Basic forest statistics – Accuracy, Precision and Bias

Prof. Dr. Christoph Kleinn, Dr. Lutz Fehrmann
Regional Course on REDD+, MRV and Monitoring
11-15 July 2011
Sokoine University of Agriculture, Morogoro, Tanzania

Main objective of this session

1. Characterize some elements of statistical sampling and monitoring from a methodological perspective.
2. Get more familiar with the terminology and concepts in statistics regarding "uncertainty".
3. Make clear that the basics are not overly complicated nor "beyond reach".
4. Discuss options how "uncertainty" can be influenced.

2

Additional resources:

- Presentation is as simple as possible, if you need more information:

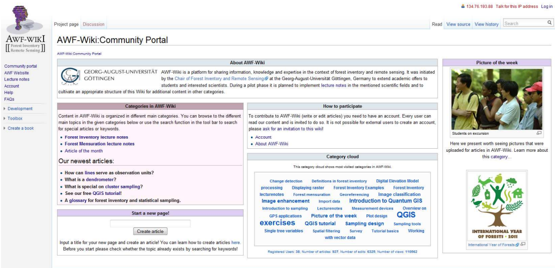
http://wiki.awf.forst.uni-goettingen.de/wiki

Or Google for „AWF-Wiki"

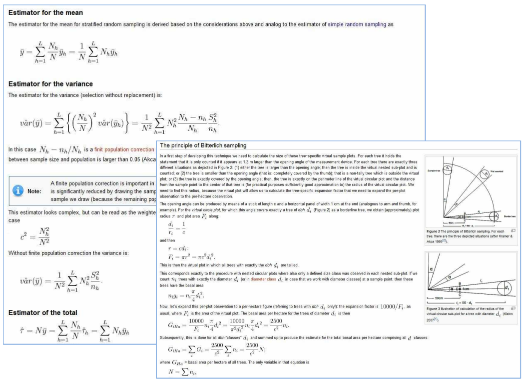
AWF-WIKI
[[Forest Inventory
Remote Sensing]]

3

AWF-Wiki



AWF-Wiki



Background

- Forest ecosystems and the forest resource are highly complex -> **it is impossible to under-stand the system without systematically generated data and information**.
  - Sound MRV and monitoring are a theoretical prerequisite for good policy formulation.
- “Sound monitoring” means from a scientific point of view: **on statistical grounds, along methodologically acknowledged paths**.

Rationale of statistical sampling

Situation:

- We are interested in knowing quantitative details of the forest resource.
- Because of the complexity and extent of the resource, we can not measure /observe everything.
- We need to devise an assessment system which efficiently combines data sources and assessment techniques such that the target data can be generated.

Data and methods

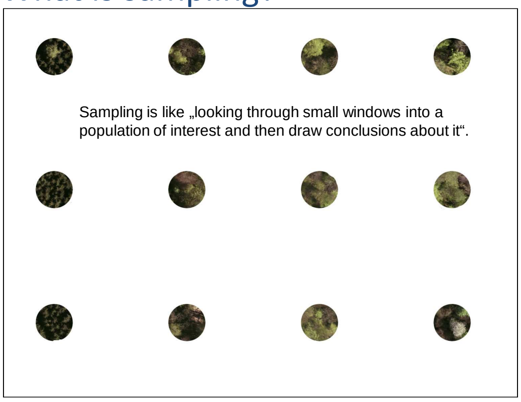
Major data sources used:

- Field data
- Remote sensing
- Existing data and maps from prior studies
- Models

Major methods applied:

- Statistical sampling
- Remote sensing image analysis
- Modelling

What is sampling?



Statistical sampling

- **These rules are defined in science by the principles of statistical sampling.**
  - In most forest inventories, sampling plays a central role:
  - Sampling (**planning/implementation/presentation**) determines largely how credible and/or reliable the results are.
- So far there are no GPG on how to plan a forest inventory (for good reason?)

10

Central questions

- As we observe a subset of the population only, there are some important questions that should be addressed:
  - **Sample size** (How many samples should be taken)?
  - The **selection** of samples (How to select)?
  - How to take **what observations**?
  - How to **calculate** the estimations?

11

Central questions

- Sampling studies are designed such that they
  a. achieve best precision (**budget is fixed**), or
  b. cause minimum cost (**precision is predefined**).
- In practice: What is more common?

12

Guiding principles in planning

- **Conformity** with objectives (regarding variables assessed and their precision).
- **Statistical soundness**.
- Economical and logistical **feasibility**.
- **Sustainability** (should be useful for future assessments).
- Overall **credibility**.

Keep it simple! (also holds for this presentation)

13

Statistical sampling

- Statistical sampling provides methodologically sound (and defendable!) results.
- Randomization is the only generally accepted selection “philosophy” for the class of sampling techniques that we deal with here (but: systematic selection is also applicable).
  - Concepts like “fairness”, objectivity”, “representativeness” should not be used as a general basis for sample selection.
  - Guided subjective or arbitrary selection is not statistical sampling! We may not apply statistical analysis techniques to such samples!

14

„Statistical estimations“

- We are interested in two types of estimations:
  - estimations of the variable(s) of interest (= **point estimates**),
  - estimations of the precision of the point estimates (**interval estimates**),
- As we observed only a small part of the population, the estimate is related to uncertainty that is expressed in form of a “confidence interval”.

15

„Statistical estimations“

- Example:
  Volume = 200 m³/ha ± 15%
  200 m³/ha ± 30 m³/ha
- The latter part of this estimate is the standard error (or half the width of the confidence interval).
- A confidence interval is a **probability statement** that gives insight into the distribution of all possible outcomes of a sampling study.

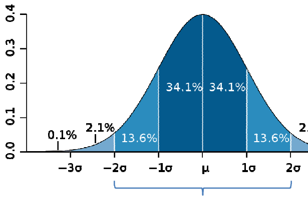
Confidence interval

- defines an upper and lower limit within which the true (population) value is expected to come to lie with a defined probability.
  - Typically with a probability of 95%, means that 5% "α-error" is accepted.
- Every (!) result from a sampling study is an estimate.
  - Also the "precision of estimation" that is the "standard error" is estimated based on the sample at hand!

Convidence interval

An estimate of 200 m³/ha ± 30 m³/ha should therefore be interpreted as:

„**With a probability of 95%** the mean volume per hectare is expected to lie in the interval 170m³ - 230m³“



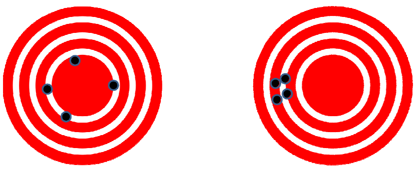
The conservativeness priniciple

- *It is good practice* to use conservative estimates ...
  - for the investor to be on the safer side
  - for the country to have an incentive to produce precise estimates.
- Applied to estimation: the lower bound of the confidence interval is being reported and evaluated.
- Therefore, an appropriate error probability needs to be defined. What are the criteria ?
- ... in forest inventory reporting, we are used to state both point and interval estimate. Then, the reader can draw his/her won conclusions ...

19

Precision and accuracy

- Throwing darts on a dartboard:



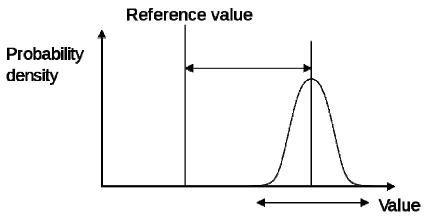
Low precision, high accuracy Low accuracy, high precision

Keep in mind: In sampling studies we have no information about the true population value (target), we just have the sample at hand (darts)!

20

Precision and accuracy

- The **accuracy** of an estimate is the degree of closeness of an estimate (e.g. mean) to the true population parameter ("**correctness**"),
- The **precision** of an estimate (reproducibility or repeatability) is the degree to which repeated estimates under unchanged conditions show the same results ("**exactness**").



21

Precision

- Obviously precision is a matter of the width of the distribution of all estimates that can be drawn from the population,
- **Standard error**: For a given population and a given sampling design, the SE gives the **standard deviation** ("width of the distribution") of all possible sample outcomes.

$$\text{est. Precision (SE)} = \frac{\text{Variability of obs. (error variance)}}{\text{Number of obs. (sample size)}}$$

22

Accuracy

- A measure of accuracy is the Bias.
- **An estimator is unbiased**, if it produces **on the average** estimations that equal the true population parameter.
- It should be noted that "Accuracy" in terms of unbiasedness is a property of an "estimator" (the formula or statistical framework that was used to calculate the estimate) rather than an estimate!

23

More on bias…

- A bias can be introduced by:
  - **systematically** biased selection (Selection bias),
  - **systematically** wrong calculations (Estimator bias),
  - **systematically** wrong observations (Observer bias).
- Obvious: the bias is a systematic error.
- We cannot quantify the bias without information about the true population parameter (remember the dartboard!).

24

More on bias…

- If observations (e.g. carbon stock per plot) are result of a model, there might be a model bias!
- **Important: a bias will not disappear with increasing sample size!**
- If precision is high, a relatively small bias may be acceptable (e.g. for cost reasons) -- However, where possible we prefer unbiased estimators!

25

What is a „good precision“?

- Very good question
  - to the decision makers,
  - to the Investors
- Examples of forest area and biomass precisions of estimation (simple standard error):

Germany NFI:	area 1%,	biomass 0.8%
Costa Rica NFMA:	area 7%,	biomass 12%
Burkina Faso (res.):	area 20%,	biomass: 20%

- Precision of estimation in forest inventory has much to do with a well designed statistical approach … … but much more with the resources allocated.

26

How to influence precision?

- In practice, we try to employ a statistically sound sampling approach, where precision is high.
- With unbiased estimators, precision is a function of (design elements of an inventory):
  - **Sample size** (NOT sampling intensity !!)
  - **Sampling design** (ancillary variables ?)
  - **Plot design**
  - (**Estimation design**)

27

Methodological structure

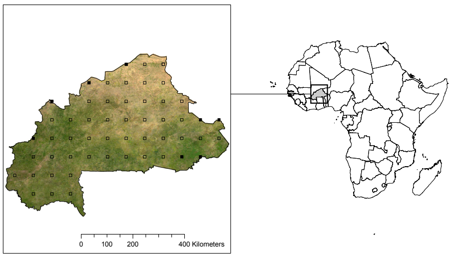
1. Sampling design
   - How to select samples (spatial distribution of sampling locations)?
2. Plot design (observation design, response design)
   - How to take what observations (what to observe at each sampling location)?
3. Estimation design
   - How to calculate estimates?

There are strong dependencies between these elements!

28

Inventory design

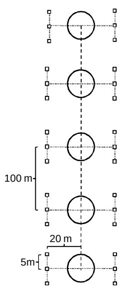
- Defines **how to select samples** (and how many) in the area of interest (population).
  - Random, systematic, stratified, …



29

Plot design (observation design)

- Defines **how an observation is assigned** to each of the selected sampling locations (=field protocol).
- Different levels of complexity:
  - cluster plots of dis-joint sub-plots
  - combining nested plots of different plot types
  - including remote-sensing based sub-plots.



30

Estimation design

- **Defines how to produce estimates** (formulae = estimators).
- Long tradition in forestry of adhering to rigorous statistical sampling; much research to develop unbiased estimators.

31

Sample size

- Is the number of **independently selected** observations!
- If the width of a confidence interval is pre-defined, the required sample size that is necessary to meet this precision can be estimated.
  - Who defines the confidence interval?

32

Sample size vs. Sampling intensity

- Sampling intensity refers to the proportion of the population (in forest inventories=Area) that is observed.
  - Usually **sampling intensity is very small** in large area inventories (German NFI: 0.001%)!
  - On the other hand, **sample size is very high** (in the order of magnitude of *n*=20.000 plots)
  - Precision of estimation for volume and forest area is about 0.8 - 1% (for the whole country)

33

Sample size vs. Sampling intensity

- How is it possible that this low intensity allows estimates with such good precision?
- Simply because intensity is not relevant, it is sample size that counts!
  - Statements like „you need to sample at least 5% of the population to obtain good results" are not helpful!

$$\text{Precision (SE)} = \frac{\text{Variability of obs. (error variance)}}{\text{Number of obs. (sample size)}}$$

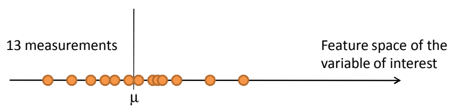
34

Assessing variability

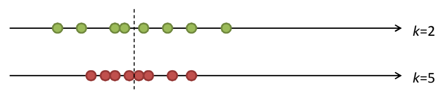
- Every natural population is characterized by a certain variability of the variable(s) of interest (imagine e.g. the variability of carbon stocks in a certain forest type).
- Only a part of this given variability is „captured" or observed inside the field plots,
- An intuitive question is therefore: **What is the consequence of capturing more or less variability inside the plots**?

35

Example



Lets assume the observation design defines that always a number of *k* measurements are combined to one observation (mean):



Only some of the (extreme) possible outcomes are shown here!

36

Example

Simple conlusions:

1. The variance (and therefore also the standard error) between observations is decreasing if more variability is captured within the observations (means: precision is increasing),

37

How can we influence precision?

- Considering the formula of the standard error it is obvious that
  - Precision is **increasing** (underproportional) **with larger sample size** *n*, and
  - Precision is **increasing** (strict proportional) **with decreasing** variance of the target variable.
- How can we influence the variance of the target variable in the population?

38

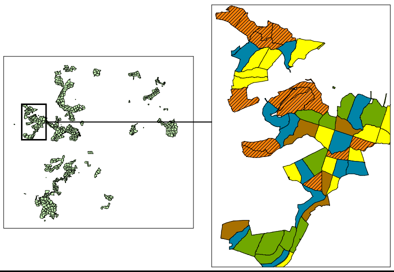
Stratification

- Stratification is a means to reduce the variance of the total population by „splitting" it in more homogeneous sub-populations (strata) that are treated independently.

39

Stratification

- If available (typically not in the REDD context), forest management data can be used to stratify according to species or age classes.



40

Stratification

- RS data are a useful ancillary information in context of stratification, but
  - They are affected by misclassification errors!
  - Example: A sample point was classified as „*Forest type B*" based on RS interpretation. In the field the crew classified it as „*Forest Type C*". What might be the reason?
    - Outdated RS data?
    - Different spatial scale?
    - Low discriminatory power of RS data for these types?

41

Stratification

- In each stratum a different inventory- and/or plot design can be applied,
  - Design elements, like a systematic sampling grid, can be (in principle) optimized independently towards the given conditions in each stratum
- But: In case of REDD the interest is mainly on changes!
  - Different sampling grids might cause problems if multiple inventory cycles are planned!

42

Stratification

- Don't mix up the statistical concept of stratified sampling (independent sampling studies in different strata) with „reporting estimates for different classes"!
  - As long as the same design (plot-, sampling design) is applied (not indipendently) a „post" stratification can be applied (means: simply querying the data for different classes).

43

Plot (observation) design

- There is **No** optimal design for all conditions (discussions on optimal plot shape and size on international level is sensless)!!!
- There are guiding principles that might help to find an efficient (but never **one** optimal) solution:
  - Practical considerations (implementation in the field, „not more than 1 day/plot", …
  - Statistical considerations: plots should encompass as much variability of the target variable as possible, …

44

Plot (observation) design

- Indicators for plot optimazation might be obtained from a pilot phase:
  - Number of trees/(cluster) plot in different strata (mean tree density/stratum)
  - Workload per (cluster) plot (feasability)
  - Co-registration of RS data (e.g. spatial arrangement of cluster subplots)
  - Intra- and inter cluster correlation
  - Resulting SE of different simulated design alternatives (varying plot size or number of subplots)

45

Plot (observation) design

- The optimazation of a plot design is a research question that can be integrated in the pilot phase of an inventory
  - Example: starting with rel. Large circular (sub)plots in the first phase (assessing e.g. 30 plots per stratum) and recording the distance of each tree to the plot center allows simulating any smaller plot size
  - Different numbers of subplots per cluster (and cluster configurations) can easaly be simulated!

46

Plot (observation) design

- Example: Country X adopted a cluster plot design that was developed for the assessment of tree resources outside forest (for a minimum diameter of 30 cm) for their REDD pilot inventory,
  - As the diameter threshold for above ground carbon stock accounting is 10 cm (IPCC), field crews have now to measure 400 trees per plot in mean!

- Does this make sense?
  - Do you think they can obtain new, important and additional information by measuring the 399th tree/plot?

47

Simulation studies

- Example: Simulation study based on data from a long term ecological research plot

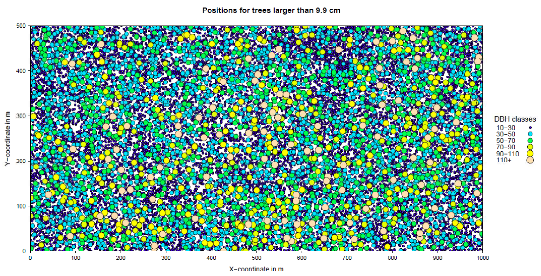
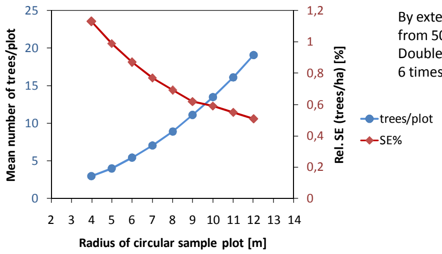


Figure: S.Schnell

48

Simulation studies

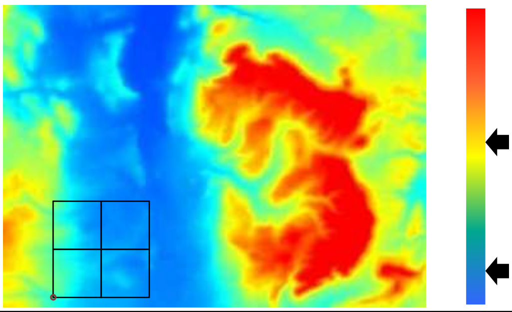
- Precision is increasing under-proportional (asymtotic) with increasing plot size,
- Number of trees is increasing over-proportional (= workload)



By extending plot size from 50m² to 500m²:
Double precision,
6 times more work!

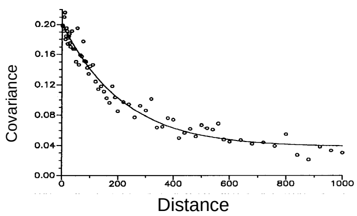
Cluster plots vs. Compact plots

- Increasing the variance inside an observation will decrease the variance between observations (= inrease precision)!



Cluster plots

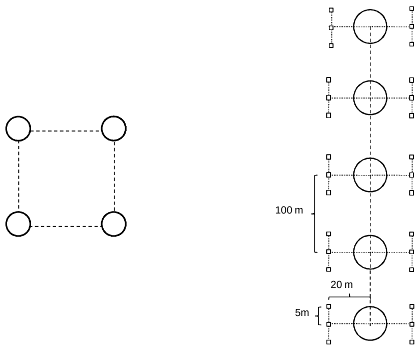
- For many natural populations we can observe a spatial autocorrelation of nearby observations:



Cluster plots

- Statistical efficiency (compared to compact plots) is dependent on the ammount of „additional information“ or variance that is captured per sample point by spreading the subplots!
  - This again is dependent on the correlation between observations on distinct subplots

Cluster plots

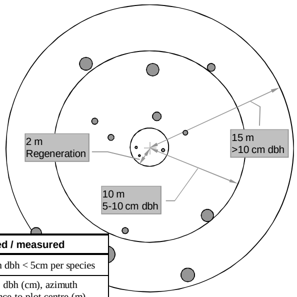


Cluster plots

- Cluster plots are a means to increase the variability of a target variable inside the plot
- It is a straight forward approach to collect as much information as possible at each sampling location
- Especially efficient if transport costs are high in relation to data assessment on the site.
- Due to the spatial extend the probability of intersecting with stratum boundaries is higher than with compact plots.
  - This is not a problem if only the part of a cluster inside the regarded stratum is considered (clusters of unequal size)

Nested plots

- Different plot size for different diameter classes:

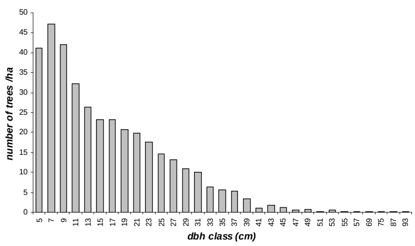


Radius	Trees recorded	Variables observed / measured
2 m	Regeneration	Number of trees with dbh < 5cm per species
10 m	Understory trees 5 - 10 cm dbh	tree number, species, dbh (cm), azimuth angle (degree), distance to plot centre (m)
15 m	Trees > 10 cm dbh	tree number, species, dbh (cm), azimuth angle (degree), distance to plot centre (m), remarks (for example "bark harvested")

55

Diameter distribution

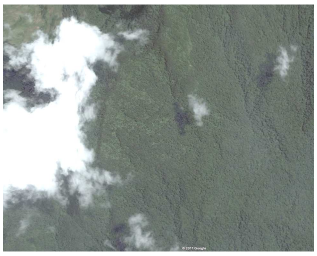
- Typical diameter distribution in a natural forest: Fixed area plots for all classes would inlude unreasonable number of small trees!



56

Non-response

- A crusial problem in many countries
  - In field observations (inaccessability)
  - In RS data (clouds, shadow, coverage,…)



57