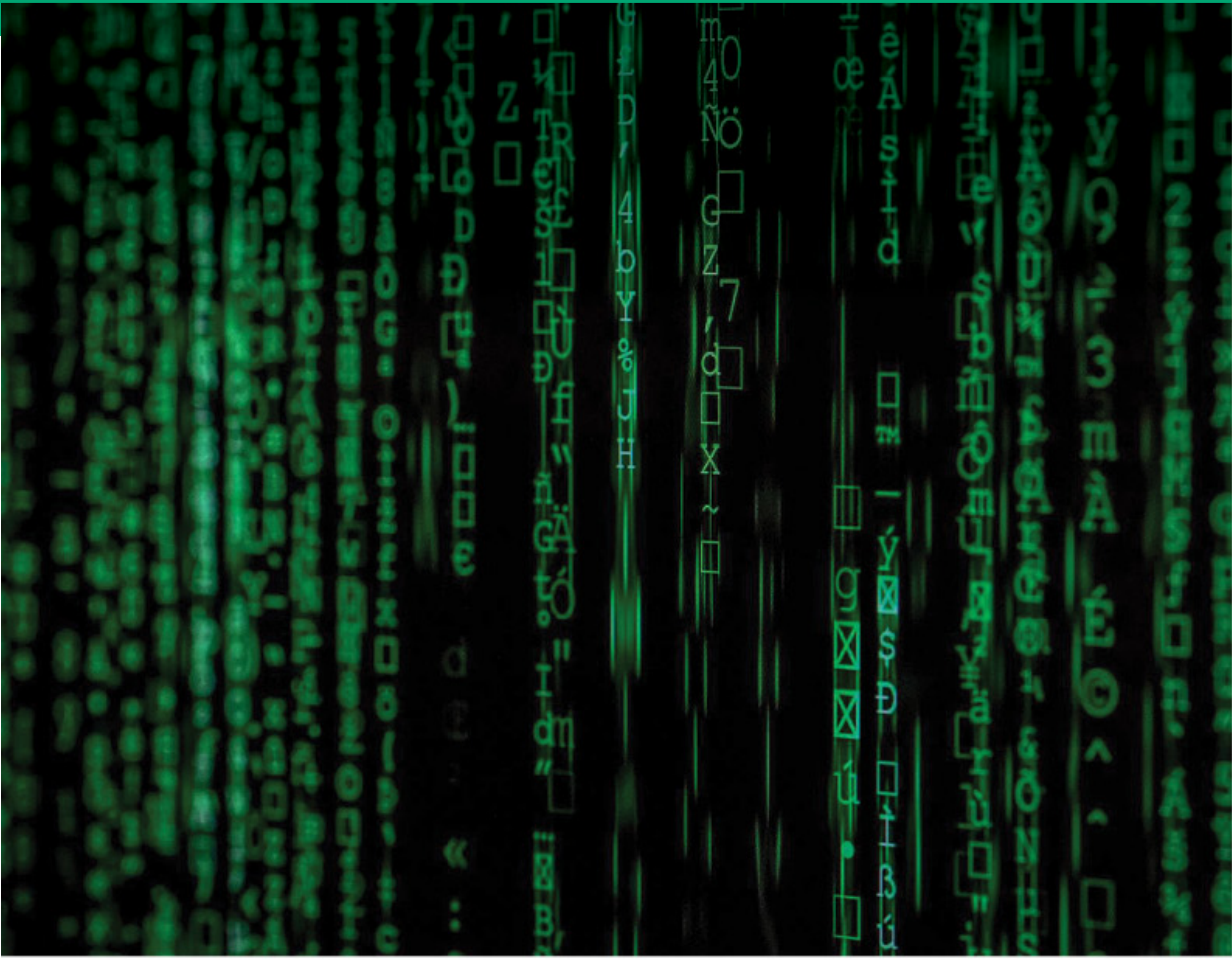


Course 7

Text-only version

Elements in data analysis

The interactive version of this lesson is available free of charge at https://elearning.fao.org/

This course offers guidance on the typical approaches and calculations used in forest data analyses and related topics.

Who is this course for?

The course is targeted mainly for those who are involved with analyzing forest monitoring data, but can be taken by anyone with an interest in the subject. Specifically, this course targets:

1. Forest technicians responsible for implementing their country's NFIs
2. National forest monitoring teams
3. Students and researchers, as curriculum material in forestry schools and academic courses
4. Youth and new generations of foresters
5. Forest data analysis practitioners

Course structure

There are four lessons in this course.

Lesson 1: Introduction to data analysis

This lesson introduces the issues that are relevant to typical data analysis after data collection and cleansing, but also need to be considered during the entire inventory planning and implementation process.

Lesson 2: Estimation

This lesson offers an overview of the process that generates results (or estimates) from the sample data. Remember, however, that this lesson gives only very basic insights – it does not cover the topic of statistical estimation exhaustively. If you are an expert who is interested in the subject at a deeper level, or deal with NFI data analyses regularly, we recommend that you supplement this lesson *with* textbooks, and/or discuss your approaches with experienced forest inventory statisticians.

Lesson 3: Statistical models in forest monitoring

This lesson provides insights into the uses of statistical models and elaborates on issues that need tobe considered when using them.

Lesson 4: Error in forest monitoring

This lesson elaborates on the various random errors as they occur along the NFI process. It also describes error propagation—how the different error sources propagate to the total error of the final result.

Lesson 5: Typical product from data analyses

This lesson discusses typical products from data analyses in forest monitoring and elaborates on the major products generated from NFI data analysis.

About the series

This course is the seventh in a series of eight self-paced courses covering various aspects of an NFI. Here's a look at the complete series.

Course	You will learn about
Course 1: Why a national forest inventory?	Goals and purpose of an NFI, and how NFIs inform policy- and decision-making in the forest sector.
Course 2: Preparing for a national forest inventory	The planning and work required to set up an efficient NFI or a National Forest Monitoring System (NFMS).
Course 3: Introduction to sampling	General aspects of sampling in forest inventories.
Course 4: Introduction to fieldwork	Considerations for fieldwork, plot-level variables and tree-level measurements.
Course 5: Data management in a national forest inventory	Information gathering and data management for NFIs.
Course 6: Quality assurance and quality control in a national forest inventory	QA and QC procedures in forest inventory data collection and management.
☞ **Course 7: Elements in data analysis**	**(You are currently studying this course)**
Course 8: National forest inventory results: Reporting and dissemination	NFI reporting and the importance of reporting in the context of REDD+ actions.

Lesson 1: introduction to data analysis

Lesson introduction

In this lesson, you will learn about topics that are relevant to typical data analysis—not only after data collection and cleansing—but those that need to be considered during the entire inventory planning and implementation process.

Learning objectives

At the end of this lesson, you will be ale to:

1. Describe the importance of data analysis in various phases of a forest inventory
2. Explain the general principles of data analysis.
3. Describe data cleansing and the considerations associated with it.

Data analysis in the various phases of a forest inventory

Although data analysis is relevant through the various phases of a forest inventory, the **actual analysis of the data** takes place **between data collection and reporting—**that is, once the data has been recorded, organized and cleaned. The final output of the data analysis is intended to satisfy the questions that were raised in the Information Needs Assessment (INA).

However, data analysis considerations are relevant to the entire inventory process because one of the overarching goals of every forest inventory is to generate a relevant and reliable database as input for the analyses. In the end, **the quality of the data co-determines the quality of the outcomes**.

Let us now look at the role of data analysis during planning and data collection.

Considerations during the planning phase

- ➲ It is necessary to ensure that all variables (that are required to produce the targeted output) are part of the inventory protocol.
- ➲ Unless it is anticipated that the data might be for future use for potentially emerging issues, it is recommended to avoid recording variables that are not needed in the analyses.
- ➲ The necessary variables need to be observed and recorded such that precision requirements can be met and results generated for the target reference units.

- The sampling design and plot design need to be defined in ways that ensure that estimators are available for statistical analysis.
- Quality assurance protocols for data are necessary. These include:
    - the organization of appropriate training measures for the field teams both before and during data collection;
    - the clear and transparent definition of data quality standards; and
    - appropriate control mechanisms.

General principles of data analysis

Data analyses in forest inventory projects follow the same principles as the whole inventory process - they need to be **methodologically well-founded**, **consistent**, **complete** and **transparently documented**.

All the steps of analysis need to be justifiable and in line with the inventory design (in terms of sampling design, plot design and models used).

Complying with information needs

Data analyses need to address all information needs that had been formulated prior to data collection and, to the greatest extent possible, issues that might have emerged in the meantime.

Analyses for future design optimization

Data analyses may also extend to methodological and practical issues that support the efficient planning of follow-up inventories. These could include longitudinal (i.e. time-following) studies, evaluating the time consumption in different inventory steps, and/or the evaluation of sampling and plot design with the goal of identifying potential optimizations in the implementation of follow-up inventories.

Double checking all analyses

When analyzing the data for the targeted outputs, it is important to double-check all results for correctness, including intermediate results.

The corresponding basic principle may be formulated according to Sutherland (1996) “Never believe your results”, which means that rather than believing, you need to be sure about and fully understand the results and the assumptions underlying them.

Any doubts–however minor–need to be followed-up, and this holds true for results that seem suspicious,as well as those that appear fully plausible and credible.

Documentation

Each analysis step needs to be documented properly in the inventory report, usually as an extra volume.

In the ideal case, the documentation needs to contain all estimators, step-by-step calculations, description of the models, conversion factors and indicator systems used, and needs to address challenges in analysis.

To conclude, remember that the documentation will be aligned with the initial methodology defined together with the sampling protocol.

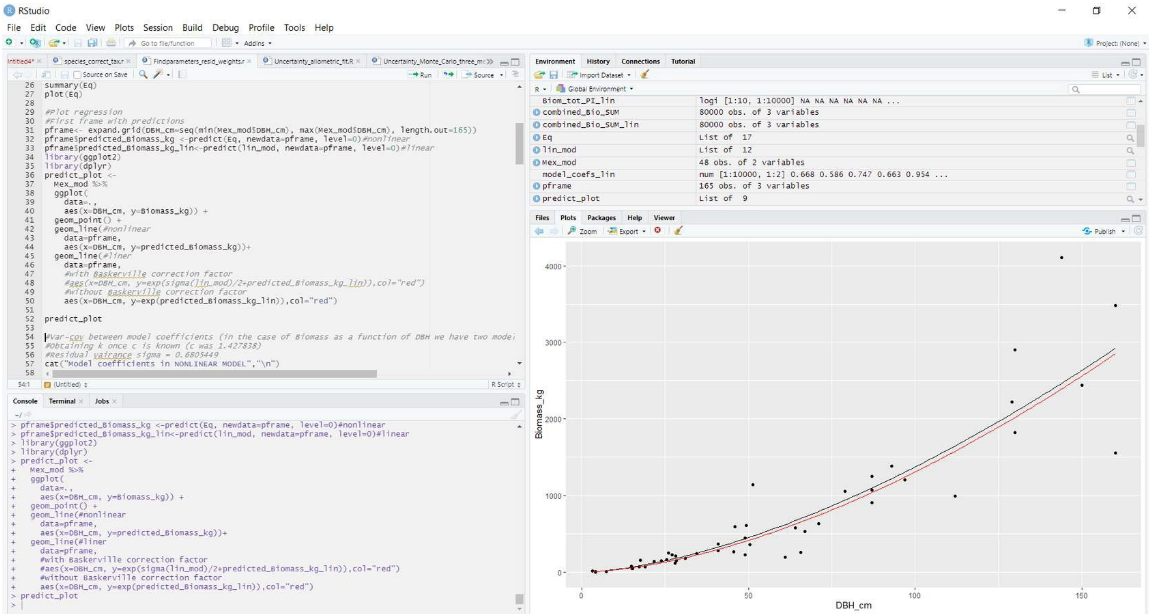
Data cleansing

Data analyses can only begin if the data is consistent and clean, and if all identifiable errors and inconsistencies are eliminated. Eventually, it is data quality that determines the final quality of the outputs.

However, sometimes data errors cannot be identified in the common cleansing process and become apparent only when the results don't look plausible. It is then necessary to revisit the data cleansing process to identify potential errors. It is useful to remember here that the lack of plausibility is not always an error, and that unexpected variabilities can always occur!

Software considerations for data analysis

The last decade has seen a rapid development of software solutions—both in general and, more specifically, for NFIs. While earlier forest inventories were evaluated by tailor-made programs, whether in a programming language or in a statistical software package, currently the trend is to develop R scripts and use already existing R packages (libraries of code specialized in conducting particular tasks) to the maximum extent possible.



```
summary(Eq)
plot(Eq)

#Plot regression
#First frame with predictions
pframe<- expand.grid(DBH_cm=seq(min(Mex_mod$DBH_cm), max(Mex_mod$DBH_cm), length.out=165))
pframe$predicted_Biomass_kg <-predict(Eq, newdata=pframe, level=0)#nonlinear
pframe$predicted_Biomass_kg_lin<-predict(lin_mod, newdata=pframe, level=0)#linear
library(ggplot2)
library(dplyr)
predict_plot <-
  Mex_mod %>%
  ggplot(
    data=.,
    aes(x=DBH_cm, y=Biomass_kg)) +
  geom_point() +
  geom_line(#nonlinear
    data=pframe,
    aes(x=DBH_cm, y=predicted_Biomass_kg))+
  geom_line(#liner
    data=pframe,
    #with Baskerville correction factor
    #aes(x=DBH_cm, y=exp(sigma(lin_mod)/2+predicted_Biomass_kg_lin)),col="red")
    #without Baskerville correction factor
    aes(x=DBH_cm, y=exp(predicted_Biomass_kg_lin)),col="red")

predict_plot

#Var-cov between model coefficients (in the case of Biomass as a function of DBH we have two mode
#Obtaining k once c is known (c was 1.427838)
#Residual vairance sigma = 0.6805449
cat("Model coefficients in NONLINEAR MODEL","\n")
```

```
~/
> pframe$predicted_Biomass_kg <-predict(Eq, newdata=pframe, level=0)#nonlinear
> pframe$predicted_Biomass_kg_lin<-predict(lin_mod, newdata=pframe, level=0)#linear
> library(ggplot2)
> library(dplyr)
> predict_plot <-
+   Mex_mod %>%
+   ggplot(
+     data=.,
+     aes(x=DBH_cm, y=Biomass_kg)) +
+   geom_point() +
+   geom_line(#nonlinear
+     data=pframe,
+     aes(x=DBH_cm, y=predicted_Biomass_kg))+
+   geom_line(#liner
+     data=pframe,
+     #with Baskerville correction factor
+     #aes(x=DBH_cm, y=exp(sigma(lin_mod)/2+predicted_Biomass_kg_lin)),col="red")
+     #without Baskerville correction factor
+     aes(x=DBH_cm, y=exp(predicted_Biomass_kg_lin)),col="red")
> predict_plot
> 
```

```
Biom_tot_PI_lin        logi [1:10, 1:10000] NA NA NA NA NA NA ...
combined_Bio_SUM       80000 obs. of 3 variables
combined_Bio_SUM_lin   80000 obs. of 3 variables
Eq                     List of  17
lin_mod                List of  12
Mex_mod                48 obs. of 2 variables
model_coefs_lin        num [1:10000, 1:2] 0.668 0.586 0.747 0.663 0.954 ...
pframe                 165 obs. of 3 variables
predict_plot           List of  9
```

Example of R script generated using RStudio, a development environment for R, a programming language for statistical computing and graphics

For smaller inventories (involving smaller data sets) the analyses may also be implemented by spreadsheet calculations (such as in Microsoft Excel), R scripts and specific software such as those developed by FAO. Skilled programmers are needed for these tasks, who can work collaboratively with inventory experts, who define the targeted outputs.

Single analysis steps or whole workflows can also be solved by implementing the statistical estimators in a suitable data model using modern Business Intelligence (BI) software or processed even directly in a database management system.

Did you know?

The role of BI software in forest data analysis

Most standard BI software does not cover the right estimators to be applied when producing results. Hence, if BI software is to be used, these estimators must be properly included as formulae. Until now, however, this has not been common in most countries. The German NFI currently calculates all estimates using SQL Server syntax directly on the database.

In any case, there is no one-size-fits-all software—but only specific procedures that can be implemented for every forest inventory that reflect the inventory design exactly. As each inventory requires new (or at least adapted) complex software, careful checks are required to ensure that the results are correct. It is a good idea to ask two different data analysts to do the same analyses independently, and then compare the results.

In some cases, these results may look good, plausible and consistent, and meet the expectations of the inventory experts, but will still be flawed. Remember that all results need to be double checked.

Summary

Before we conclude, here are the key learning points of this lesson.

- While data analysis considerations are relevant to the entire inventory process, the actual analysis of the data takes place after data collection and is a prerequisite for reporting.
- Data analyses in forest inventory projects follow the same principles as the whole inventory process—they need to be methodologically well-founded, consistent, complete and transparently documented.
- Sometimes data errors cannot be identified in the common cleansing process and become apparent only when the results don't look plausible. This entails going back to the data cleansing process to identify potential errors.
- While earlier forest inventories were evaluated by tailor made programs, the trend today (2023) is to develop R scripts and use already existing R packages (libraries of code specialized in conducting particular tasks) to the maximum extent possible.
- There is no one-size-fits-all software, but only specific procedures that can be implemented for every forest inventory that reflect the inventory design exactly.

Lesson 2: Estimation

Lesson introduction

Field data collection (and some remote sensing-based analysis) depends heavily on statistical sampling. Estimation is the process that generates results (or estimates) from sample data. As such, estimation is fundamental in data analyses.

Depending on the sampling design and the plot design used, estimation can be simple and also very complex. This lesson offers only an introduction to sample-based estimation—if you are interested to know about it in depth, please consult textbooks or discuss with forest inventory technicians.

More details on this subject are available in **Course 3: Introduction to sampling**.

Learning objectives

At the end of this lesson, you will be ale to:

1. Describe the role of estimates in NFI data analyses.
2. List some general principles of statistical estimation.
3. Explain point and interval estimates.
4. Discuss the role of auxiliary data in forest inventory estimation.

General observations on estimates

NFIs make use of various data sources—from sample-based field monitoring (that is at the core of forest monitoring), to the ever-advancing remote sensing technology.

All products of NFI data analyses are, therefore, estimates—and stem from either:

1. observations of the target variables on the field sample plots (the so- called design-based estimates); or
2. supported by auxiliary data (commonly from GIS or remote sensing; this is then called model-assisted estimates; see Lesson 3 of this course for a complete explanation); or
3. entirely based on models (model-based estimates).

Note

Remember that **all results from sampling studies are estimates**, be it means or variances, or confidence intervals, regressions and correlations. It is therefore better to use clear terminology For example, rather than concluding *The NFI analysis showed that the forest cover in the country is 43.5 percent* it is more accurate to say, *The NFI estimates the forest cover in the country to be 43.5 percent*.

Estimates are random variables. This indicates that they are contrary to fixed parametric values in the population, which are constants. All estimates follow a distribution with a mean value of that distribution (the expected value for which the point estimate is the sample-based approximation out of that distribution) and a standard deviation that describes the variability in that distribution of estimated mean values (estimated by the interval estimate).

In the hypothetical case of repeating an NFI with exactly the same design but different randomization, one would produce different numerical results of the estimations, for both the point and interval estimates.

Estimates serve to learn about the population so that the major interest is not so much in the sample data itself, but in the inference to the true population value that the sample offers. The smaller the standard error is, the closer one can assume that the estimates are—on average—to the true parametric value. In that case, one will perceive the estimate as reliable.

Because we infer from the (estimated) sample to the (true) population value, the terms design- based estimates, model-assisted estimates and model-based estimates are frequently also named design-based inference, model-assisted inference and model-based inference.

General principles in statistical estimation

When doing estimation on statistical grounds (as opposed subjective assessments) the calculations need to strictly correspond to the sampling and plot design used—which means that different experts will tend to come up with the same result. The formulas used for estimation are the estimators. There are cases in which more than one alternative estimator is available, but usually there is no choice.

A major characteristic of inventory sampling designs in NFIs is that the estimators used need to be unbiased (or at least approximately unbiased as in the case of the ratio estimator). That means that the expected value of our sampling design should be identical with the searched population value. Expected value is the value that results as a mean in the (hypothetical) case that we repeated our sampling study many times—under the same design but with different randomization.

The designs and estimators that we have presented in other courses of this series (most notably **Course 3: Introduction to sampling**) are almost all unbiased, with three notable exceptions:

1. The ratio estimator is approximately unbiased under some circumstances;
2. There is no unbiased estimator for the error variance in systematic sampling, while there are unbiased estimators for the mean; and
3. There is no unbiased approach to analyze sample plots that include the k nearest trees (not covered in this course but explained in specialized textbooks)—a plot design often used in ecological studies but not so in forest inventory.

Should one of these design elements be used in an inventory, the issues that arise from using these biased estimators should be transparently addressed. This is particularly important for systematic sampling because it is the most frequently used sampling design in NFIs. There, however, we are on the safe side when applying the estimator framework of simple random sampling (SRS) for the estimation of the error variance, because we know that such an approach yields a conservative estimate and always overestimates the true error variance (even though to an unknown extent).

Point and interval estimates: generating estimates on location and dispersion

When analyzing NFI data, we are mainly interested in point and interval estimates. In this section, we will focus on these.

The **point estimate informs about the point on the number axis where the estimate lies** (e.g. for above ground biomass on an area basis, e.g. 200 Mg/ha), and the **interval estimate informs about the estimated variability of this point estimate** (e.g. SE% = 5%). Using the terminology of descriptive statistics, we may also say: the **point estimate is a measure of location of the estimate**, while the **interval estimate is a measure of dispersion of the estimates**.

When referring to point estimates, it is often the mean value, but also the estimate of a regression coefficient (b1) or of a correlation coefficient (r), r, or of the population variance (s^2) that are point estimates.

Point estimates deliver the core information for data users. Mostly, non-experts focus their interpretation on the results on these measures of location of the estimates. It is important to reiterate that these point estimates are not the truth, but only estimates.

The fact that any point estimate will not be identical to the desired true population parameter, is not an expression of bias but an expression of the variability in sampling, or the sampling error. However, it remains unknown how far this particular estimate (derived from our one sampling study, or NFI) is, in numerical terms, from the true population parameter.

Distribution of point estimates

To allow such probabilistic inference on the true population value, one must know the distribution of the point estimates. For example, for mean values, it is known that they follow the t- distribution for small samples and the normal distribution for larger samples; where large, in statistics, is usually defined as $n \geq 30$ (for large n, the t-distribution approximates the normal distribution).

When we know that the estimated means vary according to the normal distribution around the expected value (the true mean in case of an unbiased estimator), one can use the probability densities under this normal distribution to estimate the probability that such true value is within a defined interval around the estimated mean that comes from our sampling study (NFI).

For the estimated mean, such an interval is symmetric around the estimated mean and has a standard deviation that corresponds to the standard error. Here, we need to be aware that the standard error is calculated from the estimated population variance, s^2/y.

This estimated population variance is an estimate by itself (the sample estimate of the population variance): it can be considered a point estimate (the value of the estimated population variance) that carries along an interval estimate (the variability of the population variances). One could also estimate the confidence intervals for the estimated population variance and for estimated variances, the confidence interval will be asymmetric, as estimated variances follow the (asymmetric) F distribution.

$$S_y^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n - 1}$$

Interval estimates are also relevant in NFIs because they are a measure of precision of estimation and, therefore, uncertainty, which in turn is commonly interpreted as a measure of reliability of the estimates.

Bootstrapping and jack-knifing for interval estimates in complex design

In some inventory designs where an estimator is overly complex, re-sampling is indicated. Re- sampling is a simulation in which sample data is further exploited to simulate many samples (sub- samples) and inferences are made from the corresponding results to the statistics of the whole sampling study.

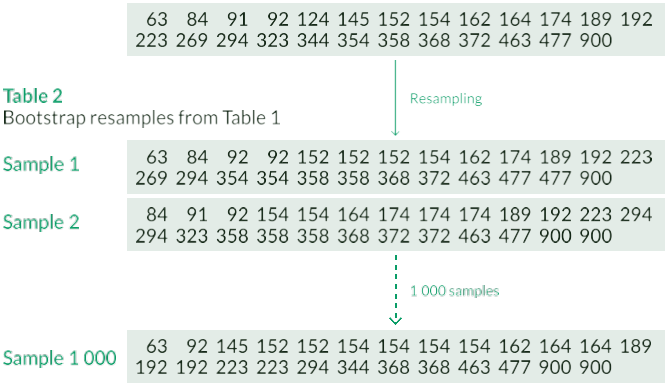
Bootstrapping is the most commonly used technique when confidence intervals are determined in complex designs where unbiased, direct estimators are not available. It goes back to 1979, when Bradley Efron coined and introduced this as a modification of the jack-knifing technique that had long before been introduced by Quenouille (1956), the so-called 'leave one out re- sampling'.

These techniques are based on simulations and not on assumptions about the parameters of a specific distribution of the estimates; they are, therefore, also called non-parametric techniques.

The idea of bootstrapping follows the above addressed idea of re-sampling from the sample of size n that had been taken. One can either do this 'with or without replacement'.

Bootstrapping 'without replacement' means that a large number of times a sub-sample of size n is taken out of the original sample of size, and those which have been selected are not placed back into the pool to be selected from (in other words, they can only be selected once).

Bootstrapping 'with replacement', however, means that the same observation may be selected several times into the bootstrap sample. Here the terminology large means that this sampling is repeated several thousands of times, say 10 000 times.



63 84 91 92 124 145 152 154 162 164 174 189 192
223 269 294 323 344 354 358 368 372 463 477 900

Resampling

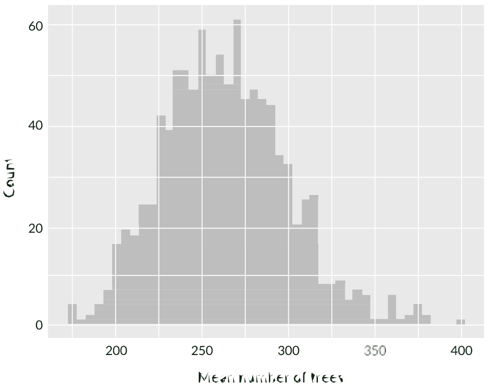
Table 2
Bootstrap resamples from Table 1

Sample 1
63 84 92 92 152 152 152 154 162 174 189 192 223
269 294 354 354 358 358 368 372 463 477 477 900

Sample 2
84 91 92 154 154 164 174 174 174 189 192 223 294
294 323 358 358 358 368 372 372 463 477 900 900

1 000 samples

Sample 1 000
63 92 145 152 152 154 154 154 154 162 164 164 189
192 192 223 223 294 344 368 368 463 477 900 900



For each one of these bootstrap samples the target statistic (e.g. the mean value) is calculated so that at the end there are, say, 10 000 bootstrapped mean values and these can be graphed as a distribution. This distribution has a mean value (which corresponds, of course, to the mean value of the original sample of size n) and a particular probability density distribution from which, for any probability, confidence intervals can be derived.

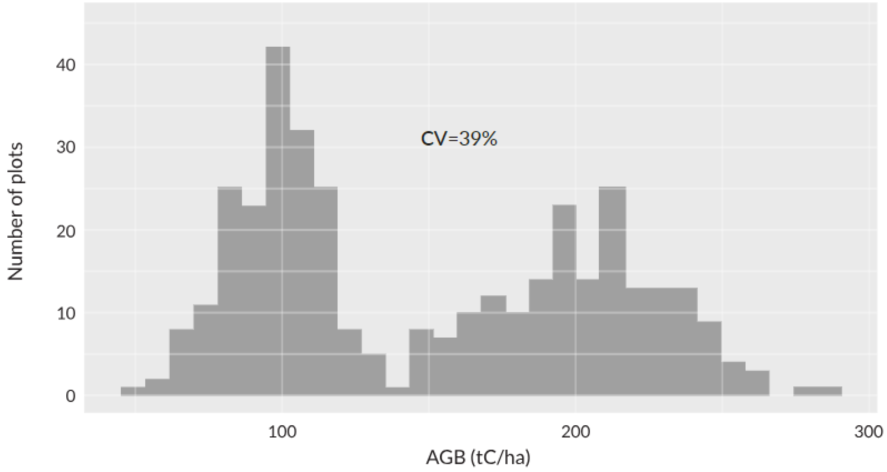
If, for example, the bounds of a 95 percent confidence interval shall be calculated, one searches the cut points where 2.5 percent of the bootstrapped means are truncated at the upper end and 2.5 percent at the lower end. The two 'cut points' are then taken as the bootstrapped upper and lower bounds of the 95 percent confidence interval.

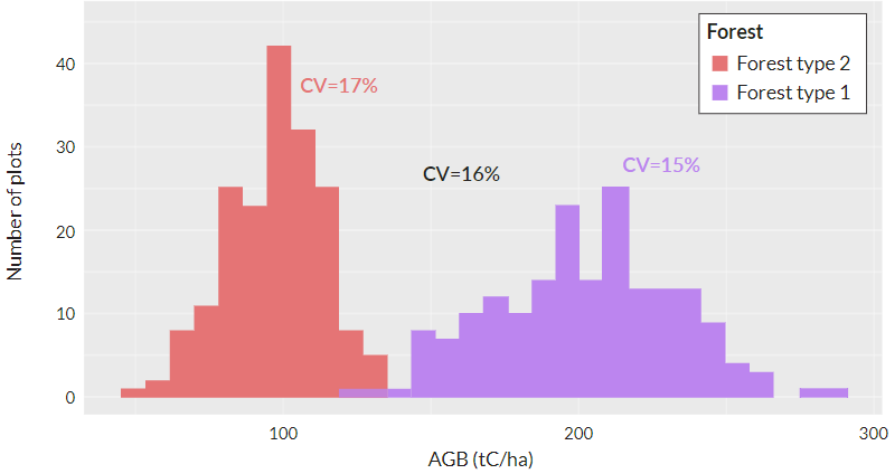
Auxiliary data in forest inventory estimation

Auxiliary data comes from the observation of auxiliary variables in some inventory designs to improve precision of estimation of target variables. Auxiliary variables are sometimes called co- variables or ancillary variables (Latin: *auxilium*=help, *ancilla*=servant). We have seen auxiliary variables so far in the ratio and regression estimators where we seized a high correlation between target and auxiliary variables to extract and integrate information from the auxiliary variable into the estimation of the target variable.

In a more general sense, when making a distinction only between target and ancillary variables, we may look at various other variables as auxiliary (supporting the analyses). This observation refers, for example, to all the topographic variables which serve for breaking down the results of our target variables into classes—for example, growing stock per elevation class or biomass per slope class. Here, such auxiliary variables define criteria for a post-stratification, allowing specific analyses and evaluating relationships between target and auxiliary variables.

Let's now look at two figures on improving error estimates through post-stratification with the help of auxiliary data.



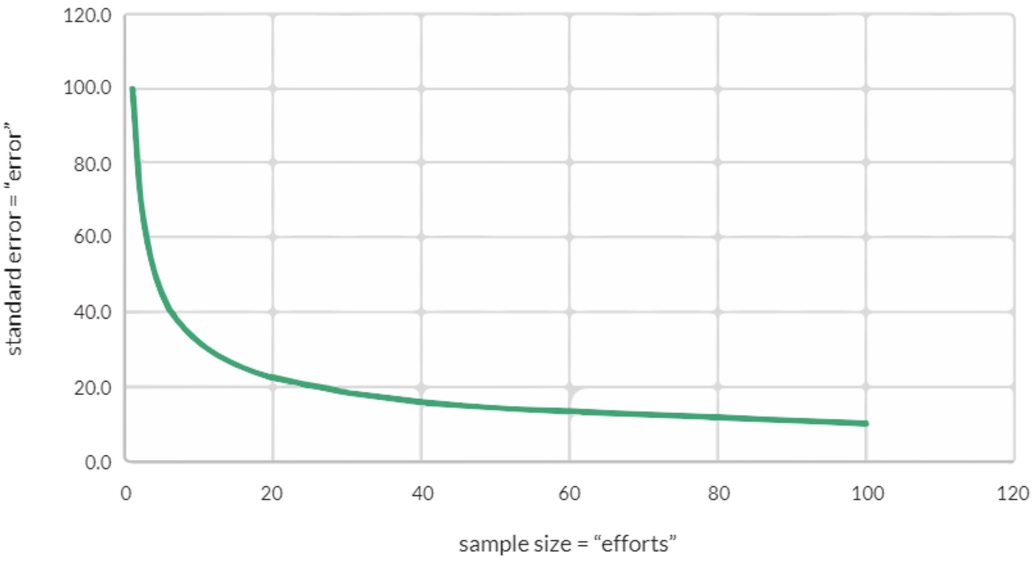


Estimates for different units of reference/sub- populations

The basic unit of reference for estimation from NFIs is the whole country. Sample size is usually defined such that the precision of estimation at country scale meets the expectations. Sample size is usually quite large and estimates will be precise, having low standard errors. Depending on the sample size, relative standard errors are in some cases less than 1 percent smaller. But such high precision occurs only when we have the whole country as unit of reference = reporting area.

Often, estimates for sub-national units—such as provinces, states or territories—are also of interest. Of course, when using the common systematic grid of sample points, sample size for these smaller units of reference will be smaller and standard errors for the corresponding estimates higher. **The smaller the unit of reference and the smaller the sample size, the less precise the estimates.** For example, in the German NFI, while the forest area for the whole country carries a relative standard error of SE%=0.7 percent with a sample size of about n=21 000 clusters, it is SE%=1.6 percent for the Federal State of Bavaria (n=2 815) and SE%=25.8 percent for the combined Federal States of Hamburg and Bremen with only n=15.

It is instructive here to consider the simple relationship between sample size and standard error in SRS, depicted in the figure below; where the basic shape of the relationship holds for all sampling designs: the marginal gain in precision for large sample sizes is small; but small changes in sample size have a much larger impact on the standard error for smaller sample sizes!



When remote sensing data are available, estimates for smaller areas may be generated with much higher precision using the so-called small area estimation, where the remote sensing data is used as support to generate estimates for (almost) arbitrarily small units of reference.

Although one can see a definition of small-area estimation in Lesson 5 of this course, we can here provide a simple example: Let's say we have a large forest area of 1 000 hectares, and we want to estimate the average tree density (number of trees per acre) in a small area of just 10 hectares within the forest. However, we only have data on 5 sample plots within the small area, which is not enough to get a precise estimate of the tree density.

Using small area estimation with remote sensing data, we can use the information from the larger forest area to generate a more precise estimate of the tree density in the small area. We can use remote sensing data, such as satellite imagery and Light Detection and Ranging (LiDAR), to derive additional information about tree density and other forest characteristics in the larger area.

We can then use this remote sensing data as support to generate an estimate for the small area of interest. For example, the remote sensing data might suggest that the average tree density in the larger forest area is 400 trees per acre. Using small area estimation, we can adjust this estimate based on the data from the small area to obtain a more accurate estimate for the small area. For example, the model might estimate that the average tree density in the small area is 450 trees per acre, with a smaller margin of error than we could obtain using the sample data alone.

Summary

Before we conclude, here are the key learning points of this lesson.

- All results from sampling studies are estimates—be it means or variances, or confidence intervals, regressions and correlations.
- Estimates serve to learn about the population: the major interest is not so much in the sample data itself, but in using these sample data for inference to the true population value.
- When doing estimation on statistical grounds, the calculations need to strictly correspond to the sampling and plot design used, which means that different experts should come to the same result; and you are not free to do the estimation with any arbitrary approach.
- In analyzing NFI samples we wish to use design-unbiased estimators, if possible. For some designs, such estimators do not exist, including the estimators for the error variance in systematic sampling.
- The point estimate is a measure of location, while the interval estimate is a measure of dispersion of a distribution. This terminology holds both for variables and or estimates.
- In some inventory designs where an estimator is overly complex, a simulation in which sample

data is further exploited to simulate many samples, that is known as “re-sampling”, may be indicated.

- Under some conditions, it is efficient to also observe auxiliary variables together with the target variables in order to improve precision of estimation of target variables.

Lesson 3: Statistical models in forest monitoring

Lesson introduction

This lesson gives an overview and insight into the uses of statistical models in NFIs and elaborates on issues that need to be considered when using them.

Learning objectives

At the end of this lesson, you will be ale to:

1. Describe the role of statistical models in forest monitoring.
2. Identify the major characteristics of statistical models.
3. Demonstrate how to build a biomass model.
4. Identify a suitable statistical model for a particular situation.
5. Explain some common reporting issues in statistical models.

What is a statistical model?

Statistical models aim to establish a quantitative relationship between a predicted variable and one or more predictor variables. In other words, by having measured/observed the predictor variable(s), the model is used to generate a value for the predicted variable.

Essentially, a statistical model predicts a value for a target variable.

Within the context of forest inventories, statistical models are used when:

1. a target variable cannot be measured in a forest inventory (e.g. biomass cannot be measured by weighing. If you fell all trees in the forest to weigh them, you don't have a forest anymore! However, it can be modelled from the measurement of other (so-called) predictor variables); or
2. a measurement is time-consuming/costly (e.g. height is time-consuming to measure, and it is often measured only for a sub-set of trees and then model-predicted for the others as a function of dbh).

Statistical models describe the relationship between data/observations of two variables, frequently observed from the same objects (such as trees). Statistical models do not serve to establish a cause-effect relationship, something which would be the goal of so-called process models: these aim to explicitly include the causes behind biological processes in order to predict specific outcomes and different situations. While such a relationship may actually exist also for

statistical models, it is not the subject of the modelling exercise, nor can a statistical model be interpreted accordingly.

Examples of statistical models used in forest monitoring

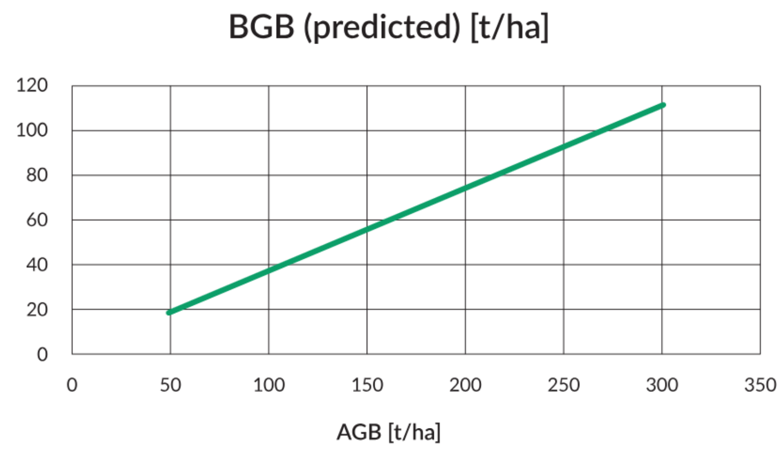
There are different types of models—of varying complexity—that are used in forest monitoring, and in some cases, it is difficult to notice that a model has been used.

Example 1: Determining basal area from diameter measurement

A very basic model, for example, is the assumption that the tree cross section at breast height will always be a circle: tree basal area is calculated along the simple circular model. Of course, realistically, no tree has a mathematically perfect circle nor whose cross section at breast height is a mathematically perfect circle, but the approximation has served reasonably well so far, and there are no better solutions.

Example 2: Conversion factors as statistical models

Other basic models are simple factors as they are frequently used to convert—for example— stem biomass to total biomass or above ground biomass (AGB) to below ground biomass (BGB). Form factors—a numerical summary indicator of a trunk's shape—are also simple models, used to determine individual tree volume. We may say that such conversion factors are just 'reduced' simple linear regression models, where the intercept is zero and the factor itself is the slope coefficient. An example is provided in the figure below.



A simple conversion factor may be considered a basic model that allows for predicting a variable from a predictor variable. The IPCC (2006, Table 4.4), for example, recommends the conversion factor of 0.37 for tropical rain forests to determine BGB from AGB. That conversion factor translates into a simple linear regression model with an intercept coefficient of zero and a slope coefficient of 0.37: BGB=0.37*AGB.

Note that in this case, the IPCC does give a source for this conversion factor but does not publish a standard error or other measures of uncertainty. Instead, it specifies, for some forest types, the range of values of the conversion factor.

Of course, considerable uncertainty arises with such simple conversion factors but in many cases, such as root biomass, it would hardly be convenient to take your own samples.

Example 3: Common regression models

Common regression models such as those used in forest inventories include:

1. predict height from *dbh* (height curves) and
2. volume biomass/carbon from dbh or from dbh and height or from dbh, height and an upper diameter (volume functions, biomass functions or carbon functions respectively).

Of course, other models are in use for specific purposes, for example, in stump inventories, predicting dbh from stump diameter (and stump height); or for buttress trees, predicting dbh from the diameter above the buttress roots.

For biomass functions, frequently the term **allometric biomass functions** or **allometric biomass**

models is used. The term allometric derives from ancient Greek and Latin, where "ἄλλος" (allos) means other and metric means measure. Allometric, therefore, means determining biomass from other variables. Following this original meaning, the term is essentially redundant when specifying a model—because it simply describes what all models do: produce the value of a variable from measurements of values of other variables.

The model types listed so far are commonly generated from research studies prior to the inventory. For specific forest inventories, biomass models are usually taken from the literature after checking their suitability for the specific inventory (discussed later in this lesson).

However, there are also models that are generated with and from the inventory data itself: a typical example is a height curve used to predict tree heights. Height measurements are time consuming and therefore expensive, so that heights are measured only on a (well-defined) subset of sample trees. A model is then built from these measurements, that is used for predicting the heights of the unmeasured trees. In this case, the height measurements will exhibit a greater variability than the predicted heights, because these predicted heights represent mean values for a given dbh class.

Example 4: Models calculated from target and auxiliary variables

Another example where models are built during the inventory implementation is when auxiliary variables are observed for use with the ratio or regression estimators, or in double sampling with the ratio or regression estimator: then, a model (either a simple ratio or a regression) is calculated from the sample plots where both variables, target and auxiliary variables, have been recorded. The model supports estimation and allows—when there is sufficient correlation between the target and auxiliary variables—for a more precise estimation of the target variable. That is, the model assists the estimation process, and therefore the estimator used is also called a model-assisted estimator.

While the estimation process in this case is supported by the model, the unbiasedness of the estimation still comes from the randomization of the sample selection, which ensures that the sample is representative of the population. However, note that the ratio estimator is only unbiased when the simple model holds. This means that the model should accurately capture the relationship between the target and auxiliary variables, and that there are no other factors that may affect the estimation process. If the simple model does not hold, the ratio estimator may be biased and the estimation results may not be accurate.

Estimation and inference may also be completely based on models—and the corresponding approaches are named model-based estimation or model-based inference. In this case, the unbiasedness of the inference is entirely dependent on the validity of the model.

A typical example here is modelling the relationship between field-observed biomass and remote sensing data. When such a model has been established, it is possible to predict forest biomass for each pixel. With such prediction, one is in the position not only to produce an estimate of biomass for the whole inventory region (by summing up the biomass predictions per pixel), but also to generate a biomass map

Note

From the relatively long list of models used in forest monitoring it becomes clear that they play a crucial role here: this is mainly because forests are complex objects to monitor, and diverse variables of interest cannot directly be observed. In order to make a monitoring system operational, it is essential that we work with model-predicted values. It is, however, important to clearly distinguish between values that are being recorded by immediate observations/measurements and those that are predicted from models.

The two major points we need to remember here are that:

- immediate observations carry only one error source: the measurement error, while model predictions carry measurement errors (of the required predictor variables) and model errors; and
- model predictions present lower variability than immediate observations: for the same set of predictor variable values the same predicted value is always generated, while in reality, for example, different trees with the same dbh may have quite different biomasses.

Major characteristics of statistical models

Let us now look at the main characteristics of statistical models.

Models predict mean values, not true values

It is important to understand that given a set of predictor variables, models do not yield the true value of a predicted variable. Predictions need to be understood as mean values. By using a model to predict, for example, tree height from dbh measurements, we assign mean heights to the sample trees: all trees with a specific dbh, say 40 cm, will have the same predicted height.

This, of course, does not correspond to the true situation where trees with the same dbh vary in height: that means that the variance of predicted heights of sample trees is always smaller than the tree heights would they have been measured.

Statistical models are based on sample observations— and the model coefficients are themselves estimates

All models are estimates by themselves—they are based on observations on a set (sample) of trees and they do not represent the parametric (true) model, but only an approximation (estimate) of it. When different field teams would take a random sample of 100 trees—each from the same inventory region but with different randomizations—to calculate a biomass model, each using the same mathematical model, they will all come up with different model coefficients.

As with the common sampling for mean values, the precision of the estimate will usually become better when the sample size is larger: for each predicted individual value or mean value, confidence intervals can be determined. To do so, of course, measures of variability of the model need to be known.

Statistical models are characterized by statistical measures

As with the sample-based estimation of mean values, it is important also for the estimated models to accompany the point estimates with interval estimates. The point estimates for regression models are the estimated regression coefficients. For each estimated regression coefficient, a standard error can be estimated, and the smaller it is, the more precise the estimate.

An important characteristic is the significance of the regression—if, for the example of a simple linear regression, the slope coefficient is not statistically significantly different from zero, we say that the regression is not significant. This means that the regression line is parallel to the x-axis and, consequently, that the predicted value for the target variable, y, will be the same for all predictor variables, x. This predicted value is the mean value for y. So instead of calculating a regression that determines a specific mean value per class of x values, one can use the overall mean, as predicted value for all x classes.

Statistical models employ variable data points

Another important statistic that characterizes the precision of prediction from the regression model is the variability of data points used for model construction around the regression line. Just as the values of one single variable vary around the mean value, the data points in a regression model vary around the regression line—where the regression line represents a **shifting mean value** that takes on different values for different x-classes. We call this variability the residual variance. When this variance of the residuals is small, the data points are closely clustered around the regression line, and we may assume that the model predictions are quite precise.

Remember that it is not only important to know the regression coefficients so that one can calculate the predicted values, but also to know the variance statistics of the model to be able to evaluate its quality. The precision of estimation from a simple linear regression model, for example is highest around the mean value of the predictor variables and becomes lower towards the ends of the range of predictor values included; beyond this range, the model should not be used, and if so, the precision will be low.

Statistical models hold for "specific conditions"

The base data used to build a model co-defines the validity of the model for a specific inventory purpose. When we refer to base data, we mean factors such as:

Geographic region	In the best case, the base data come from the geographic region where the inventory takes place. If this is not the case, one should make sure that the model is suitable (see section on Identifying suitable statistical models in this lesson).
Tree species	Some models are specific for one species or for a species group; whether they would also apply for other species would need to be checked.
Range of input variables	Models should generally be used only for values of input variables that are covered by the base data. It may be a risk, for example, to use a biomass model that has been built for values of between 30 and 150 cm also for smaller trees below 30 cm—it may be that such extrapolation

produces unplausible values as the model function is essentially not defined outside the range of the base data.

How to build your own biomass model

Building your own biomass model is usually not a task in an NFI project. One may resort to models that have been published before. Sometimes, these models were built in the framework of academic research or technical reports and are difficult to find. But it is likely that models exist that are suitable for all situations.

FAO, together with CIRAD, offer a database for hosting biomass models in the GlobAllomeTree Initiative, which may serve as reference when searching or building biomass models. An in-depth manual on the development of allometric equations is a good consultation for those interested (Picard et al. 2012).

In this section, we briefly outline the steps that you need to follow in order to build your own biomass model. The process holds equally for any other statistical model. For example, when you wish to estimate the biomass of illegally removed trees from a stump inventory, you would apply normal biomass models, and to do so you need to predict dbh from stump diameter and you may wish to build your specific model here.

Note

There are a number of models published for different tree species that allow predicting dbh from stump diameter. This is either simple factors (e.g. Bones, 1960) or regression models, some of them also including the stump height (e.g. Pond and Froese, 2014).

Note that we do not find a definition of stump diameter or of stump height in these publications not any indications on how to measure them. However, stump diameter can be very irregular (as trees are often times buttressed at very low heights) and would require clear and unambiguous definitions.

This illustrates the relevance of unambiguous definitions, not only in forest monitoring, but also when referring to input variables for statistical models.

Step 1: Start with definitions

Following good practices of forest monitoring that recommend having clear and transparently documented terminology, definitions and measurement, you should start with definitions.Definitions extend to the population from where the sample trees are to be taken, including

- the geographic definition of the precedence area of sample trees;
- a species or species group definition; and
- a definition of the range of dimensions (usually range of dbh) for which the model shall hold.

Furthermore, biomass needs to be defined in terms of biomass compartments (stem, large branches, smaller twigs, below ground, leaves, etc.) are to be considered as well as minimum diameters.

Step 2: Determine the number of sample trees to be felled

The number of sample trees to be felled needs to be determined; this is usually done according to the resources available. Because felling and weighing trees is costly, the number of trees is usually quite limited, even though it is good to work with larger numbers of sample trees to make the models predict with higher precision.

It makes a big difference to sample a large tree or a small tree, as the felling, chopping and weighing of larger trees will be over-proportionally and much more expensive. This leads to a situation where most biomass studies have many smaller trees and only a few larger trees.

This is a sort of dilemma, because the variability in biomass of smaller trees is much smaller than the variability in larger trees, and the conclusion would be that we need larger trees in order to have a better foundation for a model where the variability is larger (and where frequently the largest portion of forest stand biomass is in the large trees).

Step 3: Select the sample trees

While the general shape of biomass models is quite well known and follows some physical laws, the goal is to try to determine the shape of the biomass function over the whole range of input values (usually a dbh range). This means that one should try to select sample trees for all diameter classes—and (theoretically) more sample trees where the variability of the target variable (biomass) is largest.

Reality check

The selection of sample trees is often dominated by practical considerations such as accessibility, cutting permits, and others. This is another example in the context of forest monitoring where theory comes to head with practice, in other words, where science meets real life.

Step 4: Measure the sample trees standing

Sample trees first need to be measured standing—such as when measuring the predictor variables in a default forest inventory. For example, tree height can hardly be determined after felling—because then what we measure is tree length. Also, dbh is best measured at the standing tree because after felling it is more difficult to determine breast height.

Step 5: Fell the sample trees

Once the sample trees have been measured while standing, they need to be felled carefully, because it is necessary to ensure that all relevant biomass parts can be attributed to that sample.

Thereafter, the relevant biomass compartments need to be separated and weighed.

Reality check

While measuring and felling the sample trees, it is common practice that measurement errors are not considered nor factored into the model when determining the uncertainty measures. The measurements both at the standing and the felled sample trees are simply taken as true values.

We know, however, that this is not the case and that measurement errors may play a role, particularly when the number of sample trees is low. In NFIs, we usually do have large sample sizes and a large number of sample trees recorded; then, we may assume that the random measurement errors have a relatively small weight as the large number of observations will keep the error variance low.

Step 6: Manage sample tree data

Sample tree data will need to be managed in a database where all compartment-wise results are stored and eventually summed up to the target value(s) for each particular sample tree. In the end, and as input for further analyses, a list is produced with per-tree data of target variable and predictor variables. Such a list is the input matrix for estimation of the model coefficients.

Step 7: Identify a mathematical model that fits the data set well

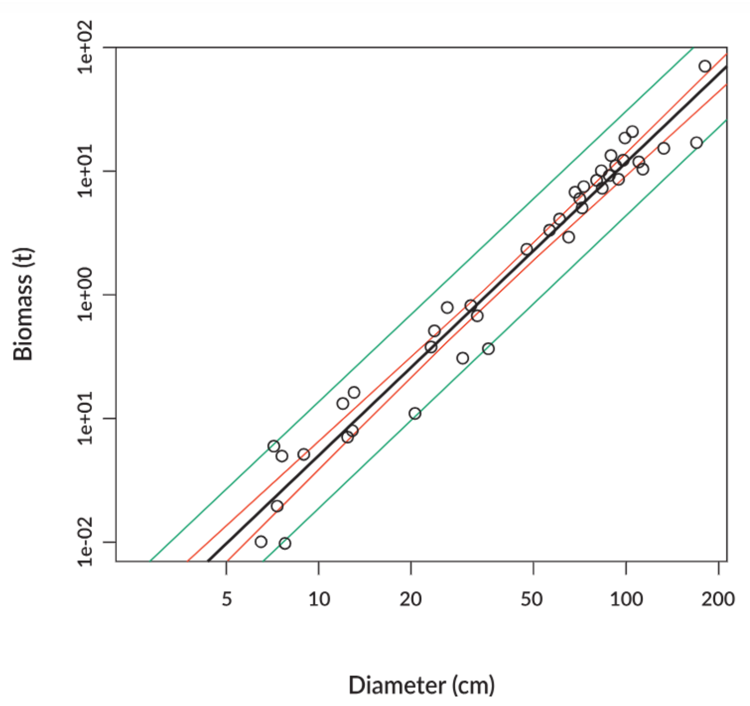
The next steps are applied statistics: a mathematical model needs to be identified that is able to fit the data set well. For biomass models, as for other models, typical mathematical models are known. It is common to compare the performance of different mathematical models and choose the one that allows the most precise predictions.

A common approach is to create a model using a random selection of 75 percent of the sample trees, and then evaluate the model's performance using the remaining 25 percent of the sample trees. Here, we consider the 25 percent check-trees as independently selected trees for model validation.

It is important here to distinguish between

1. the model uncertainty, which results from the analysis of the 75 percent of sample trees used for building the model; needless to say, the model fits quite well to this data set because this data set is the basis for the model; and
2. the prediction uncertainty, which is usually larger and refers to the prediction for trees that had not been used for model building. Consider here, that the sample trees used for building the model are a sample from the population of interest, and, of course, our sample will not be able to capture all variability present in the population but it is just an estimate.

Then, all trees recorded in an inventory belong to this set of trees that had not been used to build the model. Hence, prediction uncertainty is an important point in model development. Picard et al. (2012) provides insights into these issues.



Once you are happy with the model uncertainty and the prediction uncertainty, you may come back to the whole sample tree data set and estimate the final model coefficients from 100 percent of the sample trees.

Note

A number of statistical issues need to be considered when building the statistical model so that it makes sense to consult with a statistical modeler when building and choosing a specific model. The statistical issues include:

1. correlation between predictor variables (the so-called collinearity), and
2. a typical feature of biomass (and volume and carbon models): the variability of biomass that is small for smaller trees and large for larger trees (the so-called heteroscedasticity).

The latter affects the estimation of variances and confidence intervals for predictions of mean values and individual values—for example, for biomass models as a function of dbh only, the confidence intervals will be narrow for smaller trees and become wider with increasing dbh.

Step 8: Document the model

Once the final model has been decided, a complete and transparent reporting of the model is the last step, not only do the model and its coefficients need to be documented, but also the uncertainty characteristics, including coefficient of determination, standard error of the regression and prediction uncertainty, and other possible uncertainties.

One may also wish to make the original data set publicly available, because such data sets may be very helpful when combined with new data sets to generate more precise, more locally adapted or more generalizable models.

Identifying suitable statistical models

There are many statistical models available for forest monitoring. The IPCC, for example, offers a long list of different conversion factors and biomass functions (see, e.g. IPCC 2006 ***Guidelines Publications - IPCC-TFI*** (iges.or.jp) or their refined 2019 values ***Publications - IPCC-TFI (iges.or.jp)***).

In many cases, it is clear from the outset which model to use, because it has been used before in the same geographical or subject-matter context, or is known to perform well under the circumstances of the particular inventory project. In NFIs that extend over vast areas, and include many species and species groups, it may be adequate to apply different models depending on species group, and/or geographic region and/or site conditions.

A first step of model selection is to check which models had been in use before, the precedence of the sample tree data used for model building and to evaluate the uncertainty statistics of the models; usually, the more sample trees had been processed, the more accurate the model. Sometimes a decision needs to be made whether, for example, to use various species-specific biomass models or a general model for all species. In NFIs where the sample sizes are commonly large, the recommendation from recent research is to look at the number of sample trees used for model building rather than at the specific species for which it had been built. This means that a general model for all species based on a large number of sample trees is usually preferred over using various species-specific models built on small tree numbers each.

Quick tips!

Some rules of thumb have been identified in regard to biomass model development (McRoberts and Westfall, 2014): the number of sample trees should be at least 100, and the models should have a coefficient of determination larger than 0.95. Then, in NFIs, model errors in biomass are usually relatively small as compared to the standard error.

However, it is important to emphasize that the minor role of model errors, in NFIs with large sample sizes refers to random errors only, while potential systematic errors and biases will, of course, propagate accordingly as biases into the final estimates!

In situations where the choice of model is not clear from the outset, the task is to check the suitability and compare the performance of different models. That can only be done by a sufficiently large number of sample trees, and that is costly, particularly when dealing with biomass models because determining/measuring the biomass of sample trees is always costly.

The UNFCCC (2011) gives a basic and hands-on guidance on how to do such suitability test for models on forest aboveground biomass. A comprehensive and science-based description of such analysis of the suitability of models in general can be found in Pérez-Cruzado et al. (2015).

Reporting issues in statistical models

There are essentially two types of reporting issues when talking about statistical models in forest monitoring:

Reporting the model by itself and allowing the potential user to fully understand the precedence and characteristics of the model

Here it is important to document all details of model building transparently and completely, as mentioned before: where the sample tree data come from in terms of geographical region, sampling approach and possible restrictions, how many sample trees there were used and how distributed were they over the range of the predictor variables.

Essentially, all details of model building need to be reported that are necessary for a potential user of the model to understand it and its background. This also includes statistical measures of model accuracy, even though it is not a common default practice in NFIs where model errors are reported and propagated to the total error (also see the point below).However, the uncertainty measures are important when a potential user compares different models; then the user may tend to prefer the more accurate model.

Reporting the predictions from a model within the context of an NFI implementation

Here, model predictions are dealt with in forest inventory projects like normal observations, the corresponding model uncertainty is commonly not reported, as their contribution to the final error has been proven to be minor in empirical and theoretical studies. However, it is good practice to document which models have been used and give their source and characteristics in the inventory report.

Summary

Before we conclude, here are the key learning points of this lesson.

- Statistical models aim to establish a quantitative relationship between a predicted variable and one or more predictor variables.
- Statistical models do not serve to establish a cause-effect relationship - this would be tackled with process models, which aim to explicitly include the causalities inferred by biological processes.
- There are different types of models of varying complexity that are used in forest monitoring, and in some cases, it is difficult to notice that a model has been used.
- In many cases, it is clear from the outset which model to use because it has been used before in the same geographical or subject-matter context, or is known to perform well under the circumstances of the particular inventory project.
- In NFIs that extend over vast areas, and include many species and species groups, it may be adequate to apply different models depending on species group, and/or geographic region and/or site conditions.

Lesson 4: Errors in forest monitoring

Lesson introduction

This lesson elaborates on the types and roles of random errors as they occur along the NFI process. It also discusses error propagation—how the error sources propagate to the total error of the final result.

Learning objectives

At the end of this lesson, you will be ale to:

1. Define the term 'error' in empirical sampling studies.
2. Describe why error considerations are important in forest monitoring.
3. Explain the relationship between error and effort.
4. Understand the types and roles of errors in forest monitoring.
5. Explain how to cope with errors in forest monitoring.

General observations on errors in forest inventories

Definition of the term error in empirical studies

Forest inventories are empirical sampling studies, where it is more accurate to refer to errors as residual variability and not as mistakes. While mistakes can be avoided through careful work and a permanent quality orientation, errors cannot—we can only attempt to keep them small. The errors we are referring to here, are random in character and tend to follow the Gaussian error distribution, also known as normal distribution.

As mentioned, and contrary to random errors, systematic errors or biases can usually be avoided, because they are based on miscalibrations, the use of biased estimators, or other mistaken applications of data generating approaches. Because random errors are omnipresent, one may try to make the field teams and other data generating parties work in a careful and consistent manner to help keep the corresponding error sources small. Commonly, in statistical sampling, we refer to systematic errors as defining the accuracy, while random errors define the precision.

Did you know?

Precision and accuracy are two core terms in statistical sampling, and they are determined by systematic (bias) and random errors. Often, the term uncertainty is used in reporting when referring to errors, as it is a less technical and more accessible term. But it is also less clearly defined. Therefore, when using the term uncertainty in the context of statistical sampling, it is a good practice to clarify what is specifically meant.

Why are error considerations so important in forest monitoring?

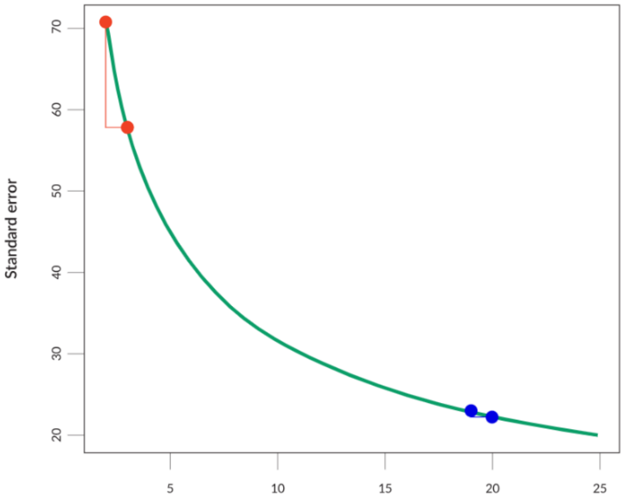
The presence and magnitude of errors are important factors contributing to much of the credibility of inventory results. If the error is 50 percent, one would have less trust in the results than for an error of 1 percent. It is, therefore, imperative to report the errors for all results, quantifying the errors that can be quantified and addressing/discussing the errors that cannot directly be quantified.

When planning the inventory design, it is always the goal to use the available resources to optimize the precision of estimation and avoid systematic errors. Therefore, it is important to understand the roles of different inventory design elements and how the corresponding error sources contribute to the overall errors. If, for example, there are additional resources available that can be used to improve the inventory design, one will need to identify how to allocate these resources, such that they contribute best to increased precision = reduced uncertainty = reduced overall error for the core target variable(s).

The relationship between errors and efforts

Standard error is often displayed as a function of effort. Effort is defined as the sample size that can be established. The marginal increase in precision is smaller with increasing precision. This means, for the same effort, you will get a smaller increase in precision if you start from an initial high level of precision. Here, the error refers to the standard error and effort refers to the sample size that can be established. Therefore, when optimizing forest inventory designs for error reduction, it may be straightforward to first look at sources of relatively large errors—even if their impact on the final error appears not the heaviest at the first glance.

To give an example: in the figure below, we assume SRS and an estimated standard deviation of s=100. The increase in sample size by 1 from n=2 to n=3 reduces the estimated standard error from SE=70.7 to SE=57.7, that is by about 18 percent. If we invest the same additional resources in absolute terms and increase the sample size by 1 from n=19 to n=20, the increase in precision = reduction of the standard error from SE=22.9 to SE=22.4, which is a relative reduction by about just 2.2 percent.



The relationship depicted in this figure holds for the standard error in SRS designs. However, one may assume that similar relationships hold for other sources of error, for example, increasing training efforts by a fixed absolute amount for better observations of variables will commonly have the largest effect for those variables whose measurement error is relatively large (such as measuring tree height), and will not have large effects when looking at variables whose measurement error is already quite small (such as measuring dbh).

What is good precision in NFIs—what error level should be targeted?

There is no generally valid rule on the target precision in NFIs. The common approach is that the resources (budget) is defined, and the inventory is designed such that the best precision is achieved under the restrictions of these resources. The achieved precision is a function of the sampling design and, particularly, of the sample size.

In some NFIs sample size is in the order of magnitude of the 10 000s so that the precision of estimation for the whole country is high—in some cases the relative standard error is under 1 percent—but in others, sample size may be two orders of magnitude lower, producing relative standard errors above 10 percent. When, however, estimates are produced for smaller sub- populations, where the sample size is much smaller, the standard error may go up and reach levels for which the reliability of the results may be compromised.

Types of errors in forest monitoring and their role

Forest monitoring systems are complex, and many people are involved in their implementation. This has the potential to increase the various sources of errors that need to be observed in planning and implementing these systems. There are three types of errors that occur in forest monitoring and all play an important role, but have varying relevance depending on the design of the inventory:

① Measurement error;

② Model error, and

③ Sampling error.

In this section, we will take a deeper look at what each of them constitutes.

Measurement error

Whenever an observation is made, this observation is subject to residual variability. When, for example, the dbh is measured with a very fine-scale caliper, say to the 5th decimal of a millimeter, repeated measurements—that were all very diligently and correctly carried out— would practically all yield different measurements. The variability of these measurements indicates the existence of measurement errors.

Such measurement errors can occur beyond quantitative variables. Categorical variables that are observed in, say, 10 different classes, can present misallocations: such measurement error would then also be referred to as classification error. Or, when observing nominal variables, such as tree species, confusion/misidentifications, also considered as measurement errors, may occur. It is important to realize that measurement errors will occur at the observation of any variable.

An interesting case is the measurement of tree height by the trigonometric principle. In fact, height is not measured, but calculated from three measurements: the horizontal distance to the tree and the angle measurements to the top and the bottom of the tree. Depending on the measurement device, horizontal distance may also be based on two measurements: slope distance and slope angle. All these individual measurements carry their specific measurement errors and the error in height comes ventually from the propagated measurement errors of all these individual measurements.

Quick tips!

In forest monitoring, there is usually not much information available about measurement errors, and measurements are used in the calculations as if they were error-free. However, for large numbers of sample trees, as usually is the case in NFIs, one can justify ignoring random measurement errors and not report them explicitly as they will be very small compared to the sampling error.

If interested, however, some small research studies may be established where different field teams make all observations on a set of sample plots. The variance of the measurements for the various variables may then be considered as an approximation of measurement/observation errors; this may extend to measurements of dbh and height, to the identification of tree species and to the number of sample trees found per plot.

Model error

Models are very frequently used in forest monitoring to establish relationships that allow predicting variables that cannot or are too laborious to be measured (also see Lesson 3 of this course). What is read from a model is, of course, not the true value of the object at stake. When, for example, the stem volume is read from a volume model as a function of dbh, then what is read as the tree's stem volume is actually the stem volume of all trees in this dbh class—and the true volume of the particular tree will deviate from that. This is what we may refer to as model error.

In forest monitoring, model errors are commonly not reported nor factored into the interval estimates. The predictions that are made from the models are taken as true values, or error- free. In models for tree variables like biomass, volume or height models, as long as the models used are based on a relatively large number of sample trees, and as long as the number of sample trees in the inventory is large, it may well be justified not to report the model errors, as they will be very small compared to the sampling error.

Sampling error

The sampling error originates from the fact that we are not observing all elements of a population but just a sample. Therefore, all results are estimates and they deviate from the true value in a way that is described by the standard error. In forest monitoring, the standard error is normally unbiasedly estimated for most sampling designs, except for systematic sampling, where we need to resort to approximations or to conservative estimates of error variance/standard error when applying the SRS framework.

The roles of these errors

All the three types of errors outlined above exist in all forest monitoring studies. Much research has been done over the past decade, particularly in the context of estimating forest biomass, to find out what the relative contributions of these errors are to the final estimate of biomass. We focus here on national forest monitoring, and among the major characteristics of NFIs is a large sample size and, consequently, a very large number of sample trees.

A most relevant finding has already been published in one of the earliest articles on propagating error sources to the final error: Gertner and Köhl (1992) coined the term **error budget** and found that in the Swiss NFI the biggest weight by far has the sampling error with about 98 percent of the total error. Model errors and measurement errors accounted only for the little percent remaining. This is an important finding because it is commonly only the sampling induced standard error that is being reported in forest inventories, while the other sources are frequently not quantified and reported.

This finding holds for NFIs with large sample sizes. In smaller inventory studies with smaller sample sizes, the relative weight of measurement and model errors may be considerable larger.

How to cope with errors in forest monitoring

While we can avoid systematic errors by careful planning and implementation, random errors still occur. Therefore, the goal is to keep these random errors (= the residual variability) small. It is among the principal goals in forest monitoring to produce estimates that carry reasonably small errors. We say 'reasonably small' because any increases in precision will cost resources and are particularly costly when the precision is already relatively high. In NFIs, the sampling error is the most important error variance component and choosing an appropriate sampling design and then an appropriate sample size are the leverages that are available to adjust for the **standard error** (i.e. **sampling error**).

Careful work, training and periodic control can help reduce **measurement errors**. Here, the major point is in avoiding systematic errors by, for example, miscalibration and to keep the field teams motivated so that they permanently maintain the ambition to produce good data; long field periods are tiring and can easily lead to a deterioration of motivation and as a consequence, of data quality.

Regarding **model errors** however, everything depends on the tightness of the statistical relationship between the target variable and predictors and on the choice of the model and the quality characteristics of the model, which are co-determined by the number of observations that underlie the model. In general, the more data used to build a model the more reliable it can be considered.

Forest inventory projects and forest monitoring programs are complex. The three types of errors addressed above may occur at any step of these systems, and the major interest is eventually in the total error in the target variables. It is intuitively clear that all errors that enter at different steps of the process will have some impact on the overall error of the target variable, that is, their impact is propagated though the different steps of the inventory process into the final target variable.

In this context of error propagation, there are a couple of points to consider:

1. **How the mechanism of error propagation works, and how the overall error is determined**. This is important for reporting the overall error as an indication of precision and overall reliability of the results.
2. **To what extent the different errors contribute to the overall error.** This is important in the context of optimizing the inventory design for follow-up inventories: one will strive to reduce those errors (at an acceptable cost) that have the biggest impact on the error of the target variables.

Basic principles of error propagation

The basic principles of error propagation are straightforward and depend on how the different input variables and their errors are linked:

A good and understandable introduction into error propagation is Taylor (1997), which covers the rules to combine random variables according to the operation used to combine them. Shorter introductory lectures that include sums, products, ratios and other operations can be found in the following locations:

Guide to Uncertainty Propagation and Error Analysis: Stony Brook Introductory Physics Labs

A Summary of Error Propagation

Propagation of Uncertainty through Mathematical Operations

What is elaborated in Taylor (1997) is an analytic approach to error propagation that can directly be applied to functions of random variables. If, however, an error propagation shall be done in a complex inventory design, where many different error sources need to be considered, such analytical error propagation becomes extremely difficult. Then, a simulation study (also called Monte-Carlo Simulation), might be more appropriate.

To conduct such a study, there must be information available on the different error components. Commonly, normal distributions in these errors are assumed. Then, the target variable (e.g. biomass) is calculated from all input data, where for each point estimate a random deviation is determined from its normally distributed error. This is repeated very often—say 10 000 times— and the variance of the resulting final values of the target variable is then the propagated total error variance. Instructive examples can be found in Molto et al. (2013), McRoberts and Westfall (2016) and Lin et al. (2023).

Both simulation and analytic calculation of error propagation allow for an evaluation of the weight that the different error components have in the error of the final estimate, so these error propagation exercises are very instructive when optimizing the inventory design.

McRoberts and Westfall (2016) give an instructive example of how a simulation study can be done when the interest is in propagating various error sources in forest monitoring to the final target variable. If an estimate of biomass is the target variable, the two authors integrated the following sources of error into their simulation:

➲ **If an estimate of biomass is the target variable**

the two authors integrated the following sources of error into their simulation: Variability of model parameters estimates, (β), in the allometric model; Variability of dbh measurements; Variability of height measurements (and other input variables to the model); Residual variability (what is predicted from the model is not the true biomass); Aggregating individual tree biomass to plot biomass; and Estimating total biomass for the study area from a sample of n plots (with a defined sampling design).

➲ **The simulation was then done as follows:**

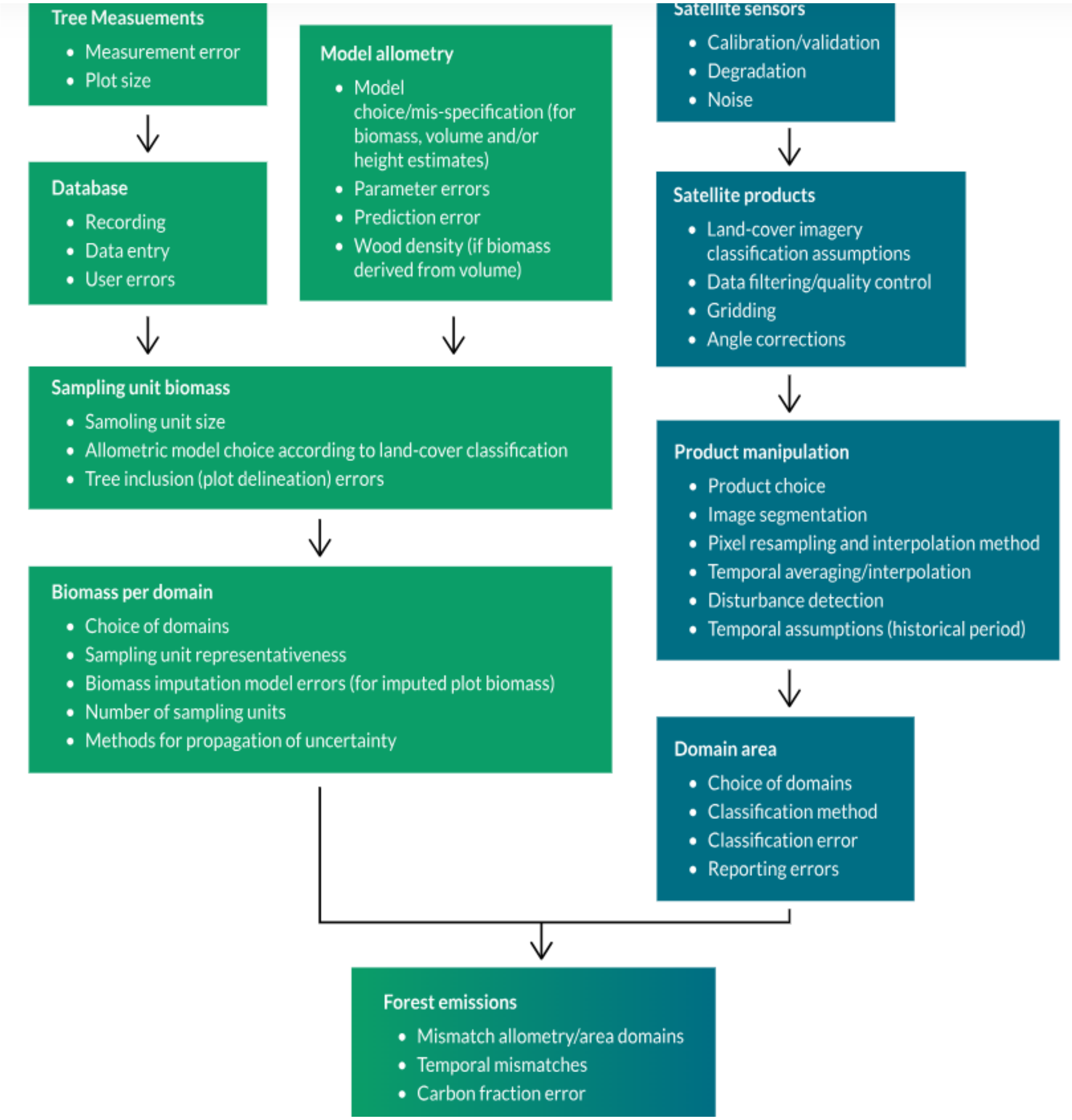
For each simulation, a set of input values is determined for the above variables, where error components randomly chosen from normally distributed errors are added to the point estimate. Then, the total biomass is determined for this particular setting. The simulation is repeated often (*m*) times, each with input values that are determined from the point estimates plus a random error component. The variability of the resulting, *m*, total biomass values is then an empirical approximation of the propagated total error.

Some closing comments on error propagation

A comprehensive list of potential sources of error typically arising in the processing chain to calculate emission factors, activity data and total carbon is reported by countries. The list contains possible concepts that are too cumbersome and out of scope of these courses.

Please note that green routes are defined exclusively by inventory data. Blue routes are followed by satellite data processing. The blue–green box stems from the combination of inventory-based emission factors and satellite-based estimates of activity data.

Flow of errors in inventory-based emission factors (in green) and satellite-based activity data (in blue)



Tree Measuements

- Measurement error
- Plot size

Database

- Recording
- Data entry
- User errors

Model allometry

- Model choice/mis-specification (for biomass, volume and/or height estimates)
- Parameter errors
- Prediction error
- Wood density (if biomass derived from volume)

Sampling unit biomass

- Samoling unit size
- Allometric model choice according to land-cover classification
- Tree inclusion (plot delineation) errors

Biomass per domain

- Choice of domains
- Sampling unit representativeness
- Biomass imputation model errors (for imputed plot biomass)
- Number of sampling units
- Methods for propagation of uncertainty

Satellite sensors

- Calibration/validation
- Degradation
- Noise

Satellite products

- Land-cover imagery classification assumptions
- Data filtering/quality control
- Gridding
- Angle corrections

Product manipulation

- Product choice
- Image segmentation
- Pixel resampling and interpolation method
- Temporal averaging/interpolation
- Disturbance detection
- Temporal assumptions (historical period)

Domain area

- Choice of domains
- Classification method
- Classification error
- Reporting errors

Forest emissions

- Mismatch allometry/area domains
- Temporal mismatches
- Carbon fraction error

Summary

Before we conclude, here are the key learning points of this lesson.

- Forest inventories are empirical sampling studies, and when we talk about errors in empirical sampling studies we refer to residual variability and not to mistakes.
- The presence and magnitude of errors are important factors contributing to much of the credibility of inventory results.
- There is no generally valid rule on the target precision in NFIs—the precision is decided by the most effective blend of resources (budget) and the inventory design.
- There are three types of errors that occur in forest monitoring and all play an important role, but have varying relevance depending on the design of the inventory: measurement error, model error and standard error.

Lesson 5: Typical products from data analyses in forest monitoring

The major goal of data analyses in NFIs is to transform NFI data into meaningful information for stakeholders and interested parties. This lesson elaborates on the major products generated from NFIs.

Learning objectives

At the end of this lesson, you will be ale to:

1. Describe potential products of NFI data analyses.
2. Identify the role of data analyses for reporting of forest inventories

Products from forest monitoring data analyses: General observations

Data analyses generate NFI results that eventually comply with the expectations stated in the information needs assessment. In this section, we will briefly address the general types of products that are expected to be produced from NFI data.

Having a clear idea of potential products from NFI data analyses in the planning phase and during the INA always helps. It is also instructive—in the INA phase—to present all potential outcomes of NFI data analyses and to narrow this down to what can be realistically expected.

Quick tips!

It is not necessary during the INA phase to consider all analysis implications—this is the business of data analysts. However, it is certainly helpful to have experts with NFI data analysis experience present in the information needs assessment so that completely unrealistic expectations can be avoided.

A typical example is the expectation that NFI data can be directly used for forest planning purposes on district or even stand level; here, an experienced inventory expert will need to clarify the possibilities and limitations of NFI data sets.

Types of products

Keeping in mind the vast array of results and products that can be generated, NFIs produce comprehensive data sets that include various analysis options. In general, it is good practice to present the results using two distinct strategies:

1. one for the stakeholder and decision maker (which needs to be technical- scientific and following the requirements expressed in the INA); and
2. one for the general public (which needs to summarize the major findings in an accessible, though precise, language).

For what concerns the technical-scientific products, we can break down the information into four categories as follows:

1. standard statistics;
2. maps
3. N FI design optimization; and
4. Use in academia.

Standard statistics

The results of NFI data analyses cannot encompass a complete list of standard results. Therefore, it is best to limit ourselves to those that are commonly produced in NFIs, along with the specific additional products that arise from the particular information needs for a specific country.

For example, in a country with a low forest cover, it may be extremely relevant to also produce results on trees outside the forest (TOF)—while in countries with a high forest cover, this tree resource may be of minor relevance (of course, analyzing data on TOF is possible only when the assessment of TOF is integrated into the inventory design).

The basic units of reference for analyses typically are the whole country and sub-national units—provinces, states, or defined eco-zones. In most cases, the sample size of NFIs does not allow to go farther down and produce estimates for smaller units, unless special scientific analysis techniques such as small-area estimations are used as an advanced approach.

Did you know?

What is small-area estimation?

NFIs commonly use systematic samples with a grid size in the range of kilometers. This is why reasonable estimates cannot be produced for relatively small units (e.g. a few km^2) because the sample size is too small.

However, ongoing research is investigating how to seize large-area information produced by NFI sampling and use it to produce results also for smaller geographical units. This may be done by linking the field observations to full cover remote sensing data, which is then used as auxiliary data to establish models that allow predicting target variables for any pixel, that is: over the whole inventory area. Then, data of the target variables are available not only at the field observed points but for any location (pixel) in the inventory region. This approach to produce results from the large area coarse field sample for any small area unit within the inventory region is called **small-area estimation**

Of course, the prediction uncertainty for the small areas depends exclusively on the quality of the model that is being derived and used, linking the large area field observations and the remote sensing data; and this depends, among other factors, on the spatial and spectral resolution of the remote sensing data, on the plot design of the field inventory, and on the quality of co- registration field and remote sensing data.

Estimates of areas include: forest area, area of forest types, areas or percent area for particular tree species, areas per management types, of protection status, of degradation status, of ownership, or topographic features. Results are then produced for each of these reporting units, for example forest area per sub-national unit, forest type areas per country and per sub-national unit, forest area in different elevations, etc. Presentation is commonly in the form of two-way tables, like for example area of, say, 5 forest types within 10 sub-national units.

	Forest specification					
Land	**Measure**	**Stocked timberland**	**Temporarily unstocked**	**Timberland**	**Unstocked forest land**	**Forest**
Baden-	[ha]	330 625	1 301	1 331	39 922	1 371 847
Württemberg	Prim	4 600	13	9264 601	372	4 620
	SE [%]	1.2	27.7	1.2	5.2	1.2
	[ha]	2 534 232	3 796	2 538 028	67 535	2 605 563
					194	2 815
					7.7	1.6
Bayern	Prim	2 795	11	2 797		
	SE [%]	1.6	33.7	1.6		
Brandenburg+	[ha]	1 096 101	2 369	1 098 470	32 378	1 130 847
Berlin	Prim	907	6	907	67	909
	SE [%]	2.7	40.8	2.7	12.8	2.7
	[ha]	845 792	7 598	853 390		
Hessen	Prim	706	19	706	40 790	894 180
					91	715
					10.8	2.9
	SE [%]	2.9	22.8	2.9		
Mecklenburg-	[ha]	538 651	2 186	540 836	17 286	558 123
Vorpommern	Prim S	2 038	19	2 041	148	2 055
	E [%]	2.1	24.0	2.1	8.8	2.0
	[ha]	1 158 459	2 985	1 161 444	43 147	1 204 591
					135	1 571
					9.2	2.4
Niedersachsen	Prim	1 552	12	1 555		
	SE [%]	2.4	30.5	2.4		
Nordrhein-	[ha]	880 082	3 997	884 059	25 452	909 511
Westfalen	Prim	861	10	863	59	867
	SE [%]	3.1	31.6	3.1	13.3	3.1
Rheinland-	[ha]	812 818	2 290	815 108	24 688	839 796
Pfalz	Prim	2 828	22	2 831	236	2 848
	SE [%]	1.5	21.7	1.5	6.5	1.4
Saarland	[ha]	101 459	783	102 242	392	102 634
					1	100
					100.0	8.0
	Prim	100	2	100		
	SE [%]	8.0	70.5	8.0		
	[ha]	517 858	2 392	520 249	12 956	533 206
					56	951
					14.5	2.9
Sachsen	Prim	943	12	946		

	SE [%]	2.9	28.8	2.9		
Sachsen-Anhalt	[ha]	493 920	9 067	502 987	29 494	532 481
	Prim	1 829	79	1 845	264	1 884
	SE [%]	2.2	12.1	2.1	6.3	2.1
Schleswig-Holstein	[ha]	168 426	199	168 626	4 787	173 412
	Prim S	775	2	776	45	778
	E [%]	3.8	70.7	3.8	15.2	3.7
	[ha]	520 944	2 799	523 743		
Thüringen	Prim	895	14	902	25 345	549 088
					118	912
					9.4	2.6
	SE [%]	2.7	26.6	2.7		
Hamburg + Bremen	[ha]	13 054	---	13 054	791	13 846
	Prim	152		15	2	15
	SE [%]	5.6		25.6	70.4	25.8
Germany (all Lander)	[ha]	11 012	41 742	11 054 162	364 962	11 419 124
	Prim	420	221	20 885	1 778	21 040
	SE [%]	20 844	8.0	0.7	2.9	0.7
		0.7				

Example of two-way table giving the forest area per Federal State in Germany ("Land") broken down into different categories of forest land. This table had been produced from the online-tool of the German NFI; not only the estimated area is given but also the estimated relative standard error SE% and the number of clusters = primary sampling units (which corresponds to the sample size per sub-national unit) that fell into the combinations of Federal State and type of forest land. It is clearly visible here that precision of estimation is a function of sample size.

Remember that when breaking down the areas, all breakdown criteria must be clearly defined so that the results can properly be interpreted along these definitions: one needs to clearly define 'forest' and contrast it to non-forest, and one needs to have clear criteria to distinguish within forest different classes of 'degradation', 'forest types', 'management types' and so on.Furthermore, it is important to consider that not all categories may be identified in the field or by using remote sensing imagery. In some cases, these categories need be taken from official documents. For example, ownership and protection status need to be extracted from cadastral maps and from maps of protected areas, respectively.

Estimates of characteristics per area, including: volume/biomass/carbon stocks per hectare, number of trees per hectare, number of large trees, regeneration density, deadwood stocks in different dimension classes, and so on.

District	Biomass (million tonnes)	MoE (%)	Carbon (million tonnes)	MoE (%)
Bumthang	80 ± 16	20	37 ± 7	20
Chhukha	91 ± 21	23	43 ± 10	23
Dagana	50 ± 8	15	24 ± 4	15
Gasa	7±2	31	3 ± 1	31
Haa	57 ± 10	18	27 ± 5	18
Lhuntse	77 ± 19	24	36 ±9	24
Mongar	95 ± 18	19	45 ±9	19
Paro	30 ± 8	27	14 ±4	28
Pemagatshel	18 ± 4	21	8±2	21
Punakha	54 ± 9	36	25 ± 9	36
Samdrup Jongkhar	72 ± 14	20	34 ± 7	20
Samtse	20 ± 5	22	10 ± 2	22
Sarpang	37 ± 6	17	18 ±3	17
Thimphu	48 ± 25	51	23 ± 12	52
Trashigang	96 ± 16	16	45 ± 7	16
Trashiyangtse	41 ± 20	48	19 ± 9	48
Trongsa	52 ± 11	21	25 ±5	21
Tsirang	29 ± 11	39	14 ± 5	39
Wangduephodrang	91 ± 16	18	43 ± 8	18
Zhemgang	56 ± 7	13	26 ±4	13

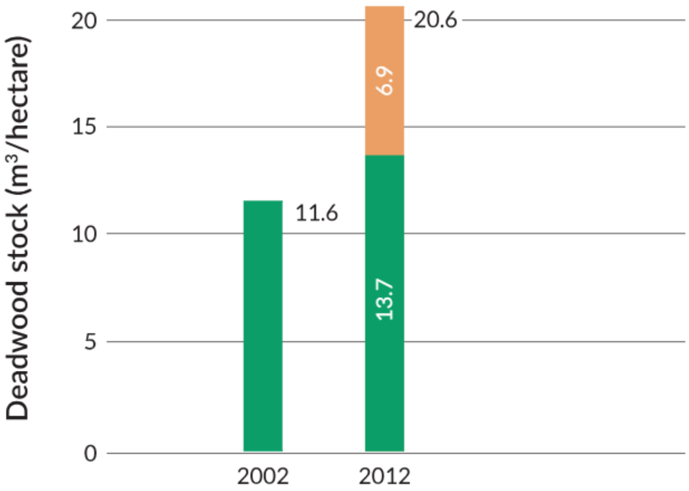
Again, we may illustrate this with an example of a one-way table giving the total mass in biomass and carbon per Bhutan (DFPS. 2019). The relative standard error (computed from the confidence interval) is also given for every district.

Estimates of changes in the target variables, when the analyses refer to repeated inventories. When analyzing these results, it is important to observe whether the definitions might have changed. It may happen that the analysis reveals a change—but that part of the change can be attributed to changes in definitions.

One example is given here for the changes in estimates of deadwood stocks in the German NFI between 2002 and 2012. A very large change in estimates of deadwood stocks resulted from the analysis. Part of this unexpectedly large change in estimates was due to the adaptation of the minimum diameter of recorded deadwood pieces from formerly 20 cm to the international IPCC standard of 10 cm.

In this case, one may easily analyze which portion of the change can be attributed to the change in definition, because all required information was in the data (for application of the old definition they just needed to leave out all deadwood pieces with a diameter smaller than 20 cm). This would result to be more difficult when analyzing other modifications of definitions like the change of minimum crown cover in the definition of forest.

It is important then, that the analysis makes clear what the components of these changes are: in this case, maintaining the old definition, the change would be an increase of 2.1 m³/ha from 11.6 m³ to 13.7 m³, but in the graph it is shown as 4 times as much (9 m/ha from 11.6 m³ to 20.6 m³) because of the modification towards a more inclusive definition (BMEL 2014).



If the analysis also produces **estimates on trees outside the forest**, this will need to refer to different non-forest land-use types, presenting essentially the per-area results, but for the non- forest land-use types.

It is important to repeat here again that, for all point estimates, the analysis should also produce interval estimates (standard error or confidence intervals) so that the analysis shows the status estimate as well as the uncertainty of such estimate.

Confidence intervals are, of course, also of crucial interest for change estimates. If the value zero is contained within the upper and lower confidence intervals, one may assume that the changes not be significant. Here, of course, when deriving statements about statistical significance, the interpretation of the confidence intervals must take into account that it is usually systematic sampling that is used in NFIs.

Maps

Maps are frequently used to present NFI results and, often for many non-experts, they are more convincing than statistics. Full cover continuous maps can only be produced when area-wide remote sensing imagery has been analyzed and the respective models developed.

Forest/non-forest maps are a base product that, accompanied also by biomass maps that are of great interest. As with all other forest inventory products, the inventory analysts should stress that maps may come with inaccuracies just like all products pertaining to empirical studies. As a result, these uncertainties should be properly documented and reported together with the maps.

Example of a regional map showing the predicted distribution of growing stock over a region in India (courtesy Dr. Paul Magdon from an FAO consultancy to the Forest Survey of India). Sentinel- 2 multispectral imagery was used as carrier data and n=170 field plots were available for model building. The following figure informs about the quality of the map by giving some variability measures of the underlying model; assuming here, as is commonly the case, that the field- determined values of growing stock had been generated free of observation or model errors.

Validation of the model used in in the map above for the Sentinel 2- based model of growing stock. Explained variance: 34.02 percent. Observe that this model does not follow the 1:1 line and that there is considerable variability. The map gives, as so often, an impression of the spatial pattern of the target variable, but carries considerable uncertainty. Specific interpretations at high resolutions should, therefore, be avoided.

If remote sensing data are not used in an NFI, maps can only be generated at the spatial resolution of the systematic sample grid used. Such maps cannot produce a continuous representation of a target variable but only give a rough idea about spatial distributions at a quite coarse scale. Typically then, one information is given per sample point—and these are sometimes at a distance of various kilometers.

Using NFI data for future NFI design optimization

Analysis of NFI data may also be beneficial towards an optimization of the inventory design for the planning of subsequent inventories. Of course, in doing so, one must be cautious when changing the NFI design between subsequent cycles as consistency is necessary in time series. Still, design adaptations carried out from time to time may increase efficiency, and, frequently, new target variables need to be integrated so that the NFIs can meaningfully respond to newly emerging issues.

Also, time consumption may be optimized by introducing new measurement devices. The control measurements might be re-organized and target accuracies newly defined. It is also important to note that new technologies may imply reductions in the size of field teams.

In this context, it may be interesting to check whether a reduction of the number of sub-plots in a cluster-plot design would lead to a significant reduction in precision of estimation. It may turn out that for some target variables a smaller number of subplots would be enough. Such an optimization analysis can easily be implemented by carrying out the data analysis for a smaller number of sub-plots per cluster, thus reducing the plot size per selected sampling unit.

Using NFI data in academics

The main (and default) task of NFI data analysis is to generate the core results that the stakeholders and decision makers demanded in the information needs assessment. However, NFI data are also a great source for many other purposes such as research and academic teaching; sometimes, NFIs are the only projects that generate a base of scientifically sound data over the forests or even landscapes of an entire country.

Many topics can be analyzed from NFI data, including methodological questions (such as evaluations on optimizing inventory designs or on the application or adaptation of models) and subject matter questions (yield issues, comparison of species compositions and locations, drivers of forest loss, and so on). These types of analyses are not a standard task of the inventory team but require that the databases are made available to researchers.

Academic research that uses NFI data contributes also to educate future NFI experts so that NFI planners and analysts should proactively foster the use of NFI data in academia, in research and teaching (Liang and Gamarra 2020). In particular in longer term forest monitoring systems, valuable time series are available that allow monitoring the forest development and the sustainability of the national forest policies.

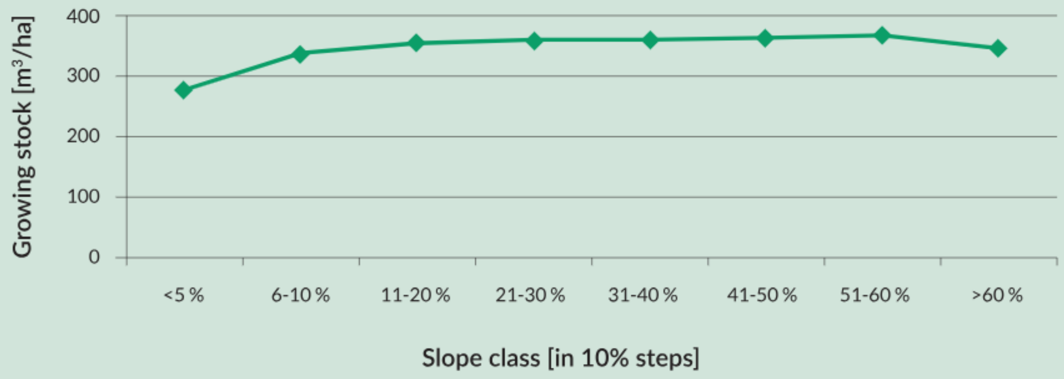
There are research studies that use **data from repeated NFIs to update yield tables** (e.g. Staupendahl and Schmidt 2016) or that **identify the suitability of tree species under climate change** (e.g. Prasad et al. 2020).

Two examples of specific research studies are the **estimation of the forest edge length** from the German NFI data at different scales that the plot design offered in Kleinn et al. 2011; and the **comparison of the stocks of trees outside the forest** (TOF) from 12 FAO-supported NFIs in the Global South (Schnell et al. 2015).

Researchers may be interested in the large area and long-term data sets from national forest monitoring programs when they wish to examine specific methodological or subject-matter questions. For example, systematic samples of cluster plots contain information about landscape fragmentation: if the forests are more fragmented, there will be more intersections with the cluster plots and a smaller number of cluster plots is fully contained within or outside the forest, not having intersections. From the number and proportion of intersecting clusters, one may derive estimates of the general fragmentation status—as, for example, presented by Kleinn (2000) for a large area inventory in Costa Rica—or estimates of the forest edge length for the whole region or sub-national reference units, as was done in Kleinn et al. (2011) for Germany.

Did you know?

In a lecture given by one of the authors, a student raised the question whether growing stock is higher at greater slope classes because the tree crowns get more light and the surface area is bigger than in the plane. The lecturer had access to the NFI database of the German NFI, that allows for flexible analyses by combining variables. Linking growing stock as a response variable and slope classes as categories, the graph below could quickly be produced, allowing for a preliminary answer to that question.



Growing stock [m³/ha] over slope classes [%] – a graph rapidly produced by analysis of data from the German NFI responding to a student´s question: Does growing stock tend to be higher at steeper slopes (at least up to an upper slope limit) because the terrain surface area is larger?

Major characteristics of products from data analyses

The major features that characterize data analyses are essentially the same that also characterize reporting and that have been formulated in general terms as the guiding principles in the Enhanced Transparency Framework (UNFCCC 2020): **transparency**, **accuracy**, **consistency**, **completeness**, **comparability—**and all this manifested in a comprehensive documentation. It is essentially the combination of these features that makes data analyses from NFIs credible for the stakeholders and users of the data.

One always needs to be aware that NFI results enter into the domain of forest-related policies in a country and that there are different interests at stake: not all stakeholders may be happy with the results for multiple reasons.

Then, it is of utmost relevance that:

- the data analysis is 'waterproof' and correct and can be defended on the basis of the complete and transparent documentation; and
- the interpretation of the flndings is compatible with the results of the analyses. Interpretation from different actors may then vary under the same results and statistics, based on particular ideas and values of different interest groups—but this is then outside of the reach of the data analyst.

The role of data analyses for reporting of forest inventories

It probably has become clear now that data analysis predates reporting. Data analysis takes place between data collection/data management and reporting. Hence, when performing data analyses it is important to look at both sides: where the data comes from (i.e. data collection/data management) and how the outcomes of the analysis intend to be used and processed (i.e. reporting).

It is thus imperative that data analysis and reporting are closely interlinked and, if different experts are working in these domains, they need to work closely together. Consequently, early and intermediate results may be discussed, interpreted and compared so that potential inconsistencies may be detected early. This would be particularly useful as such inconsistencies may very well be an expression of unexpected results or, more simply, they may be caused by mistakes in calculations or mistakes in data collection.

It is sometimes underestimated how time-consuming data analyses are with all the cross- checking for data quality and consistency of results—and eventually complying with all expectations expressed in the INA.

Summary

Before we conclude, here are the key learning points of this lesson.

- It is helpful to have a clear idea of potential products from NFI data analyses in the planning phase and during the information needs assessment (INA).
- Forest data analysis information can be used to generate standard statistics and can support to generate remote sensing-based maps and can also be used to optimize future NFI designs, and for research and academia.
- Maps are a common and convincing presentation of NFI results; often and for many more easily accessible and more convincing than statistics.
- Data analysis needs to look at both the data source (as part of data collection/data management) and where the outcomes of the analysis shall be used and processed (reporting).