/D^2 that is approximately constant (Figure 6.15). Fitting the simple linear regression yields:

```
              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  ***
```

with $R^2 = 0.3106$ and a residual standard deviation of 0.0003985 tonnes cm^{-2}. The model is written: $B/D^2 = 1.124 \times 10^{-3} - 3.84 \times 10^{-3} D^{-1}$, or if we return to the starting variables: $B = -3.84 \times 10^{-3} D + 1.124 \times 10^{-3} D^2$. We now need to check that this model is strictly identical to the polynomial regression of B against D shown in red line 12 with weighting proportional to $1/D^4$. The plot of the residuals against the fitted values and the quantile-quantile plot of the residuals are shown in Figure 6.16.

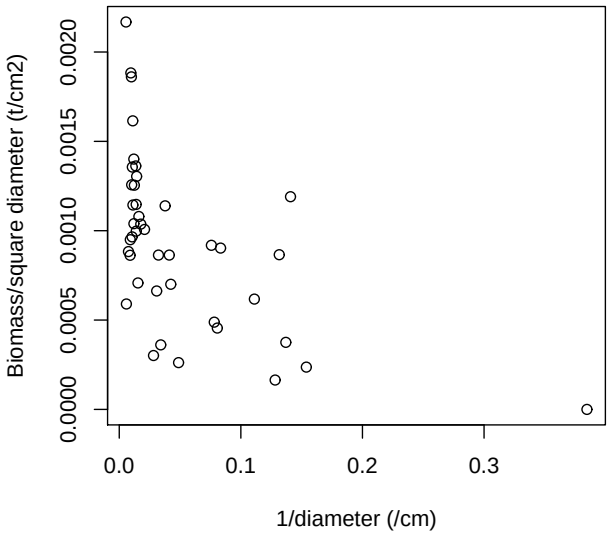


Figure 6.15 – *Scatter plot of biomass divided by the square of the dbh (tonnes in* cm^{-2}*) against the inverse of the dbh (*cm^{-1}*) for 42 trees measured in Ghana by Henry et al. (2010).*

6.2 Fitting a non-linear model

Let us now address the more general case of fitting a non-linear model. This model is written:

$$Y = f(X_1, \ldots, X_p; \theta) + \varepsilon$$

where Y is the response variable, $X_1, \ldots, X_p$ are the effect variables, θ is the vector of all the model coefficients, ε is the residual error, and f is a function. If f is linear in relation to the coefficients θ, this brings us back to the previously studied linear model. We will henceforth no longer make any a priori hypotheses concerning the linearity of f in relation to coefficients θ. As previously, we assume that the residuals are independent and that they follow a centered normal distribution. By contrast, we do not make any a priori hypothesis concerning their variance. $\mathrm{E}(\varepsilon) = 0$ means that $\mathrm{E}(Y) = f(X_1, \ldots, X_p; \theta)$. This is why we can say that f defines the mean model (i.e. for Y). Let us write:

$$\mathrm{Var}(\varepsilon) = g(X_1, \ldots, X_p; \vartheta)$$

where g is a function and ϑ a set of parameters. As $\mathrm{Var}(Y) = \mathrm{Var}(\varepsilon)$, we can say that g defines the variance model. Function g may take various forms, but for biomass or volume data it is generally a power function of a variable that characterizes tree size (typically dbh). Without loss of generality, we can put forward that this effect variable is X_1, and therefore:

$$g(X_1, \ldots, X_p; \vartheta) \equiv (kX_1^c)^2$$

where $\vartheta \equiv (k, c)$, $k > 0$ and $c \geq 0$.

Interpreting the results of fitting a non-linear model is fundamentally the same as for the linear model. The difference between the linear model and the non-linear model, in addition to their properties, lies in the manner by which model coefficients are estimated. Two particular approaches are used: (*i*) exponent c is fixed in advance; (*ii*) exponent c is a parameter to be estimated in the same manner as the model's other parameters.

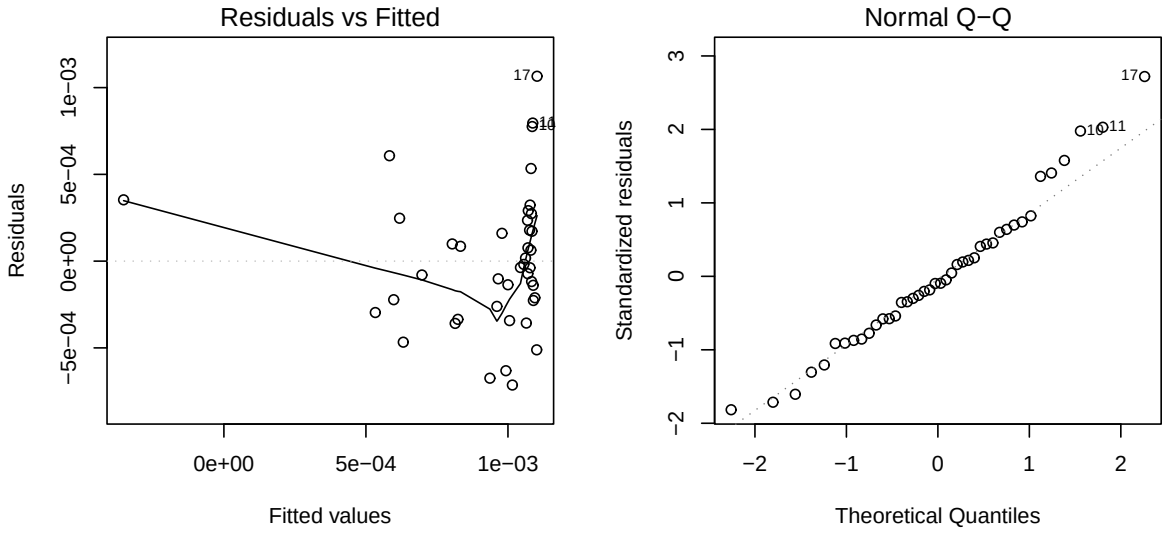


Figure 6.16 – *Residuals plotted against fitted values (left) and quantile–quantile plot (right) of the residuals of the simple linear regression of B/D^2 against $1/D$ fitted for the 42 trees measured by Henry et al. (2010) in Ghana.*

6.2.1 Exponent known

Let us first consider the case where the exponent c of the variance model is known in advance. Here, the least squares method can again be used to fit the model. The weighted sum of squares corresponds to:

$$\mathrm{SSE}(\theta) = \sum_{i=1}^{n} w_i \, \varepsilon_i^2 = \sum_{i=1}^{n} w_i \, [Y_i - f(X_{i1}, \ldots, X_{ip}; \theta)]^2$$

where the weights are inversely proportional to the variance of the residuals:

$$w_i = \frac{1}{X_{i1}^{2c}} \propto \frac{1}{\mathrm{Var}(\varepsilon_i)}$$

As previously, the estimator of the model's coefficients corresponds to the value of θ that minimizes the weighted sum of squares:

$$\hat{\theta} = \arg\min_{\theta} \mathrm{SSE}(\theta) = \arg\min_{\theta} \left\{ \sum_{i=1}^{n} \frac{1}{X_{i1}^{2c}} [Y_i - f(X_{i1}, \ldots, X_{ip}; \theta)]^2 \right\}$$

In the particular case where the residuals have a constant variance (i.e. $c = 0$), the weighted least squares method is simplified to the ordinary least squares method (all weights w_i are 1), but the principle behind the calculations remains the same. The estimator θ is obtained by resolving

$$\frac{\partial \mathrm{SSE}}{\partial \theta}(\hat{\theta}) = 0 \tag{6.18}$$

with the constraint $(\partial^2 \mathrm{SSE}/\partial \theta^2) > 0$ to ensure that this is indeed a minimum, not a maximum. In the previous case of the linear model, resolving (6.18) yielded an explicit expression for the estimator $\hat{\theta}$. This is not the case for the general case of the non-linear model: there is no explicit expression for $\hat{\theta}$. The sum of squares must therefore be minimized using a numerical algorithm. We will examine this point in depth in section 6.2.3.

A priori value for the exponent

The *a priori* value of exponent c is obtained in the non-linear case in the same manner as for the linear case (see page 123): by trial and error, by dividing X_1 into classes and estimating the variance of Y for each class, or by minimizing Furnival's index (see p.154).

17

Weighted non-linear regression between B and D

The graphical exploration (red lines 2 and 5) showed that the relation between biomass B and dbh D was of the power type, with the variance of the biomass increasing with dbh:

$$B = aD^b + \varepsilon$$

where

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

We saw in red line 11 that the conditional standard deviation of the biomass derived from the dbh was proportional to the square of the dbh: $c = 2$. We can therefore fit a non-linear regression by the weighted least squares method using a weighting that is inversely proportional to 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)
```

The non-linear regression is fitted using the `nls` command which calls on the start values of the coefficients. These start values are contained in the `start` object and are calculated by re-transforming the coefficients of the linear regression on the log-transformed data. Fitting the non-linear regression by the weighted least squares method gives:

```
     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  ***
```

with a residual standard deviation $k = 0.0003598$ tonnes cm^{-2}. The model is therefore written: $B = 2.492 \times 10^{-4} D^{2.346}$. Let us now return to the linear regression fitted to log-transformed data (red line 7) which was written: $\ln(B) = -8.42722 + 2.36104 \ln(D)$. If we return naively to the starting data by applying the exponential function (we will see in § 7.2.4 why this is naive), the model becomes: $B = \exp(-8.42722) \times D^{2.36104} = 2.188 \times 10^{-4} D^{2.36104}$. The model fitted by non-linear regression and the model fitted by linear regression on log-transformed data are therefore very similar.

18

Weighted non-linear regression between B and D^2H

We have already fitted a power model $B = a(D^2H)^b$ by simple linear regression on log-transformed data (red line 8). Let us now fit this model directly by non-linear regression:

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

where

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

In order to take account of the heteroscedasticity, and considering that the conditional standard deviation of the biomass derived from the diameter is proportional to D^2 (red line 11), we can fit this non-linear model by the weighted least squares method using a weighting inversely proportional to D^4:

```
start <- coef(lm(log(Btot)~I(log(dbh^2*heig)),data=dat[dat$Btot>0,]))
start[1] <- exp(start[1])
names(start) <- c("a","b")
m <- nls(Btot~a*(dbh^2*heig)^b,data=dat,start=start,weights=1/dat$dbh^4)
summary(m)
```

As previously (red line 17), the `nls` command calls on start values for the coefficients and these are obtained from the coefficients of the multiple regression on log-transformed data. The result of the fitting is as follows:

```
     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  ***
```

with a residual standard deviation $k = 0.0003325$ tonnes cm^{-2}. The model is therefore written: $B = 7.885 \times 10^{-5}(D^2H)^{0.9154}$. Let us now return to the linear regression fitted to log-transformed data (red line 8), which was written: $\ln(B) = -8.99427 + 0.87238 \ln(D^2H)$. If we return naively to the starting data by applying the exponential function, this model becomes: $B = \exp(-8.99427) \times D^{0.87238} = 1.241 \times 10^{-4} D^{0.87238}$. The model fitted by non-linear regression and the model fitted by linear regression on log-transformed data are therefore very similar.

19

Weighted non-linear regression between B, D and H

We have already fitted a power model $B = aD^{b_1}H^{b_2}$ by multiple linear regression on log-transformed data (red line 10). Let us now fit this model directly by non-linear regression:

$$B = aD^{b_1}H^{b_2} + \varepsilon$$

where

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

In order to take account of the heteroscedasticity, and considering that the conditional standard deviation of the biomass derived from dbh is proportional to D^2 (red line 11), we can fit this non-linear model by the weighted least squares method using a weighting inversely proportional to D^4:

```
start <- coef(lm(log(Btot)~I(log(dbh))+I(log(heig)),data=dat[dat$Btot>0,]))
start[1] <- exp(start[1])
names(start) <- c("a","b1","b2")
m <- nls(Btot~a*dbh^b1*heig^b2,data=dat,start=start,weights=1/dat$dbh^4)
summary(m)
```

As previously (red line 17), the `nls` command calls on start values for the coefficients and these are obtained from the coefficients of the multiple regression on log-transformed data. The result of the fitting is as follows:

```
     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  *
```

with a residual standard deviation $k = 0.0003356$ tonnes cm^{-2}. The model is therefore written: $B = 1.003 \times 10^{-4} D^{1.923} H^{0.7435}$. The model is similar to that fitted by multiple regression on log-transformed data (red line 10). But coefficient a is estimated with less precision here than by the multiple regression on log-transformed data.

6.2.2 Estimating the exponent

Let us now consider the case where the exponent c needs to be estimated at the same time as the model's other parameters. This case includes the linear regression with variance model that we mentioned in section 6.1.4. In this case the least squares method is no longer valid. We are therefore obliged to use another fitting method: the maximum likelihood method. The likelihood of an observation $(X_{i1}, \ldots, X_{ip}, Y_i)$ is the probability density of observing $(X_{i1}, \ldots, X_{ip}, Y_i)$ in the specified model. The probability density of a normal distribution of expectation μ and standard deviation σ is:

$$\phi(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

As Y_i follows a normal distribution of expectation $f(X_{i1}, \ldots, X_{ip}; \theta)$ and standard deviation kX_{i1}^c, the likelihood of the ith observation is:

$$\frac{1}{kX_{i1}^c\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{Y_i - f(X_1, \ldots, X_p; \theta)}{kX_{i1}^c}\right)^2\right]$$

As the observations are independent, their joint likelihood is the product of the likelihoods of each observation. The likelihood of the sample of n observations is therefore:

$$\begin{aligned} \ell(\theta, k, c) &= \prod_{i=1}^{n} \frac{1}{kX_{i1}^c\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{Y_i - f(X_1, \ldots, X_p; \theta)}{kX_{i1}^c}\right)^2\right] \\ &= \frac{1}{(k\sqrt{2\pi})^n} \frac{1}{(\prod_{i=1}^{n} X_{i1})^c} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{Y_i - f(X_1, \ldots, X_p; \theta)}{kX_{i1}^c}\right)^2\right] \end{aligned} \tag{6.19}$$

This likelihood is considered to be a function of parameters θ, k and c.

The better the values of parameters θ, k and c the higher the probability of the observations being obtained in the model corresponding to these parameter values. In other words, the best values for parameters θ, k and c are those that maximize the likelihood of the observations. The corresponding estimator is by definition the maximum likelihood estimator, and is written:

$$(\hat{\theta}, \hat{k}, \hat{c}) = \arg\max_{(\theta,k,c)} \ell(\theta, k, c) = \arg\max_{(\theta,k,c)} \ln[\ell(\theta, k, c)]$$

where the last equality stems from the fact that a function and its logarithm reach their maximum for the same values of their argument. The logarithm of the likelihood, which we

call the log-likelihood and which we write $\mathcal{L}$, is easier to calculate than the likelihood, and therefore, for our calculations, it is the log-likelihood we will be seeking to maximize. In the present case, the log-likelihood is written:

$$\begin{aligned}\mathcal{L}(\theta,\, k,\, c) &= \ln[\ell(\theta,\, k,\, c)] \\ &= -n\ln(k\sqrt{2\pi}) - c\sum_{i=1}^{n}\ln(X_{i1}) - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{Y_i - f(X_1,\,\ldots,\,X_p;\theta)}{kX_{i1}^c}\right)^2 \\ &= -\frac{1}{2}\sum_{i=1}^{n}\left[\left(\frac{Y_i - f(X_1,\,\ldots,\,X_p;\theta)}{kX_{i1}^c}\right)^2 + \ln(2\pi) + \ln(k^2X_i^{2c})\right]\end{aligned} \tag{6.20}$$

To obtain the maximum likelihood estimators of the parameters, we would need to calculate the partial derivatives of the log-likelihood with respect to these parameters, and look for the values where they cancel each other out (while ensuring that the second derivatives are indeed negative). In the general case, there is no analytical solution to this problem. In the same manner as previously for the sum of squares, we will need to use a numerical algorithm to maximize the log-likelihood. We can show that the maximum likelihood method yields a coefficients estimator that is asymptotically (i.e. when the number n of observations tends toward infinity) the best. We can also show that in the case of the linear model, the least squares estimator and the maximum likelihood estimator are the same.

20

Non-linear regression between B and D with variance model

Let us look again at the non-linear regression between biomass and dbh (see red line 17), but this time consider that exponent c in the variance model is a parameter that needs to be estimated like the others. The model is written in the same fashion as before (red line 17):

$$B = aD^b + \varepsilon$$

where

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

but is fitted by the maximum likelihood method:

```
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)
```

The model is fitted by the `nlme` command[1] which, like the `nls` command (red line 17), requires `start` values for the coefficients. These start values are calculated as in red line 17. The result of the fitting is as follows:

[1] The `nlme` command in fact serves to fit mixed effect non-linear models. The `nlreg` command is used to fit non-linear models with variance model, but we have obtained abnormal results with this command (version 3.1-96) which explains why we prefer to use the `nlme` command here, even though there is no mixed effect in the models considered here.

```
        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
```

with an estimated value of exponent $c = 2.090814$. This estimated value is very similar to that evaluated for the weighted non-linear regression ($c = 2$, see in red line 11). The fitted model is therefore: $B = 2.445 \times 10^{-4} D^{2.35105}$, which is very similar to the model fitted by weighted non-linear regression (red line 17).

21

Non-linear regression between B and D^2H with variance model

Let us look again at the non-linear regression between biomass and D^2H (see red line 18), but this time consider that exponent c in the variance model is a parameter that needs to be estimated like the others. The model is written in the same fashion as before (red line 18):

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

where

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

but is fitted by the maximum likelihood method:

```
start <- coef(lm(log(Btot)~I(log(dbh^2*heig)),data=dat[dat$Btot>0,]))
start[1] <- exp(start[1])
names(start) <- c("a","b")
library(nlme)
m <- nlme(Btot~a*(dbh^2*heig)^b,data=cbind(dat,g="a"),fixed=a+b~1,start=start,
groups=~g,weights=varPower(form=~dbh))
summary(m)
```

The model is fitted by the `nlme` command, which, like the `nls` command (red line 17), requires start values for the coefficients. These `start` values are calculated as in red line 17. The result of the fitting is as follows:

```
        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
```

with an estimated value of exponent $c = 2.042586$. This estimated value is very similar to that evaluated for the weighted non-linear regression ($c = 2$, see red line 11). The fitted model is therefore: $B = 8.19 \times 10^{-5} (D^2H)^{0.9122144}$, which is very similar to the model fitted by weighted non-linear regression (red line 18).

22

Non-linear regression between B, D and H with variance model

Let us look again at the non-linear regression between biomass, dbh and height (see red line 19):

$$B = aD^{b_1}H^{b_2} + \varepsilon$$

where

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

but this time consider that exponent c in the variance model is a parameter that needs to be estimated like the others. The fitting by maximum likelihood is as follows:

```
library(nlme)
start <- coef(lm(log(Btot)~I(log(dbh))+I(log(heig)),data=dat[dat$Btot>0,]))
start[1] <- exp(start[1])
names(start) <- c("a","b1","b2")
m <- nlme(Btot~a*dbh^b1*heig^b2,data=cbind(dat,g="a"),fixed=a+b1+b2~1,
start=start,groups=~g,weights=varPower(form=~dbh))
summary(m)
```

and, as previously, requires the provision of `start` values for the coefficients. The fitting yields:

```
        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
```

with an estimated value of exponent $c = 2.055553$. This estimated value is very similar to that evaluated for the weighted non-linear regression ($c = 2$, see red line 11). The fitted model is therefore: $B = 1.109 \times 10^{-4} D^{1.9434876} H^{0.6926256}$, which is very similar to the model fitted by weighted non-linear regression (red line 19).

23

Non-linear regression between B and a polynomial of $\ln(D)$

Previously (red line 9), we used multiple regression to fit a model between $\ln(B)$ and a polynomial of $\ln(D)$. If we look again at the start variable, the model is written:

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

where

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

We will now directly fit this non-linear model by maximum likelihood (such that exponent c is estimated at the same time as the model's other parameters). For a third-order polynomial, the fitting is obtained by:

```
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)
```

and results in:

```
        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
```

with an estimated value of exponent $c = 2.099938$. This result is very similar to that obtained by multiple regression on log-transformed data (red line 9).

6.2.3 Numerical optimization

If, in the case of a non-linear model, we are looking to minimize the sum of squares (when exponent c is known) or maximize the log-likelihood (when exponent c needs to be estimated) we need to use a numerical optimization algorithm. As maximizing the log-likelihood is equivalent to minimizing the opposite of the log-likelihood, we shall, in all that follows, consider only the problem of minimizing a function in a multidimensional space. A multitude of optimization algorithms have been developed (Press *et al.*, 2007, chapitre 10) but the aim here is not to present a review of them all. What simply needs to be recognized at this stage is that these algorithms are iterative and require parameter start values. Based on this start value and each iteration, the algorithm shifts in the parameter space while looking to minimize the target function (i.e. the sum of squares or minus the log-likelihood). The target function may be represented as a hypersurface in the parameter space (Figure 6.17). Each position in this space corresponds to a value of the parameters. A lump in this surface corresponds to a local maximum of the target function, whereas a dip in the surface corresponds to a local minimum. The aim is to determine the overall minimum, i.e. the deepest dip. The position of this dip corresponds to the estimated value of the parameters. If the algorithm gives the position of a dip that is not the deepest, the estimation of the parameters is false.

Descending algorithm

The simplest optimization algorithm consists in calculating successive positions, i.e. successive parameter values, by descending the surface defined by the target function along the line that has the steepest slope (Figure 6.17A). This algorithm leads to one dip in the surface, but nothing says that this dip is the deepest for the surface may have several watersheds with several dips. Depending on the starting position, the algorithm will converge toward one dip or another (Figure 6.17B). Also, even two starting positions very close together may be located on either side of a crest separating the two watersheds, and will therefore result in different dips, i.e. in different parameter estimations. The only case where this algorithm gives the correct parameter estimation regardless of starting value is when the surface has only a single dip, i.e. when the target function is convex. This in particular is the case for the linear model, but generally not for the non-linear model.

Improving algorithms in the case of local minima

More subtle algorithms have been developed than those descending along the steepest slope. For example, these may include the possibility to climb out of a dip into which the algorithm has temporarily converged in order to determine whether there might not be a deeper dip in the neighborhood. But no algorithm, even the most subtle, offers the certainty that it has converged to the deepest dip. Thus, any numerical optimization algorithm (i) may

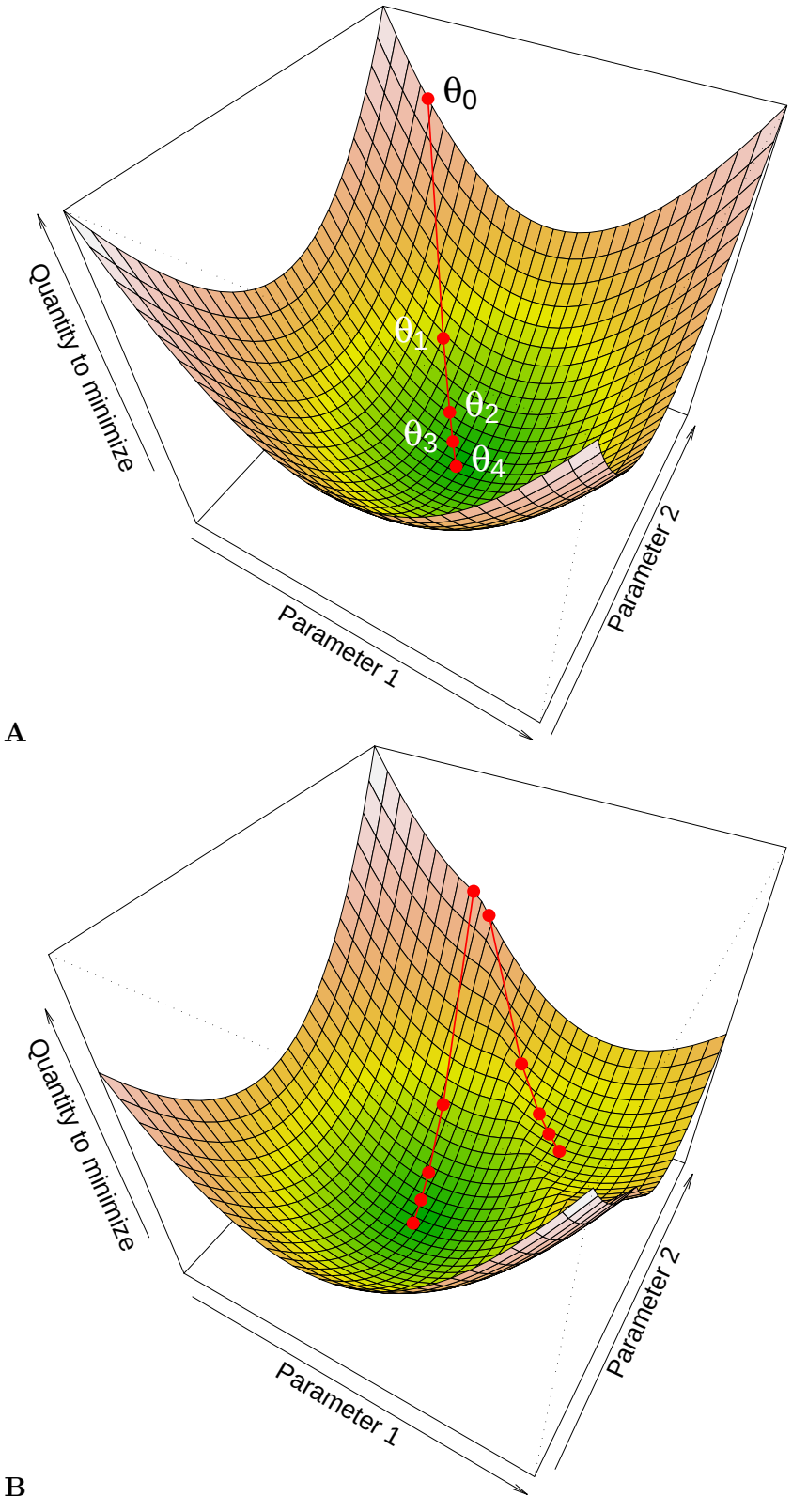


Figure 6.17 – *Representation of the target function (i.e. the quantity to be minimized) as a surface in parameter space. Each position in this space corresponds to a value of the parameters. The successive values θ_1, θ_2, ..., for the parameters are obtained from a start value θ_0 by descending the surface along the line with the steepest slope.(A) The surface has a single watershed. (B) The surface has several watersheds.*

be trapped by a local minimum instead of converging to the overall minimum, and (*ii*) is sensitive to the indicated start position, which in part determines the final position toward which the algorithm will converge.

If we now return to the issue at hand, this means (*i*) fitting a non-linear model could yield erroneous parameter estimations, and (*ii*) selecting parameter start values for the algorithm is a sensitive issue. Herein lies the main drawback of fitting a non-linear model, and if it is to be avoided, the parameter start values must be carefully selected and, above all, several values must be tested.

Selecting parameter start values

When mean model f can be transformed into a linear relation between the response variable Y and the effect variables $X_1, \ldots, X_p$, a start value for the coefficients may be obtained by fitting a linear regression on the transformed variables without considering the possible heteroscedasticity of the residuals. Let us for example consider a power type biomass table:

$$B = aD^{b_1}H^{b_2}\rho^{b_3} + \varepsilon \tag{6.21}$$

where

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

The power model for the expectation of B can be rendered linear by log-transforming the variables: $\ln(B) = a' + b_1 \ln(D) + b_2 \ln(H) + b_3 \ln(\rho)$. But this transformation is incompatible with the additivity of the errors in the model (6.21). In other words, the multiple regression of the response variable $\ln(B)$ against the effect variables $\ln(D)$, $\ln(H)$ and $\ln(\rho)$:

$$\ln(B) = a' + b_1 \ln(D) + b_2 \ln(H) + b_3 \ln(\rho) + \varepsilon' \tag{6.22}$$

where $\varepsilon' \sim \mathcal{N}(0, \sigma)$, is not a model equivalent to (6.21), even when the residuals ε' of this model have a constant variance. Even though models (6.21) and (6.22) are not mathematically equivalent, the coefficients of (6.22) — estimated by multiple regression — may serve as start values for the numerical algorithm that estimates the coefficients of (6.21). If we write as $x^{(0)}$ the start value of parameter x for the numerical optimization algorithm, we thus obtain:

$$a^{(0)} = \exp(\hat{a}'), \quad b_i^{(0)} = \hat{b}_i, \quad k^{(0)} = \hat{\sigma}, \quad c^{(0)} = 0$$

Sometimes, the mean model cannot be rendered linear. An example of this is the following biomass table used for trees in a plantation (Saint-André *et al.*, 2005):

$$B = a + [b_0 + b_1 T + b_2 \exp(-b_3 T)]D^2 H + \varepsilon$$

where T is the age of the plantation and $\varepsilon \sim \mathcal{N}(0, kD^c)$, which has a mean model that cannot be rendered linear. In this case, parameter start values must be selected empirically. For this current example we could for instance take:

$$a^{(0)} = \hat{a}, \quad b_0^{(0)} + b_2^{(0)} = \hat{b}_0, \quad b_1^{(0)} = \hat{b}_1, \quad b_3^{(0)} = 0, \quad k^{(0)} = \hat{\sigma}, \quad c^{(0)} = 0$$

where $\hat{a}$, $\hat{b}_0$, $\hat{b}_1$ and $\hat{\sigma}$ are estimated values for the coefficients and residual standard deviations of the multiple regression of B against D^2H and D^2HT. Selecting parameter start values does not mean that several start values should not be tested. Therefore, when fitting a non-linear model with a numerical optimization algorithm, it is essential to test several parameter start values to ensure that the estimations are stable.

6.3 Selecting variables and models

When we look to construct a volume or biomass table, the graphical exploration of the data (chapter 5) generally yields several possible model forms. We could fit all these potentially interesting models, but ultimately, which of all these fitted models should be recommended to the user? Selecting variables and selecting models aims to determine which is the "best" possible expression of the model among all those fitted.

6.3.1 Selecting variables

Let us take the example of a biomass table we are looking to construct from a dataset that includes tree diameter (dbh) and height, and wood specific density. By working on log-transformed data, and given the variables they include, we may fit the following models:

$$\begin{aligned}
\ln(B) &= a_0 + a_1 \ln(D) + \varepsilon \\
\ln(B) &= a_0 + a_2 \ln(H) + \varepsilon \\
\ln(B) &= a_0 + a_3 \ln(\rho) + \varepsilon \\
\ln(B) &= a_0 + a_1 \ln(D) + a_2 \ln(H) + \varepsilon \\
\ln(B) &= a_0 + a_1 \ln(D) + a_3 \ln(\rho) + \varepsilon \\
\ln(B) &= a_0 + a_2 \ln(H) + a_3 \ln(\rho) + \varepsilon \\
\ln(B) &= a_0 + a_1 \ln(D) + a_2 \ln(H) + a_3 \ln(\rho) + \varepsilon
\end{aligned}$$

The *complete* model (the last in the above list) is that which includes all the effect variables available. All the other models may be considered to be subsets of the complete model, but where certain effect variables have been employed and other set aside. Selecting the variables aims to choose — from among the effect variables of the complete model — those that should be kept and those that may be set aside as they contribute little to the prediction of the response variable. In other words, in this example, selecting the variables would consist in choosing the best model from among the seven models envisaged for $\ln(B)$.

Given that there are p effect variables X_1, X_2, ..., X_p, there are $2^p - 1$ models that include all or some of these effect variables. Selecting the variables consists in choosing the "best" combination of effect variables from all those available. This means firstly that we must have criterion that can be used to evaluate the quality of a model. We have already seen (p.116) that R^2 is a poor criterion for evaluating the quality of one model in comparison with that of another as it increases automatically with the number of effect variables, and this regardless of whether these provide information useful for predicting the response variable or not. A better criterion for selecting effect variables is the residual variance estimator, which is linked to R^2 by the relation:

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

where S_Y^2 is the empirical variance of the response variable.

Several methods may be used to select the best combination of effect variables. If p is not too high, we can review all the $2^p - 1$ possible models exhaustively. When p is too high, a step by step method may be used to select the variables. Step-by-step methods proceed by the successive elimination or successive addition of effect variables. The descending method consists in eliminating the least significant of the p variables. The regression is then recalculated and the process repeated until a stop criterion is satisfied (e.g. when all model

coefficients are significantly different from zero). The ascending method proceeds in the opposite direction: we start with the best single-variable regression and add the variable that increases R^2 the most until the stop criterion is satisfied.

The so-called *stepwise* method is a further improvement upon the previous algorithm that consists in performing an additional Fisher's significance test at each step such as not to introduce a non-significant variable and possibly eliminate variables that have already been introduced but are no longer informative given the last variable selected. The algorithm stops when no more variables can be added or withdrawn. The different step-by-step selection methods do not necessarily give the same result, but the "stepwise" method would appear to be the best. They do not, however, safeguard from the untimely removal of variables that are actually significant, which may well bias the result. And in connection with this, it should be recalled that if we know (for biological reasons) why a variable should be included in a model (e.g. wood specific density), it is not because a statistical test declares it non-significant that it should be rejected (because of the test's type II error).

24

Selecting variables

Let us select the variables $\ln(D)$, $[\ln(D)]^2$, $[\ln(D)]^3$, $\ln(H)$ to predict the log of the biomass. The complete model is therefore:

$$\ln(B) = a_0 + a_1 \ln(D) + a_2[\ln(D)]^2 + a_3[\ln(D)]^3 + a_4 \ln(H) + \varepsilon$$

where

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

Variables are selected in R using the `step` command applied to the fitted complete model:

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

which yields:

```
                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(heig))    1.01874      0.17950    5.675  1.59e-06  ***
```

The variables selected are therefore $[\ln(D)]^2$ and $\ln(H)$. The model finally retained is therefore: $\ln(B) = -6.50202 + 0.23756[\ln(D)]^2 + 1.01874 \ln(H)$, with a residual standard deviation of 0.3994 and $R^2 = 0.974$.

6.3.2 Selecting models

Given two competitor models that predict the same response variable to within one transformation of a variable, which should we choose? Let us look at a few different cases before answering this question.

Nested models

The simplest case is where the two models to be compared are nested. A model is *nested* in another if the two predict the same response variable and if we can move from the second to the first by removing one or several effect variables. For example, the biomass table $B = a_0 + a_1 D + \varepsilon$ is nested in $B = a_0 + a_1 D + a_2 D^2 H + \varepsilon$ since we can move from the second to the first by deleting $D^2 H$ from the effect variables. Likewise, the model $B = a_0 + a_2 D^2 H + \varepsilon$ is nested in $B = a_0 + a_1 D + a_2 D^2 H + \varepsilon$ since we can move from the second to the first by deleting D from the effect variables. By contrast, the model $B = a_0 + a_1 D + \varepsilon$ is not nested in $B = a_0 + a_2 D^2 H + \varepsilon$. Let p be the number of effect variables in the complete model and $p' < p$ be the number of effect variables in the nested model. Without loss of generality, we can write the complete model as:

$$Y = f(X_1, \ldots, X_{p'}, X_{p'+1}, \ldots, X_p; \theta_0, \theta_1) + \varepsilon \tag{6.23}$$

where (θ_0, θ_1) is the vector of the coefficients associated with the complete model and θ_0 is the vector of the coefficients associated with the nested model, which may be obtained by setting $\theta_1 = \mathbf{0}$. In the particular case of the linear model, the complete model is obtained as the sum of the nested model and additional terms:

$$\underbrace{\underbrace{Y = a_0 + a_1 X_1 + \ldots + a_{p'} X_{p'}}_{\text{nested model}} + a_{p'+1} X_{p'+1} + \ldots + a_p X_p + \varepsilon}_{\text{complete model}} \tag{6.24}$$

where $\theta_0 = (a_0, \ldots, a_{p'})$ and $\theta_1 = (a_{p'+1}, \ldots, a_p)$.

In the case of nested models, a statistical test can be used to test one of the models against the other. The null hypothesis of this test is that $\theta_1 = \mathbf{0}$, i.e. the additional terms are not significant, which can also be expressed as: the nested model is better than the complete model. If the p-value of this test proves to be below the significance level (typically 5 %), then the null hypothesis is rejected, i.e. the complete model is best. Conversely, if the p-value is above the significance threshold, the nested model is considered to be the best.

In the case of the linear model (6.24), the test statistic is a ratio of the mean squares, which under the null hypothesis follows Fisher's distribution. This is the same type of test as that used to test the overall significance of a multiple regression, or that used in the "stepwise" method of selecting variables. In the general case of the non-linear model (6.23), the test statistic is a likelihood ratio, such that $-2\log(\text{likelihood ratio})$ under the null hypothesis follows a χ^2 distribution.

25

Testing nested models: $\ln(D)$

In red line 24 the variable $[\ln(D)]^2$ was selected with $\ln(H)$ as effect variable of $\ln(B)$ but not $\ln(D)$. Model $\ln(B) = a_0 + a_1 \ln(D) + a_2[\ln(D)]^2 + a_4 \ln(H)$, which includes the additional term $\ln(D)$, can be compared with the model $\ln(B) = a_0 + a_2[\ln(D)]^2 + a_4 \ln(H)$ using the nested models test. In R, the `anova` command can be used to test a nested model, with the first argument being the nested model and the second being the complete model:

```
comp <- lm(log(Btot)~I(log(dbh))+I(log(dbh)^2)+I(log(heig)),data=dat[dat$Btot>0,])
nest <- lm(log(Btot)~I(log(dbh)^2)+I(log(heig)),data=dat[dat$Btot>0,])
anova(nest,comp)
```

The test gives the following result:

```
  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
```

The p-value is 0.3169, therefore greater than 5 %. The nested model (without $\ln(D)$) is therefore selected rather than the complete model.

26

Testing nested models: $\ln(H)$

The model $\ln(B) = -8.42722 + 2.36104\ln(D)$ was obtained in red line 7 whereas red line 10 gave the model $\ln(B) = -8.9050 + 1.8654\ln(D) + 0.7083\ln(H)$. As the first is nested in the second, we can test for which is the best. The command

```
comp <- lm(log(Btot)~I(log(dbh))+I(log(heig)),data=dat[dat$Btot>0,])
nest <- lm(log(Btot)~I(log(dbh)),data=dat[dat$Btot>0,])
anova(nest,comp)
```

yields:

```
  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  **
```

As the p-value is less than 5 %, the complete model (including $\ln(H)$ as effect variable) is selected rather than the nested model.

Models with the same response variable

When we want to compare two models that have the same response variable but are not nested, we can no longer use a statistical test. For example, we cannot use the above-mentioned test to compare $B = a_0 + a_1 D + \varepsilon$ and $B = a_0 + a_2 D^2 H + \varepsilon$. In this case, we must use an information criterion (Bozdogan, 1987; Burnham & Anderson, 2002, 2004). There are several, suited to different contexts. The most widespread are the "Bayesian information criterion", or BIC, and above all the Akaike (1974) information criterion (or AIC). The AIC is expressed as:

$$\mathrm{AIC} = -2\ln\ell(\hat{\theta}) + 2q$$

where $\ell(\hat{\theta})$ if the model's likelihood, i.e. the likelihood of the sample for the values estimated from model parameters (see equation 6.19), and q is the number of free parameters estimated. In particular, in the case of a multiple regression against p effect variables, $q = p + 1$ (i.e. the p coefficients associated with the p effect variables plus the y-intercept). The coefficient -2 in front of the log-likelihood in the AIC expression is identical to that in the test statistic on the likelihood ratio in the case of nested models. Given two models with the same number of parameters, the best model is that with the highest likelihood, therefore that with the smallest AIC. At equal likelihoods, the best model is that with the fewest parameters (in accordance with the principle of Occam's razor), and therefore is once more that with the smallest AIC. When all is said and done, the best model is that with the smallest value of AIC.

The BIC is similar in expression to the AIC, but with a term that penalizes more strongly the number of parameters:

$$\text{BIC} = -2\ln \ell(\hat{\theta}) + q\ln(n)$$

where n is the number of observations. Here again, the best model is that with the smallest value of BIC. When fitting volume or biomass tables, AIC is used rather than BIC as a model selection criterion.

27

Selecting models with B as response variable

The following models with B as response variable were fitted:

- red line 12 or 16: $B = -3.840 \times 10^{-3} D + 1.124 \times 10^{-3} D^2$
- red line 14: $B = -3.319456 \times 10^{-3} D + 1.067068 \times 10^{-3} D^2$
- red line 17: $B = 2.492 \times 10^{-4} D^{2.346}$
- red line 20: $B = 2.445 \times 10^{-4} D^{2.35105}$
- red line 11 ou 15: $B = 2.747 \times 10^{-5} D^2 H$
- red line 13: $B = 2.740688 \times 10^{-5} D^2 H$
- red line 18: $B = 7.885 \times 10^{-5} (D^2 H)^{0.9154}$
- red line 21: $B = 8.19 \times 10^{-5} (D^2 H)^{0.9122144}$
- red line 19: $B = 1.003 \times 10^{-4} D^{1.923} H^{0.7435}$
- red line 22: $B = 1.109 \times 10^{-4} D^{1.9434876} H^{0.6926256}$

The models in red lines 12, 14, 11 and 13 are fitted by linear regression while the others are fitted by non-linear regression. These models have five distinct forms, with two fitting methods for each: using a weighted regression by the weighted least squares method (red lines 12, 17, 11, 18 and 19) or using a regression with variance model by the maximum likelihood method (red lines 14, 20, 13, 21 and 22). The predictions made by these different models are shown in Figure 6.18. Let `m` be one of the fitted models with dbh as the only entry. A plot of the predictions made by this model may be obtained as follows:

```
with(dat,plot(dbh,Btot,xlab="Dbh(cm)",ylab="Biomass (t)"))
D <- seq(par("usr")[1],par("usr")[2],length=200)
lines(D,predict(m,newdata=data.frame(dbh=D)),col="red")
```

For a model `m` that has dbh and height as entries, its predictions may be obtained as follows:

```
D <- seq(0,180,length=20)
H <- seq(0,61,length=20)
B <- matrix(predict(m,newdata=expand.grid(dbh=D,height=H)),length(D))
```

and a plot of the biomass response surface against diameter and height may be obtained by:

```
M <- persp(D,H,B,xlab="Dbh (cm)",ylab="Height (m)",zlab="Biomass (t)",
ticktype="detailed")
points(trans3d(dat$dbh,dat$heig,dat$Btot,M))
```

Table 6.1 – *AIC values for 10 biomass tables fitted to data from 42 trees measured in Ghana by Henry et al. (2010). These 10 tables directly predict biomass.*

Red line	Entry	Fitting* method	Command R	AIC
12	D	WLS	`lm`	76.71133
14	D	ML	`nlme`	83.09157
17	D	WLS	`nls`	24 809.75727
20	D	ML	`nlme`	75.00927
11	D^2H	WLS	`lm`	65.15002
13	D^2H	ML	`nlme`	69.09644
18	D^2H	WLS	`nls`	24 797.53706
21	D^2H	ML	`nlme`	69.24482
19	D, H	WLS	`nls`	24 802.91248
22	D, H	ML	`nlme`	76.80204

*WLS = weighted least squares, ML = maximum likelihood

Given a fitted model `m`, its AIC may be calculated by the command:

```
AIC(m)
```

AIC values for the 10 models listed above are given in Table 6.1. This table illustrates a problem that arises with several statistical software packages, including R: when we maximize the log-likelihood (6.20), the constants (such as $-n\ln(2\pi)/2$) no longer play any role. The constant we use to calculate the log-likelihood, and by consequence AIC, is therefore a matter of convention, and different constants are used depending on the calculation. Thus, in Table 6.1, we can see that the values of AIC in models fitted by the `nls` command are substantially higher than those in the other models: it is not that these models are substantially worse than the others, it is simply that the `nls` command uses a constant different from the others to calculate the log-likelihood. Therefore, with R, it should be remembered that the values of AIC should be compared only for models that have been fitted using the same command. In our present case, if we compare the two models that were fitted with the `lm` command, the best (i.e. that with the smallest AIC) is that which has D^2H as effect variable (red line 11). If we compare the five models fitted with the `nlme` command, the best is again that with D^2H as effect variable (red line 13). And if we compare the three models fitted with the `nls` command, the best is yet again that with D^2H as effect variable (red line 18). It may therefore be concluded that whatever fitting method is used, the biomass table that uses D^2H as effect variable is the best.

28

Selecting models with $\ln(B)$ as response variable

The following models with $\ln(B)$ as response variable were fitted:

- red line 7 or 9: $\ln(B) = -8.42722 + 2.36104\ln(D)$
- red line 8: $\ln(B) = -8.99427 + 0.87238\ln(D^2H)$

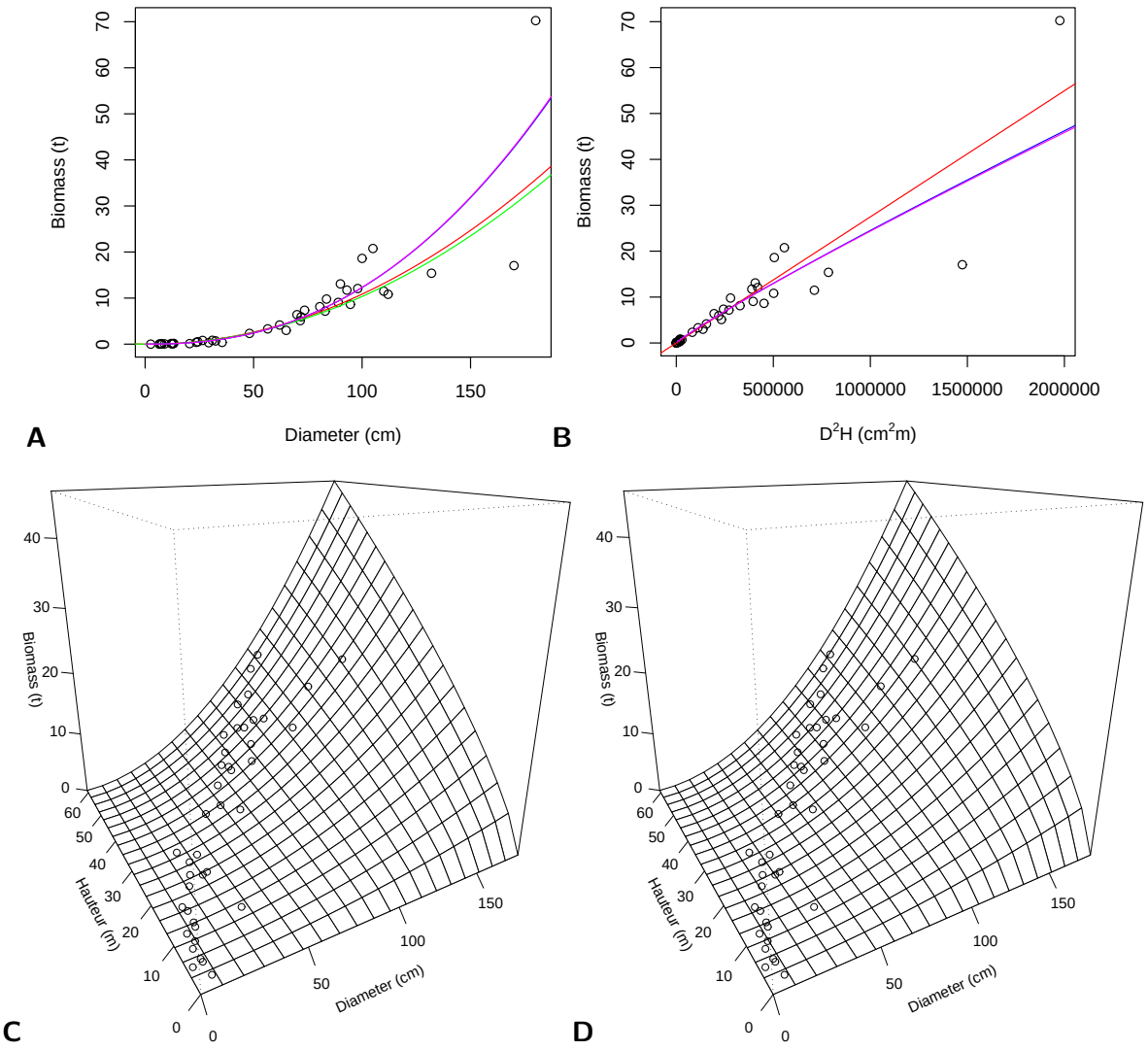


Figure 6.18 – *Biomass predictions by different tables fitted to data from 42 trees measured in Ghana by Henry et al. (2010).The data are shown as points. (A) Tables with dbh as only entry, corresponding to red lines 12 (red), 14 (green), 17 (blue) and 20 (magenta). (B) Tables with D^2H as the only effect variable, corresponding to red lines 11 (red), 13 (green), 18 (blue) and 21 (magenta). (C) Table corresponds to red line 19. (D) Table corresponds to red line 22.*

Table 6.2 – *AIC values for four biomass tables fitted to data from 42 trees measured in Ghana by Henry et al. (2010). These four tables predict the log of the biomass and are all fitted using linear regression by the ordinary least squares (OLS) method.*

Red line	Entry	Fitting method	Command R	AIC
7	D	OLS	`lm`	56.97923
8	D^2H	OLS	`lm`	46.87780
10	D, H	OLS	`lm`	48.21367
24	D, H	OLS	`lm`	45.96998

- red line 10: $\ln(B) = -8.9050 + 1.8654\ln(D) + 0.7083\ln(H)$
- red line 24: $\ln(B) = -6.50202 + 0.23756[\ln(D)]^2 + 1.01874\ln(H)$

All these models were fitted using linear regression by the ordinary least squares method. A plot of the predictions on a log scale for model `m` with dbh as only entry may be obtained by the following command:

```
with(dat,plot(dbh,Btot,xlab="Dbh (cm)",ylab="Biomass (t)",log="xy"))
D <- 10^par("usr")[1:2]
lines(D,exp(predict(m1,newdata=data.frame(dbh=D))))
```

For a model that uses both dbh and height as entries, a plot on a log scale may be obtained by the command:

```
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,heig=H)),length(D))
M <- persp(log(D),log(H),B,xlab="log(Diameter) (cm)",ylab="log(height) (m)",zlab=
"log(Biomass) (t)",ticktype="detailed")
points(trans3d(log(dat$dbh),log(dat$heig),log(dat$Btot),M))
```

Predictions of $\ln(B)$ by the four models are given in Figure 6.19. Given a fitted model `m`, its AIC may be calculated by the command:

```
AIC(m)
```

Table 6.2 gives AIC values for the four models. As all four models were fitted by the same `lm` command, the AIC values are directly comparable. The best model, i.e. that with the smallest AIC, is the fourth (red line 24). It should also be noted that an AIC-based classification of these models is entirely consistent with the results of the nested models tests performed previously (red lines 25 and 26).

Models with different response variables

The more general case is when we want to compare two models that do not have the same response variable because one is a transform of the other. For example, the models $B = aD^b + \varepsilon$ and $\ln(B) = a + b\ln(D) + \varepsilon$ both predict biomass, but the response variable is B in one case and $\ln(B)$ in the other. In this case, we cannot use the AIC or BIC

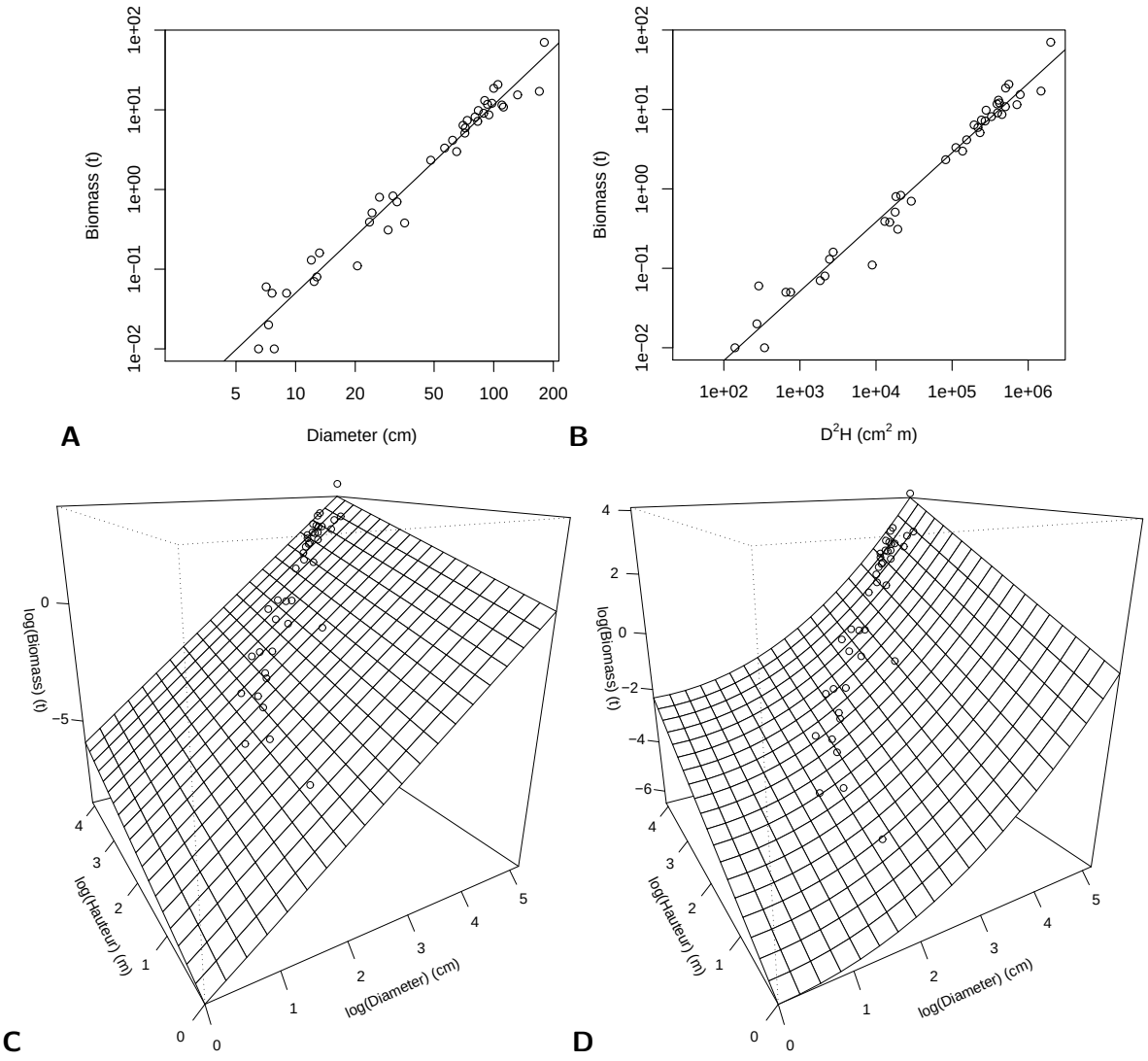


Figure 6.19 – *Biomass predictions by different tables fitted to data from 42 trees measured in Ghana by Henry et al. (2010). The data are shown as points. (A) Model in red line 7. (B) Model in red line 8. (C) Model in red line 10. (D) Model in red line 24.*

to compare the models. By contrast, Furnival (1961) index can in this case be used to compare the models. The model with the smallest Furnival's index is considered to be the best (Parresol, 1999).

Furnival's index is defined only for a model whose residual error ε has a variance that is assumed to be constant: $\text{Var}(\varepsilon) = \sigma^2$. By contrast, no constraint is imposed on the form of the transformation of the variable linking the modeled response variable Y to the variable of interest (volume or biomass). Let us consider the case of a biomass model (that can immediately be transposed into a volume model) and let ψ be this variable transformation: $Y = \psi(B)$. Furnival's index is defined by:

$$F = \frac{\hat{\sigma}}{\sqrt[n]{\prod_{i=1}^{n} \psi'(B_i)}} = \exp\Big(-\frac{1}{n} \sum_{i=1}^{n} \ln[\psi'(B_i)] \Big)\, \hat{\sigma}$$

where $\hat{\sigma}$ if the estimation of the residual standard deviation of the fitted model and B_i is the biomass of the ith tree measured. When there is no transformation of variables, ψ is the identity function and Furnival's index F is then equal to the residual standard deviation $\hat{\sigma}$. The most common variable transformation is the log transformation: $\psi(B) = \ln(B)$ and $\psi'(B) = 1/B$, in which case Furnival's index is:

$$F_{\ln} = \hat{\sigma} \sqrt[n]{\prod_{i=1}^{n} B_i} = \exp\Big(\frac{1}{n} \sum_{i=1}^{n} \ln(B_i) \Big)\, \hat{\sigma}$$

For linear regressions where the residual variance is assumed to be proportional to a power of an effect variable X_1, a trick can nevertheless be used to define Furnival's index. The linear regression:

$$Y = a_0 + a_1 X_1 + a_2 X_2 + \ldots + a_p X_p + \varepsilon \tag{6.25}$$

where $\text{Var}(\varepsilon) = (k X_1^c)^2$; is strictly identical to the linear regression (see p.132):

$$Y' = a_0 X_1^{-c} + a_1 X_1^{1-c} + a_2 X_2 X_1^{-c} + \ldots + a_p X_p X_1^{-c} + \varepsilon' \tag{6.26}$$

where $Y' = Y X_1^{-c}$, $\varepsilon' = \varepsilon X_1^{-c}$ and $\text{Var}(\varepsilon') = k^2$. As model (6.26) has constant residual variance, its Furnival's index is defined. By extension, we can define the Furnival index of model (6.25) as being the Furnival index of model (6.26). If $Y = \psi(B)$, then $Y' = X_1^{-c}\psi(B)$, such that Furnival's index is:

$$F = \frac{\hat{k}}{\sqrt[n]{\prod_{i=1}^{n} X_{i1}^{-c} \psi'(B_i)}} = \exp\Big(\frac{1}{n} \sum_{i=1}^{n} \{ c \ln(X_{i1}) - \ln[\psi'(B_i)] \} \Big)\, \hat{k}$$

Therefore, Furnival's index can be used to select the value of exponent c in a weighted regression (see p.123).

6.3.3 Choosing a fitting method

Let us return to the method used to fit a volume or biomass model. Many solutions may be available to fit a model. Let us for instance consider the biomass model:

$$B = a \rho^{b_1} D^{b_2} H^{b_3} + \varepsilon$$

where

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

This model could be adjusted as a non-linear model (*i*) by the weighted least squares method (c fixed in advance) or (*ii*) by the maximum likelihood method (c not fixed in advance). If we apply a log transformation to the data, we could (*iii*) fit the multiple regression:

$$\ln(B) = a' + b_1 \ln(\rho) + b_2 \ln(D) + b_3 \ln(H) + \varepsilon$$

where

$$\varepsilon \sim \mathcal{N}(0,\, \sigma)$$

Thus, for a given model that predicts biomass as a power of effect variables, we have three possible fitting methods. Obviously, methods (*i*)–(*ii*) and (*iii*) are based on different hypotheses concerning the structure of the residual errors: additive error with regard to B in cases (*i*) and (*ii*), multiplicative error with regard to B in case (*iii*). But both these error types reflect data heteroscedasticity, and therefore fitting methods (*i*), (*ii*) and (*iii*) all have a chance of being valid.

As another example, let us consider the biomass table:

$$B = \exp\{a_0 + a_1 \ln(D) + a_2[\ln(D)]^2 + a_3[\ln(D)]^3 + a_4 \ln(\rho)\} + \varepsilon$$

where

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

Here again, we can (*i*) fit a non-linear model by the least squares method (specifying c in advance), (*ii*) fit a non-linear model by the maximum likelihood method (estimating c), or (*ii*) fit a multiple regression on log-transformed data:

$$\ln(B) = a_0 + a_1 \ln(D) + a_2[\ln(D)]^2 + a_3[\ln(D)]^3 + a_4 \ln(\rho) + \varepsilon$$

where

$$\varepsilon \sim \mathcal{N}(0,\, \sigma)$$

Here again, the structure of the errors is not the same in the three cases, but all can correctly reflect the heteroscedasticity of the biomass. In most cases, the different fitting methods give very similar results in terms of predictions. Should any doubt persist regarding the most appropriate fitting method, then model selection methods may be employed to decide. But in practice the choice of a particular fitting method results rather from the importance given to the respective advantages and drawbacks of each method. Multiple regression has the drawback of imposing constraints on the form of the residuals, and has less flexibility for the mean model. By contrast, it has the advantage of offering an explicit expression for the estimators of the model's coefficients: there is therefore no risk of having erroneous estimators for the coefficients. The non-linear model has the advantage of not imposing any constraints on the mean model or the variance model. Its drawback is that it does not have an explicit expression for parameter estimators: there is therefore the risk of having erroneous parameter estimators.

29

Power model fitting methods

We have already taken a look at three methods for fitting the power model $B = aD^b$:

1. using a simple linear regression on log-transformed data (red line 7): $\ln(B) = -8.42722 + 2.36104 \ln(D)$, if we "naively" apply the exponential inverse transformation;

2. using a weighted non-linear regression (red line 17): $B = 2.492 \times 10^{-4} D^{2.346}$;
3. using a non-linear regression with variance model (red line 20): $B = 2.445 \times 10^{-4} D^{2.35105}$.

The predictions given by these three fittings of the same model are illustrated in Figure 6.20. The differences can be seen to be minimal and well below prediction precision, as we will see later (§ 7.2).

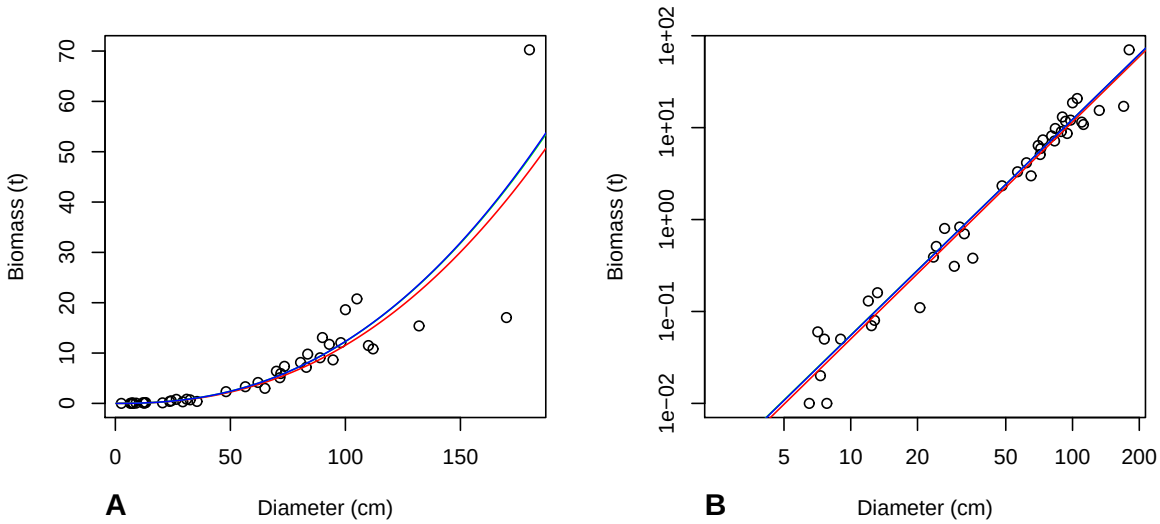


Figure 6.20 – *Biomass predictions by the same power table fitted in three different ways to data from 42 trees measured in Ghana by Henry et al. (2010). The data are shown as points. The red line is the linear regression on log-transformed data (red line 7). The green line (practically superimposed by the blue line) is the fitting by weighted non-linear regression (red line 17). The blue color shows the fitting by non-linear regression with variance model (red line 20). (A) No data transformation. (B) On a log scale.*

6.4 Stratification and aggregation factors

Up until now we have considered that the dataset used to fit the volume or biomass model is homogeneous. In reality, the dataset may be drawn from measurements made under different conditions, or may have been generated by merging several separate datasets. Covariables are generally used to describe this dataset heterogeneity. For example, a covariable may indicate the type of forest in which the measurements were made (deciduous, semi-deciduous, evergreen, etc.), the type of soil, or year planted (if a plantation), etc. A crucial covariable for multispecific data is tree species. Initially, all these covariables that can explain the heterogeneity of a dataset were considered as qualitative variables (or factors). The various modalities of these factors define strata, and a well constituted dataset is one that is sampled in relation to previously identified strata (see § 2.2.3). How can these qualitative covariables be taken into account in a volume or biomass model? Is it valid to analyze the dataset in an overall manner? Or should we analyze the data subsets corresponding to each stratum separately? These are the questions we will be addressing here (§ 6.4.1).

Also, biomass measurements are made separately for each compartment in the tree (see chapter 3). Therefore, in addition to an estimation of its total biomass, we also have, for each tree in the sample, an estimation of the biomass of its foliage, the biomass of its truck, of its large branches, of its small branches, etc. How can we take account of these different compartments when establishing biomass models? We will also be addressing this question (§ 6.4.2).

6.4.1 Stratifying data

Let us now consider that there are qualitative covariables that stratify the dataset into S strata. As each stratum corresponds to a cross between the modalities of the qualitative covariables (in the context of experimental plans we speak of *treatment* rather than stratum) we will not consider each covariable separately. For example, if one covariable indicates the type of forest with three modalities (let us say: deciduous forest, semi-deciduous forest and evergreen forest) and another covariable indicates the type of soil with three modalities (let us say: sandy soil, clay soil, loamy soil), the cross between these two covariables gives $S = 3 \times 3 = 9$ strata (deciduous forest on sandy soil, deciduous forest on clay soil, etc.). We are not looking to analyze the forest type effect separately, or the soil type effect separately. Also, if certain combinations of the covariable modalities are not represented in the dataset, this reduces the number of strata. For example, if there are no evergreen forests on loamy soil, the number of strata $S = 8$ instead of 9.

Therefore, when stratifying a dataset, one strategy consists in fitting a model separately for each stratum. In the case of a multiple regression, this would be written:

$$Y_s = a_{0s} + a_{1s}X_{1s} + a_{2s}X_{2s} + \ldots + a_{ps}X_{ps} + \varepsilon_s$$

where

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

where $(Y_s,\, X_{1s},\, \ldots,\, X_{ps})$ designates an observation relative to stratum s, for $s = 1,\, \ldots,\, S$. There are now $S \times (p+1)$ coefficients to be estimated. An alternative strategy consists in analyzing the dataset as a whole, by fitting a model such as:

$$Y_s = a_{0s} + a_{1s}X_{1s} + a_{2s}X_{2s} + \ldots + a_{ps}X_{ps} + \varepsilon \tag{6.27}$$

where

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

This model differs only in the structure of the error. This type of model is called an analysis of *covariance*. It assumes that all the residuals have the same variance, not only within each stratum, but also from one stratum to another. An analysis of covariance can test whether there is a stratum effect on the response variable, alone or in interaction with each of the effect variables $X_1,\, \ldots,\, X_p$. Testing the main effect of the stratification is equivalent to testing the null hypothesis $a_{01} = a_{02} = \ldots = a_{0S}$. The test statistic is a mean squares ratio which, under the null hypothesis, follows Fisher's distribution. Testing the effect of the interaction between the stratification and the jth effect variable is equivalent to testing the null hypothesis $a_{j1} = a_{j2} = \ldots = a_{jS}$. As previously, the test statistic is a mean squares ratio which, under the null hypothesis, follows Fisher's distribution.

The advantage of testing these effects is that each time one proves not to be significant, the S coefficients $a_{j1},\, a_{j2},\, \ldots,\, a_{jS}$ to be estimated can be replaced by a single common coefficient a_j. Let us imagine, for example, that in analysis of covariance (6.27), the main

effect of the stratum is not significant, and neither is the interaction between the stratum and the first p' effect variables (where $p' < p$). In this case the model to be fitted is:

$$Y_s = a_0 + a_1 X_{1s} + \ldots + a_{p'} X_{p's} + a_{p'+1,s} X_{p'+1,s} + \ldots + a_{ps} X_{ps} + \varepsilon$$

where $\varepsilon \sim \mathcal{N}(0, \sigma)$. This model now includes "only" $p' + 1 + (p - p')S$ coefficients to be estimated, instead of $(p+1)S$ coefficients if we fit the model separately for each stratum. As all the observations serve to estimate common coefficients $a_0, \ldots, a_{p'}$, these are estimated more precisely than if we fit the model separately for each stratum.

This principle of an analysis of covariance can be immediately extended to the case of a non-linear model. Here again, we can test whether or not the coefficients are significantly different between the strata, and if necessary estimate a common coefficient for all the strata.

30

Specific biomass model

In red line 8, we used simple linear regression to fit a power model with D^2H as effect variable to log-transformed data: $\ln(B) = a + b \ln(D^2H)$. We can now include species-related information in this model to test whether the coefficients a and b differ from one species to another. The model corresponds to an analysis of covariance:

$$\ln(B_s) = a_s + b_s \ln(D_s^2 H_s) + \varepsilon$$

where

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

where index s designates the species. This model is fitted by the command:

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

To test whether the coefficients a and b differ from one species to another, we can use the command:

```
anova(m)
```

which yields:

```
                           Df  Sum Sq Mean Sq  F value    Pr(>F)
species                    15 117.667   7.844  98.4396 1.647e-13 ***
I(log(dbh^2*heig))          1 112.689 112.689 1414.1228 < 2.2e-16 ***
species:I(log(dbh^2*heig))  7   0.942   0.135    1.6879    0.1785
Residuals                  17   1.355   0.080
```

The first line in the table tests for a species effect, i.e. whether the y-intercept a_s differs from one species to another. The null hypothesis of this test is that there is no difference between species: $a_1 = a_2 = \ldots = a_S$, where $S = 16$ is the number of species. The test statistic is given in the "F value" column. As the p-value of the test is less than 5 %, it may be concluded that the y-intercept of the model is significantly different from one species to another. The second line in the table tests for an effect of the D^2H, variable, i.e. whether the mean slope associated with this variable is significantly different from zero. The third line in the table tests whether the slope-species interaction is significant, i.e. whether slope b_s differs from one species to another. The null hypothesis is that there is no difference between species: $b_1 = b_2 = \ldots = b_S$. The p-value here is 0.1785, therefore greater than 5 %: there is therefore no significant slope difference between the species.

This leads us to fit the following model:

$$\ln(B_s) = a_s + b\ln(D_s^2 H_s) + \varepsilon \tag{6.28}$$

which considers that slope b is the same for all species. The relevant command is:

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

and yields:

```
                    Df  Sum Sq Mean Sq F value    Pr(>F)
species             15 117.667   7.844   81.99 < 2.2e-16 ***
I(log(dbh^2*heig))   1 112.689 112.689 1177.81 < 2.2e-16 ***
Residuals           24   2.296   0.096
```

Model coefficients may be obtained by the command:

```
summary(m)
```

which yields:

```
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                     -9.00359    0.45144 -19.944   <2e-16 ***
speciesAubrevillea kerstingii   -0.54634    0.43784  -1.248   0.2241
speciesCecropia peltata         -0.77688    0.36261  -2.142   0.0425 *
speciesCeiba pentandra          -0.70841    0.38048  -1.862   0.0749 .
speciesCola nitida              -0.46428    0.44476  -1.044   0.3069
speciesDaniellia thurifera       0.04685    0.46413   0.101   0.9204
speciesDialium aubrevilliei     -0.15626    0.43757  -0.357   0.7241
speciesDrypetes chevalieri       0.04953    0.45395   0.109   0.9140
speciesGarcinia epunctata        1.09645    0.47318   2.317   0.0293 *
speciesGuarea cedrata           -0.45255    0.38460  -1.177   0.2509
speciesHeritiera utilis         -0.26865    0.32663  -0.822   0.4189
speciesNauclea diderrichii      -0.55464    0.35759  -1.551   0.1340
speciesNesogordonia papaverifera -0.47817   0.44335  -1.079   0.2915
speciesPiptadeniastrum africanum -0.17956   0.35718  -0.503   0.6197
speciesStrombosia glaucescens    0.06333    0.39597   0.160   0.8743
speciesTieghemella heckelii     -0.09104    0.33908  -0.268   0.7906
I(log(dbh^2*heig))               0.89985    0.02622  34.319   <2e-16 ***
```

The last line in this table gives the value of the slope: $b = 0.89985$. The other lines give y-intercept values for the 16 species. By convention, R proceeds in the following manner when specifying these values: the first line in the table gives the y-intercept for the first species in alphabetical order. As the first species in alphabetical order is *Afzelia bella*, the y-intercept for *Afzelia bella* is therefore $a_1 = -9.00359$. The subsequent lines give the *difference* $a_s - a_1$ between the y-intercepts of the species indicated and the y-intercept of *Afzelia bella*. Thus, the y-intercept of *Aubrevillea kerstingii* is: $a_2 = a_1 - 0.54634 = -9.00359 - 0.54634 =$

-9.54993. Finally, the expression for the specific table corresponds to:

$$\ln(B) = 0.89985\ln(D^2H) - \begin{cases} 9.00359 & \text{for } \textit{Afzelia bella} \\ 9.54993 & \text{for } \textit{Aubrevillea kerstingii} \\ 9.78047 & \text{for } \textit{Cecropia peltata} \\ 9.71200 & \text{for } \textit{Ceiba pentandra} \\ 9.46786 & \text{for } \textit{Cola nitida} \\ 8.95674 & \text{for } \textit{Daniellia thurifera} \\ 9.15985 & \text{for } \textit{Dialium aubrevilliei} \\ 8.95406 & \text{for } \textit{Drypetes chevalieri} \\ 7.90713 & \text{for } \textit{Garcinia epunctata} \\ 9.45614 & \text{for } \textit{Guarea cedrata} \\ 9.27223 & \text{for } \textit{Heritiera utilis} \\ 9.55823 & \text{for } \textit{Nauclea diderrichii} \\ 9.48176 & \text{for } \textit{Nesogordonia papaverifera} \\ 9.18315 & \text{for } \textit{Piptadeniastrum africanum} \\ 8.94026 & \text{for } \textit{Strombosia glaucescens} \\ 9.09462 & \text{for } \textit{Tieghemella heckelii} \end{cases}$$

For a given diameter and height, the species with the highest biomass is *Garcinia epunctata* while that with the lowest biomass is *Cecropia peltata*. The model's residual standard deviation is $\hat{\sigma} = 0.3093$ and $R^2 = 0.9901$.

Case of a numerical covariable

Up until now we have considered that the covariables defining the stratification are qualitative factors. But in certain cases these covariables can also be interpreted as numerical variables. For instance, let us consider a biomass model for plantations (Saint-André *et al.*, 2005). The year the plantation was set up (or, which comes to the same thing, the age of the trees), may be used as a stratification covariable. This year or age may be seen indifferently as a qualitative variable (cohort of trees of the same age) or as a numerical value. More generally, any numerical variable may be seen as a qualitative variable if it is divided into classes. In the case of age, we could also consider plantations aged 0 to 5 years as a stratum, plantations aged 5 to 10 years as another stratum, plantations aged 10 to 20 years as a third stratum, etc.. The advantage of dividing covariable Z into classes and considering it as a qualitative variable is that this allows us to model the relation between Z and response variable Y without any *a priori* constraint on the form of this relation. By contrast, if we consider Z to be a numerical variable, we are obliged *a priori* to set the form of the relation between Y and Z (linear relation, polynomial relation, exponential relation, power relation, etc.). The disadvantage of dividing Z into classes and considering this covariable as being qualitative is that the division introduces an element of randomness. Also, the covariance model that uses classes of Z (qualitative covariables) will generally have more parameters that need to be estimated than the model that considers Z to be a numerical covariable.

It is customary in modeling to play on this dual interpretation of numerical variables. When covariable Z is numerical (e.g. tree age), it is advisable to proceed in two steps (as explained in § 5.1.1):

1. consider Z as a qualitative variable (if necessary after division into classes) and fit a covariance model which will provide a picture of the form of the relation between Z and the model's coefficients;

2. model this relation using an appropriate expression then return to the fitting of a linear or non-linear model, but considering Z to be a numerical variable.

If we return to the example of the age of the trees in a plantation: let us assume that age Z has been divided into S age classes. The first step consists typically of an analysis of covariance (assuming that the model has been rendered linear):

$$Y_s = a_{0s} + a_{1s}X_{1s} + a_{2s}X_{2s} + \ldots + a_{ps}X_{ps} + \varepsilon$$

where $s = 1, \ldots, S$. Let Z_s be the median age of age class s. We then draw scatter plots of a_{0s} against Z_s, a_{1s} against Z_s, $\ldots$, a_{ps} against Z_s and for each scatter plot we look for the form of the relation that fits the plot. Let us imagine that a_{0s} varies in a linear manner with Z_s, that a_{1s} varies in a exponential manner with Z_s, that a_{2s} varies in a power manner with Z_s, and that the coefficients a_{3s} to a_{ps} do not vary in relation to Z_s (which besides can be formally tested). In this particular case we will then, as the second step, fit the following non-linear model:

$$Y = \underbrace{b_0 + b_1 Z}_{a_{0s}} + \underbrace{b_2 \exp(-b_3 Z)}_{a_{1s}} X_1 + \underbrace{b_4 Z^{b_5}}_{a_{2s}} X_2 + a_3 X_3 + \ldots + a_p X_p + \varepsilon$$

where age Z is now considered as a numerical variable. Such a model involving a numerical effect covariable is called a *parameterized* model (here by age).

Ordinal covariables warrant a special remark. An ordinal variable is a qualitative variable that establishes an order. For example, the month of the year is a qualitative variable that establishes a chronological order. The type of soil along a soil fertility gradient is also an ordinal variable. Ordinal variables are generally treated as fully fledged qualitative variables, but in this case the order information they contain is lost. An alternative consists in numbering the ordered modalities of the ordinal variable using integers, then considering the ordinal variable as a numerical variable. For example, in the case of the months of the year, we can set January = 1, February = 2, etc. This approach is meaningful only if the differences between the integers correctly reflect the differences between the modalities of the ordinal variable. For instance, if we set 1 = January 2011 up to 12 = December 2011, we will set 1 = January 2012 if the response is cyclically seasonal, whereas we will set 13 = January 2012 if the response presents as a continuous trend. In the case of three soil types along a fertility gradient, we will set 1 = poorest soil, 2 = soil of intermediate fertility, and 3 = richest soil if we consider the fertility difference between two soils induces a response that is proportional to this difference, but we will set 1 = poorest soil, 4 = soil of intermediate fertility, and 9 = richest soil if we consider that the response is proportional to the square of the fertility difference.

Special case of the species

In the case of multispecific datasets, the species is a stratification covariable that warrants special attention. If the dataset includes only a few species (less than about 10), and there are sufficient observations per species (see § 2.2.1), then species may be considered to be a stratification covariable like any other. The model in this case would be divided into S specific models, or the models could be grouped together based on the allometric resemblance of the species.

If the dataset contains many species, or if only a few observations are available for certain species, it is difficult to treat the species as a stratification covariable. A solution in this case consists in using species *functional traits*. Functional traits are defined here,

a little abusively, as the numerical variables that characterize the species (Díaz & Cabido, 1997; Rösch *et al.*, 1997; Lavorel & Garnier, 2002; see Violle *et al.*, 2007 for a more rigorous definition). The most widely used trait in biomass models is wood density. If we decide to use functional traits to represent species, these traits intervene as effect variables in the model in the same way as the effect variables that characterize the tree, for example its dbh or height. The single-entry (dbh) monospecific biomass model of the power type, which in its linear form may be written:

$$\ln(B) = a_0 + a_1 \ln(D) + \varepsilon$$

will thus, in the multispecific case, become a double-entry biomass model:

$$\ln(B) = a_0 + a_1 \ln(D) + a_2 \ln(\rho) + \varepsilon$$

if we decide to use wood density ρ to represent the specific effect.

31

Specific wood density-dependent biomass model

In red line 30, species-related information was taken into account in the model $\ln(B) = a + b \ln(D^2 H)$ through a qualitative covariable. We can now capture this information through specific wood density ρ. The fitted model is therefore:

$$\ln(B) = a_0 + a_1 \ln(D^2 H) + a_2 \ln(\rho) + \varepsilon \tag{6.29}$$

where

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

As wood density in the dataset was measured for each individual, we must start by calculating mean wood density for each species:

```
dm <- tapply(dat$dens,dat$species,mean)
dat <- cbind(dat,dmoy=dm[as.character(dat$species)])
```

The `dat` dataset now contains an additional variable `dmoy` that gives specific wood density. The model is fitted by the command:

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

which yields:

```
                    Estimate  Std.  Error  t value  Pr(>|t|)
 (Intercept)        -8.38900      0.26452  -31.714   < 2e-16  ***
 I(log(dbh^2*heig))  0.85715      0.02031   42.205   < 2e-16  ***
 I(log(dmoy))        0.72864      0.17720    4.112  0.000202  ***
```

with a residual standard deviation of 0.3442 and $R^2 = 0.9806$. The model is therefore written: $\ln(B) = -8.38900 + 0.85715 \ln(D^2 H) + 0.72864 \ln(\rho)$. Is it better to take account of the species through wood density, as we have just done, or by constructing specific models as in red line 30? To answer this question we need to compare model (6.28) to the model (6.29) using the AIC:

```
AIC(m)
```

which yields AIC = 34.17859 for specific model (6.28) and AIC = 33.78733 for the model (6.29) that uses wood density. The latter therefore is slightly better, but the difference in AIC is nevertheless minor.

To better take account of wood density variations within a tree, it is possible to analyze inter- and intra-specific variations rather than use a mean density based on the hypothesis that wood density is the same from the pith to the bark and from the top to the bottom of trees (see chapter 1). Wood density can be modeled by taking account of factors such as species, functional group, tree dimensions, and radial or vertical position in the tree. An initial comparison may be made using Friedman's analysis of variance test, followed by Tuckey's highly significant difference (HSD) test. These tests will indicate the variables that most influence wood density. This wood density may then be modeled based on these variables (Henry *et al.*, 2010).

32

Individual wood density-dependent biomass model

In red line 31, wood density ρ was defined for the species by calculating the mean of individual densities for trees in the same species. Let us now fit a biomass model based on individual wood density measurements in order to take account of inter-individual variations in wood density in a given species. The fitted model is:

$$\ln(B) = a_0 + a_1 \ln(D) + a_2 \ln(\rho) + \varepsilon$$

where

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

where ρ unlike in red line 31, is here the *individual* measurement of wood density. The model is fitted by the command:

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

which yields:

```
               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  ***
```

with a residual standard deviation of 0.3052 and $R^2 = 0.9848$. The model is written: $\ln(B) = -7.76644 + 2.35272 \ln(D) + 1.00717 \ln(\rho)$. According to this model, biomass is dependent upon individual wood density by the term $\rho^{1.00717}$, i.e. practically ρ. By way of a comparison, model (6.29) was dependent upon specific wood density by the term $\rho^{0.72864}$. From a biological standpoint, the exponent 1.00717 is more satisfactory than the exponent 0.72864 since it means that the biomass is the product of a volume (that depends solely on tree dimensions) and a density. The difference between the two exponents may be attributed to inter-individual variations in wood density within the species. But the model based on individual wood density has little practical usefulness as it would require a wood density measurement in every tree whose biomass we are seeking to predict.

6.4.2 Tree compartments

Tree biomass is determined separately for each compartment in the tree (stump, trunk, large branches, small branches, leaves, etc.). Total above-ground biomass is the sum of these compartments. The approach already presented for fitting a model could be followed for each compartment separately. In this case we would construct a model for leaf biomass, a model for the biomass of the large branches, etc.. This approach integrates dataset stratification. Thus, we could initially fit a model for each compartment and each stratum; as a second step, and depending on the differences found between strata, we could aggregate the strata and/or parameterize the model in order to construct a model by compartment for all the strata. But the approach would not end there. We could also continue the data integration in order to move toward a smaller number of more integrating models.

Compartment additivity

As total above-ground biomass is the sum of the biomasses of the different compartments, it could be imagined that the best model for predicting above-ground biomass is the sum of the models predicting the biomass of each compartment. In fact, because of the correlations between the biomasses of the different compartment, this is not the case (Cunia & Briggs, 1984, 1985a; Parresol, 1999). Also, certain model families are not stable by addition. This in particular is the case for power models: the sum of two power functions is not a power function. If we have fitted a power model for each compartment:

$$\begin{aligned} B^{\text{stump}} &= a_1 D^{b_1} \\ B^{\text{trunk}} &= a_2 D^{b_2} \\ B^{\text{large branches}} &= a_3 D^{b_3} \\ B^{\text{small branches}} &= a_4 D^{b_4} \\ B^{\text{leaves}} &= a_5 D^{b_5} \end{aligned}$$

The sum $B^{\text{above-ground}} = B^{\text{stump}} + B^{\text{trunk}} + B^{\text{large branches}} + B^{\text{small branches}} + B^{\text{leaves}} = \sum_{m=1}^{5} a_m D^{b_m}$ is not a power function of dbh. By contrast, polynomial models are stable by addition.

Fitting a multivariate model

In order to take account of the correlations between the biomasses of the different compartments, we can fit the models relative to the different compartments in a simultaneous manner rather than separately. This last step in model integration requires us to redefine the response variable. As we are looking to predict simultaneously the biomasses of different compartments, we are no longer dealing with a response variable but a *response vector* $\mathbf{Y}$. The length of this vector is equal to the number M of compartments. For example, if the response variable is the biomass,

$$\mathbf{Y} = \begin{bmatrix} B^{\text{above-ground}} \\ B^{\text{stump}} \\ B^{\text{trunk}} \\ B^{\text{large branches}} \\ B^{\text{small branches}} \\ B^{\text{leaves}} \end{bmatrix}$$

If the response variable is the log of the biomass,

$$\mathbf{Y} = \begin{bmatrix} \ln(B^{\text{above-ground}}) \\ \ln(B^{\text{stump}}) \\ \ln(B^{\text{trunk}}) \\ \ln(B^{\text{large branches}}) \\ \ln(B^{\text{small branches}}) \\ \ln(B^{\text{leaves}}) \end{bmatrix}$$

Let Y_m be the response variable of the mth compartment (where $m = 1, \ldots, M$). Without loss of generality, we can consider that all the compartments have the same set X_1, X_2, $\ldots$, X_p of effect variables (if a variable is not included in the prediction of a compartment, the corresponding coefficient can simply be set at zero). A model that predicts a response vector rather than a response variable is a multivariate model. An observation used to fit a multivariate model consists of a vector $(Y_1, \ldots, Y_M, X_1, \ldots, X_p)$ of length $M + p$. The residual of a multivariate model is a vector $\boldsymbol{\varepsilon}$ of length M, equal to the difference between the observed response vector and the predicted response vector.

The expression of an M-variate model differs from the M univariate models corresponding to the different compartments only by the structure of the residual error; the structure of the mean model does not change. Let us consider the general case of a non-linear model. If the M univariate models are:

$$Y_m = f_m(X_1, \ldots, X_p; \theta_m) + \varepsilon_m \tag{6.30}$$

for $m = 1, \ldots, M$, then the multivariate model is written:

$$\mathbf{Y} = \mathbf{F}(X_1, \ldots, X_p; \boldsymbol{\theta}) + \boldsymbol{\varepsilon}$$

where $\mathbf{Y} = {}^{\mathrm{t}}[Y_1, \ldots, Y_M]$, $\boldsymbol{\theta} = {}^{\mathrm{t}}[\theta_1, \ldots, \theta_M]$, and

$$\mathbf{F}(X_1, \ldots, X_p; \boldsymbol{\theta}) = \begin{bmatrix} f_1(X_1, \ldots, X_p; \theta_1) \\ \vdots \\ f_m(X_1, \ldots, X_p; \theta_m) \\ \vdots \\ f_M(X_1, \ldots, X_p; \theta_M) \end{bmatrix} \tag{6.31}$$

The residual vector $\boldsymbol{\varepsilon}$ now follows a centered multinormal distribution, with a variance-covariance matrix of:

$$\mathrm{Var}(\boldsymbol{\varepsilon}) \equiv \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \zeta_{12} & \cdots & \zeta_{1M} \\ \zeta_{21} & \sigma_2^2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \zeta_{M-1,M} \\ \zeta_{M1} & \cdots & \zeta_{M,M-1} & \sigma_M^2 \end{bmatrix}$$

Matrix $\boldsymbol{\Sigma}$ is symmetrical with M lines and M columns, such that $\sigma_m^2 = \mathrm{Var}(\varepsilon_m)$ is the residual variance of the biomass of the mth compartment and $\zeta_{ml} = \zeta_{lm}$ is the residual covariance between the biomass of the mth compartment and that of the lth compartment. Like in the univariate case, two residuals corresponding to two different observations are assumed to be independent: $\boldsymbol{\varepsilon}_i$ is independent of $\boldsymbol{\varepsilon}_j$ pour $i \neq j$. The difference arises from the fact that the different compartments are no longer assumed to be independent one from another.

A multivariate model such as (6.31) is fitted based on the same principles as univariate models (6.30). If the variance-covariance matrix $\mathbf{\Sigma}$ is diagonal (i.e. $\zeta_{ml} = 0$, $\forall m, l$), then the fitting of the multivariate model (6.31) is equivalent to the separate fitting of the M univariate models (6.30). In the case of a linear model, estimated values for coefficients θ_1, θ_2, ..., θ_M resulting from fitting the M-variate linear model are the same as the values obtained by the separate fitting of the M univariate linear models (on condition that the same effect variables X_1, ..., X_p are used in all cases) (Muller & Stewart, 2006, chapter 3). But the significance tests associated with the coefficients do not give the same results in the two cases. If the different compartments are sufficiently correlated one with the other, the simultaneous fitting of all the compartments by the multivariate model (6.31) will yield a more precise estimation of model coefficients, and therefore more precise biomass predictions.

Harmonizing a model

In certain cases, particularly energy wood, we are looking to predict the dry biomass of the trunk at different cross-cut diameters. For instance, we want to predict simultaneously the total biomass B of the trunk, the biomass B_7 of the trunk to the small-end diameter of 7 cm, and the biomass B_{10} of the trunk to the small-end diameter of 10 cm. We could consider the entire trunk, the trunk to a cross-cut of 7 cm, and the trunk to a cross-cut of 10 cm as three different compartments and apply the same fitting principles as presented in the previous section. In fact, the problem is more complex because, unlike the trunk and leaf compartments which are separate, the compartments defined by different cross-cut diameters are nested one within the others: $B = B_7+$ biomass of the section to a small-end diameter of 7 cm, and $B_7 = B_{10}+$ biomass of the section from the 10 cm to the 7 cm small-end cross-cuts. Thus, the multivariate model that predicts vector (B, B_7, B_{10}) must ensure that $B > B_7 > B_{10}$ across the entire valid range of the model. The process consisting of constraining the multivariate model in such a manner that it predicts the biomasses of the different compartments while checking the logic of their nesting is called model *harmonization* (Parresol, 1999). Jacobs & Cunia (1980) and Cunia & Briggs (1985b) have put forward solutions to this problem in the form of equations relating the coefficients in the models used for the different compartments. In this case we must fit an M-variate model (if there are M cross-cut diameters) while ensuring that the coefficients θ_1, ..., θ_M corresponding to the M cross-cut diameters satisfy a number of equations linking them together. When the coefficients in the multivariate model are estimated by maximum likelihood, their numerical estimation is in fact a problem of constrained optimization.

If predicting the volume or biomass of a stem, an alternative to the volume or biomass model is to use stem profile integration (Parresol & Thomas, 1989; Parresol, 1999). Let $P(h)$ be a stem profile, i.e. a plot of trunk transversal section area against height h from the ground (h also represents the length when a stem is followed from its large to its small end) (Maguire & Batista, 1996; Dean & Roxburgh, 2006; Metcalf *et al.*, 2009). If the section through the stem is roughly circular, the diameter of the tree at height h may be calculated as: $D(h) = \sqrt{4P(h)/\pi}$. The biomass of the trunk to a cross-cut diameter D is calculated by integrating the stem profile from the ground ($h = 0$) to height $P^{-1}(\frac{\pi}{4}D^2)$ corresponds to this diameter:

$$B_D = \int_0^{P^{-1}(\frac{\pi}{4}D^2)} \rho(h)\ P(h)\,\mathrm{d}h$$

where $\rho(h)$ is wood density at height h. Stem volume to cross-cut diameter D is calculated in the same manner, except that ρ is replaced by 1. The stem profile approach has the

advantage that model harmonization here is automatic. But it is an approach that is conceptually different from volume and biomass models, with specific fitting problems (Fang & Bailey, 1999; Parresol, 1999), and is outside the scope of this guide. It should be noted that when dealing with very large trees, for which it is almost impossible to measure biomass directly, the stem profile approach is a relevant alternative (Van Pelt, 2001; Dean *et al.*, 2003; Dean, 2003; Dean & Roxburgh, 2006; Sillett *et al.*, 2010).

7 Uses and prediction

Once a volume or biomass model has been fitted, several uses may be made of its predictions. Most commonly, it is used to predict the volume or biomass of trees whose volume or biomass has not been measured. This is *prediction* in the strictest sense (§ 7.2–7.4). Sometimes, tree volume or biomass will also have been measured in addition to table entry variables. In this case, when a dataset is available *independently* of that used to fit the model, and which contains both the model's response variable and its effect variables, this can be used to *validate* the model (§ 7.1). When validation criteria are applied to the same dataset that served to calibrate the model, we speak of model *verification* We will not dwell here on model verification as this is already implicit in the analysis of the fitted model's residuals. Finally, if we are in possession of models that were developed prior to a new fitted model, we may also wish to compare or combine the models (§ 7.5).

Model valid range

Before using a model, we must check that the characteristics of the tree whose volume or biomass we wish to predict fall within the *valid range* of the model (Rykiel, 1996). If a volume or biomass model has been fitted for trees of dbh between $D_{\min}$ and $D_{\max}$, in principle it cannot be used to predict the volume or biomass of a tree whose dbh is less than $D_{\min}$ or more than $D_{\max}$. The same applies for all model entries. But not all models are subject to the same errors when extrapolating outside their valid range. Power models can generally be extrapolated outside their valid range and still yield reliable results as they are based on a fractal allometric model that is invariant on all scales (Zianis & Mencuccini, 2004). By contrast, polynomial models often behave abnormally outside their valid range (e.g. predicting negative values), particularly when using high-order polynomials.

7.1 Validating a model

Validating a model consists in comparing its predictions with observations independent of those used for its fitting (Rykiel, 1996). Let $(Y'_i, X'_{i1}, \ldots, X'_{ip})$ with $i = 1, \ldots, n'$ be a dataset of observations independent of that used to fit a model f, where $X'_{i1}, \ldots, X'_{ip}$ are the effect variables and Y'_i is the response variable, i.e. the volume, biomass or a transform

of one of these quantities. Let

$$\hat{Y}'_i = f(X'_{i1}, \ldots, X'_{ip}; \hat{\theta})$$

be the predicted value of the response variable for the ith observation, where $\hat{\theta}$ are the estimated values of the model's parameters. The validation consists in comparing the predicted values $\hat{Y}'_i$ with the observed values Y'_i.

7.1.1 Validation criteria

Several criteria, which are the counterparts of the criteria used to evaluate the quality of model fitting, may be used to compare predictions with observations (Schlaegel, 1982; Parresol, 1999; Tedeschi, 2006), in particular:

- bias: $\sum_{i=1}^{n'} |Y'_i - \hat{Y}'_i|$
- sum of squares of the residuals: $\mathrm{SSE} = \sum_{i=1}^{n'} (Y'_i - \hat{Y}'_i)^2$
- residual variance: $s^2 = \mathrm{SSE}/(n' - p)$
- fitted residual error: $\mathrm{SSE}/(n' - 2p)$
- R^2 of the regression: $R^2 = 1 - s^2/\mathrm{Var}(Y')$
- Akaike information criterion: $\mathrm{AIC} = n' \ln(s^2) + n' \ln(1 - p/n') + 2p$

where $\mathrm{Var}(Y')$ is the empirical variance of Y' and p is the number of freely estimated parameters in the model. The first two criteria correspond to two distinct norms of the difference between the vector $(Y'_1, \ldots, Y'_{n'})$ of the observations and the vector $(\hat{Y}'_1, \ldots, \hat{Y}'_{n'})$ of the predictions: norm L^1 for the bias and norm L^2 for the sum of squares. Any other norm is also valid. The last three criteria involve the number of parameters used in the model, and are therefore more appropriate for comparing different models.

7.1.2 Cross validation

If no independent dataset is available, it is tempting to split the calibration dataset into two subsets: one to fit the model, the other to validate the model. Given that volume or biomass datasets are costly, and often fairly small, we do not recommend this practice when constructing volume or biomass tables. By contrast, in this case, we do recommend the use of *cross validation* (Efron & Tibshirani, 1993, chapitre 17).

A "K fold" cross validation consists in dividing the dataset into K subsets of approximately equal size and using each subset, one after the other, as validation datasets, with the model being fitted on the $K - 1$ subsets remaining. The "K fold" cross validation pseudo-algorithm is as follows:

1. Split the dataset $\mathcal{S}_n \equiv \{(Y_i, X_{i1}, \ldots, X_{ip}): i = 1, \ldots, n\}$ into K data subsets $\mathcal{S}_n^{(1)}$, $\ldots$, $\mathcal{S}_n^{(K)}$ of approximately equal size (i.e. with about n/K observations in each data subset, totaling n).
2. For k from 1 to K:
   (a) fit the model on the dataset deprived of its kth subset, i.e. on $\mathcal{S}_n \backslash \mathcal{S}_n^{(k)} = \mathcal{S}_n^{(1)} \cup \ldots \cup \mathcal{S}_n^{(k-1)} \cup \mathcal{S}_n^{(k+1)} \cup \ldots \cup \mathcal{S}_n^{(K)}$;

(b) calculate a validation criterion (see § 7.1.1) for this fitted model by using the remaining dataset $\mathcal{S}_n^{(k)}$ as validation dataset; let C_k be the value of this criterion calculated for $\mathcal{S}_n^{(k)}$.

3. Calculate the mean $(\sum_{k=1}^{K} C_k)/K$ of the K validation criteria calculated.

The fact that there is no overlap between the data used for model fitting and the data used to calculate the validation criterion means that the approach is valid. Cross validation needs more calculations than a simple validation, but has the advantage of making the most of all the observations available for model fitting.

A special case of a "K fold" cross validation is when K is equal to the number n of observations available in the dataset. This method is also called "leave-one-out" cross validation and is conceptually similar to the Jack-knife technique (Efron & Tibshirani, 1993). The principle consists in fitting the model on $n-1$ observations and calculating the residual error for the observation that was left out. It is used when analyzing the residuals to quantify the influence of the observations (and in particular is the basis of the calculation of Cook's distance, see Saporta, 1990).

7.2 Predicting the volume or biomass of a tree

Making a prediction based on model f consists in using the known values of the effect variables $X_1, \ldots, X_p$, to calculate the predicted value $\hat{Y}$. A prediction does not end with the calculation of

$$\hat{Y} = f(X_1, \ldots, X_p; \hat{\theta})$$

The estimator $\hat{\theta}$ of model parameters is a random vector whose distribution stems from the distribution of the observations used to fit the model. Any model prediction $\hat{Y}$ is therefore itself a random variable whose distribution stems from the distribution of the observations used to fit the model. If we are to express this intrinsic variability of the prediction, we must associate it with an uncertainty indicator such as the standard deviation of the prediction or its 95 % confidence interval.

Several confidence intervals are available depending on whether we are predicting the volume or biomass of a tree selected at random in the stand, or the volume or biomass of the mean tree in the stand. We will now describe in detail the analytical expressions of these confidence intervals for a linear model (§ 7.2.1), then for a non-linear model (§ 7.2.2). We will then present expressions for these confidence intervals that are similar but simpler to compute (§ 7.2.3), before focusing on the case of transformed variables (§ 7.2.4).

7.2.1 Prediction: linear model

Prediction by simple linear regression

Let $\hat{a}$ be the estimated y-intercept of a linear regression, and $\hat{b}$ be its estimated slope. The prediction of $\hat{Y}$ the response variable may be written in two different ways:

$$\hat{Y} = \hat{a} + \hat{b}X \tag{7.1}$$

$$\hat{Y} = \hat{a} + \hat{b}X + \varepsilon \tag{7.2}$$

In both cases the expectation of $\hat{Y}$ is the same as $\mathrm{E}(\varepsilon) = 0$. By contrast, the variance of $\hat{Y}$ is not the same: it is higher in the second than in the first. These two equations may be interpreted as follows. Let us assume that effect variable X is dbh and that the response

variable Y is the biomass. The number of trees in the whole forest with a given dbh X (to within measurements errors) is immeasurably huge. If we were able to measure the biomass of all these trees with the same dbh, we would obtain variable values oscillating around a certain mean value. If we seek to predict this mean biomass (i.e. the mean of all the trees with dbh X), then equation (7.1) of the prediction is valid. By contrast, if we seek to predict the biomass of a tree selected at random from all the trees with dbh X, then equation (7.2) of the prediction is valid. The variability of the prediction is greater for (7.2) than for (7.1) because here the variability of the mean biomass prediction is supplemented by the differences in biomass between the trees.

This means that two methods can be used to calculate a confidence interval for a prediction. There is a confidence interval for the prediction of the mean of Y, and a confidence interval for the prediction of an individual selected at random from the population in which the mean of Y was calculated. The second confidence interval is wider than the first.

In the case of a simple linear regression, it can be shown (Saporta, 1990, p.373–374) that the confidence interval at a significance level of α for predicting the mean (7.1) is:

$$\hat{a} + \hat{b}X \pm t_{n-2}\,\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(X-\bar{X})^2}{nS_X^2}} \tag{7.3}$$

while the confidence interval at a significance level of α for predicting a tree selected at random (7.2) is:

$$\hat{a} + \hat{b}X \pm t_{n-2}\,\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(X-\bar{X})^2}{nS_X^2}} \tag{7.4}$$

where t_{n-2} is the quantile $1-\alpha/2$ of a Student's distribution with $n-2$ degrees of freedom, $\bar{X} = (\sum_{i=1}^{n} X_i)/n$ is the mean of the observed values of X in the dataset that served to fit the model, and $S_X^2 = [\sum_{i=1}^{n}(X_i - \bar{X})^2]/n$ is the empirical variance of the observed values of X in the dataset that served to fit the model.

These expressions call for several remarks. The first is that the difference between the bounds of confidence interval (7.4) for a tree selected at random and the bounds of confidence interval (7.3) for the mean tree is approximately $t_{n-2}\hat{\sigma}$. This difference reflects the difference between equations (7.2) and (7.1), which is due to the residual term ε whose estimated standard deviation is $\hat{\sigma}$.

The second is that the width of the confidence interval is not constant, but varies with X. The confidence interval is narrowest when $X = \bar{X}$ but widens as X moves away from $\bar{X}$.

The third remark is that in order to calculate the confidence interval of a prediction based on a linear regression, we must be in possession of — if not the original data that served to fit the model — at least the mean $\bar{X}$ of the effect variable and its empirical standard deviation S_X. If the original data that served to fit the model are not available, and the values of $\bar{X}$ and S_X have not been documented, we cannot accurately calculate the confidence interval.

33

Confidence interval of $\ln(B)$ predicted by $\ln(D)$

Let us return to the simple linear regression between $\ln(B)$ and $\ln(D)$ that was fitted in red line 7. Let `m` be the object containing the fitted model (see red line 7). The confidence intervals may be calculated using the `predict` command. For example, for a tree with a dbh of 20 cm, the 95 % confidence interval for the mean tree is obtained by the command:

```
predict(m,newdata=data.frame(dbh=20),interval="confidence",level=0.95)
```

which yields:

```
          fit        lwr        upr
1  -1.354183  -1.533487  -1.174879
```

Thus, the fitted model is $\ln(B) = -1.354183$ with a 95 % confidence interval of -1.533487 to -1.174879. For a tree with a dbh of 20 cm selected at random, the confidence interval may be obtained by the command:

```
predict(m,newdata=data.frame(dbh=20),interval="prediction",level=0.95)
```

which yields:

```
          fit        lwr         upr
1  -1.354183  -2.305672  -0.4026948
```

Figure 7.1 shows the confidence intervals for the entire range of the data.

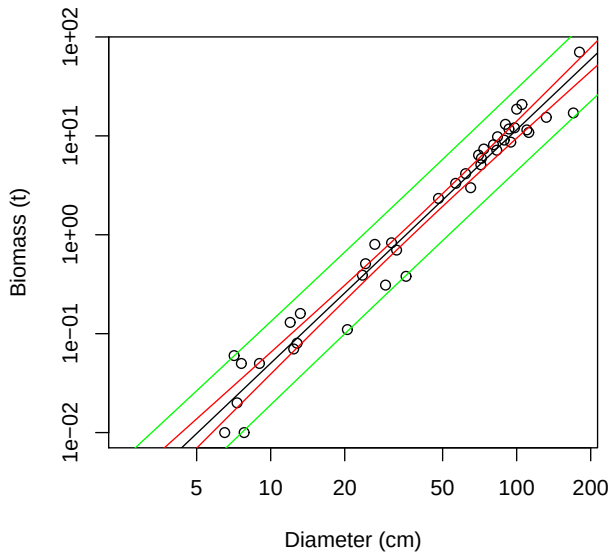


Figure 7.1 – *Biomass against dbh (on a log scale) for 42 trees measured in Ghana by Henry et al. (2010) (points), prediction (solid line) of the simple linear regression of* $\ln(B)$ *against* $\ln(D)$*, and confidence intervals of this prediction for a tree selected at random (green line) and for the mean tree (red line).*

Prediction by multiple regression

The principles laid down in the case of the linear regression also apply to the multiple regression. The confidence interval can be expressed in two different ways: one for the prediction of the mean tree, the other for the prediction of a tree selected at random.

In the case of a multiple regression with estimated coefficients $\hat{\mathbf{a}} = {}^{\mathrm{t}}[\hat{a}_0, \hat{a}_1, \hat{a}_2, \ldots, \hat{a}_p]$, the predicted value $\hat{Y}$ of the response variable for a tree whose effect variables are $\mathbf{x} = {}^{\mathrm{t}}[1, X_1, X_2, \ldots, X_p]$, is:

$$\hat{Y} = {}^{\mathrm{t}}\mathbf{x}\,\hat{\mathbf{a}}$$

and the confidence interval (with a significance level of α) of this prediction is (Saporta, 1990, p.387):

- for the prediction of the mean tree:

$$ {}^{t}\mathbf{x}\,\hat{\mathbf{a}} \pm t_{n-p-1}\,\hat{\sigma}\sqrt{{}^{t}\mathbf{x}({}^{t}\mathbf{X}\mathbf{X})^{-1}\mathbf{x}} \tag{7.5} $$

- for the prediction of a tree selected at random:

$$ {}^{t}\mathbf{x}\,\hat{\mathbf{a}} \pm t_{n-p-1}\,\hat{\sigma}\sqrt{1+{}^{t}\mathbf{x}({}^{t}\mathbf{X}\mathbf{X})^{-1}\mathbf{x}} \tag{7.6} $$

where $\mathbf{X}$ is is the design matrix constructed using the data that served to fit the multiple regression. In order to calculate the confidence interval of the predictions, we need to be in possession of — if not the original data that served to fit the model — at least the matrix $({}^{t}\mathbf{X}\mathbf{X})^{-1}$. It should be noted that the variance of the prediction in the case (7.6) of a tree selected at random is made up of two terms: a term $\hat{\sigma}^2$ representing the residual error and a term $\hat{\sigma}^2\ {}^{t}\mathbf{x}({}^{t}\mathbf{X}\mathbf{X})^{-1}\mathbf{x}$ representing the variability induced by the estimation of the model's coefficients. When estimating the mean tree, the first term disappears, only the second remains.

34

Confidence interval of $\ln(B)$ predicted by $\ln(D)$ and $\ln(H)$

Let us return to the multiple linear regression between $\ln(B)$, $\ln(D)$ and $\ln(H)$ that was fitted in red line 10. Let `m` be the object containing the fitted model (see red line 10). The confidence intervals may be calculated using the `predict` command. For example, for a tree with a dbh of 20 cm and a height of 20 m, the 95 % confidence interval for the mean tree is obtained by the command:

```
predict(m,newdata=data.frame(dbh=20,haut=20),interval="confidence",level=0.95)
```

which yields:

```
          fit        lwr        upr
1  -1.195004  -1.380798  -1.009211
```

Thus, the model predicts $\ln(B) = -1.195004$ with a 95 % confidence interval of -1.380798 to -1.009211. For a tree with a dbh of 20 cm and a height of 20 m selected at random, the confidence interval is obtained by the command:

```
predict(m,newdata=data.frame(dbh=20,haut=20),interval="prediction",level=0.95)
```

which yields:

```
          fit        lwr         upr
1  -1.195004  -2.046408  -0.3436006
```

7.2.2 Prediction: case of a non-linear model

In the general case of a non-linear model as defined by

$$Y = f(X_1, \ldots, X_p; \theta) + \varepsilon$$

where

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

unlike the case of the linear model, there is no exact explicit expression for the confidence intervals of the predictions. But the δ-method can be used to obtain an approximate (and asymptotically exact) expression of these confidence intervals (Serfling, 1980). As before, there are two confidence intervals:

- a confidence interval for the prediction of the mean tree:

$$f(X_1, \ldots, X_p; \hat{\theta}) \pm t_{n-q} \sqrt{{}^{\mathrm{t}}[\mathrm{d}_\theta f(\hat{\theta})]\ \hat{\mathbf{\Sigma}}_{\hat{\theta}}\ [\mathrm{d}_\theta f(\hat{\theta})]} \tag{7.7}$$

- a confidence interval for the prediction of a tree selected at random:

$$f(X_1, \ldots, X_p; \hat{\theta}) \pm t_{n-q} \sqrt{\hat{k}^2 X_1^{2\hat{c}} + {}^{\mathrm{t}}[\mathrm{d}_\theta f(\hat{\theta})]\ \hat{\mathbf{\Sigma}}_{\hat{\theta}}\ [\mathrm{d}_\theta f(\hat{\theta})]} \tag{7.8}$$

where q is the number of coefficients in the model (i.e. the length of vector θ), $\mathrm{d}_\theta f(\hat{\theta})$ is the value in $\theta = \hat{\theta}$ of the differential of f respect to model coefficients, and $\hat{\mathbf{\Sigma}}_{\hat{\theta}}$ is an estimation in $\theta = \hat{\theta}$ of the variance-covariance matrix $\mathbf{\Sigma}_\theta$ of the estimator of θ. The differential of f with respect to model coefficients is the vector of length q:

$$\mathrm{d}_\theta f(\theta) = {}^{\mathrm{t}}\left[\left(\frac{\partial f(X_1, \ldots, X_p; \theta)}{\partial \theta_1}\right), \ \ldots, \ \left(\frac{\partial f(X_1, \ldots, X_p; \theta)}{\partial \theta_q}\right)\right]$$

where θ_i is the ith element of vector θ, If using the estimator of the maximum likelihood of θ, we can show that asymptotically, when $n \to \infty$ (Saporta, 1990, p.301):

$$\mathbf{\Sigma}_\theta \underset{n\to\infty}{\sim} \mathbf{I}_n(\theta)^{-1} = \frac{1}{n}\,\mathbf{I}_1(\theta)^{-1}$$

where $\mathbf{I}_n(\theta)$ is the Fisher information matrix provided by a sample of size n on the vector of parameters θ. This Fisher information matrix has q lines and q columns and is calculated from the second derivative of the log-likelihood of the sample:

$$\mathbf{I}_n(\theta) = -\mathrm{E}\left[\frac{\partial^2 \mathcal{L}(\theta)}{\partial \theta^2}\right]$$

An approximate estimation of the variance-covariance matrix of the parameters is therefore:

$$\hat{\mathbf{\Sigma}}_{\hat{\theta}} = -\left[\left(\frac{\partial^2 \mathcal{L}(\theta)}{\partial \theta^2}\right)\Big|_{\theta=\hat{\theta}}\right]^{-1}$$

In practice, the algorithm that numerically optimizes the log-likelihood of the sample at the same time gives a numerical estimation of the second derivative ($\partial^2\mathcal{L}/\partial\theta^2$). This provides an immediate numerical estimation of $\hat{\mathbf{\Sigma}}_{\hat{\theta}}$.

As before, the variance of the predictions in the case (7.8) of a tree selected at random is made up of two terms: a term $(\hat{k}X_1^{\hat{c}})^2$ representing the residual error and a term ${}^{\mathrm{t}}[\mathrm{d}_\theta f(\hat{\theta})]\ \hat{\mathbf{\Sigma}}_{\hat{\theta}}\ [\mathrm{d}_\theta f(\hat{\theta})]$ representing the variability induced by the estimation of the model's coefficients. When estimating the mean tree, the first term disappears, only the second remains.

7.2.3 Approximated confidence intervals

The exact calculation of prediction confidence intervals requires information (matrix $\mathbf{X}$ for a linear model, variance-covariance matrix $\hat{\mathbf{\Sigma}}_{\hat{\theta}}$ for a non-linear model) that is only rarely given in papers on volume or biomass tables. Generally, these papers provide only the number n of observations used to fit the model and the residual standard deviation $\hat{\sigma}$ (linear model) or $\hat{k}$ and $\hat{c}$ (non-linear model). Sometimes, even this basic information about the fitting is missing. If $\mathbf{X}$ (linear model) or $\hat{\mathbf{\Sigma}}_{\hat{\theta}}$ (non-linear model) is not provided, it is impossible to use the formulas given above to calculate confidence intervals. In this case we must use an approximate method.

Residual error alone

Very often, only the residual standard deviation $\hat{\sigma}$ (linear model) or $\hat{k}$ and $\hat{c}$ (non-linear model) is given. In this case, an approximate confidence interval with a significance level of α may be determined:

- in the case of a linear regression:

$$(a_0 + a_1 X_1 + \ldots + a_p X_p) \pm q_{1-\alpha/2}\, \hat{\sigma} \tag{7.9}$$

- in the case of a non-linear model:

$$f(X_1, \ldots, X_p; \theta) \pm q_{1-\alpha/2}\, \hat{k} X_1^{\hat{c}} \tag{7.10}$$

where $q_{1-\alpha/2}$ is the quantile $1-\alpha/2$ of the standard normal distribution. This confidence interval is a direct retranscription of the relation $Y = a_0 + a_1 X_1 + \ldots + a_p X_p + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \hat{\sigma})$ (linear case) or $Y = f(X_1, \ldots, X_p; \theta) + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \hat{k} X_1^{\hat{c}})$ (non-linear case), where we have deliberately written the model's coefficients without a circumflex to underline that these are fixed values. These relations therefore assume implicatively that the model's coefficients are exactly known and that the only source of variability is the residual error. In other words, these approximate confidence intervals may be interpreted as follows: these confidence intervals (7.9) (linear case) and (7.10) (non-linear case) are those that would be obtained for the prediction of a tree selected at *random* if sample size was *infinite*. This is confirmed when $n \to \infty$, t_{n-p-1} tends toward $q_{1-\alpha/2}$ and the matrix $({}^{\mathrm{t}}\mathbf{X}\mathbf{X})^{-1}$ in (7.6) tends toward the null matrix (whose coefficients are all zero). Thus, confidence interval (7.9) is indeed the bound of confidence interval (7.6) when $n \to \infty$. The same may be said of (7.8) and (7.10).

Confidence interval for the mean tree

If an estimation $\hat{\mathbf{\Sigma}}$ is given of the variance-covariance matrix of the parameters, a confidence interval at a significance level of α for the prediction of the mean tree corresponds to:

- for a linear model:

$$(\hat{a}_0 + \hat{a}_1 X_1 + \ldots + \hat{a}_p X_p) \pm q_{1-\alpha/2} \sqrt{{}^{\mathrm{t}}\mathbf{x}\hat{\mathbf{\Sigma}}\mathbf{x}} \tag{7.11}$$

  where $\mathbf{x}$ is the vector ${}^{\mathrm{t}}[X_1, \ldots, X_p]$,

- for a non-linear model:

$$f(X_1, \ldots, X_p; \hat{\theta}) \pm q_{1-\alpha/2} \sqrt{{}^{\mathrm{t}}[\mathrm{d}_\theta f(\hat{\theta})]\ \hat{\mathbf{\Sigma}}\ [\mathrm{d}_\theta f(\hat{\theta})]} \tag{7.12}$$

These confidence intervals consider that all the prediction's variability stems from the estimation of the model's coefficients. Also, these confidence intervals are a direct retranscription of the fact that the model's coefficients follow a multinormal distribution whose mean is their true value and that has a variance-covariance matrix $\hat{\mathbf{\Sigma}}$ for in the linear case, if $\hat{\mathbf{a}} = {}^{\mathrm{t}}[\hat{a}_1, \ldots, \hat{a}_p]$ follows a multinormal distribution of mean ${}^{\mathrm{t}}[a_1, \ldots, a_p]$ and variance-covariance matrix $\hat{\mathbf{\Sigma}}$, then the linear combination ${}^{\mathrm{t}}\mathbf{x}\,\hat{\mathbf{a}}$ follows a normal distribution of mean ${}^{\mathrm{t}}\mathbf{x}\,\mathbf{a}$ and variance ${}^{\mathrm{t}}\mathbf{x}\hat{\mathbf{\Sigma}}\mathbf{x}$ (Saporta, 1990, p.85).

In the case of a linear model, we can show that the variance-covariance matrix of the estimator of the model's coefficients is (Saporta, 1990, p.380): $\mathbf{\Sigma} = \sigma^2({}^{\mathrm{t}}\mathbf{X}\mathbf{X})^{-1}$. Thus, an estimation of this variance-covariance matrix is: $\hat{\mathbf{\Sigma}} = \hat{\sigma}^2({}^{\mathrm{t}}\mathbf{X}\mathbf{X})^{-1}$. By integrating this expression in (7.11), we create an expression similar to (7.5). Similarly, in the non-linear case, confidence interval (7.12) is an approximation of (7.7).

In the non-linear case (7.12), if we want to avoid calculating the partial derivatives of f, we could use the Monte Carlo method. This method is based on a simulation that consists in making Q samplings of coefficients θ in accordance with a multinormal distribution of mean $\hat{\theta}$ and variance-covariance matrix $\hat{\mathbf{\Sigma}}$, calculating the prediction for each of the simulated values, then calculating the empirical confidence interval of these Q predictions. This method is known in the literature as providing "population prediction intervals" (Bolker, 2008; Paine *et al.*, 2012). The pseudo-algorithm is as follows:

1. For k of 1 to Q:
   (a) draw a vector $\hat{\theta}^{(k)}$ following a multinormal distribution of mean $\hat{\theta}$ and variance-covariance matrix $\hat{\mathbf{\Sigma}}$;
   (b) calculate the prediction $\hat{Y}^{(k)} = f(X_1, \ldots, X_p; \hat{\theta}^{(k)})$.
2. The confidence interval of the prediction is the empirical confidence interval of the Q values $\hat{Y}^{(1)}, \ldots, \hat{Y}^{(Q)}$.

Very often we do not know the variance-covariance matrix $\hat{\mathbf{\Sigma}}$, but we do have an estimation of the standard deviations of the coefficients. Let $\mathrm{Var}(\hat{a}_i) = \Sigma_i$ (linear case) or $\mathrm{Var}(\hat{\theta}_i) = \Sigma_i$ (non-linear case) be the variance of the model's ith coefficient. In this case we will ignore the correlation between the model's coefficients and will approximate the variance-covariance matrix of the coefficients by a diagonal matrix:

$$\hat{\mathbf{\Sigma}} \approx \begin{bmatrix} \hat{\Sigma}_1 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \hat{\Sigma}_p \end{bmatrix}$$

Confidence interval for a tree selected at random

The error resulting from the estimation of the model's coefficients, as described in the last section, can be cumulated with the residual error described in the section before last, in order to construct a prediction confidence interval for a tree selected at random. These are the variances of the predictions that are added one to the other. The confidence interval with a significance level of α is therefore:

- for a linear model:

$$(\hat{a}_0 + \hat{a}_1 X_1 + \ldots + \hat{a}_p X_p) \pm q_{1-\alpha/2}\sqrt{\hat{\sigma}^2 + {}^{\mathrm{t}}\mathbf{x}\hat{\mathbf{\Sigma}}\mathbf{x}}$$

  which is an approximation of (7.6),

- for a non-linear model:

$$f(X_1, \ldots, X_p; \hat{\theta}) \pm q_{1-\alpha/2} \sqrt{\hat{k}^2 X_1^{2\hat{c}} + {}^{\mathrm{t}}[\mathrm{d}_\theta f(\hat{\theta})]\ \hat{\boldsymbol{\Sigma}}\ [\mathrm{d}_\theta f(\hat{\theta})]}$$

  which is an approximation of (7.8).

As before, if we want to avoid a great many calculations, we could employ the Monte Carlo method, using the following pseudo-algorithm:

1. For k of 1 to Q:
   (a) draw a vector $\hat{\theta}^{(k)}$ following a multinormal distribution of mean $\hat{\theta}$ and variance-covariance matrix $\hat{\boldsymbol{\Sigma}}$;
   (b) draw a residual $\hat{\varepsilon}^{(k)}$ following a centered normal distribution of standard deviation $\hat{\sigma}$ (linear case) or $\hat{k} X_1^{\hat{c}}$ (non-linear case);
   (c) calculate the prediction $\hat{Y}^{(k)} = f(X_1, \ldots, X_p; \hat{\theta}^{(k)}) + \hat{\varepsilon}^{(k)}$.
2. The confidence interval of the prediction is the empirical confidence interval of the Q values $\hat{Y}^{(1)}, \ldots, \hat{Y}^{(Q)}$.

Confidence interval with measurement uncertainties

Fitting volume and biomass models assumes that the effect variables $X_1, \ldots, X_p$ are known exactly. In actual fact, this hypothesis is only an approximation as these variables are measured and are therefore subject to measurement error. Warning: measurement error should not be confused with the residual error of the response variable: the first is related to the measuring instrument and in principle may be rendered as small as we wish by using increasingly precise measuring instruments. The second reflects the intrinsic biological variability between individuals. We can take account of the impact of measurement error on the prediction by including it in the prediction confidence interval. Thus, the effect variables $X_1, \ldots, X_p$ are no longer considered as fixed, but as being contained within a distribution. Typically, to predict the volume or biomass of a tree with characteristics $X_1, \ldots, X_p$, we will consider that the ith characteristic is distributed according to a normal distribution of mean X_i and standard deviation τ_i. Typically, if X_i is a dbh, then τ_i will be 3–5 mm; if X_i is a height, τ_i will be 3 % of X_i for $X_i \leq 15$ m and 1 m for $X_i > 15$ m.

It is difficult to calculate an explicit expression for the prediction confidence interval when the effect variables are considered to be random since this means we must calculate the variances of the products of these random variables, some of which are correlated one with the other. The δ-method offers an approximate analytical solution (Serfling, 1980). Or, more simply, we can again use a Monte Carlo method. In this case, the pseudo-algorithm becomes:

1. For k of 1 to Q:
   (a) for i of 1 to p, draw $\hat{X}_i^{(k)}$ following a normal distribution of mean X_i and standard deviation τ_i;
   (b) draw a vector $\hat{\theta}^{(k)}$ following a multinormal distribution of mean $\hat{\theta}$ and variance-covariance matrix $\hat{\boldsymbol{\Sigma}}$;
   (c) draw a residual $\hat{\varepsilon}^{(k)}$ following a centered normal distribution of standard deviation $\hat{\sigma}$ (linear case) or $\hat{k} X_1^{\hat{c}}$ (non-linear case);
   (d) calculate the prediction $\hat{Y}^{(k)} = f(\hat{X}_1^{(k)}, \ldots, \hat{X}_p^{(k)}; \hat{\theta}^{(k)}) + \hat{\varepsilon}^{(k)}$.

2. The confidence interval of the prediction is the empirical confidence interval of the Q values $\hat{Y}^{(1)}, \ldots, \hat{Y}^{(Q)}$.

This confidence interval corresponds to the prediction for a tree selected at random. To obtain the confidence interval for the mean tree, simply apply the same pseudo-algorithm but replaced step (c) by:

(...)

(c) with $\hat{\varepsilon}^{(k)} = 0$;

(...)

7.2.4 Inverse variables transformation

We saw in section 6.1.5 how a variable transformation can render linear a model that initially did not satisfy the hypotheses of the linear model. Variable transformation affects both the mean and the residual error. The same may be said of the inverse transformation, with implications on the calculation of prediction expectation. The log transformation is that most commonly used, but other types of transformation are also available.

Log transformation

Let us first consider the case of the log transformation of volume or biomass values, which is by far the most common case for volume and biomass models. Let us assume that a log transformation has been applied to biomass B to fit a linear model using effect variables $X_1, \ldots, X_p$:

$$Y = \ln(B) = a_0 + a_1 X_1 + \ldots + a_p X_p + \varepsilon \tag{7.13}$$

where

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

This is equivalent to saying that $\ln(B)$ follows a normal distribution of mean $a_0 + a_1 X_1 + \ldots + a_p X_p$ and standard deviation σ or of saying, by definition, that B follows a log-normal distribution of parameters $a_0 + a_1 X_1 + \ldots + a_p X_p$ and σ. The expectation of this log-normal distribution is:

$$\mathrm{E}(B) = \exp\left(a_0 + a_1 X_1 + \ldots + a_p X_p + \frac{\sigma^2}{2}\right)$$

Compared to the inverse model of (7.13) which corresponds to $B = \exp(a_0 + a_1 X_1 + \ldots + a_p X_p)$, the inverse transformation of the residual error introduces bias into the prediction which can be corrected by multiplying the prediction $\exp(a_0 + a_1 X_1 + \ldots + a_p X_p)$ by an correction factor (Parresol, 1999):

$$\mathrm{CF} = \exp\left(\frac{\hat{\sigma}^2}{2}\right) \tag{7.14}$$

We must however take care as the biomass models reported in the literature, and fitted after log transformation of the biomass, sometimes include this factor and sometimes do not.

If the decimal $\log_{10}$ has been used for variable transformation rather than the natural log, the correction factor is:

$$\mathrm{CF} = \exp\left[\frac{(\hat{\sigma}\ln 10)^2}{2}\right] \approx \exp\left(\frac{\hat{\sigma}^2}{0.3772}\right)$$

35

Correction factor for predicted biomass

Let us return to the biomass model in red line 31 fitted by multiple regression on log-transformed data:

$$\ln(B) = -8.38900 + 0.85715 \ln(D^2 H) + 0.72864 \ln(\rho)$$

If we consider the starting data but use the exponential function (without taking account of the correction factor), we obtain an under-estimated prediction: $B = \exp(-8.38900) \times (D^H)^{0.85715} \rho^{0.72864} = 2.274 \times 10^{-4} (D^2 H)^{0.85715} \rho^{0.72864}$. Let `m` be the object containing the fitted model (red line 31). The adjustment factor $\mathrm{CF} = \exp(\hat{\sigma}^2/2)$ is obtained by the command:

```
exp(summary(m)$sigma^2/2)
```

and here is 1.061035. The correct model is therefore: $B = 2.412 \times 10^{-4} (D^2 H)^{0.85715} \rho^{0.72864}$.

Any transformation

In the general case, let ψ be a variable transformation of the biomass (or volume) such that the response variable $Y = \psi(B)$ may be predicted by a linear regression against the effect variables $X_1, \ldots, X_p$. We shall assume that ψ is derivable and invertible. As $\psi(B)$ follows a normal distribution of mean $a_0 + a_1 X_1 + \ldots + a_p X_p$ and standard deviation σ, $B = \psi^{-1}[\psi(B)]$ has an expectation of (Saporta, 1990, p.26):

$$\mathrm{E}(B) = \int \psi^{-1}(x)\, \phi(x)\, \mathrm{d}x \tag{7.15}$$

where ϕ is the probability density of the normal distribution of mean $a_0 + a_1 X_1 + \ldots + a_p X_p$ and standard deviation σ. This expectation is generally different from $\psi^{-1}(a_0 + a_1 X_1 + \ldots + a_p X_p)$: the variable transformation induces prediction bias when we return to the starting variable by inverse transformation. The drawback of formula (7.15) is that it requires the calculation of an integral.

When the residual standard deviation σ is small, the δ-method (Serfling, 1980) provides an approximate expression of this prediction bias:

$$\begin{aligned} \mathrm{E(B)} &\simeq \psi^{-1}[\mathrm{E}(Y)] + \frac{1}{2}\mathrm{Var}(Y)\, (\psi^{-1})''[\mathrm{E}(Y)] \\ &\simeq \psi^{-1}(a_0 + a_1 X_1 + \ldots + a_p X_p) + \frac{\sigma^2}{2}\, (\psi^{-1})''(a_0 + a_1 X_1 + \ldots + a_p X_p) \end{aligned}$$

"Smearing" estimation

The "smearing" estimation method is a nonparametric method used to correct prediction bias when an inverse transformation is used on the response variable of a linear model (Duan, 1983; Taylor, 1986; Manning & Mullahy, 2001). Given that we can rewrite biomass (or volume) expectation equation (7.15) as:

$$\begin{aligned} \mathrm{E}(B) &= \int \psi^{-1}(x)\, \phi_0(x - a_0 - a_1 X1 - \ldots - a_p X_p)\, \mathrm{d}x \\ &= \int \psi^{-1}(x + a_0 + a_1 X1 + \ldots + a_p X_p)\, \mathrm{d}\Phi_0(x) \end{aligned}$$

where ϕ_0 (respectively Φ_0) is the probability density (respectively the distribution function) of the centered normal distribution of standard deviation σ, the smearing method consists in replacing Φ_0 by the empirical distribution function of the residuals from model fitting, i.e.:

$$\begin{aligned} B_{\text{smearing}} &= \int \psi^{-1}(x + a_0 + a_1 X_1 + \ldots + a_p X_p) \times \frac{1}{n} \sum_{i=1}^{n} \delta(x - \hat{\varepsilon}_i)\, \mathrm{d}x \\ &= \frac{1}{n} \sum_{i=1}^{n} \psi^{-1}(a_0 + a_1 X_1 + \ldots + a_p X_p + \hat{\varepsilon}_i) \end{aligned}$$

where δ is Dirac's function with zero mass and $\hat{\varepsilon}_i$ is the residual of the fitted model for the ith observation. This method of correcting the prediction bias has the advantage of being both very general and easy to calculate. It has the drawback that the residuals $\hat{\varepsilon}_i$ of model fitting need to be known. This is not a problem when you yourself fit a model to the data, but is a problem when using published volume or biomass tables that are not accompanied by the residuals.

In the particular case of the log transformation, ψ^{-1} is the exponential function, and therefore the smearing estimation of the biomass corresponds to: $\exp(a_0 + a_1 X_1 + \ldots + a_p X_p) \times \mathrm{CF}_{\text{smearing}}$, where the smearing correction factor is:

$$\mathrm{CF}_{\text{smearing}} = \frac{1}{n} \sum_{i=1}^{n} \exp(\hat{\varepsilon}_i)$$

Given that $\hat{\sigma}^2 = (\sum_{i=1}^{n} \hat{\varepsilon}_i^2)/(n - p - 1)$, the smearing correction factor is different from correction factor (7.14). But, on condition that $\hat{\sigma} \to 0$, the two adjustment coefficients are equivalent.

36

"Smearing" estimation of biomass

Let us again return to the biomass model in red line 31 fitted by multiple regression on log-transformed data:

$$\ln(B) = -8.38900 + 0.85715 \ln(D^2 H) + 0.72864 \ln(\rho)$$

The smearing correction factor may be obtained by the command:

```
mean(exp(residuals(m)))
```

where `m` is the object containing the fitted model, and in this example is 1.059859. By way of a comparison, the previously calculated correction factor (red line 35) was 1.061035.

7.3 Predicting the volume or biomass of a stand

If we want to predict the volume or biomass of a stand using biomass tables, we cannot measure model entries for all the trees in the stand, only for a sample. The volume or biomass of the trees in this sample will be calculated using the model, then extrapolated to the entire stand. Predicting the volume or biomass of a stand therefore includes two sources of variability: one related to the model's individual prediction, the other to the sampling of the trees in the stand. If we rigorously take account of these two sources of variability when

making predictions on a stand scale, this will cause complex double sampling problems as mentioned in sections 2.1.2 and 2.3 (Parresol, 1999).

The problem is slightly less complex when the sample of trees used to construct the model is independent of the sample of trees in which the entries were measured. In this case we can consider that the prediction error stemming from the model is independent of the sampling error. Let us assume that n sampling plots of unit area A were selected in a stand whose total area is $\mathcal{A}$. Let N_i be the number of trees found in the ith plot ($i = 1, \ldots, n$) and let $X_{ij1}, \ldots, X_{ijp}$ be the p effect variables measured on the jth tree of the ith plot ($j = 1, \ldots, N_i$). Cunia (1965, 1987b) considered the particular case where biomass is predicted by multiple regression based on the p effect variables. The estimation of stand biomass is in this case:

$$\begin{aligned}\hat{B} &= \frac{\mathcal{A}}{n}\sum_{i=1}^{n}\frac{1}{A}\sum_{j=1}^{N_i}(\hat{a}_0 + \hat{a}_1 X_{ij1} + \ldots + \hat{a}_p X_{ijp}) \\ &= \hat{a}_0\Big(\frac{\mathcal{A}}{nA}\sum_{i=1}^{n} N_i\Big) + \hat{a}_1\Big(\frac{\mathcal{A}}{nA}\sum_{i=1}^{n}\sum_{j=1}^{N_i} X_{ij1}\Big) + \ldots + \hat{a}_p\Big(\frac{\mathcal{A}}{nA}\sum_{i=1}^{n}\sum_{j=1}^{N_i} X_{ijp}\Big)\end{aligned}$$

where $\hat{a}_0, \ldots, \hat{a}_p$ are the estimated coefficients of the regression. If $X_0^\star = (\mathcal{A}/nA)\sum_{i=1}^{n} N_i$ and for each $k = 1, \ldots, p$,

$$X_k^\star = \frac{\mathcal{A}}{nA}\sum_{i=1}^{n}\sum_{j=1}^{N_i} X_{ijk}$$

The estimated biomass of the stand is therefore

$$\hat{B} = \hat{a}_0 X_0^\star + \hat{a}_1 X_1^\star + \ldots + \hat{a}_p X_p^\star$$

Interestingly, the variability of $\hat{\mathbf{a}} = {}^{\mathrm{t}}[\hat{a}_0, \ldots, \hat{a}_p]$ depends entirely on model fitting, not on stand sampling, while on the other hand, the variability of $\mathbf{x} = {}^{\mathrm{t}}[X_0^\star, \ldots, X_p^\star]$ depends entirely on the sampling, not on the model. Given that these two errors are independent,

$$\mathrm{E}(\hat{B}) = \mathrm{E}({}^{\mathrm{t}}\hat{\mathbf{a}}\mathbf{x}) = {}^{\mathrm{t}}\mathrm{E}(\hat{\mathbf{a}})\,\mathrm{E}(\mathbf{x})$$

and

$$\mathrm{Var}(\hat{B}) = {}^{\mathrm{t}}\mathbf{a}\boldsymbol{\Sigma}_{\mathbf{x}}\mathbf{a} + {}^{\mathrm{t}}\mathbf{x}\boldsymbol{\Sigma}_{\mathbf{a}}\mathbf{x}$$

where $\boldsymbol{\Sigma}_{\mathbf{a}}$ is the $(p+1)\times(p+1)$ variance-covariance matrix of the model's coefficients while $\boldsymbol{\Sigma}_{\mathbf{x}}$ is the $(p+1)\times(p+1)$ variance-covariance matrix of the sample of $\mathbf{x}$. The first matrix is derived from model fitting whereas the second is derived from stand sampling. Thus, the error for the biomass prediction of a stand is the sum of these two terms, one of which is related to the prediction error of the model and the other to stand sampling error.

More generally, the principle is exactly the same as when we considered on page 180, an uncertainty related to the measurement of the effect variables $X_1, \ldots, X_p$. A measurement error is not of the same nature as a sampling error. But from a mathematical standpoint, the calculations are the same: in both cases this means considering that the effect variables $X_1, \ldots, X_p$ are random, not fixed. We may therefore, in this general case, use a Monte Carlo method to estimate stand biomass. The pseudo-algorithm of this Monte Carlo method is the same as that used previously (see p.180):

1. For k from 1 to Q, where Q is the number of Monte Carlo iterations:
   (a) for i from 1 to p, draw $\hat{X}_i^{(k)}$ following a distribution corresponding to the variability of the stand sample (this distribution depends on the type of sampling conducted, the size and number of plots inventoried, etc.);

(b) draw a vector $\hat{\theta}^{(k)}$ following a multinormal distribution of mean $\hat{\theta}$ and variance-covariance matrix $\hat{\mathbf{\Sigma}}$;

(c) calculate the prediction $\hat{Y}^{(k)} = f(\hat{X}_1^{(k)}, \ldots, \hat{X}_p^{(k)}; \hat{\theta}^{(k)})$.

2. The confidence interval of the prediction is the empirical confidence interval of the Q values $\hat{Y}^{(1)}, \ldots, \hat{Y}^{(Q)}$.

7.4 Expanding and converting volume and biomass models

It may happen that we have a model which predicts a variable that is not exactly the one we need, but is closely related to it. For example, we may have a model that predicts the dry biomass of the trunk, but we are looking to predict the total above-ground biomass of the tree. Or, we may have a model that predicts the volume of the trunk, but we are looking to predict its dry biomass. Rather than deciding not to use a model simply because it does not give us exactly what we want, we can use it by default and adjust it using a factor. We can use *conversion* factors to convert volume to biomass (and *vice versa*), *expansion* factors to extrapolate a part to the whole, or combinations of the two. Using this approach Henry *et al.* (2011) put forward three methods to obtain the total biomass:

- The biomass of the trunk is the product of its volume and specific wood density ρ;
- above-ground biomass is the product of trunk biomass and a biomass expansion factor (BEF);
- above-ground biomass is the product of trunk volume and a biomass expansion and conversion factor (BCEF = BEF$\times\rho$).

Tables of these different conversion and expansion factors are available. These values are often very variable as they implicitly include difference sources of variability. Although the default table may be very precise, we often loose the benefit of this precision when using an expansion or correction factor as the prediction error cumulates all the sources of error involved in its calculation.

If we have a table that uses height as entry value, but do not have any height information, we may use a corollary model that predicts height from the entry values we do have (typically a height-dbh relation). Like for conversion and expansion factors, this introduces an additional source of error.

7.5 Arbitrating between different models

If we want to predict the volume or biomass of given trees, we often have a choice between different models. For instance, several models may have been fitted in different places for a given species. Or we may have a local model and a pan-tropical model. Arbitrating between the different models available is not always an easy task (Henry *et al.*, 2011). For example, should we choose a specific, local model fitted using little data (and therefore unbiased *a priori* but with high prediction variability) or a pan-tropical, multispecific model fitted with a host of data (therefore potentially biased but with low prediction variability)? This shows how many criteria are potentially involved when making a choice: model quality (the range of its validity, its capacity to extrapolate predictions, etc.), its specificity (with local, monospecific models at one extreme and pan-tropical, multispecific models at the other), the size of the dataset used for model fitting (and therefore implicitly the variability of

its predictions). Arbitrating between different models should not be confused with model selection, as described in section 6.3.2 for when selecting models their coefficients are not yet known, and when we estimate these coefficients we are looking for the model that best fits the data. When arbitrating between models we are dealing with models already fitted and whose coefficients are known.

Arbitration between different models often takes place without any biomass or volume data. But the case we will focus on here is when we have a reference dataset $\mathcal{S}_n$, with n observations of the response variable (volume of biomass) and the effect variables.

7.5.1 Comparing using validation criteria

When we have a dataset $\mathcal{S}_n$, we can compare the different models available on the basis of the validation criteria described in section 7.1.1, by using $\mathcal{S}_n$ as the validation dataset. Given that the models do not necessarily have the same number p of parameters, and in line with the parsimony principle, we should prefer the validation criteria that depend on p in such a manner as to penalize those models with many parameters.

When comparing a particular, supposedly "best" candidate model with different competitors, we can compare its predictions with those of its competitors. To do this we look to see whether or not the predictions made by its competitors are contained within the confidence interval at a significance level of α of the candidate model.

7.5.2 Choosing a model

A model may be chosen with respect to a "true" model f that although unknown may be assumed to exist. Let M be the number of models available. Let us note as $\hat{f}_m$ the function of the p effect variables that predicts volume or biomass based on the mth model. This function is random as it depends on coefficients that are estimated and therefore have their own distribution. The distribution of $\hat{f}_m$ therefore describes the variability of the predictions based on the model m, as described in section 7.2. The M models may have very different forms: the $\hat{f}_1$ model may correspond to a power function, the $\hat{f}_2$ model to a polynomial function, etc. We will also assume that there is a function f of the p effect variables that describes the "true" relation between the response variable (volume or biomass) and these effect variables. We do not know this "true" relation. We do not know what form it has. But each of the M models may be considered to be an approximation of the "true" relation f.

In the models selection theory (Massart, 2007), the difference between the relation f and a model $\hat{f}_m$ is quantified by a function γ that we call the *loss* function. For example, the loss function may be the norm L^2 of the difference between f and $\hat{f}_m$:

$$\gamma(f,\, \hat{f}_m) = \int_{x_1} \dots \int_{x_p} [f(x_1,\, \dots,\, x_p) - \hat{f}_m(x_1,\, \dots,\, x_p)]^2 \, \mathrm{d}x_1 \dots \mathrm{d}x_p$$

The expectation of the loss with respect to the distribution of $\hat{f}_m$ is called the *risk* (written R) i.e. when integrating on the variability of model predictions:

$$R = \mathrm{E}[\gamma(f,\, \hat{f}_m)]$$

The best of the M models available is that which minimizes the risk. The problem is that we do not know the true relation f, therefore, this "best" model is also unknown. In the models selection theory this best model is called an *oracle*. The model finally chosen will be that such that the risk of the oracle is bounded for a vast family of functions f. Intuitively,

the model chosen is that where the difference between it and the "true" relation is limited whatever this "true" relation may be (within the limits of a range of realistic possibilities). We will not go into further detail here concerning this topic as it is outside the scope of this guide.

7.5.3 Bayesian model averaging

Rather than choose one model from the M available, with the danger of not choosing the "best", an alternative consists in combining the M competitor models to form a new model. This is called "Bayesian model averaging" and although it has been very widely used for weather forecasting models (Raftery *et al.*, 2005; Furrer *et al.*, 2007; Berliner & Kim, 2008; Smith *et al.*, 2009) it is little used for forestry models (Li *et al.*, 2008; Picard *et al.*, 2012). Let $\mathcal{S}_n = \{(Y_i, X_{i1}, \ldots, X_{ip}), i = 1, \ldots, p\}$ be a reference dataset with n observations of the response variable Y and the p effect variables. Bayesian model averaging considers that the distribution of the response variable Y is a mixture of M distributions:

$$g(Y|X_1, \ldots, X_p) = \sum_{m=1}^{M} w_m \, g_m(Y|X_1, \ldots, X_p)$$

where g is the distribution density of Y, g_m is the conditional distribution density of Y with model m being the "best", and w_m is the weight of the mth model in the mixture, that we may interpret as the *a posteriori* probability that the mth model is the "best". The *a posteriori* w_m probabilities are reflections of the quality of the fitting of the models to the data, and their sum is one: $\sum_{m=1}^{M} w_m = 1$.

In the same manner as for the model selection mentioned in the previous section, Bayesian model averaging assumes that there is a "true" (though unknown) relation between the response variable and the p effect variables, and that each model differs from this "true" relation according to a normal distribution of standard deviation σ_m. In other words, density g_m is the density of the normal distribution of mean $f_m(x_1, \ldots, x_p)$ and standard deviation σ_m, where f_m is the function of p variables corresponding to the mth model. Thus,

$$g(Y|X_1, \ldots, X_p) = \sum_{m=1}^{M} w_m \, \phi(Y; f_m(x_1, \ldots, x_p), \, \sigma_m)$$

where $\phi(\cdot; \mu, \sigma)$ is the probability density of the normal distribution of expectation μ and standard deviation σ. The model f_{mean} resulting from the combination of the M competitor models is defined as the expectation of the mixture model, i.e.:

$$f_{\text{mean}}(X_1, \ldots, X_p) = \mathrm{E}(Y|X_1, \ldots, X_p) = \sum_{m=1}^{M} w_m \, f_m(X_1, \ldots, X_p)$$

Thus, the model resulting from the combination of the M competitor models is a weighted mean of these M models, with the weight of model m being the *a posteriori* probability that model m is the best. We can also calculate the variance of the predictions made by model f_{mean} that is a combination of the M competitor models:

$$\begin{aligned} \mathrm{Var}(Y|X_1, \ldots, X_p) \;=\; & \sum_{m=1}^{M} w_m \Big[f_m(X_1, \ldots, X_p) - \sum_{l=1}^{M} w_l \, f_l(X_1, \ldots, X_p) \Big]^2 \\ & + \sum_{m=1}^{M} w_m \, \sigma_m^2 \end{aligned}$$

The first term corresponds to between-model variance and expresses prediction variability from one model to another; the second term corresponds to within-model variance, and expresses the conditional error of the prediction, with this model being the best.

If we want to use model f_{mean} instead of the M models $f_1, \ldots, f_M$, we now have to estimate the weights $w_1, \ldots, w_M$ and the within-model standard deviations $\sigma_1, \ldots, \sigma_M$. These $2M$ parameters are estimated from reference dataset $\mathcal{S}_n$ using an EM algorithm (Dempster *et al.*, 1977; McLachlan & Krishnan, 2008). The EM algorithm introduces latent variables z_{im} such that z_{im} is the *a posteriori* probability that model m is the best model for observation i of $\mathcal{S}_n$. The latent variables z_{im} take values of between 0 and 1. The EM algorithm is iterative and alternates between two steps at each iteration: step E (as in "expectation") and step M (as "maximisation"). The EM algorithms is as follows:

1. Choose start values $w_1^{(0)}, \ldots, w_M^{(0)}, \sigma_1^{(0)}, \ldots, \sigma_M^{(0)}$ for the $2M$ parameters to be estimated.

2. Alternate the two steps:

   (a) step E: calculate the value of z_{im} at iteration j sbased on the values of the parameters at iteration $j-1$:

   $$z_{im}^{(j)} = \frac{w_m^{(j-1)}\, \phi[Y_i; f_m(X_{i1}, \ldots, X_{ip}), \sigma_m^{(j-1)}]}{\sum_{k=1}^{M} w_k^{(j-1)}\, \phi[Y_i; f_k(X_{i1}, \ldots, X_{ip}), \sigma_k^{(j-1)}]}$$

   (b) step M: estimate the parameters at iteration j using current values of z_{im} as weight, i.e.:

   $$\begin{aligned} w_m^{(j)} &= \frac{1}{n}\sum_{i=1}^{n} z_{im}^{(j)} \\ {\sigma_m^{(j)}}^2 &= \frac{\sum_{i=1}^{n} z_{im}^{(j)}[Y_i - f_m(X_{i1}, \ldots, X_{ip})]^2}{\sum_{i=1}^{n} z_{im}^{(j)}} \end{aligned}$$

   such as $\sum_{m=1}^{M} |w_m^{(j)} - w_m^{(j-1)}| + \sum_{m=1}^{M} |\sigma_m^{(j)} - \sigma_m^{(j-1)}|$ is larger than a fixed infitesimal threshold (for example 10^{-6}).

3. The estimated value of w_m is $w_m^{(j)}$ and the estimated value of σ_m is $\sigma_m^{(j)}$.

Conclusions and recommendations

The methods used to estimate tree volume and biomass are constantly evolving as the search is always on for estimations that are as close as possible to reality. Volume and biomass estimation models have undergone different changes in different ecological areas. In dry tropical areas, the problem of obtaining sufficient supplies of firewood dates backs to ancient times, and here primarily allometric equations have been developed to quantify fuelwood. In wet tropical areas, where felling is mainly for timber, equations have been developed primarily to estimate volume. Today, climate change is of increasing concern and interest is being shown in biomass models in both wet and dry forests.

Biomass measurements are destined to increase in coming years to meet the needs of carbon stock estimations and understand the contribution made by terrestial ecosystems to the carbon cycle. Experience acquired in volume estimations has shown that two or three thousand observations are needed to estimate the trunk volume of *one* given species with sufficient precision to cover the variability across the geographic range of its distribution (CTFT, 1989). By comparison, the biomass model established by Chave *et al.* (2005), which is one of the most widely used biomass models today, was calibrated using 2410 observations. It is a pan-tropical model that covers all species and all ecological zones in dry and wet areas! The similarity between these two sample sizes — despite variability that differs by several orders of magnitude — shows that, with regard to biomass measurements, there is still considerable scope for progress in exploring the full extent of natural variability. And this particularly given that biomass, which involves all the compartments of a tree, probably has far greater intrinsic variability than simply the volume of the trunk.

The reliability of biomass estimations can be enhanced by increasing the number of observations available. But measuring the above-ground biomass of a tree requires a far greater effort than measuring the volume of its trunk. And the effort is even greater still when root biomass is included. At present, it is very unlikely that vast campaigns will be funded to measure the above-ground and below-ground biomass of trees. Like for Chave *et al.* (2005), the construction of new allometric equations must therefore be based on compilations of datasets collected at different locations by independent teams. Standardized biomass measurement methods, and model-fitting statistics capable of including additional information through effect covariables are therefore crucial if we are to improve tree biomass estimations in the years ahead. Experiments on regular stands (effects of ontogeny, plantation density, soil fertility or fertilizers, and more generally the effects of silviculture) will facilitate the construction of these generic models.

Contrary to other guides already available, we sought here to cover the entire process of constructing allometric equations, from field to prediction, and including model fitting. But we do not claim to have covered all possible situations. Many cases will require the development of specific methods. Large trees with buttresses, for instance, pose special problems for any prediction of their biomass — starting with the difficulty that their diameter at breast height, which is the first entry variable for most models, is not measurable. Hollow trees, strangler fig, bamboo and large epiphytes are all species with particularities that will pose problems when applying the methods put forward in this guide. New dendrometric

methods will very likely have to be developed to tackle these specific cases. Tools such as 3D modeling, photogrammetry, radar and lasers, on the ground or in the air, will facilitate or revolutionize biomass estimation methods and perhaps ultimately replace the power saw and weighing scales.

Statistical methods are also evolving. A comparison of the report by Whraton & Cunia (1987) with the fitting methods used today in forestry clearly shows just how much progress has been made in the use of increasingly cutting-edge statistical methods, and that we have attempted to explain in this guide. The consideration of within-stem variability could become common practice in the future when fitting biomass models.

Improvements in measurements and model fitting methods, and the increasing use of field measurements, will improve the processes of scientific research and tree biomass estimations only if the models and methods produced are made available in a transparent manner. A great deal of data are confined to libraries and are never published in scientific journals or on the Internet. Also, for a country that does not possess biomass data for some of its ecological areas, data obtained by neighboring countries or in identical ecological areas may not be easily accessible. We therefore encourage forestry players to identify the data already available for ecological areas or countries of interest. These data may be integrated into databases and serve to identify gaps. These gaps can then be plugged through field measurements conducted in accordance with the advice and red lines given in this guide.

If we are to guarantee continuity in the efforts made to improve estimations, a data archiving system will have to be set up as a starting point to improve future estimations. A robust archiving system should help reduce the efforts made by future workers to understand and recalculate current estimations. Also, it is important to underline that the methods set up must be consistent over time. This guide describes different measuring methods, but it is preferable to adopt one single method that can be easily reproduced and is less dependent upon financial, technological and human factors. Should an alternative method be developed for practical reasons, this should be reported and be made accessible such that the next guide will include the full diversity of the methodologies possible. Finally, it is always best to adopt methods that are both simple and reproducible.

Bibliography

AFNOR. 1985. Bois – détermination de la masse volumique. Tech. Rep. NF B51-005, AFNOR. 62

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 T. Automat. Contr.*, 19(6): 716–723. 150

Alder, D. 1980. *Estimation des volumes et accroissement des peuplements forestiers – Vol. 2. Étude et prévision de la production.* Études FAO : forêts No. 22/2. Rome, FAO. 194 pp. 25

Andrews, J.A. & 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. & de Carvalho, J.A. 1999. Comparison of formulae for biomass content determination in a tropical rain forest site in the state of Pará, Brazil. *For. Ecol. Manag.*, 117(1-3): 43–52. 101

Archibald, S. & 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.* Oxford, UK, Pergamon Press. 506 pp. 24, 25

Augusto, L., Meredieu, C., Bert, D., Trichet, P., Porté, A., Bosc, A., Lagane, F., Loustau, D., Pellerin, S., Danjon, F., Ranger, J. & Gelpe, J. 2008. Improving models of forest nutrient export with equations that predict the nutrient concentration of tree compartments. *Ann. For. Sci.*, 65(8): 808. 24

Basuki, T.M., van Laake, P.E., Skidmore, A.K. & Hussin, Y.A. 2009. Allometric equations for estimating the above-ground biomass in tropical lowland *Dipterocarp* forests. *For. Ecol. Manag.*, 257(8): 1684–1694. 101

Batho, A. & García, O. 2006. De Perthuis and the origins of site index: a historical note. *For. Biometry Model. Inform. Sci.*, 1: 1–10. 24

Becking, J.H. 1953. Einige gesichtspunkte für die durchführung von vergleichenden durchforstungsversuchen in gleichälterigen beständen. *In 11^{e} Congrès de l'Union Internationale des Instituts de Recherches Forestiers, Rome, 1953 : comptes rendus.* IUFRO, pp. 580–582. 211

Bellefontaine, R., Petit, S., Pain-Orcet, M., Deleporte, P. & Bertault, J.G. 2001. *Les arbres hors forêt : vers une meilleure prise en compte.* Cahier FAO Conservation No. 35. Rome, FAO. 214 pp. 33

Bergès, L., Nepveu, G. & Franc, A. 2008. Effects of ecological factors on radial growth and wood density components of sessile oak (*Quercus petraea* Liebl.) in Northern France. *For. Ecol. Manag.*, 255(3-4): 567–579. 24, 27

Berliner, L.M. & Kim, Y. 2008. Bayesian design and analysis for superensemble-based climate forecasting. *J. Climate*, 21(9): 1891–1910. 187

Bloom, A.J., Chapin, F.S. & Mooney, H.A. 1985. Resource fimitation in plants–an economic analogy. *Annu. Rev. Ecol. Syst.*, 16: 363–392. 27

Bohlman, S. & O'Brien, S. 2006. Allometry, adult stature and regeneration requirement of 65 tree species on Barro Colorado Island, Panama. *J. Trop. Ecol.*, 22(2): 123–136. 23

Bolker, B. 2008. *Ecological Models and Data in R.* Princeton, NJ, Princeton University Press. 179

Bontemps, J.D., Hervé, J.C. & Dhôte, J.F. 2009. Long-term changes in forest productivity: a consistent assessment in even-aged stands. *For. Sci.*, 55(6): 549–564. 26

Bontemps, J.D., Hervé, J.C., Leban, J.M. & Dhôte, J.F. 2011. Nitrogen footprint in a long-term observation of forest growth over the twentieth century. *Trees–Struct. Funct.*, 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. 47, 48

Bouchon, J. 1974. Les tarifs de cubage. Tech. rep., ENGREF, Nancy, France. 30

Bouriaud, O., Leban, J.M., Bert, D. & 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. & Draper, N.R. 1987. *Empirical Model Building and Response Surfaces.* Wiley series in probability and mathematical statistics. New York, NY, Wiley. 669 pp. 41

Bozdogan, H. 1987. Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, 52(3): 345–370. 150

Bradley, P.N. 1988. Survey of woody biomass on farms in western Kenya. *Ambio*, 17(1): 40–48. 29

Brown, I.F., Martinelli, L.A., Thomas, W.W., Moreira, M.Z., Victoria, R.A. & Ferreira, C.A.C. 1995. Uncertainty in the biomass of Amazonian forests: An example from Rondônia, Brazil. *For. Ecol. Manag.*, 75(1-3): 175–189. 41

Brown, S. 1997. *Estimating Biomass and Biomass Change of Tropical Forests: a Primer.* FAO Forestry Paper No. 134. Rome, FAO. 65 pp. 43, 101

Brown, S., Gillespie, A.J.R. & Lugo, A.E. 1989. Biomass estimation methods for tropical forests with applications to forest inventory data. *For. Sci.*, 35(4): 881–902. 101, 120

Burdon, R.D., Kibblewhite, R.P., Walker, J.C.F., Megraw, E.R. & 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*. *For. Sci.*, 50(4): 399–415. 27

Burnham, K.P. & Anderson, D.R. 2004. Multimodel inference: understanding AIC and BIC in model selection. *Sociol. Method. Res.*, 33(2): 261–304. 150

Burnham, K.P. & Anderson, D.R. 2002. *Model Selection and Multimodel Inference. A Practical Information-Theoretic Approach*. New York, NY, Springer Science+Business Media, Inc., 2nd edn. 488 pp. 150

Cailliez, F. 1980. *Forest volume estimation and yield prediction. Volume estimation*, *Études FAO forêts*, vol. 1. Rome, FAO. 98 pp. 30

Cairns, M.A., Brown, S., Helmer, E.H. & Baumgardner, G.A. 1997. Root biomass allocation in the world's upland forests. *Oecologia*, 111(1): 1–11. 27

Calama, R., Barbeito, I., Pardos, M., del Río, M. & Montero, G. 2008. Adapting a model for even-aged *Pinus pinea* L. stands to complex multi-aged structures. *For. Ecol. Manag.*, 256(6): 1390–1399. 28

Cavaignac, S., Nguyen Thé, N., Melun, F. & Bouvet, A. 2012. élaboration d'un modèle de croissance pour l'Eucalyptus gundal. *FCBA INFO*, p. 16. 26

Charru, M., Seynave, I., Morneau, F. & 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. *For. Ecol. Manag.*, 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. & Yamakura, T. 2005. Tree allometry and improved estimation of carbon stocks and balance in tropical forests. *Oecologia*, 145(1): 87–99. 101, 189

Chave, J., Coomes, D., Jansen, S., Lewis, S.L., Swenson, N.G. & Zanne, A.E. 2009. Towards a worldwide wood economics spectrum. *Ecol. Lett.*, 12(4): 351–366. 24

Chave, J., Riéra, B. & Dubois, M.A. 2001. Estimation of biomass in a neotropical forest of French Guiana: spatial and temporal variability. *J. Trop. Ecol.*, 17(1): 79–96. 101

Chave, J., Condit, R., Aguilar, S., Hernandez, A., Lao, S. & Perez, R. 2004. Error propagation and scaling for tropical forest biomass estimates. *Philos. Trans. R. Soc. Lond., B Biol. Sci.*, 359(1443): 409–420. 40, 46, 50

Chave, J., Condit, R., Lao, S., Caspersen, J.P., Foster, R.B. & Hubbell, S.P. 2003. Spatial and temporal variation of biomass in a tropical forest: results from a large census plot in Panama. *J. Ecol.*, 91(2): 240–252. 48, 50

Cochran, W.G. 1977. *Sampling Techniques*. Wiley Series in Probability and Mathematical Statistics. New York, NY, John Wiley & Sons, 3rd edn. 428 pp. 33, 38, 39, 42

Colin-Belgrand, M., Ranger, J. & Bouchon, J. 1996. Internal nutrient translocation in chesnut tree stemwood: III. Dynamics across an age series of *Castanea sativa* (Miller). *Ann. Bot.*, 78(6): 729–740. 24

Cotta, H. 1804. *Principes fondamentaux de la science forestière*. Paris, Bouchard-Huzard. 495 pp. 30

Courbaud, B., Goreaud, F., Dreyfus, P. & 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. *For. Ecol. Manag.*, 145(1): 15–28. 28

Cressie, N. 1993. *Statistics for Spatial Data*. Wiley Series in Probability and Mathematical Statistics. New York, NY, John Wiley & Sons, 2nd edn. 900 pp. 47

CTFT. 1989. *Mémento du forestier*. Paris, France, Ministère de la Coopération et du Développement, 3rd edn. 1266 pp. 33, 40, 41, 43, 47, 189

Cunia, T. 1964. Weighted least squares method and construction of volume tables. *For. Sci.*, 10(2): 180–191. 30, 120

Cunia, T. 1965. Some theory on reliability of volume estimates in a forest inventory sample. *For. Sci.*, 11(1): 115–128. 184

Cunia, T. 1987a. Construction of tree biomass tables by linear regression techniques. *In* E.H. Whraton & 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.* Broomall, PA, USA, USDA Forest Service, Northeastern Forest Experiment Station, General Technical Report no. NE-117, pp. 27–36. 120

Cunia, T. 1987b. Error of forest inventory estimates: its main components. *In* E.H. Whraton & 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.* Broomall, PA, USA, USDA Forest Service, Northeastern Forest Experiment Station, General Technical Report no. NE-117, pp. 1–14. 34, 39, 46, 184

Cunia, T. 1987c. An optimization model for subsampling trees for biomass measurement. *In* E.H. Whraton & 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.* Broomall, PA, USA, USDA Forest Service, Northeastern Forest Experiment Station, General Technical Report no. NE-117, pp. 109–118. 34, 39, 46, 48

Cunia, T. 1987d. An optimization model to calculate the number of sample trees and plots. *In* E.H. Whraton & T. Cunia, eds., *Estimating tree biomass regressions and their error. Proceedings of the workshop on tree biomass regression functions and their contri-*

bution to the error of forest inventory estimates, May 26–30, 1986, Syracuse, N.Y. – Part E. Broomall, PA, USA, USDA Forest Service, Northeastern Forest Experiment Station, General Technical Report no. NE-117, pp. 15–24. 34, 39, 46, 48

Cunia, T. & Briggs, R.D. 1984. Forcing additivity of biomass tables: some empirical results. *Can. J. For. Res.*, 14: 376–384. 166

Cunia, T. & Briggs, R.D. 1985a. Forcing additivity of biomass tables: use of the generalized least squares method. *Can. J. For. Res.*, 15: 23–28. 166

Cunia, T. & Briggs, R.D. 1985b. Harmonizing biomass tables by generalized least squares. *Can. J. For. Res.*, 15: 331–340. 168

de Vries, P.G. 1986. *Sampling Theory for Forest Inventory – A Teach-Yourself Course.* Berlin, Springer-Verlag. 399 pp. 38, 46

Dean, C. 2003. Calculation of wood volume and stem taper using terrestrial single-image close-range photogrammetry and contemporary software tools. *Silva Fenn.*, 37(3): 359–380. 169

Dean, C. & Roxburgh, S. 2006. Improving visualisation of mature, high-carbon sequestering forests. *For. Biometry Model. Inform. Sci.*, 1: 48–69. 168, 169

Dean, C., Roxburgh, S. & Mackey, B. 2003. Growth modelling of *Eucalyptus regnans* for carbon accounting at the landscape scale. *In* A. Amaro, D. Reed & P. Soares, eds., *Modelling Forest Systems.* Wallingford, UK, CAB International Publishing, pp. 27–39. 169

Deans, J.D., Moran, J. & Grace, J. 1996. Biomass relationships for tree species in regenerating semi-deciduous tropical moist forest in Cameroon. *For. Ecol. Manag.*, 88(3): 215–225. 41

Decourt, N. 1973. Production primaire, production utile : méthodes d'évaluation, indices de productivité. *Ann. Sci. For.*, 30(3): 219–238. 25

Deleuze, C., Blaudez, D. & Hervé, J.C. 1996. Fitting a hyperbolic model for height versus girth relationship in spruce stands. Spacing effects. *Ann. Sci. For.*, 53(1): 93–111. 26

Dempster, A.P., Laird, N.M. & Rubin, D.B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Stat. Soc. B*, 39(1): 1–38. 188

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. 26

Dhôte, J.F. 1991. Modélisation de la croissance des peuplements réguliers de hêtre : dynamique des hiérarchies sociales et facteurs de production. *Ann. Sci. For.*, 48(4): 389–416. 24

Dhôte, J.F. 1996. A model of even-aged beech stands productivity with process-based interpretations. *Ann. Sci. For.*, 53(1): 1–20. 25, 26

Díaz, S. & Cabido, M. 1997. Plant functional types and ecosystem function in relation to global change. *J. Veg. Sci.*, 8: 463–474. 164

Dietz, J. & 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. & Clark, J.S. 2008. Capturing diversity and interspecific variability in allometries: A hierarchical approach. *For. Ecol. Manag.*, 256(11): 1939–1948. 23

Djomo, A.N., Ibrahima, A., Saborowski, J. & Gravenhorst, G. 2010. Allometric equations for biomass estimations in Cameroon and pan moist tropical equations including biomass data from Africa. *For. Ecol. Manag.*, 260(10): 1873–1885. 45, 101

Dong, J., Kaufmann, R.K., Myneni, R.B., Tucker, C.J., Kauppi, P.E., Liski, J., Buermann, W., Alexeyev, V. & Hughes, M.K. 2003. Remote sensing estimates of boreal and temperate forest woody biomass: carbon pools, sources, and sinks. *Remote Sens. Environ.*, 84: 393–410. 29

Dreyfus, P. 2012. Joint simulation of stand dynamics and landscape evolution using a tree-level model for mixed uneven-aged forests. *Ann. For. Sci.*, 69(2): 283–303. 28

Duan, N. 1983. Smearing estimate: a nonparametric retransformation method. *J. Am. Statist. Assoc.*, 78(383): 605–610. 31, 182

Durbin, J. & Watson, G.S. 1971. Testing for serial correlation in least squares regression. III. *Biometrika*, 58(1): 1–19. 111

Ebuy Alipade, J., Lokombé Dimandja, J.P., Ponette, Q., Sonwa, D. & Picard, N. 2011. Biomass equation for predicting tree aboveground biomass at Yangambi, DRC. *J. Trop. For. Sci.*, 23(2): 125–132. 41

Efron, B. & Tibshirani, R.J. 1993. *An Introduction to the Bootstrap.* Monographs on Statistics and Applied Probability No. 57. New York, NY, Chapman & Hall. 436 pp. 172, 173

Eichhorn, F. 1904. Beziehungen zwischen Bestandshöhe und Bestandsmasse. *Allgemeine Forst- und Jagdzeitung*, 80: 45–49. 25, 212

Enquist, B.J., Brown, J.H. & West, G.B. 1998. Allometric scaling of plant energetics and population density. *Nature*, 395(6698): 163–165. 23, 101

Enquist, B.J., West, G.B., Charnov, E.L. & Brown, J.H. 1999. Allometric scaling of production and life-history variation in vascular plants. *Nature*, 401(6756): 907–911. 23, 101

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. & 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, USA. 28

Fairfield Smith, H. 1938. An empirical law describing heterogeneity in the yields of agricultural crops. *J. Agric. Sci.*, 28: 1–23. 47, 49

Fang, Z. & Bailey, R.L. 1999. Compatible volume and taper models with coefficients for tropical species on Hainan island in southern China. *For. Sci.*, 45(1): 85–100. 169

FAO. 2006. *Global Forest Resources Assessment 2005. Progress towards sustainable forest management*, *FAO Forestry Paper*, vol. 147. Rome, Food and Agriculture Organization of the United Nations. 28

Favrichon, V. 1998. Modeling the dynamics and species composition of tropical mixed-species uneven-aged natural forest: effects of alternative cutting regimes. *For. Sci.*, 44(1): 113–124. 28

Fonweban, J.N. & Houllier, F. 1997. Tarif de peuplement et modèle de production pour *Eucalyptus saligna* au Cameroun. *Bois For. Trop.*, 253: 21–36. 92

Fournier-Djimbi, M. 1998. Le matériau bois : structure, propriétés, technologie. Cours, ENGREF, Département de foresterie rurale et tropicale, Montpellier, France. 62

Franc, A., Gourlet-Fleury, S. & Picard, N. 2000. *Introduction à la modélisation des forêts hétérogènes*. Nancy, France, ENGREF. 312 pp. 28, 101

Furnival, G.M. 1961. An index for comparing equations used in constructing volume tables. *For. Sci.*, 7(4): 337–341. 123, 156

Furrer, R., Knutti, R., Sain, S.R., Nychka, D.W. & Meehl, G.A. 2007. Spatial patterns of probabilistic temperature change projections from a multivariate Bayesian analysis. *Geophys. Res. Lett.*, 34(L06711): 1–4. 187

Gambill, C.W., Wiant, H. V., J. & Yandle, D.O. 1985. Optimum plot size and BAF. *For. Sci.*, 31(3): 587–594. 48, 49

García, O. 2003. Dimensionality reduction in growth models: an example. *For. Biometry Model. Inform. Sci.*, 1: 1–15. 25

García, O. 2011. Dynamical implications of the variability representation in site-index modelling. *Eur. J. For. Res.*, 130(4): 671–675. 24, 25

Gayon, J. 2000. History of the concept of allometry. *Am. Zool.*, 40(5): 748–758. 23

Gehring, C., Park, S. & Denich, M. 2004. Liana allometric biomass equations for Amazonian primary and secondary forest. *For. Ecol. Manag.*, 195: 69–83. 31

Genet, A., Wernsdörfer, H., Jonard, M., Pretzsch, H., Rauch, M., Ponette, Q., Nys, C., Legout, A., Ranger, J., Vallet, P. & Saint-André, L. 2011. Ontogeny partly explains the apparent heterogeneity of published biomass equations for *Fagus sylvatica* in central Europe. *For. Ecol. Manag.*, 261(7): 1188–1202. 27, 53

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. & Thomas, D.W. 2006. A standard protocol for liana censuses. *Biotropica*, 38(2): 256–261. 31

Gerwing, J.J. & Farias, D.L. 2000. Integrating liana abundance and forest stature into an estimate of total aboveground biomass for an eastern Amazonian forest. *J. Trop. Ecol.*, 16(3): 327–335. 31

Gibbs, H.K., Brown, S., Niles, J.O. & Foley, J.A. 2007. Monitoring and estimating tropical forest carbon stocks: making REDD a reality. *Environ. Res. Lett.*, 2(4): 1–13. Doi:10.1088/1748-9326/2/4/045023. 28

Gomat, H.Y., Deleporte, P., Moukini, R., Mialounguila, G., Ognouabi, N., Saya, R.A., Vigneron, P. & Saint-André, L. 2011. What factors influence the stem taper of *Eucalyptus*: growth, environmental conditions, or genetics? *Ann. For. Sci.*, 68(1): 109–120. 27

Gonzalez, P., Asner, G.P., Battles, J.J., Lefsky, M.A., Waring, K.M. & Palace, M. 2010. Forest carbon densities and uncertainties from Lidar, QuickBird, and field measurements in California. *Remote Sens. Environ.*, 114(7): 1561–1575. 29

Gould, S.J. 1979. An allometric interpretation of species-area curves. The meaning of the coefficient. *Am. Nat.*, 114(3): 335–343. 101

Gould, S.J. 1966. Allometry and size in ontogeny and phylogeny. *Biol. Rev.*, 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. *Am. Nat.*, 105(942): 113–136. 23

Goupy, J. 1999. *Plans d'expériences pour surfaces de réponse*. Paris, Dunod. 409 pp. 41

Gourlet-Fleury, S. & Houllier, F. 2000. Modelling diameter increment in a lowland evergreen rain forest in French Guiana. *For. Ecol. Manag.*, 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. & Picard, N. 2011. Environmental filtering of dense-wooded species controls above-ground biomass stored in African moist forests. *J. Ecol.*, 99(4): 981–990. 27, 62

Gregoire, T.G. & Dyer, M.E. 1989. Model fitting under patterned heterogeneity of variance. *For. Sci.*, 35(1): 105–125. 30

Guilley, E., Hervé, J.C. & Nepveu, G. 2004. The influence of site quality, silviculture and region on wood density mixed model in *Quercus petraea* Liebl. *For. Ecol. Manag.*, 189(1-3): 111–121. 24, 27

Hairiah, K., Sitompul, S.M., van Noordwijk, M. & Palm, C.A. 2001. *Methods for sampling carbon stocks above and below ground*. ASB Lecture Note No. 4B. Bogor, Indonesia, International Centre for Research in Agroforestry (ICRAF). 32 pp. 30

Härdle, W. & Simar, L. 2003. *Applied Multivariate Statistical Analysis*. Berlin, Springer-Verlag. 496 pp. 96

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. 211

Hawthorne, W. 1995. *Ecological Profiles of Ghanaian Forest Trees*. Tropical Forestry Paper No. 29. Oxford, UK, Oxford Forestry Institute, Department of Plant Sciences, University of Oxford. 71

Hebert, J., Rondeux, J. & Laurent, C. 1988. Comparaison par simulation de 3 types d'unités d'échantillonnage en futaies feuillues de hêtre (*Fagus silvatica* l.). *Ann. Sci. For.*, 45(3): 209–221. 48

Henry, M., Besnard, A., Asante, W.A., Eshun, J., Adu-Bredu, S., Valentini, R., Bernoux, M. & Saint-André, L. 2010. Wood density, phytomass variations within and among trees, and allometric equations in a tropical rainforest of Africa. *For. Ecol. Manag.*, 260(8): 1375–1388. 7, 8, 9, 13, 24, 31, 68, 84, 86, 87, 92, 93, 98, 99, 101, 102, 113, 114, 118, 120, 125, 126, 127, 134, 136, 137, 152, 153, 154, 155, 158, 165, 175

Henry, M., Picard, N., Trotta, C., Manlay, R., Valentini, R., Bernoux, M. & Saint-André, L. 2011. Estimating tree biomass of sub-Saharan African forests: a review of available allometric equations. *Silva Fenn.*, 45(3B): 477–569. 40, 100, 185

Hitchcock, H.C.I. & McDonnell, J.P. 1979. Biomass measurement: a synthesis of the literature. *In Proceedings of IUFRO workshop on forest resource inventories, July 23-26, 1979*. Fort Collins, Colorado, USA, SAF-IUFRO, pp. 544–595. 30

Hoeting, J.A., Madigan, D., Raftery, A.E. & Volinsky, C.T. 1999. Bayesian model averaging: a tutorial. *Stat. Sci.*, 14(4): 382–417. 132

Hofstad, O. 2005. Review of biomass and volume functions for individual trees and shrubs in Southeast Africa. *J. Trop. For. Sci.*, 17(1): 151–162. 100

Holmgren, P., Masakha, E.J. & 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. 29

Huxley, J.S. 1924. Constant differential growth-ratios and their significance. *Nature*, 114: 895–896. 23

Ikonen, V.P., Kellomäki, S., Väisänenet, H. & Peltola, H. 2006. Modelling the distribution of diameter growth along the stem in Scots pine. *Trees–Struct. Funct.*, 20(3): 391–402. 27

Jackson, R.B., Canadell, J., Ehleringer, J.R., Mooney, H.A., Sala, O.E. & Schulze, E.D. 1996. A global analysis of root distributions for terrestrial biomes. *Oecologia*, 108(3): 389–411. 27

Jacobs, M.W. & Cunia, T. 1980. Use of dummy variables to harmonize tree biomass tables. *Can. J. For. Res.*, 10: 483–490. 168

Johnson, F.A. & Hixon, H.J. 1952. The most efficient size and shape of plot to use for cruising in old growth Douglas-fir timber. *J. For.*, 50: 17–20. 47

Keller, M., Palace, M. & Hurtt, G. 2001. Biomass estimation in the Tapajos National Forest, Brazil. Examination of sampling and allometric uncertainties. *For. Ecol. Manag.*, 154(3): 371–382. 47

Kelly, J.F. & 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, LA, USA. 30

Ketterings, Q.M., Coe, R., van Noordwijk, M., Ambagau, Y. & Palm, C.A. 2001. Reducing uncertainty in the use of allometric biomass equations for predicting above-ground tree biomass in mixed secondary forests. *For. Ecol. Manag.*, 146(1-3): 199–209. 45, 101

King, D.A. 1996. Allometry and life history of tropical trees. *J. Trop. Ecol.*, 12: 25–44. 23

Knapic, S., Louzada, J.L. & Pereira, H. 2011. Variation in wood density components within and between *Quercus faginea* trees. *Can. J. For. Res.*, 41(6): 1212–1219. 24

Kozak, A. 1970. Methods for ensuring additivity of biomass components by regression analysis. *For. Chron.*, 46(5): 402–405. 31

Lahti, T. & 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 No. 27. Rome, Italie, FAO. 208 pp. 46

Larson, P.R. 1963. Stem form development of forest trees. *For. Sci. Monog.*, 5: 1–42. 27

Lavorel, S. & Garnier, E. 2002. Predicting changes in community composition and ecosystem functioning from plant traits: revisiting the Holy Grail. *Funct. Ecol.*, 16: 545–556. 164

Lefsky, M.A., Cohen, W.B., Harding, D.J., Parker, G.G., Acker, S.A. & Gower, S.T. 2002. Lidar remote sensing of above-ground biomass in three biomes. *Global Ecol. Biogeogr.*, 11(5): 393–399. 29

Levillain, J., Thongo M'Bou, A., Deleporte, P., Saint-André, L. & Jourdan, C. 2011. Is the simple auger coring method reliable for below-ground standing biomass estimation in *Eucalyptus* forest plantations? *Ann. Bot.*, 108(1): 221–230. 71, 72, 74

Li, Y., Andersen, H.E. & McGaughey, R. 2008. A comparison of statistical methods for estimating forest biomass from light detection and ranging data. *West. J. Appl. For.*, 23(4): 223–231. 187

Loetsch, F. & Haller, K.E. 1973. *Forest Inventory. Statistics of Forest Inventory and Information from Aerial Photographs*, vol. 1. Munchen, BLV Verlagsgesellschaft mbH. 436 pp. 46

Louppe, D., Koua, M. & Coulibaly, A. 1994. Tarifs de cubage pour *Afzelia africana* Smith en forêt de Badénou (nord Côte d'Ivoire). Tech. rep., Institut des Forêts (IDEFOR), département foresterie, Côte d'Ivoire. 92

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, VA, USA. 30

Magnus, J.R. & Neudecker, H. 2007. *Matrix Differential Calculus with Applications in Statistics and Econometrics*. Wiley series in probability and statistics. Chichester, UK, John Wiley and Sons, 3rd edn. 450 pp. 116, 121

Magnussen, S., Kleinn, C. & Picard, N. 2008a. Two new density estimators for distance sampling. *Eur. J. For. Res.*, 127(3): 213–224. 50

Magnussen, S., Picard, N. & Kleinn, C. 2008b. A gamma-poisson distribution of the point to the k nearest event distance. *For. Sci.*, 54(4): 429–441. 50

Maguire, D.A. & Batista, J.L.F. 1996. Sapwood taper models and implied sapwood volume and foliage profiles for coastal Douglas-fir. *Can. J. For. Res.*, 26: 849–863. 168

Maniatis, D., Saint-André, L., Temmerman, M., Malhi, Y. & Beeckman, H. 2011. The potential of using xylarium wood samples for wood density calculations: a comparison of approaches for volume measurements. *iForest – Biogeosci. For.*, 4: 150–159. 65

Manning, W.G. & Mullahy, J. 2001. Estimating log models: to transform or not to transform? *J. Health Econ.*, 20: 461–494. 182

Martinez-Yrizar, A., Sarukhan, J., Perez-Jimenez, A., Rincon, E., Maass, J.M., Solis-Magallanes, A. & Cervantes, L. 1992. Above-ground phytomass of a tropical deciduous forest on the coast of Jalisco, México. *J. Trop. Ecol.*, 8: 87–96. 101

Massart, P. 2007. *Concentration Inequalities and Model Selection. École d'Été de Probabilités de Saint-Flour XXXIII – 2003*. Lecture Notes in Mathematics No. 1896. Berlin Heidelberg, Springer-Verlag. 335 pp. 186

McCarthy, M.C. & Enquist, B.J. 2007. Consistency between an allometric approach and optimal partitioning theory in global patterns of plant biomass allocation. *Funct. Ecol.*, 21(4): 713–720. 27

McLachlan, G.J. & Krishnan, T. 2008. *The EM Algorithm and Extensions*. Wiley Series in Probability and Statistics. Hoboken, NJ, John Wiley & Sons, 2nd edn. 360 pp. 188

Meredieu, C., Perret, S. & Dreyfus, P. 2003. Modelling dominant height growth: effect of stand density. *In* A. Amaro, D. Reed & 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*. Wallingford, UK, CAB International Publishing, pp. 111–121. 26

Metcalf, C.J.E., Clark, J.S. & Clark, D.A. 2009. Tree growth inference and prediction when the point of measurement changes: modelling around buttresses in tropical forests. *J.*

Trop. Ecol., 25(1): 1–12. 168

Mokany, K., Raison, R.J. & Prokushkin, A.S. 2006. Critical analysis of root : shoot ratios in terrestrial biomes. *Global Change Biol.*, 12(1): 84–96. 27

Monreal, C.M., Etchevers, J.D., Acosta, M., Hidalgo, C., Padilla, J., López, R.M., Jiménez, L. & Velázquez, A. 2005. A method for measuring above- and below-ground C stocks in hillside landscapes. *Can. J. Soil Sci.*, 85(Special Issue): 523–530. 30

Muller, K.E. & Stewart, P.W. 2006. *Linear Model Theory. Univariate, Multivariate and Mixed Models*. Wiley series in probability and statistics. Hoboken, NJ, John Wiley & Sons. 410 pp. 168

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. & Ashton, P. 2006. Testing metabolic ecology theory for allometric scaling of tree size, growth and mortality in tropical forests. *Ecol. Lett.*, 9(5): 575–588. 24, 101

Myers, R.H. & Montgomery, D.C. 2002. *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*. Wiley series in probability and statistics. New York, NY, Wiley. 824 pp. 41

Namaalwa, J., Eid, T. & Sankhayan, P. 2005. A multi-species density-dependent matrix growth model for the dry woodlands of Uganda. *For. Ecol. Manag.*, 213(1-3): 312–327. 28

Návar, J. 2009. Allometric equations for tree species and carbon stocks for forests of northwestern Mexico. *For. Ecol. Manag.*, 257(2): 427–434. 101

Návar, J., Méndez, E. & Dale, V. 2002. Estimating stand biomass in the Tamaulipan thornscrub of northeastern Mexico. *Ann. For. Sci.*, 59(8): 813–821. 31

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. & Roupsard, O. 2008. Fruit development, not GPP, drives seasonal variation in NPP in a tropical palm plantation. *Tree Physiology*, 28(11): 1661–1674. 72

Nelson, B.W., Mesquita, R., Pereira, L.G., Garcia Aquino de Souza, J.S., Teixeira Batista, G. & Bovino Couto, L. 1999. Allometric regressions for improved estimate of secondary forest biomass in the central Amazon. *For. Ecol. Manag.*, 117(1-3): 149 –167. 101

Ngomanda, A., Moundounga Mavouroulou, Q., Engone Obiang, N.L., Midoko Iponga, D., Mavoungou, J.F., Lépengué, N., Picard, N. & Mbatchi, B. 2012. Derivation of diameter measurements for buttressed trees, an example from Gabon.

J. Trop. Ecol., 28(3): 299–302. 132

Nicolini, É., Chanson, B. & Bonne, F. 2001. Stem growth and epicormic branch formation in understorey beech trees (*Fagus sylvatica* L.). *Ann. Bot.*, 87(6): 737–750. 27

Nogueira, E.M., Fearnside, P.M., Nelson, B.W., Barbosa, R.I. & Keizer, E.W.H. 2008. Estimates of forest biomass in the Brazilian Amazon: New allometric equations and adjustments to biomass from wood-volume inventories. *For. Ecol. Manag.*, 256(11): 1853–1867. 101

Nogueira, E.M., Nelson, B.W. & Fearnside, P.M. 2006. Volume and biomass of trees in central Amazonia: influence of irregularly shaped and hollow trunks. *For. Ecol. Manag.*, 227(1-2): 14–21. 32

Paine, C.E.T., Marthews, T.R., Vogt, D.R., Purves, D., Rees, M., Hector, A. & Turnbull, L.A. 2012. How to fit nonlinear plant growth models and calculate growth rates: an update for ecologists. *Method. Ecol. Evol.*, 3(2): 245–256. 179

Pardé, J. 1980. Forest biomass. *Forestry Abstracts*, 41(8): 343–362. 30

Pardé, J. & Bouchon, J. 1988. *Dendrométrie*. Nancy, France, ENGREF, 2nd edn. 328 pp. 25, 30, 35, 40, 41, 43

Parresol, B.R. 1993. Modeling multiplicative error variance: an example predicting tree diameter from stump dimensions in baldcypress. *For. Sci.*, 39(4): 670–679. 30

Parresol, B.R. 1999. Assessing tree and stand biomass: a review with examples and critical comparisons. *For. Sci.*, 45(4): 573–593. 30, 31, 46, 120, 156, 166, 168, 169, 172, 181, 184

Parresol, B.R. 2001. Additivity of nonlinear biomass equations. *Can. J. For. Res.*, 31(5): 865–878. 30

Parresol, B.R. & Thomas, C.E. 1989. A density-integral approach to estimating stem biomass. *For. Ecol. Manag.*, 26: 285–297. 168

Patenaude, G., Hill, R.A., Milne, R., Gaveau, D.L.A., Briggs, B.B.J. & Dawson, T.P. 2004. Quantifying forest above ground carbon content using LiDAR remote sensing. *Remote Sens. Environ.*, 93(3): 368–380. 29

Pearson, T. & Brown, S. 2005. Guide de mesure et de suivi du carbone dans les forêts et prairies herbeuses. Report, Winrock International, Arlington, VA, USA. 30

Peng, C. 2000. Growth and yield models for uneven-aged stands: past, present and future. *For. Ecol. Manag.*, 132(2-3): 259–279. 28

Perot, T., Goreaud, F., Ginisty, C. & Dhôte, J.F. 2010. A model bridging distance-dependent and distance-independent tree models to simulate the growth of mixed forests. *Ann. For. Sci.*, 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

(ITCF), Paris. 96

Picard, N. & Bar-Hen, A. 2007. Estimation of the density of a clustered point pattern using a distance method. *Environ. Ecol. Stat.*, 14(4): 341–353. 47, 50

Picard, N. & Favier, C. 2011. A point-process model for variance-occupancy-abundance relationships. *Am. Nat.*, 178(3): 383–396. 47

Picard, N. & Franc, A. 2001. Aggregation of an individual-based space-dependent model of forest dynamics into distribution-based and space-independent models. *Ecol. Model.*, 145(1): 69–84. 28

Picard, N., Kouyaté, A.M. & Dessard, H. 2005. Tree density estimations using a distance method in mali savanna. *For. Sci.*, 51(1): 7–18. 50

Picard, N., Sylla, M.L. & Nouvellet, Y. 2004. Relationship between plot size and the variance of the density estimator in West African savannas. *Can. J. For. Res.*, 34(10): 2018–2026. 47

Picard, N., Henry, M., Mortier, F., Trotta, C. & Saint-André, L. 2012. Using Bayesian model averaging to predict tree aboveground biomass. *For. Sci.*, 58(1): 15–23. 187

Picard, N., Yalibanda, Y., Namkosserena, S. & Baya, F. 2008. Estimating the stock recovery rate using matrix models. *For. Ecol. Manag.*, 255(10): 3597–3605. 28

Ponce-Hernandez, R., Koohafkan, P. & Antoine, J. 2004. *Assessing carbon stocks and modelling win-win scenarios of carbon sequestration through land-use changes.* Rome, Food and Agriculture Organization of the United Nations (FAO). 156 pp. 30

Porté, A. & Bartelink, H.H. 2002. Modelling mixed forest growth: a review of models for forest management. *Ecol. Model.*, 150(1-2): 141–188. 28

Press, W.H., Teukolsky, S.A., Vetterling, W.T. & Flannery, B.P. 2007. *Numerical Recipes: The Art of Scientific Computing.* Cambridge, UK, Cambridge University Press, 3rd edn. 1235 pp. 144

Preßler, M.R. 1864. *Das Gesetz der Stammbildung.* Leipzig, Germany, Arnoldische Buchhandlung. 153 pp. 212

Pretzsch, H. 2009. *Forest Dynamics, Growth and Yield: From Measurement to Model.* Berlin, Springer-Verlag. 664 pp. 24

Pukkala, T., Lähde, E. & Laiho, O. 2009. Growth and yieldmodels for unevensizedforest stands in Finland. *For. Ecol. Manag.*, 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. 31

R Development Core Team. 2005. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. 21

Raftery, A.E., Gneiting, T., Balabdaoui, F. & Polakowski, M. 2005. Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Weather Rev.*, 133(5): 1155–1174. 187

Reed, D.D. & Green, E.J. 1985. A method of forcing additivity of biomass tables when using nonlinear models. *Can. J. For. Res.*, 15(6): 1184–1187. 31

Reinecke, L.H. 1933. Perfecting a stand-density index for even-aged forests. *J. Agr. Res.*, 46(7): 627–638. 212

Reyes, G., Brown, S., Chapman, J. & Lugo, A.E. 1992. Wood densities of tropical tree species. General Technical Report SO-88, USDA Forest Service, Southern Forest Experiment Station, New Orleans, Louisiana, USA. 62

Rivoire, M., Genet, A., Didier, S., Nys, C., Legout, A., Longuetaud, F., Cornu, E., Freyburger, C., Motz, A., Bouxiero, N. & Saint-André, L. 2009. Protocole d'acquisition de données volume-biomasse-minéralomasse, Bure. Rapport technique, INRA, Nancy, France. 53

Rösch, H., Van Rooyen, M.W. & Theron, G.K. 1997. Predicting competitive interactions between pioneer plant species by using plant traits. *J. Veg. Sci.*, 8: 489–494. 164

Russell, C. 1983. *Nutrient cycling and productivity of native and plantation forests at Jari Florestal, Para, Brazil.* Ph.D. thesis, University of Georgia, Athens, GA, USA. 41

Rutishauser, E., Wagner, F., Herault, B., Nicolini, E.A. & Blanc, L. 2010. Contrasting above-ground biomass balance in a Neotropical rain forest. *J. Veg. Sci.*, 21: 672–682. 50

Rykiel, E.J.J. 1996. Testing ecological models: the meaning of validation. *Ecol. Model.*, 90: 229–244. 171

Saatchi, S.S., Houghton, R.A., Dos Santos Alvalá, R.C., Soares, J.V. & Yu, Y. 2007. Distribution of aboveground live biomass in the Amazon basin. *Global Change Biol.*, 13(4): 816–837. 29

Saint-André, L., Laclau, J.P., Bouillet, J.P., Deleporte, P., Miabala, A., Ognouabi, N., Baillères, H., Nouvellon, Y. & Moukini, R. 2002a. 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

Saint-André, L., Laclau, J.P., Deleporte, P., Ranger, J., Gouma, R., Saya, A. & Joffre, R. 2002b. A generic model to describe the dynamics of nutrient concentrations within stemwood across an age series of a eucalyptus hybrid. *Ann. Bot.*, 90(1): 65–76. 24, 57

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. & 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*. Bogor, Indonesia, CIFOR, pp. 173–189. 25

Saint-André, L., Leban, J.M., Houllier, F. & 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. *Ann. For. Sci.*, 56(2): 121–132. 27

Saint-André, L., Thongo M'Bou, A., Mabiala, A., Mouvondy, W., Jourdan, C., Roupsard, O., Deleporte, P., Hamel, O. & Nouvellon, Y. 2005. Age-related equations for above- and below-ground biomass of a *Eucalyptus* hybrid in Congo. *For. Ecol. Manag.*, 205(1-3): 199–214. 31, 43, 53, 146, 162

Saporta, G. 1990. *Probabilités, analyse des données et statistique*. Paris, Technip. 493 pp. 35, 38, 41, 46, 173, 174, 176, 177, 179, 182

Savage, V.M., Deeds, E.J. & Fontana, W. 2008. Sizing up allometric scaling theory. *PLoS Comput. Biol.*, 4(9): e1000171. 27

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*. Clemson, SC, Belle W. Baruch Forest Science Institute, Clemson University, pp. 95–112. 172

Schnitzer, S.A., DeWalt, S.J. & Chave, J. 2006. Censusing and measuring lianas: a quantitative comparison of the common methods. *Biotropica*, 38(5): 581–591. 31

Schnitzer, S.A., Rutishauser, S. & Aguilar, S. 2008. Supplemental protocol for liana censuses. *For. Ecol. Manag.*, 255: 1044–1049. 31

Schreuder, H.T., Banyard, S.G. & Brink, G.E. 1987. Comparison of three sampling methods in estimating stand parameters for a tropical forest. *For. Ecol. Manag.*, 21(1-2): 119–127. 48

Schreuder, H.T., Gregoire, T.G. & Wood, G.B. 1993. *Sampling methods for multiressource forest inventory*. New York, NY, Wiley & Sons. 446 pp. 46, 50

Serfling, R.J. 1980. *Approximation Theorems of Mathematical Statistics*. Wiley Series in Probability and Statistics. New York, NY, John Wiley & Sons. 371 pp. 177, 180, 182

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*. Bethesda, MD, USA, Society of American Foresters, pp. 1–13. 26

Shinozaki, K., Yoda, K., Hozumi, K. & Kira, T. 1964a. A quantitative analysis of plant form - the pipe model theory. I. Basic analyses. *Jpn. J. Ecol.*, 14: 97–104. 23, 27

Shinozaki, K., Yoda, K., Hozumi, K. & Kira, T. 1964b. A quantitative analysis of plant form - the pipe model theory. II. Further evidence of the theory and its application on forest ecology. *Jpn. J. Ecol.*, 14: 133–139. 23, 27

Shiver, B.D. & Borders, B.E. 1996. *Sampling techniques for forest ressource inventory.* New York, NY, Wiley & Sons. 356 pp. 38, 39, 46

Sillett, S.C., Van Pelt, R., Koch, G.W., Ambrose, A.R., Carroll, A.L., Antoine, M.E. & Mifsud, B.M. 2010. Increasing wood production through old age in tall trees. *For. Ecol. Manag.*, 259(5): 976–994. 169

Skovsgaard, J.P. & 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, 25

Smith, R.L., Tebaldi, C., Nychka, D. & Mearns, L.O. 2009. Bayesian modeling of uncertainty in ensembles of climate models. *J. Am. Statist. Assoc.*, 104(485): 97–116. 187

Soares, P. & Tomé, M.. 2002. Height-diameter equation for first rotation eucalypt plantations in Portugal. *For. Ecol. Manag.*, 166(1-3): 99–109. 26

St.-Onge, B., Hu, Y. & Vega, C. 2008. Mapping the height and above-ground biomass of a mixed forest using lidar and stereo Ikonos images. *Int. J. Remote Sens.*, 29(5): 1277–1294. 29

Stoyan, D. & Stoyan, H. 1994. *Fractals, Random Shapes and Point Fields.* Chichester, UK, John Wiley & Sons. 390 pp. 47

Tateno, R., Hishi, T. & 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. *For. Ecol. Manag.*, 193(3): 297–306. 27

Taylor, J.M.G. 1986. The retransformed mean after a fitted power transformation. *J. Am. Statist. Assoc.*, 81: 114–118. 31, 182

Tedeschi, L.O. 2006. Assessment of the adequacy of mathematical models. *Agr. Syst.*, 89(2-3): 225–247. 172

Thompson, S.K. 1992. *Sampling.* Wiley Series in Probability and Mathematical Statistics. New York, NY, John Wiley & Sons. 343 pp. 38, 39

Thornley, J.H. 1972. A balanced quantitative model for root: shoot ratios in vegetative plants. *Ann. Bot.*, 36(2): 431–441. 27

Tomé, M., Barreiro, S., Paulo, J.A. & Tomé, J. 2006. Age-independent difference equations for modelling tree and stand growth. *Can. J. For. Res.*, 36(7): 1621–1630. 28

Valinger, E. 1992. Effects of thinning and nitrogen fertilization on stem growth and stem form of *Pinus sylvestris* trees. *Scand. J. For. Res.*, 7(1-4): 219–228. 27

Vallet, P., Dhôte, J.F., Le Moguédec, G., Ravart, M. & Pignard, G. 2006. Development of total aboveground volume equations for seven important forest tree species in France. *For. Ecol. Manag.*, 229(1-3): 98–110. 27

Vallet, P. & 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. *J. Veg. Sci.*, 22(5): 932–942. 28

van Breugel, M., Ransijn, J., Craven, D., Bongers, F. & Hall, J.S. 2011. Estimating carbon stock in secondary forests: Decisions and uncertainties associated with allometric biomass models. *For. Ecol. Manag.*, 262(8): 1648–1657. 40, 43, 46, 50

Van Pelt, R. 2001. *Forest Giants of the Pacific Coast.* Vancouver, Canada, Global Forest Society. 169

Vanclay, J.K. 1994. *Modelling Forest Growth and Yield – Applications to Mixed Tropical Forests.* Wallingford, UK, CAB International Publishing. 312 pp. 28

Vanclay, J.K. 2009. Tree diameter, height and stocking in even-aged forests. *Ann. For. Sci.*, 66(7): 702. 26

Verzelen, N., Picard, N. & Gourlet-Fleury, S. 2006. Approximating spatial interactions in a model of forest dynamics as a means of understanding spatial patterns. *Ecol. Complex.*, 3(3): 209–218. 28

Violle, C., Navas, M.L., Vile, D., Kazakou, E., Fortunel, C., Hummel, I. & Garnier, E. 2007. Let the concept of trait be functional! *Oikos*, 116: 882–892. 164

Wagner, F., Rutishauser, E., Blanc, L. & Herault, B. 2010. Effects of plot size and census interval on descriptors of forest structure and dynamics. *Biotropica*, 42(6): 664–671. 47, 48, 50

Weiskittel, A.R., Hann, D.W., Hibbs, D.E., Lam, T.Y. & Bluhm, A.A. 2009. Modeling top height growth of red alder plantations. *For. Ecol. Manag.*, 258(3): 323–331. 25

West, G.B., Brown, J.H. & Enquist, B.J. 1997. A general model for the origin of allometric scaling laws in biology. *Science*, 276: 122–126. 23, 101

West, G.B., Brown, J.H. & Enquist, B.J. 1999. A general model for the structure and allometry of plant vascular systems. *Nature*, 400(6745): 664–667. 23, 27, 101

West, P.W. 2009. *Tree and Forest Measurement.* Berlin, Springer-Verlag, 2nd edn. 191 pp. 46, 50

White, J.F. & Gould, S.J. 1965. Interpretation of the coefficient in the allometric equation. *Am. Nat.*, 99(904): 5–18. 23

Whraton, E.H. & 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. Broomall, PA, USA, USDA Forest Service, Northeastern Forest Experiment Station. 50, 120, 190

Yamakura, T., Hagihara, A., Sukardjo, S. & Ogawa, H. 1986. Tree size in a mature dipterocarp forest stand in Sebulu, East Kalimantan, Indonesia. *Southeast Asian Stud.*, 23(4): 452–478. 101

Zeide, B. 1980. Plot size optimization. *For. Sci.*, 26(2): 251–257. 48, 49

Zianis, D. & Mencuccini, M. 2004. On simplifying allometric analyses of forest biomass. *For. Ecol. Manag.*, 187(2-3): 311–332. 23, 171

Zianis, D., Muukkonen, P., Mäkipää, R. & Mencuccini, M. 2005. *Biomass and Stem Volume Equations for Tree Species in Europe.* Silva Fennica Monographs No. 4. Vantaa, Finland, The Finnish Society of Forest Science and The Finnish Forest Research Institute. 63 pp. 23, 40, 100

Glossary

This glossary provides definitions for certain terms used in this guide that are either unusual or used to convey a meaning other than their usual meaning.

Additivity. Property of a system of allometric equations fitted to different compartments of a tree and to a tree in its entirety such that the sum of the predictions for each of the compartments indeed corresponds to the prediction for the entire tree.

Aliquot. Part extracted from a tree compartment, the measurement of which can be used to measure the compartment by a rule of three.

Allometry. Statistical relation on a population scale between two size characteristics of the individuals in this population. This relation in most cases has a power form. Example: there is allometry in vertebrates between adult body mass and brain size.

Biomass. Mass of living or dead organic matter in an organism, expressed as mass of dry matter. For a tree, this is expressed in kg. By extension, the biomass of an area is the sum of the biomasses of the organisms found in the area. This is measured in kg per unit area.

White noise. In probabilities, a random process generating random variables that are all independent one of the other.

Compartment. Part of a tree, generally determined such that the organs within a compartment have similar densities (ratio of dry biomass to fresh volume) (leaves, trunk, large branches, etc. are compartments).

Covariance. Quantity that measures the simultaneous variation of two random variables, i.e. the covariance becomes increasingly positive for each pair of values that differ from their mean in this same direction, and increasingly negative for each pair of values that differ from their mean in opposite directions. The covariance of a random variable and of this same random variable gives the variance.

Doublet. Collection of two numerical values.

Fractal. Object whose structure does not change with scale changes.

Hart-Becking index. In forest sciences, an index established by Hart (1928) and Becking (1953) that measures the storage of a stand based on the average distance between trees and the dominant height in the stand. This index is calculated as a ratio between the average distance between trees over dominant height, multiplied by 100.

Heteroscedasticity. The opposite of homoscedasticity, i.e. when the variance of the residual error of a model is not constant (and typically varies with one of the model's effect variables).

Homoscedasticity. Property by which the variance of the residual error of a model is constant. Homoscedasticity is one of the conditions required for fitting a linear model.

Eichhorn's rule. In forest sciences, an empirical rule laid down by Eichhorn (1904) which stipules: the specific volume of an even-aged, single-species, closed-cover stand depends only on its dominant height. This is Eichhorn's second rule; Eichhorn's first rule stipules: the dominant height of an even-aged, single-species, closed-cover stand depends only on age, species and station conditions.

Pressler's rule. In forest sciences, an empirical rule laid down by Preßler (1864) which stipules: the basal area increment is constant from the stump up to the base of the functional part of the crown.

Nutrient content. Quantity of mineral elements in the biomass.

Dirac's function. Distribution (in the statistical sense of the term) concentrated on a value x_0 of a continuous random variable (i.e. the probability of the random variable being $< x$ is 0 for $x < x_0$ and 1 for $x > x_0$).

Monte Carlo. Said of a method that aims to calculate a numerical value by simulating a random process.

Social position. For a tree, the position of its crown in the canopy, which determines its hierarchy with respect to competitors for light (we also speak of social status). The distinction is often made between dominant, co-dominant and dominated trees.

Reinecke Density Index (RDI). In forest sciences, an index established by Reinecke (1933) that measures the storage of a stand based on the number of stems per hectare (= stand density) and tree mean basal area (= mean quadratic diameter). This index is calculated as a ratio between stand density and maximal density such as determined from the mean quadratic diameter by the self-thinning line.

Ordinal variable. Variable that takes discrete values such that the different modalities of these discrete values may be ordered. For example, the month of the year is an ordinal value (months may be arranged in chronological order).

Variance. Quantity that measures the dispersion of a random variable around its mean value. It is calculated as the mean of the sum of squares.

Lexicon of mathematical symbols

Latin symbols

a estimated value of a coefficient in a predictive model

A area of a sampling plot

$\mathcal{A}$ area of a stand

b estimated value of a coefficient in a predictive model

B biomass of an aliquot, a compartment (trunk, branches, leaves...), of a tree or stand

CV_X coefficient of variation of a variable X

c exponent of a power expression

C definition 1: circumference of a tree; definition 2: cost of sampling; definition 3: model validation criterion

D diameter of a tree at breast height (dbh)

D_0 dbh of the dominant tree in a stand

E precision of the estimation of a variable

f a function linking a response variable to one or more effect variables

F Furnival's index

g a function

G basal area of a tree or a stand

h a height between zero (the ground) and the height H of the tree

H height of a tree

H_0 height of the dominant tree in the stand

$\mathbf{I}_n$ Fisher's information matrix for a sample of size n

k multiplier of a power expression

K number of subsets in a cross validation

ℓ likelihood of a sample

$\mathcal{L}$ log-likelihood of a sample

L length of a log

M definition 1: number of biomass compartments in a tree; definition 2: number of competitor models for predicting the same response variable

n sample size

N definition 1: total number of sampling units (tree or plot) in a stand; definition 2: stand density (number of stems per hectare)

$\mathcal{N}$ the normal distribution (also called a Gaussian or Laplace-Gauss distribution)

p number of effect variables in a model (not including the y-intercept)

P stem profile (plot giving the area of a cross-cut through the trunk at different heights)

q definition 1: number of estimated parameters in a model; definition 2: quantile of the standard normal distribution

Q number of Monte Carlo iterations

R definition 1: a model's determination coefficient; definition 2 (in the model selection theory): a risk; definition 3: radius of a log

S number of strata in a stratification

S_X empirical standard deviation of a variable X

$\mathcal{S}_n$ dataset containing n observations

t_n quantile of a Student's distribution with n degrees of freedom

T age of a plantation

V volume of a log, tree or stand

w definition 1: weight of an observation in a weighted regression; definition 2: weight of a model in a mixture of models

X a variable (generally the effect variable of a model)

$\mathbf{x}$ a vector of effect variables

$\mathbf{X}$ design matrix for a linear model

Y a variable (generally the response variable of a model)

$\mathbf{Y}$ response vector of a multivariate model

z a latent variable for the EM algorithm

Z a variable (generally a covariable defining the stratification of a dataset)

Greek symbols

α definition 1: (unknown) "true" value of a coefficient in a predictive model; definition 2: confidence threshold for a confidence interval (generally 5 %)

β (unknown) "true" value of a coefficient in a predictive model

γ loss function (in the model selection theory)

δ Dirac's function

Δ a difference in value for a given variable

ε residual error of a predictive model

$\boldsymbol{\varepsilon}$ vector of the residual errors of a multivariate model

ζ residual covariance between two compartments

η volumetric shrinkage coefficient

θ a set of model parameters

$\boldsymbol{\theta}$ a vector of parameters in a multivariate model

ϑ a set of parameters

μ expectation of a random variable = (unknown) "true" mean of a variable to be estimated

ξ Box-Cox transformation parameter

ρ wood density

σ standard deviation of the residual error of a predictive model

$\boldsymbol{\Sigma}$ variance-covariance matrix of a multinormal distribution

τ (unknown) "true" standard deviation of a variable to be estimated

ϕ probability density of a normal distribution

ψ function defining a variable transformation

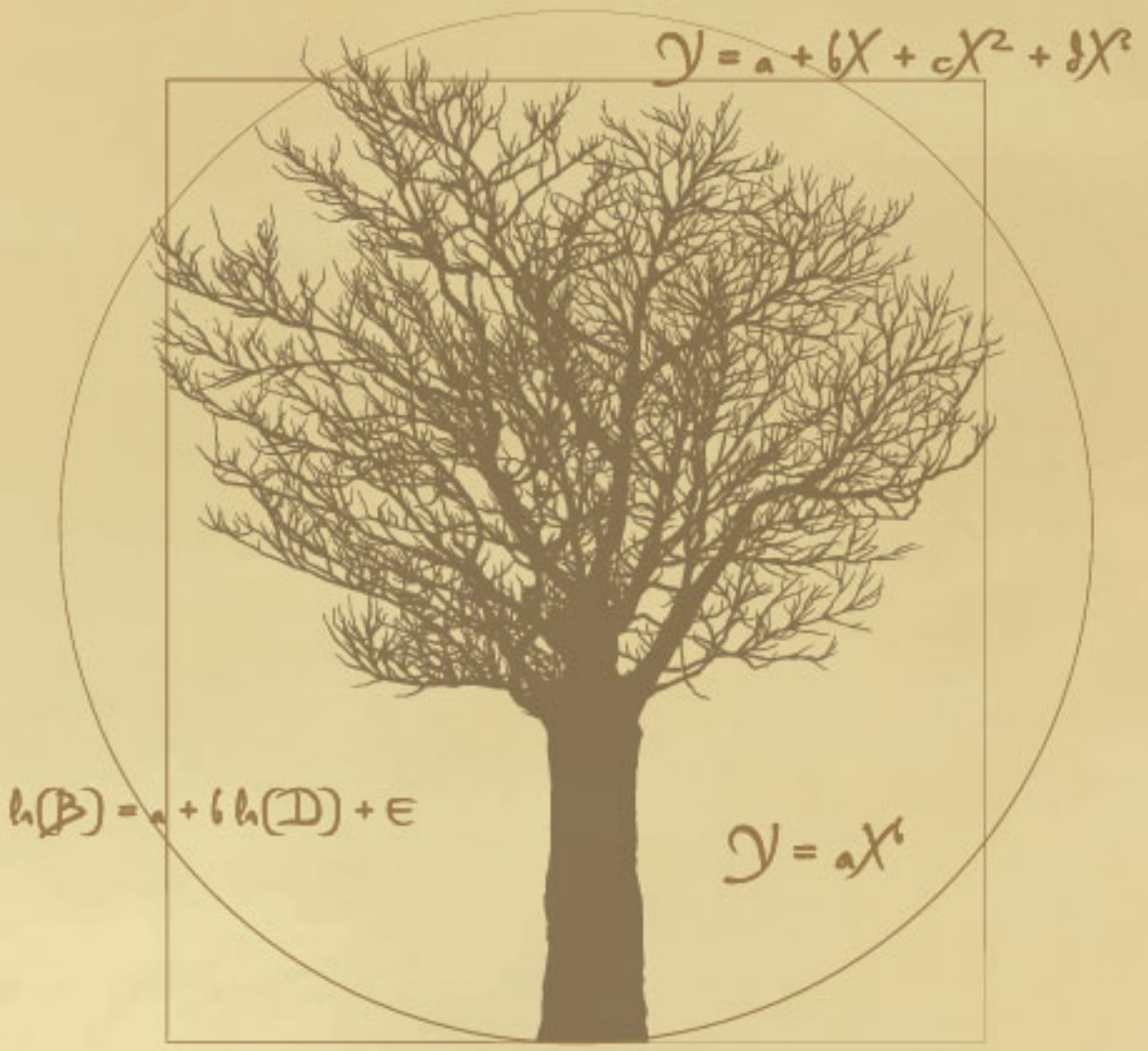
χ water content

χ_0 fiber saturation point

ω a proportion (for example, the proportion of fresh biomass in a log)

Non-alphabetical symbols

$\varnothing$ diameter of a tree, a log, a branch or a root



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