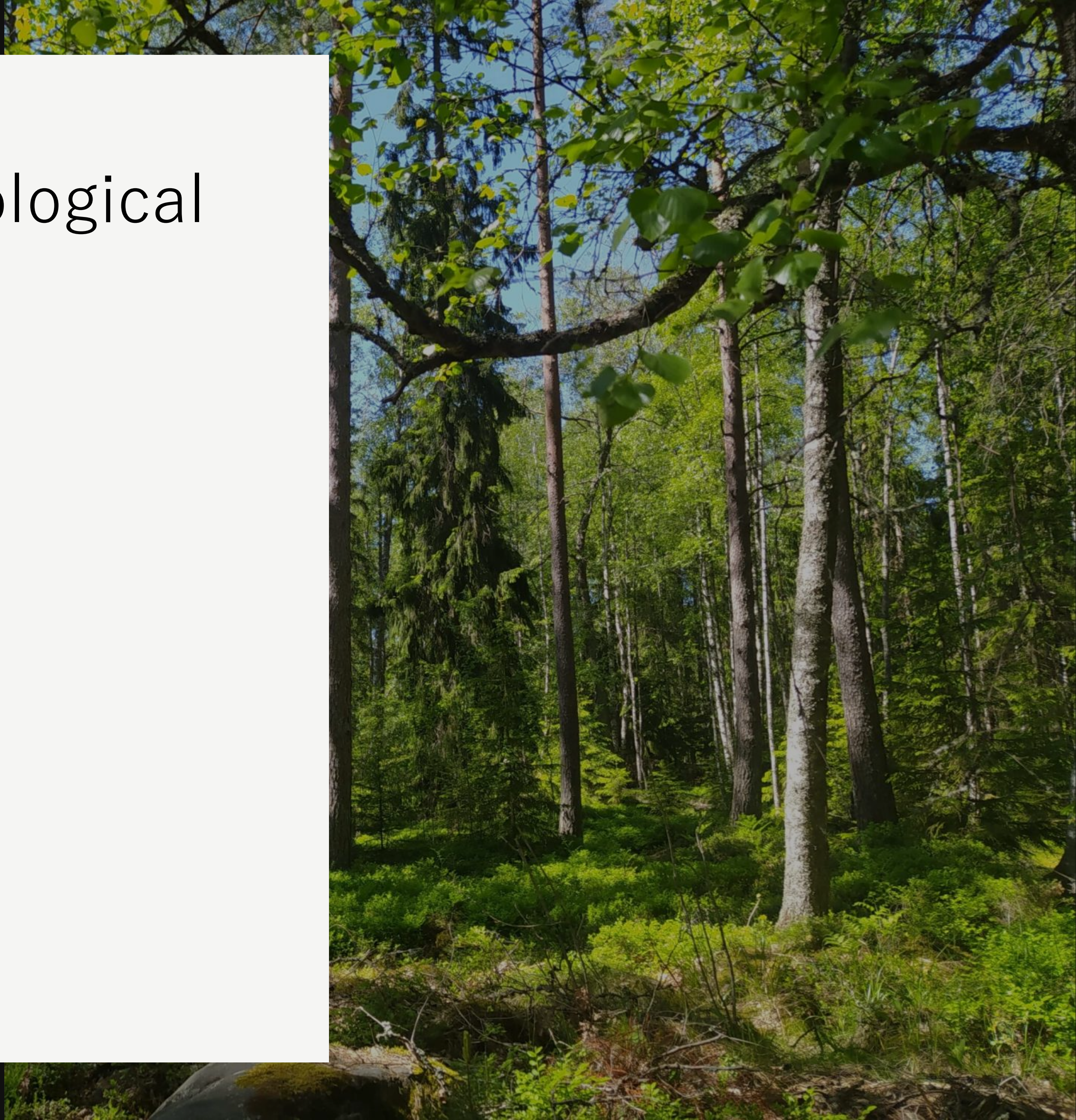
FW 599 Special Topics: Multivariate Analysis of Ecological Data in R

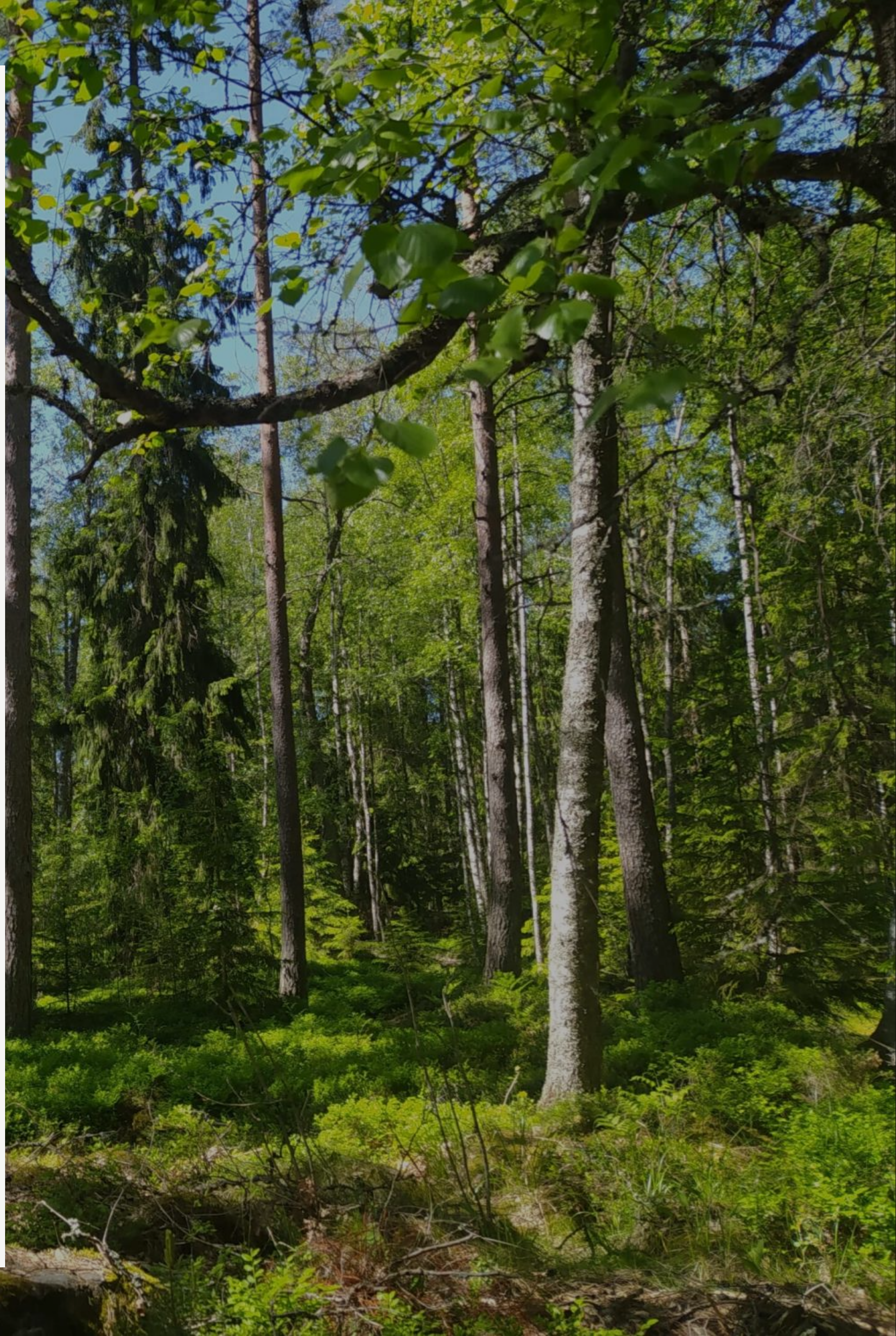
Lecture 2: Data Transformation and Standardization

Thursday, October 3, 2024



Lecture 2: Data Transformation and Standardization

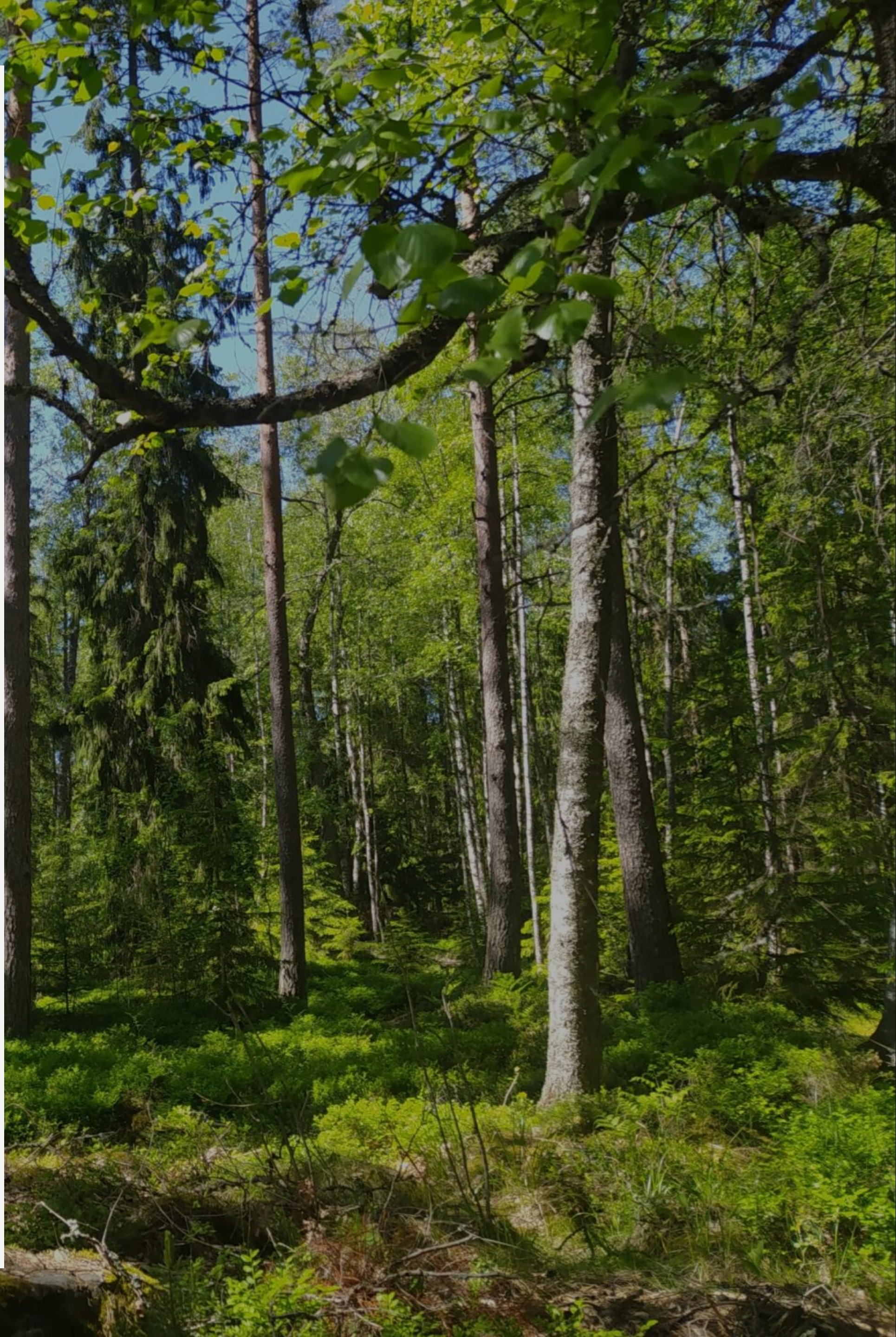
- Transformation
- Standardization
- Univariate Metrics of Ecological Diversity



Recap: Data Distributions

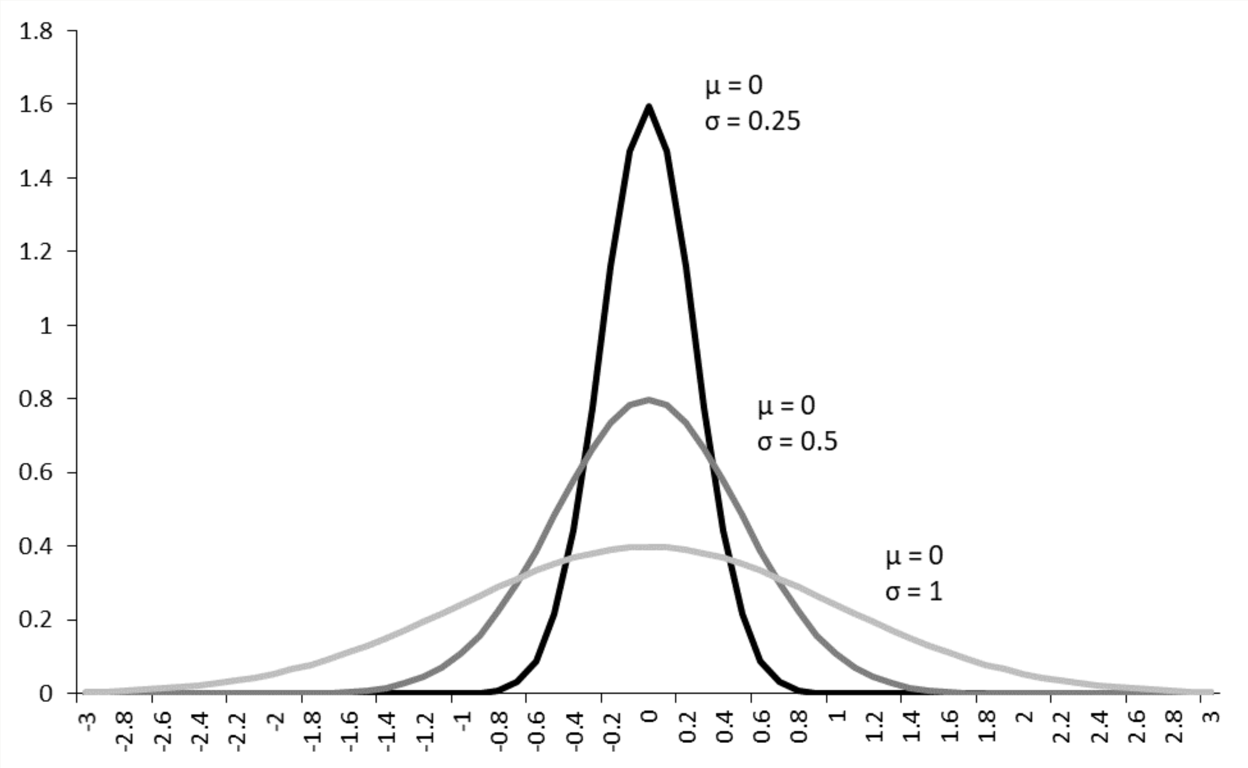
Exploratory Analysis: Data Distributions

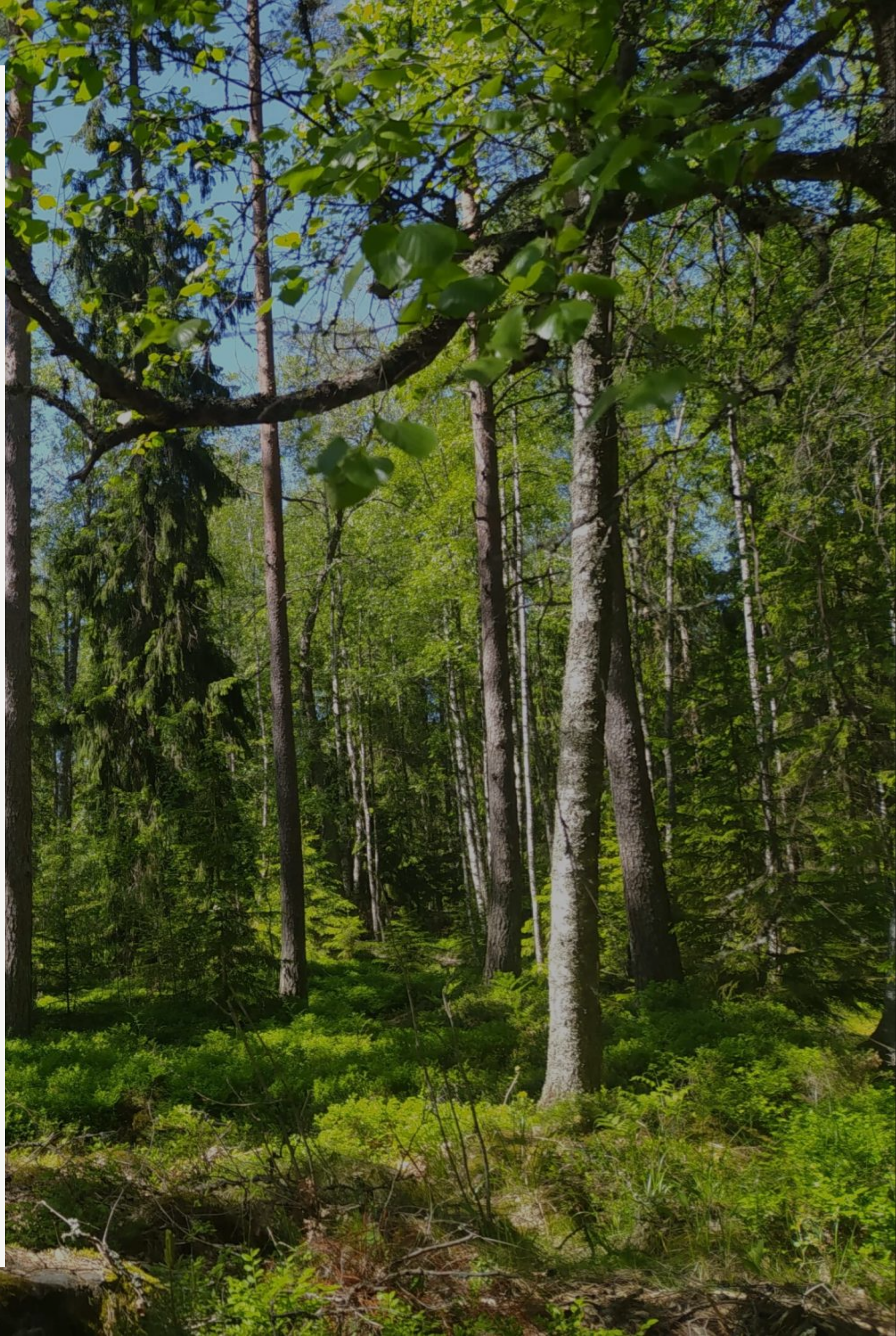
Distribution	**Characteristics**	**Suited For…**
Normal	Symmetrical, bell-shaped	Environmental variables, trait measurements
Poisson	Right-skewed, mean = variance	Integer/count data
Binomial	Can be symmetric or skewed	Presence/absence
Negative Binomial	Right-skewed, over-dispersed counts	Aggregated counts (i.e, N per unit)
Log-Normal	Right-skewed, log-transformed normal	Species abundance
Gamma	Right-skewed, flexible shape	Environmental variables
Beta	Flexible shapes, bounded on [0,1]	Proportional data
Uniform	Constant probability over interval	Indicative of complete randomness



Exploratory Analysis: Data Distributions

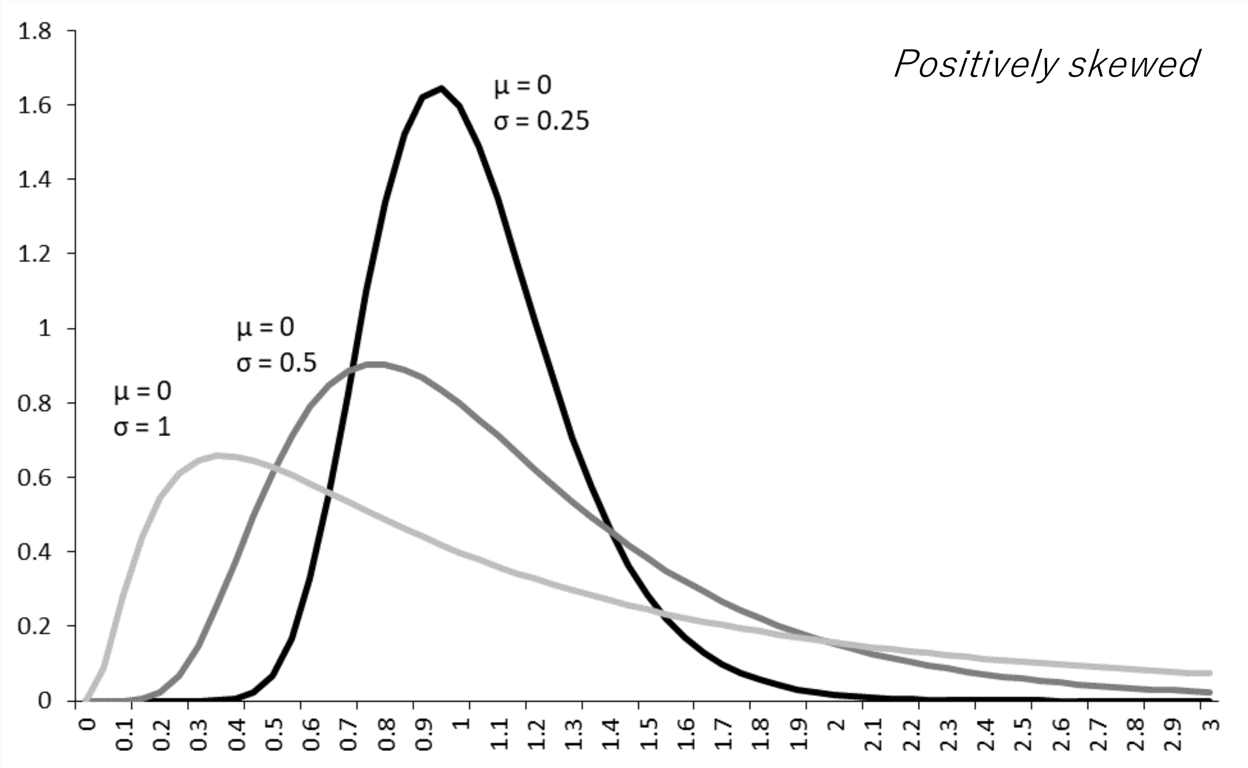
Normal/Log-Normal

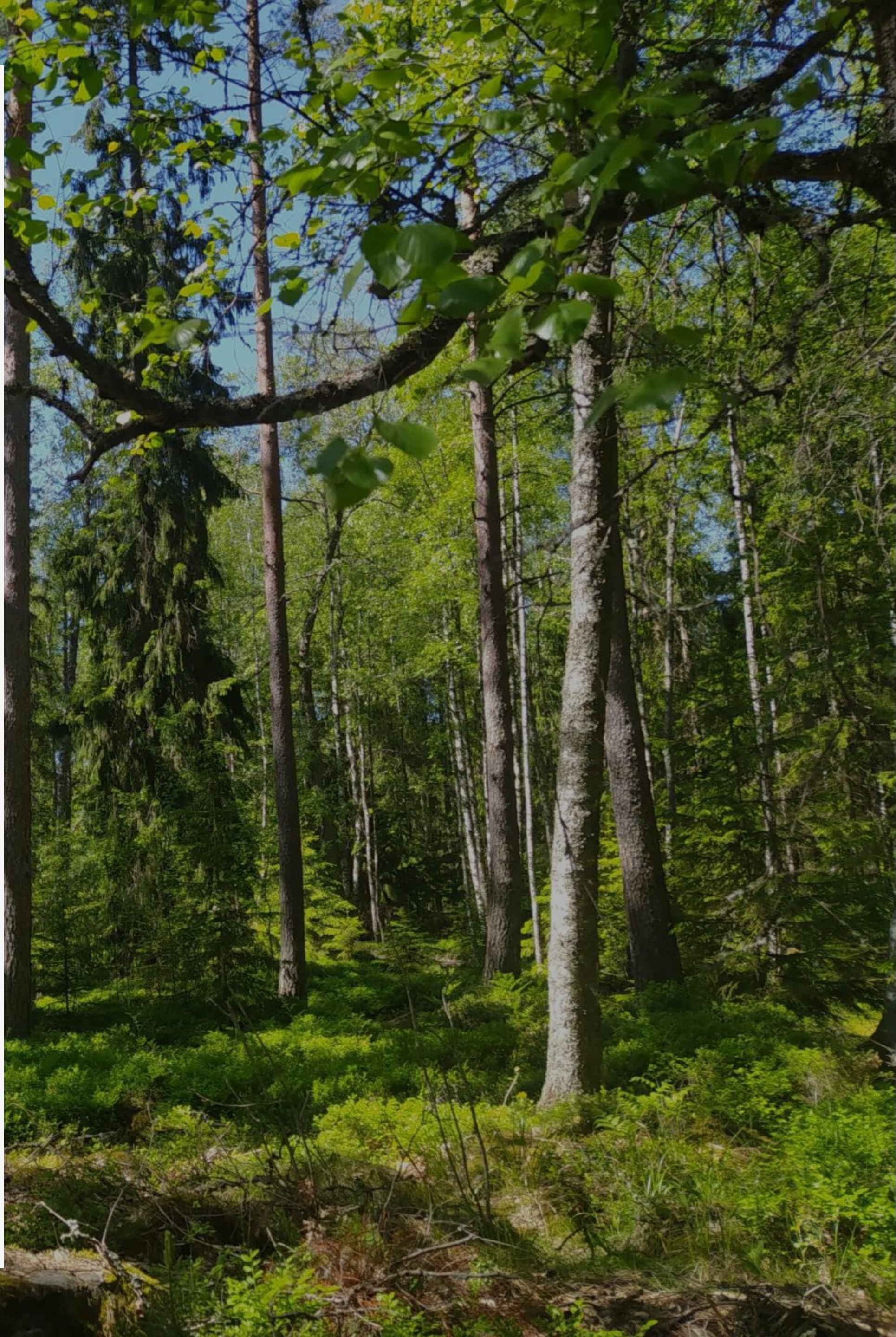




Exploratory Analysis: Data Distributions

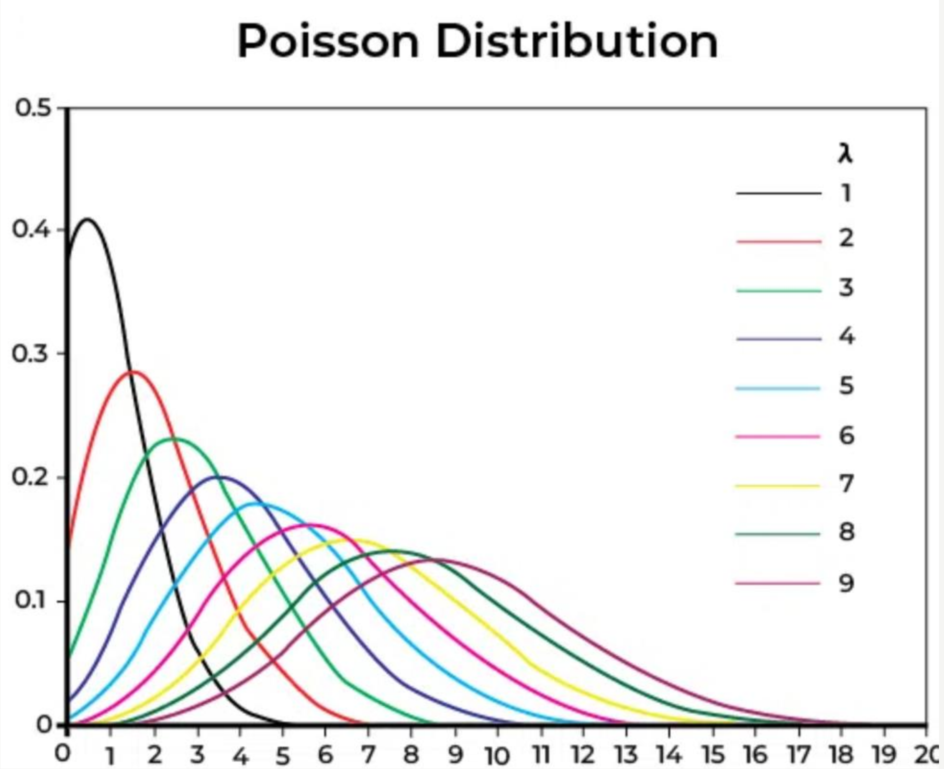
Normal/Log-Normal

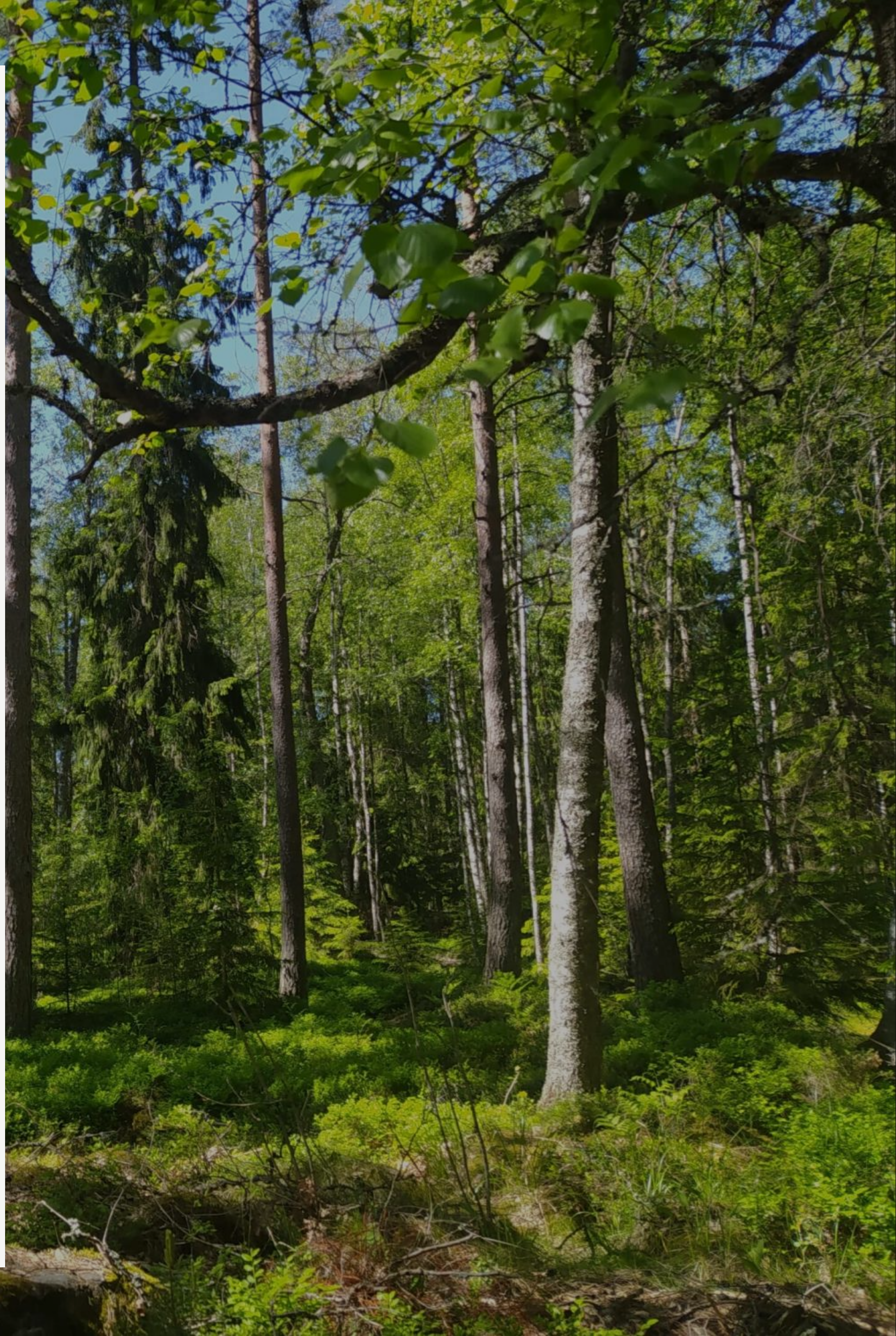




Exploratory Analysis: Data Distributions

Poisson – count data. *How many times is an event likely to occur over a given period/area?*





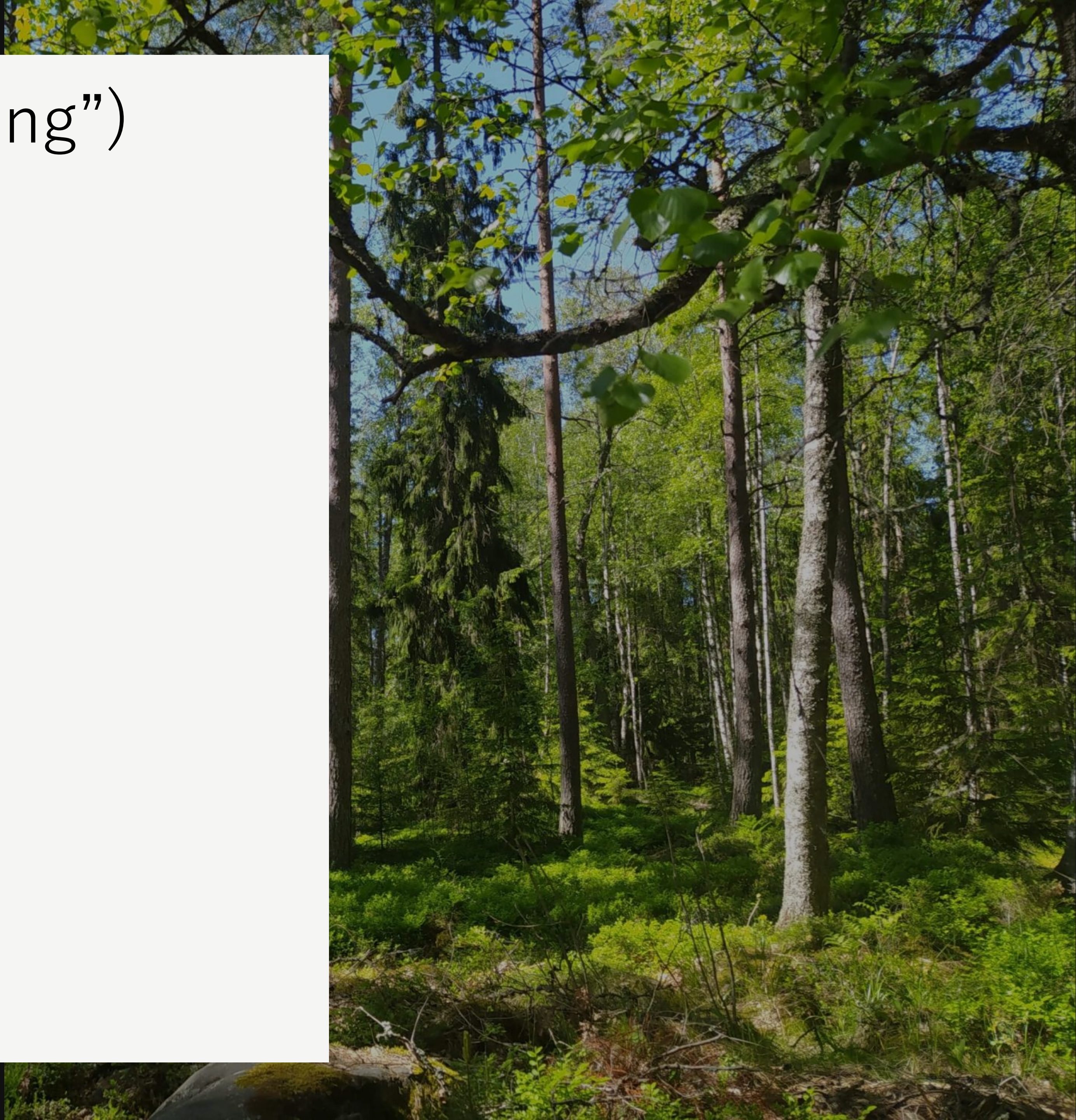
Exploratory Analysis: Homogeneity of Variance

Homoscedasticity means that the variances of the error terms are equal for all observations

The variance in a sampled population remains the same, regardless of the mean

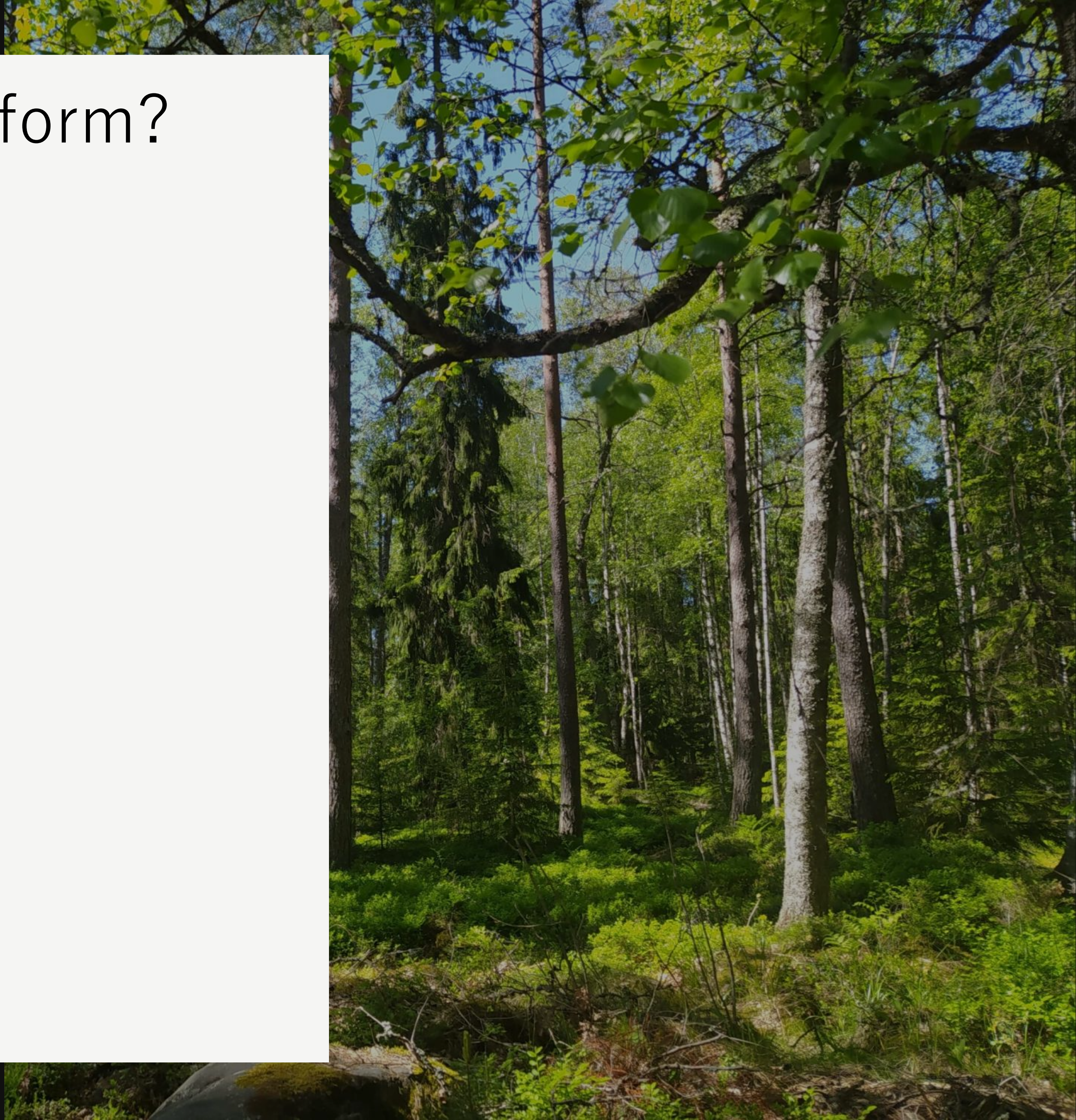
This means that Poisson distributed data are not homoscedastic!

Transformation (a.k.a “Coding”)



Transformation: Why Transform?

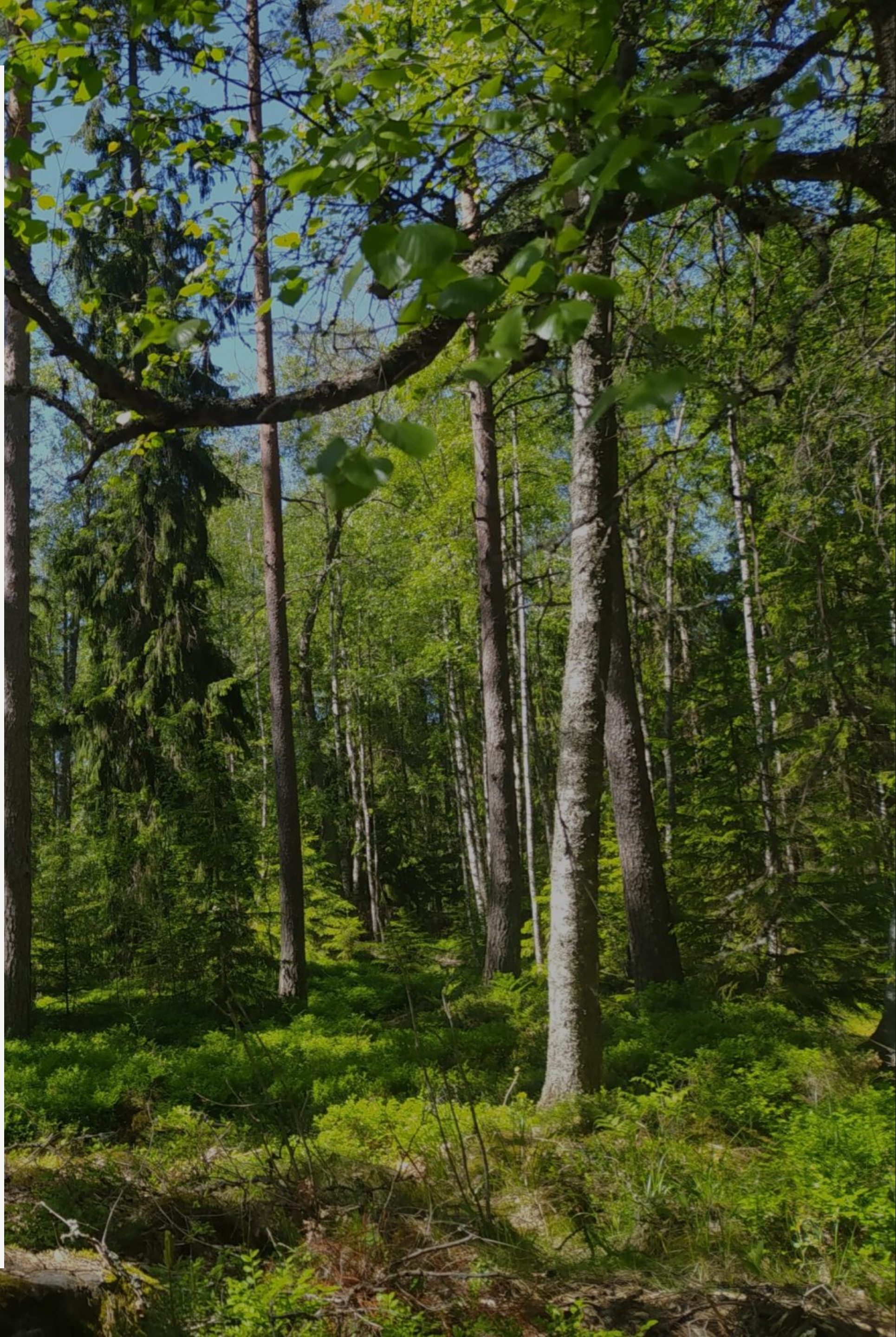
- Ensure normality
- Stabilize variance
- Handle outliers



Transformation: Why Transform?

- **Ensure normality**
- Stabilize variance
- Handle outliers

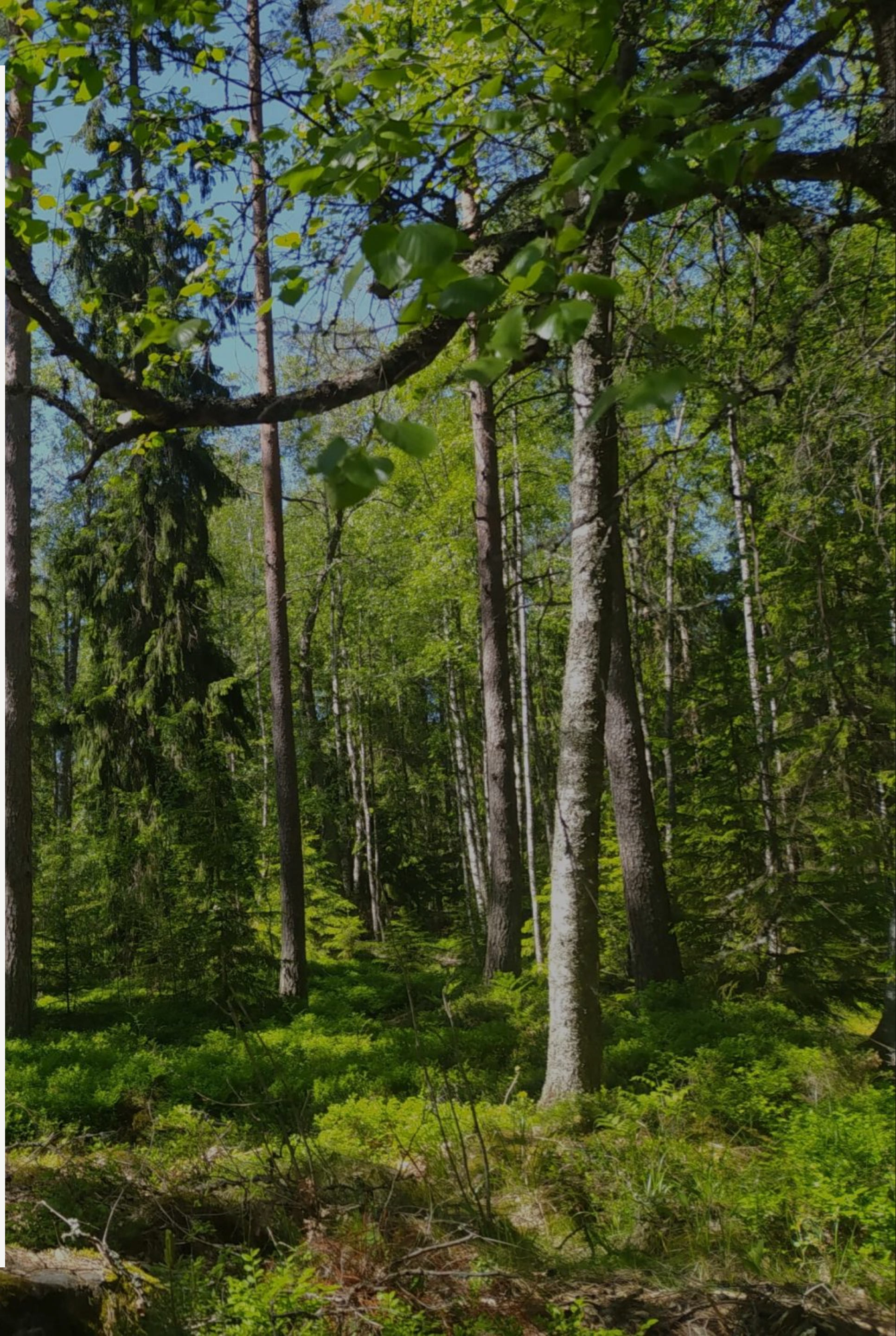
Some multivariate techniques (e.g., PCA, MANOVA) assume data are normally distributed.



Transformation: Why Transform?

- Ensure normality
- **Stabilize variance**
- Handle outliers

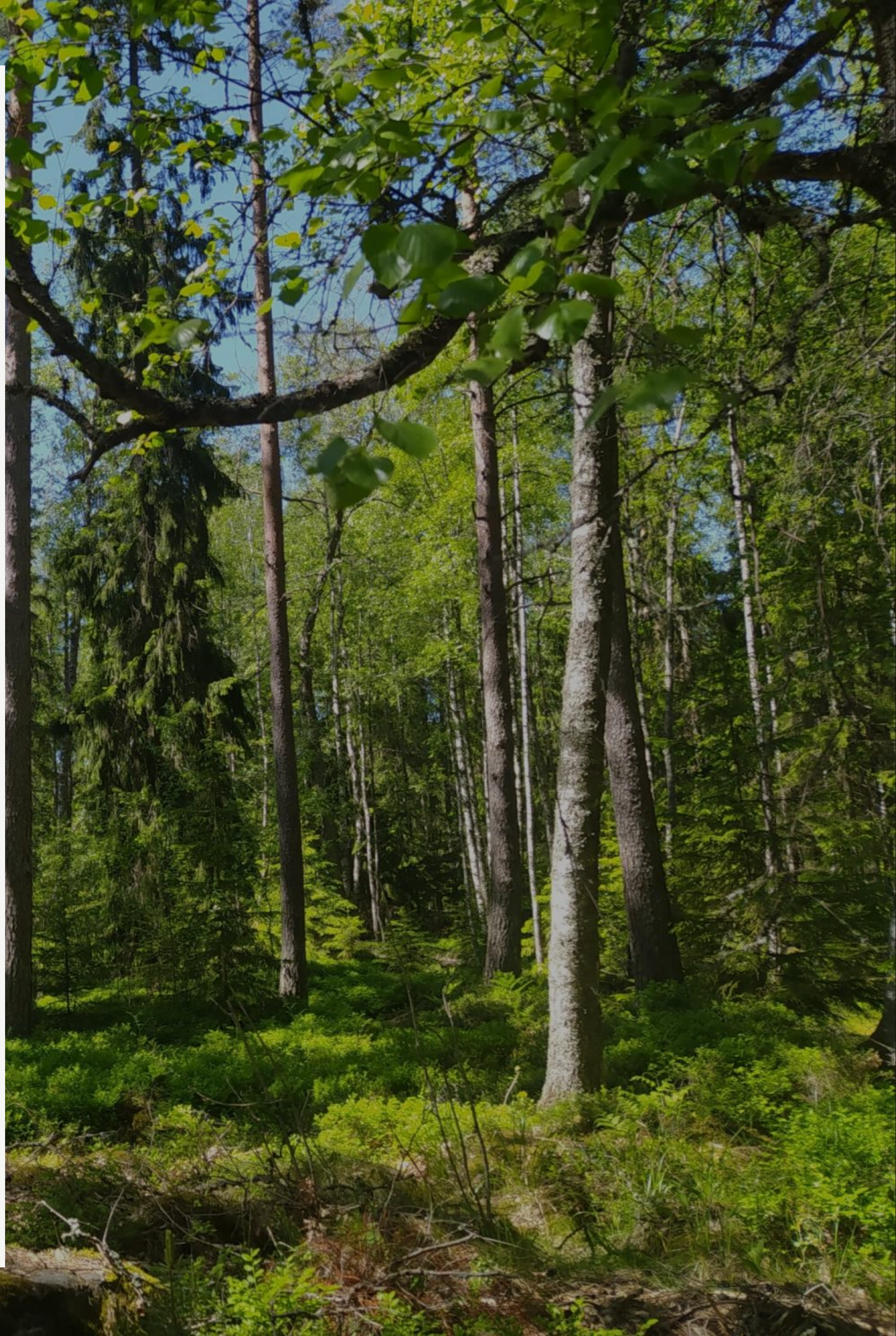
Like normality, homoscedasticity is a key assumption of many statistical methods.



Transformation: Why Transform?

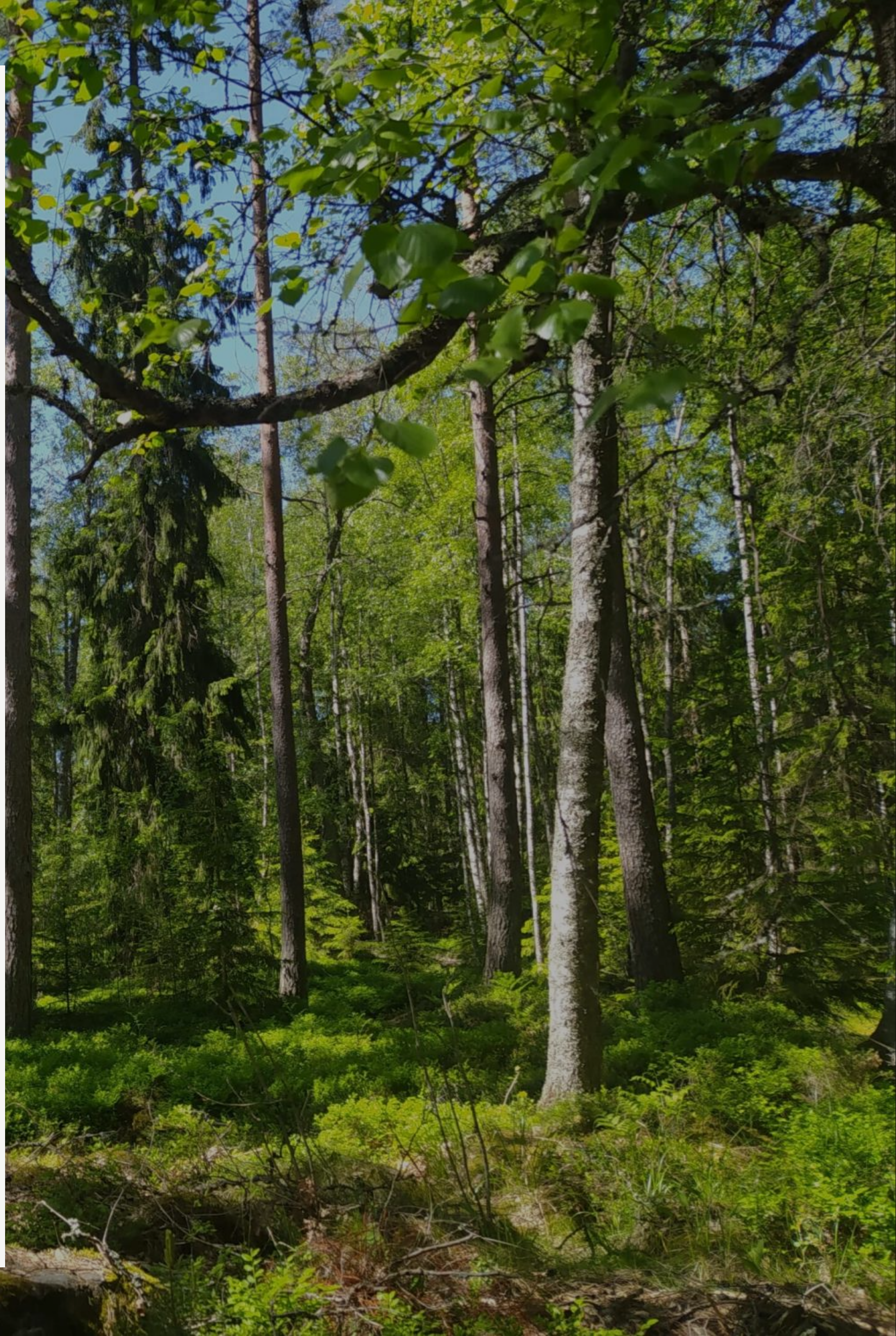
- Ensure normality
- Stabilize variance
- **Handle outliers**

Outliers can disproportionately influence the results of multivariate analyses



Transformation: Types of Transformations

- Logarithmic
- Square (N) Root
- Box-Cox
- Logit
- Angular/Arcsine
- Dummy Coding or Rank Transformation

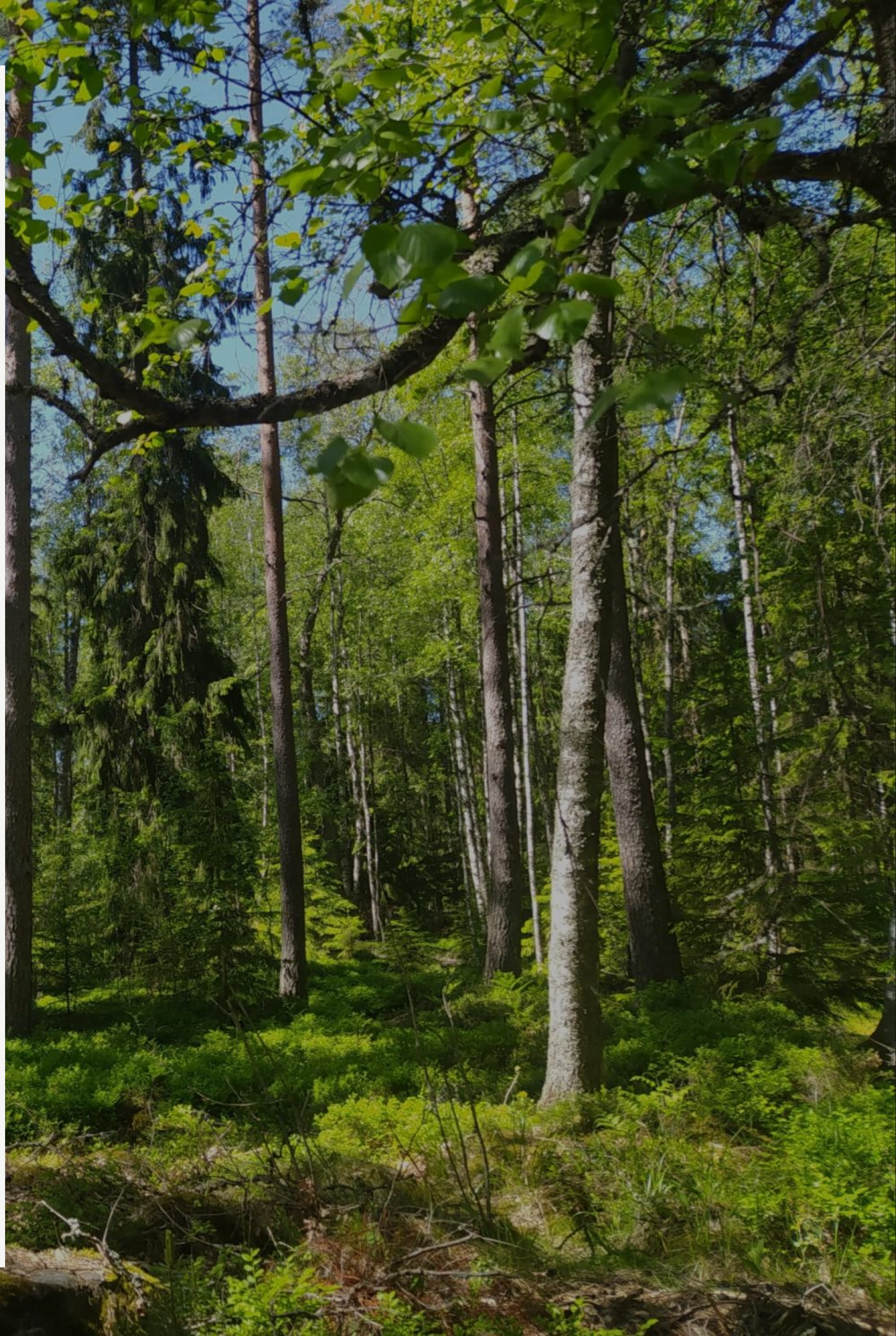


Transformation: Log Transformation

$$\mathbf{Y'} = \log(\mathbf{Y})$$

- Reduces right-skewness
- Stabilizes variance
- Linearizes exponential growth patterns

Use for: species abundance, biomass



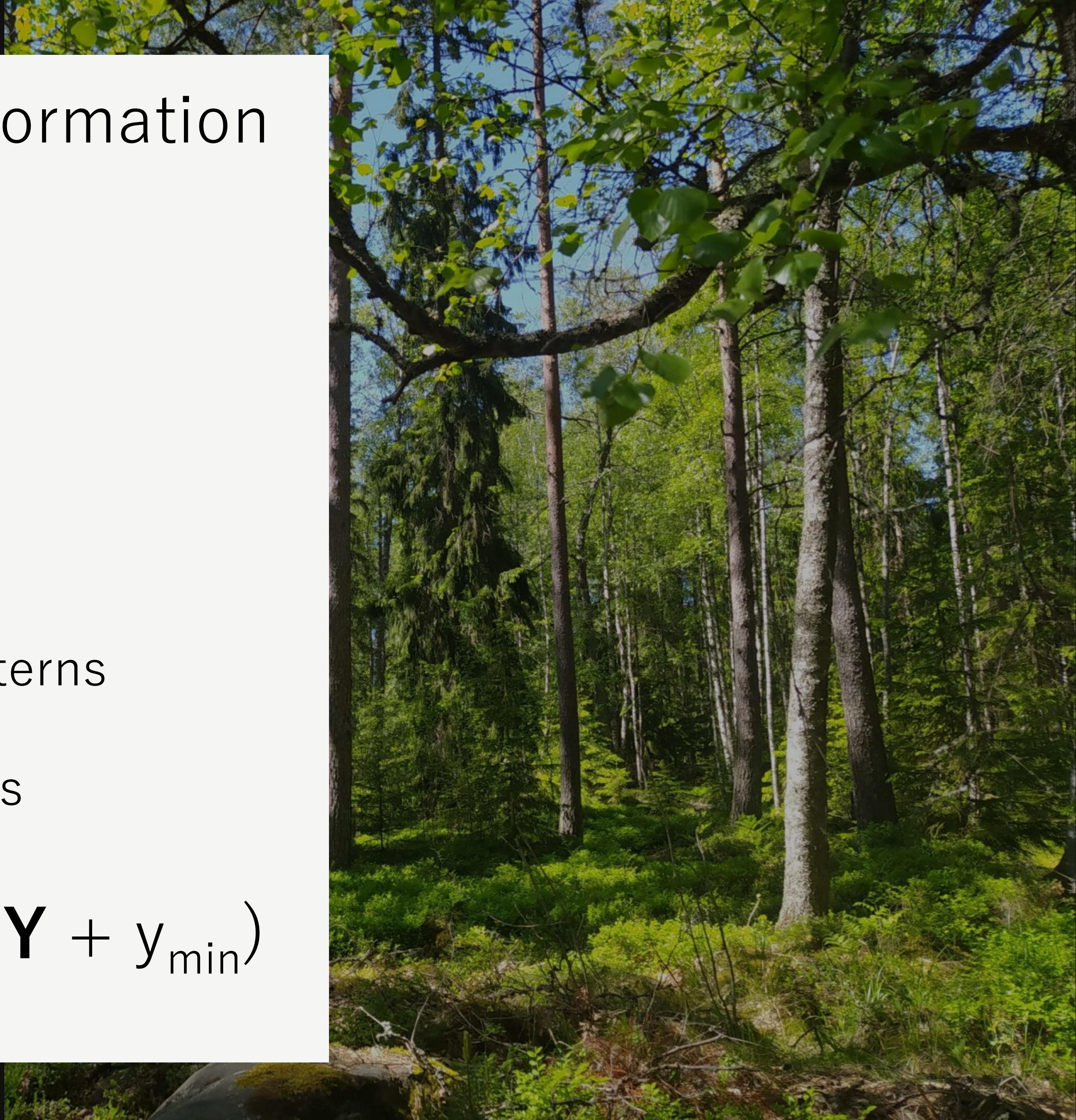
Transformation: Log Transformation

$$\mathbf{Y'} = \log(\mathbf{Y})$$

- Reduces right-skewness
- Stabilizes variance
- Linearizes exponential growth patterns

Use for: species abundance, biomass

$$\mathbf{Y'} = \log(\mathbf{Y} + 1) \qquad \mathbf{Y'} = \log(\mathbf{Y} + y_{min})$$

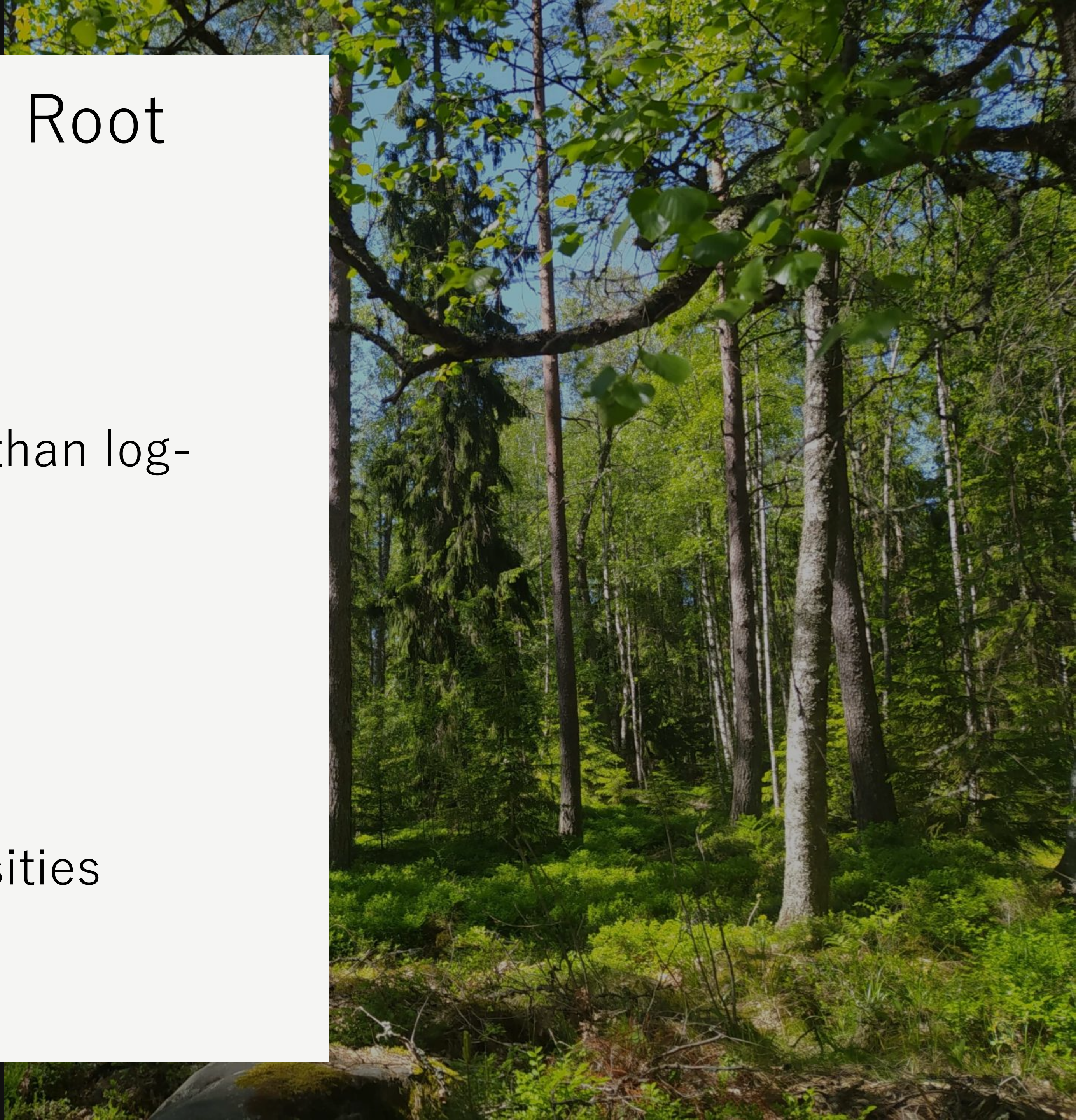


Transformation: Square (N) Root Transformation

$$\mathbf{Y'} = \mathbf{Y}^{1/2}$$

- Reduces right-skewness (less so than log-transforming)
- Stabilizes variance
- Handles count data well

Use for: count data, population densities



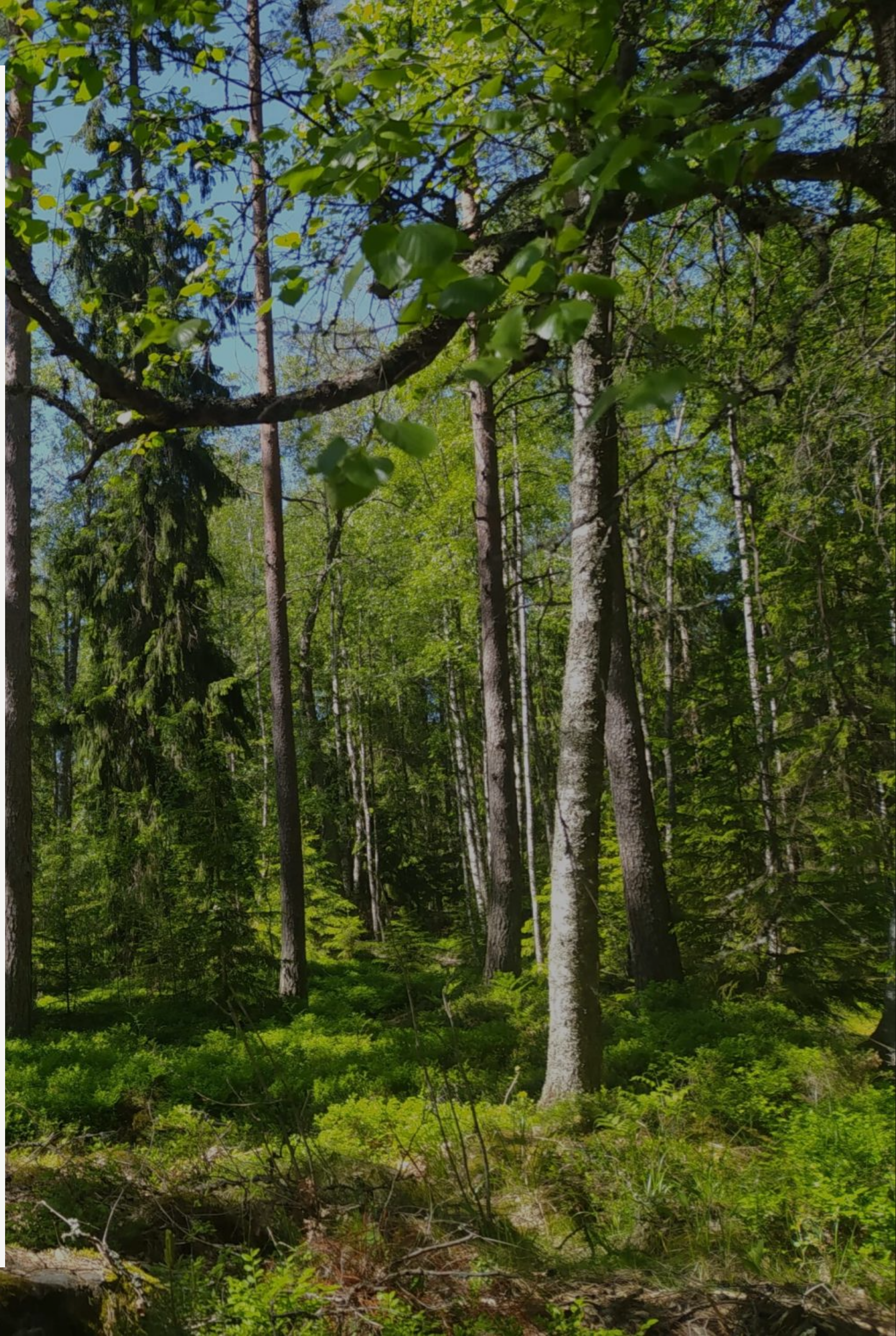
Transformation: Square (N) Root Transformation

$$\mathbf{Y'} = \mathbf{Y}^{1/N}$$

- Reduces right-skewness (less so than log-transforming)
- Stabilizes variance
- Handles count data well

Use for: count data, population densities

Use larger N for higher counts, greater skew



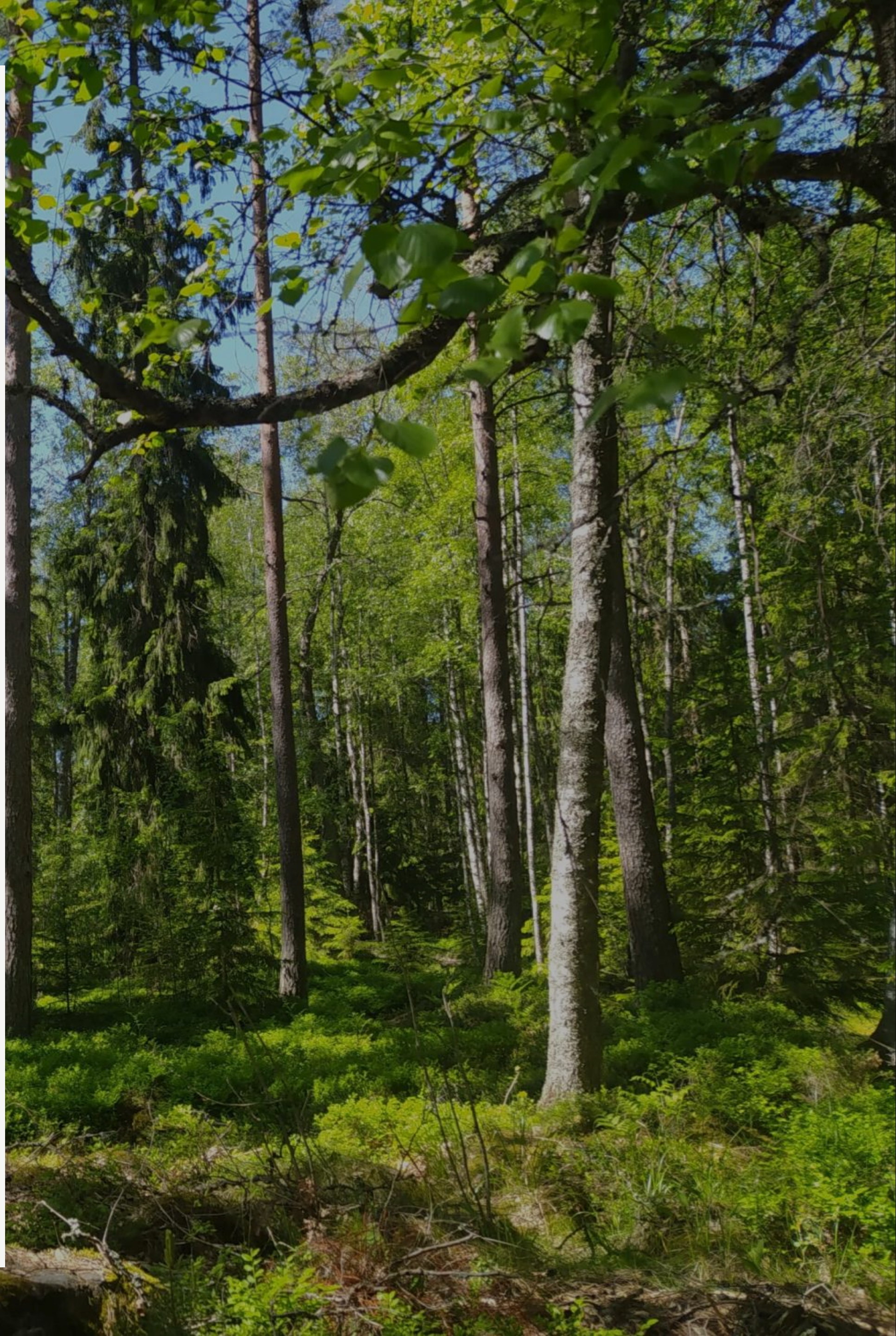
Transformation: Box-Cox Transformation

$$\mathbf{Y'} = \frac{\mathbf{Y}^{\lambda} - 1}{\lambda} \qquad \text{for } \lambda \neq 0$$

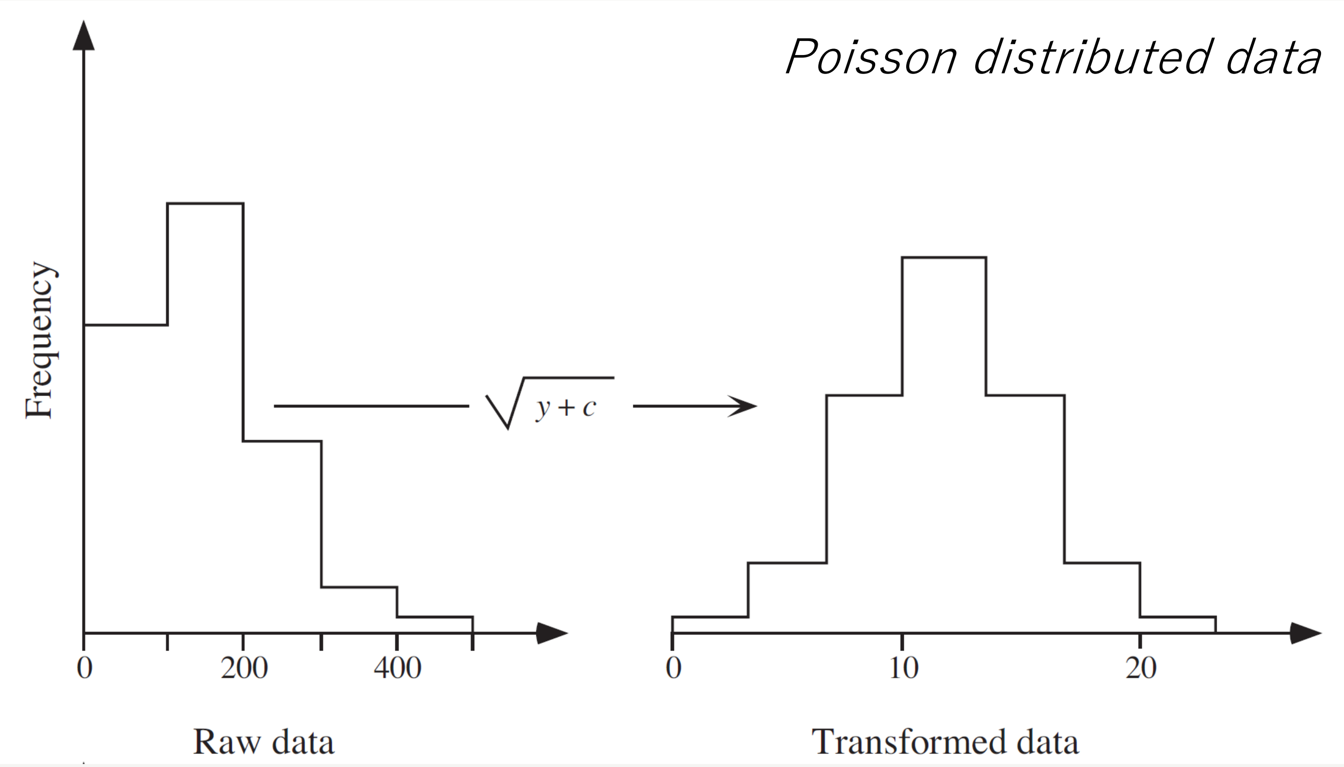
$$\mathbf{Y'} = \log(\mathbf{Y}) \qquad \text{for } \lambda = 0$$

- Normalizes data
- Stabilizes variance
- Can be used to identify optimal power transformation

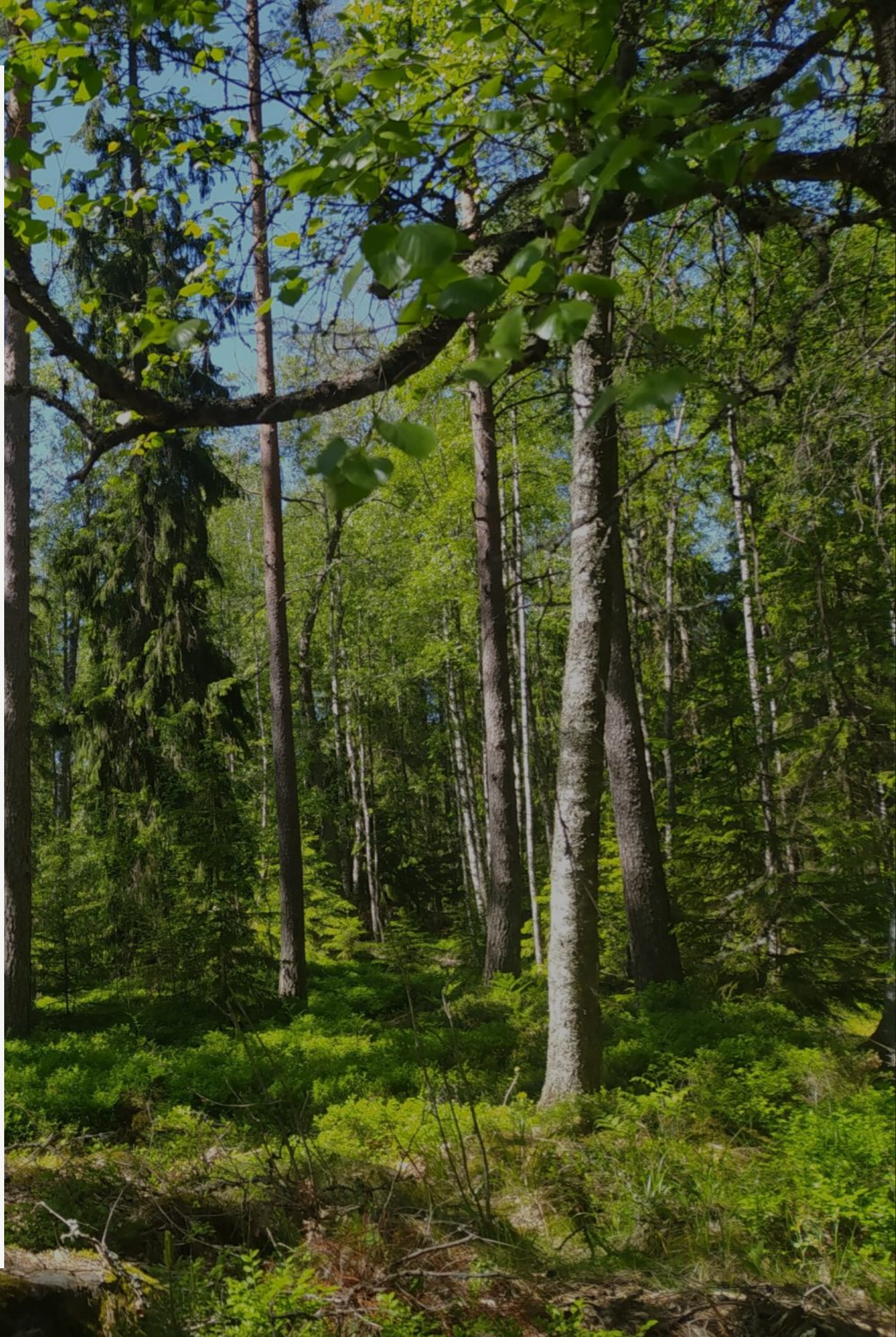
Use for: environmental measures (e.g., concentrations), biomass



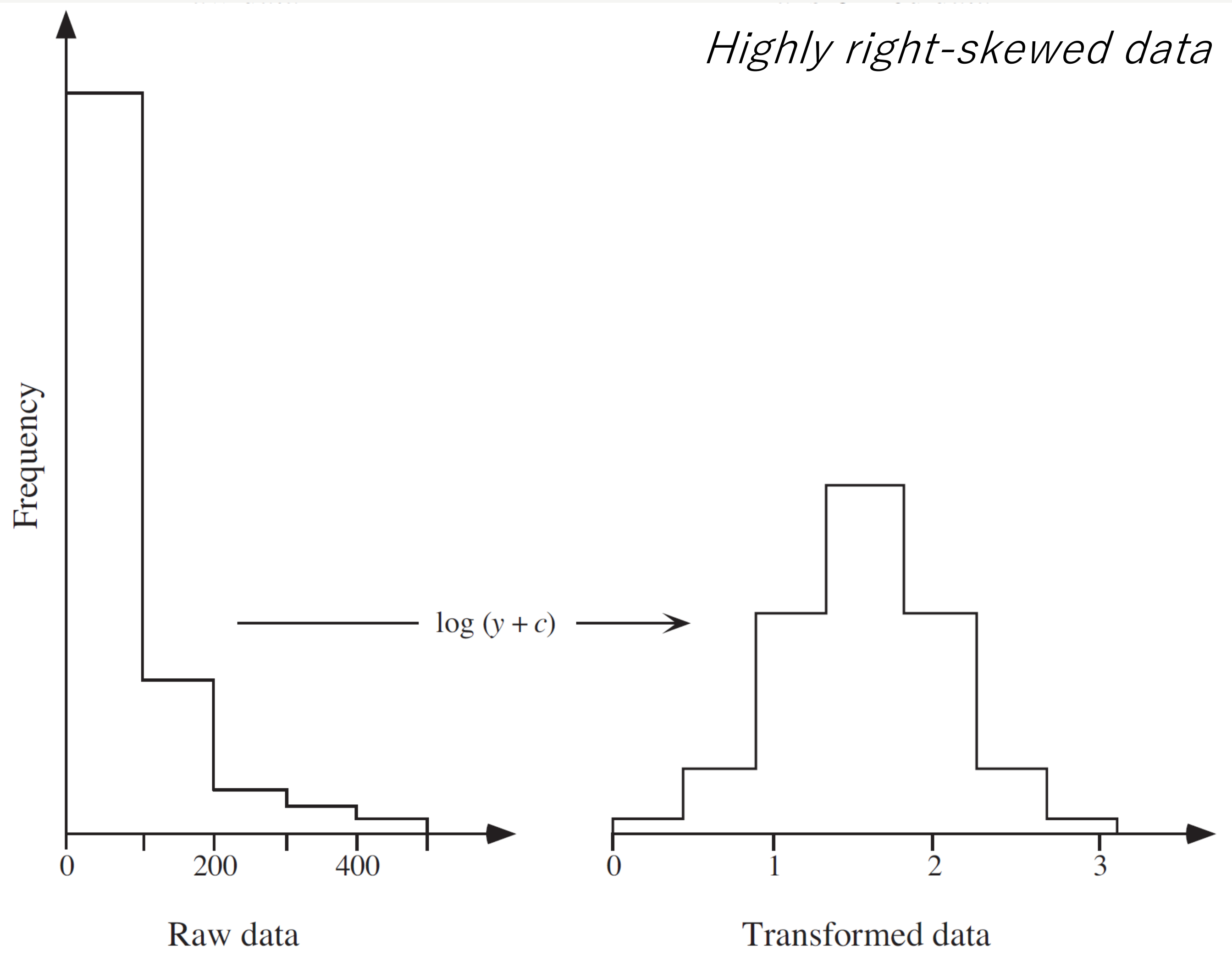
Transformation: Comparison



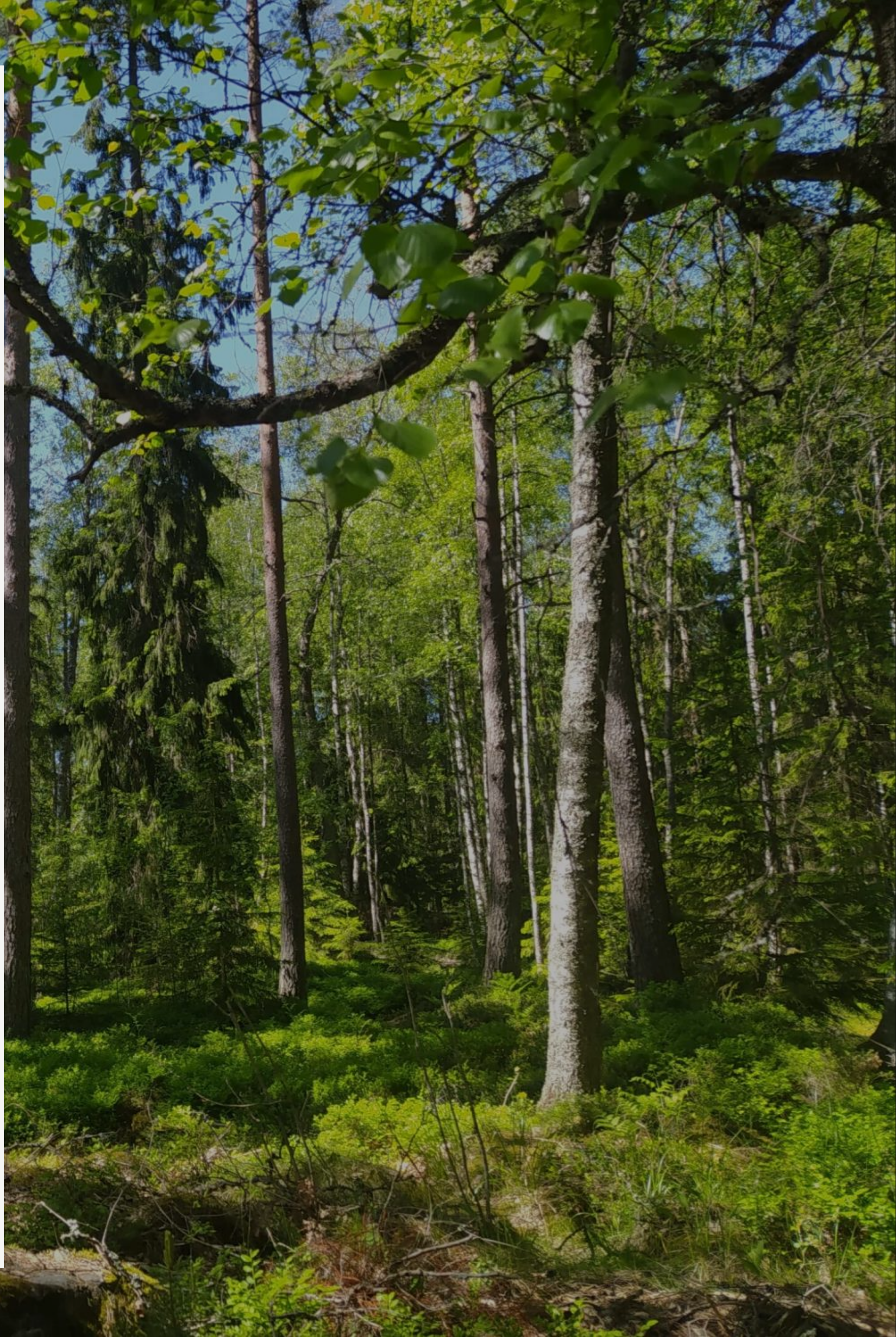
Legendre and Legendre



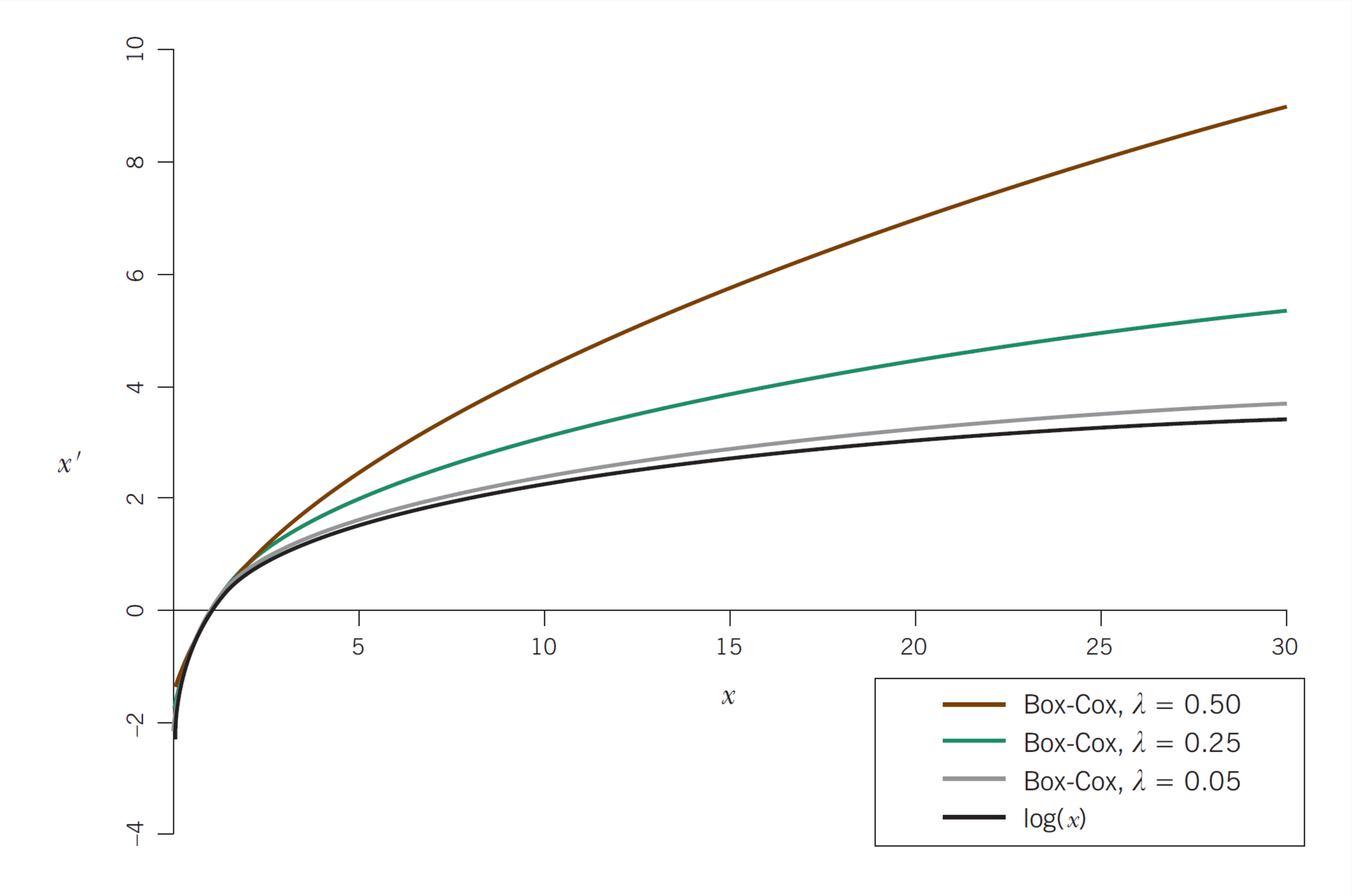
Transformation: Comparison



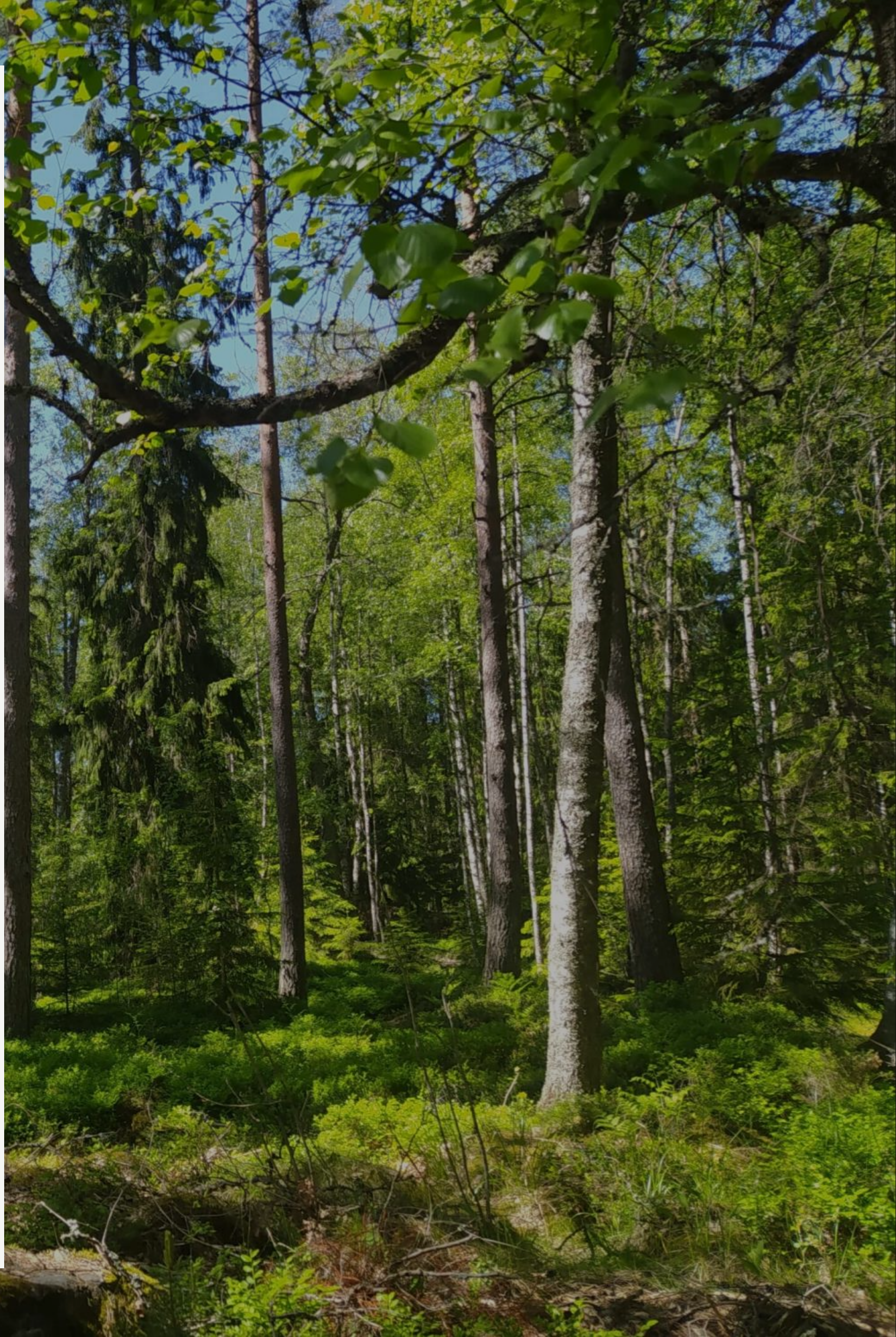
Legendre and Legendre



Transformation: Comparison



Greenacre and Primicerio

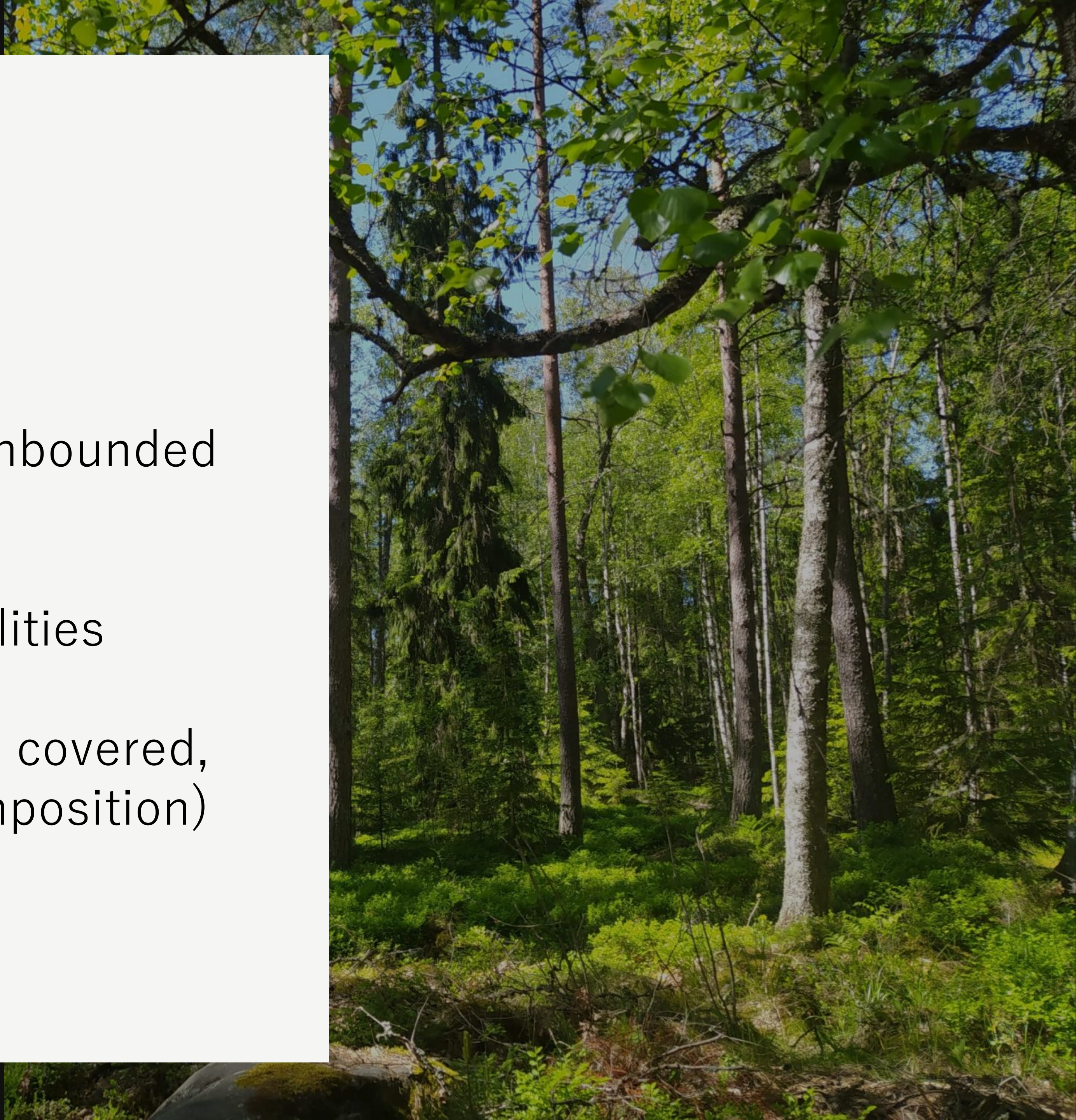


Transformation: Logit Transformation

$$\mathbf{Y'} = \log\left(\frac{\mathbf{Y}}{\mathbf{1 - Y}}\right)$$

- Transforms proportional data to unbounded data
- Normalizes proportions or probabilities

Use for: proportional data (e.g., area covered, survival rates, species or dietary composition)

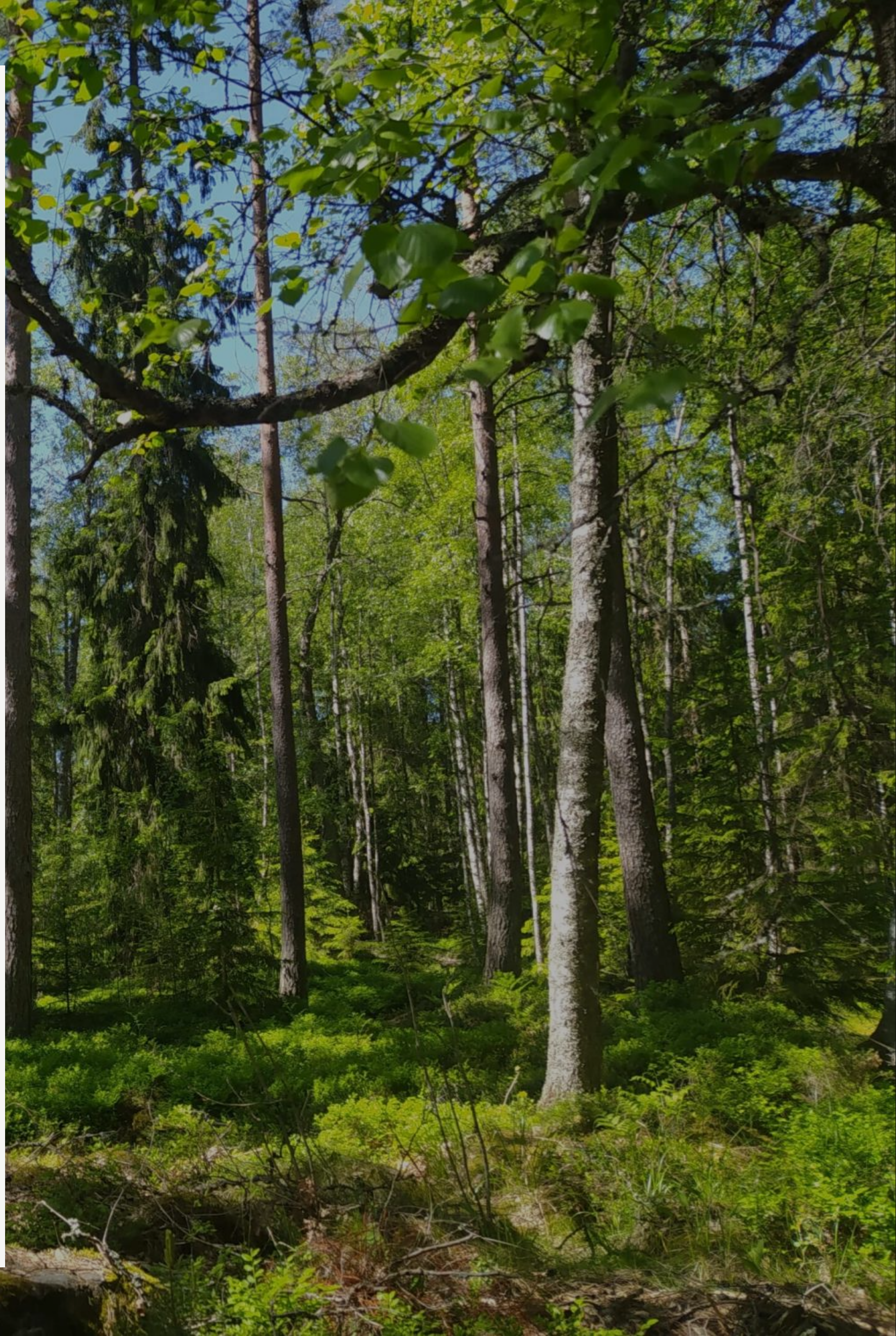


Transformation: Angular/Arcsine Transformation

$$\mathbf{Y'} = \arcsin(\sqrt{\mathbf{Y}})$$

- Transforms proportional data to standardize variance

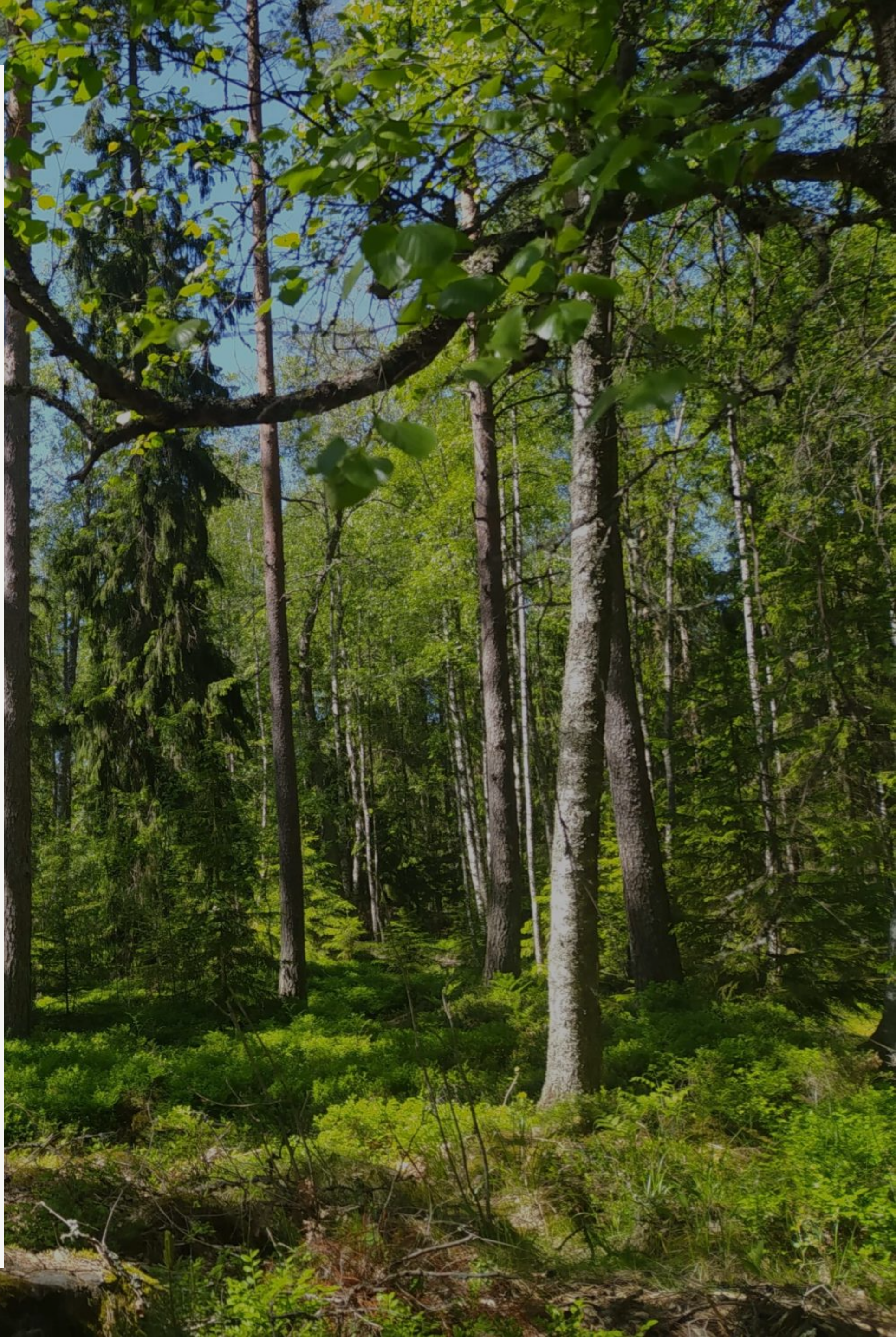
Use for: proportional data (e.g., area covered, survival rates, species or dietary composition)



Transformation: Dummy Coding

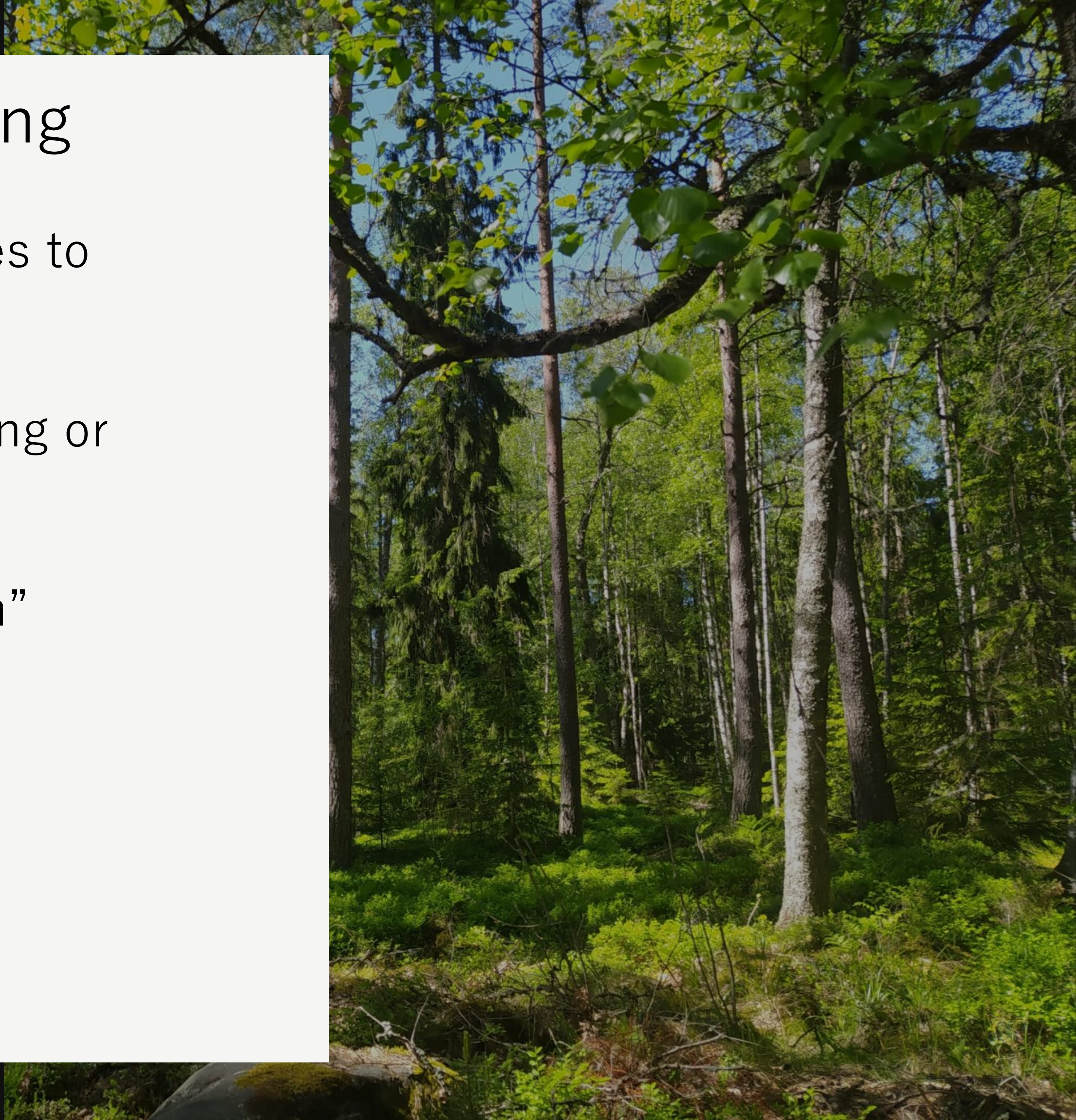
- Transforms categorical variables into numerical format
- Allows for inclusion of categorical data in statistical models
- Each category of the categorical variable is represented by a separate binary variable

Species	Species B	Species C
A	0	0
B	1	0
C	0	1
A	0	0
B	1	0



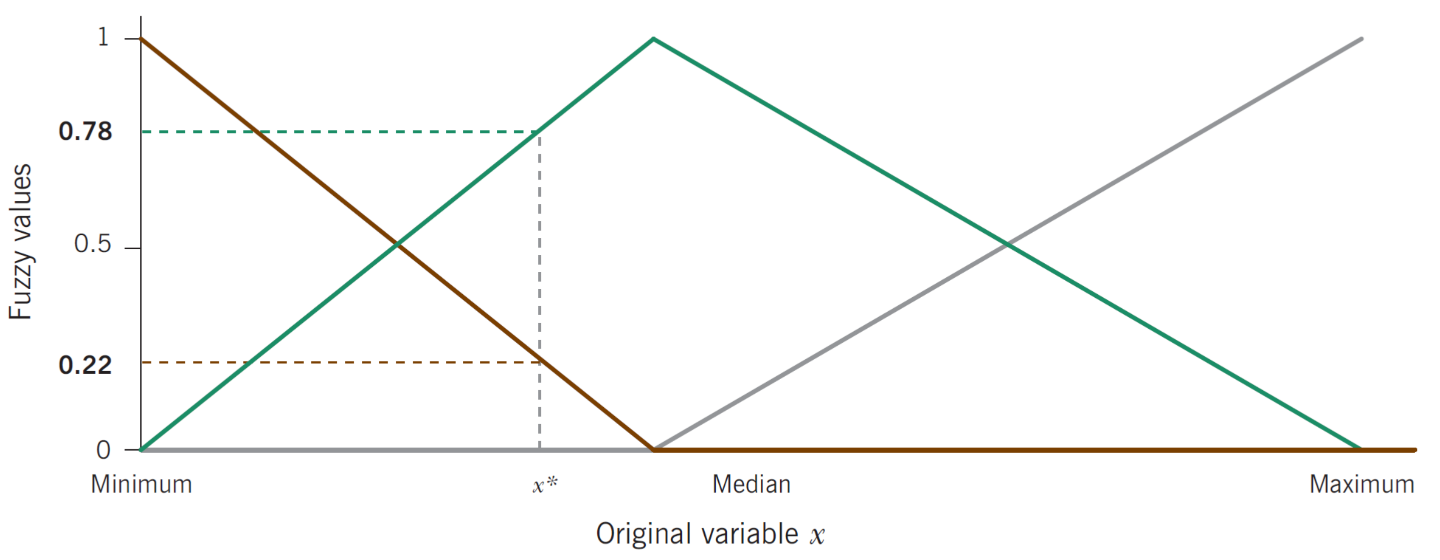
Transformation: Fuzzy Coding

- Assigns partial membership values to categorical variables
- Ideal for scenarios with overlapping or ambiguous categories
- Relies on a “membership function”

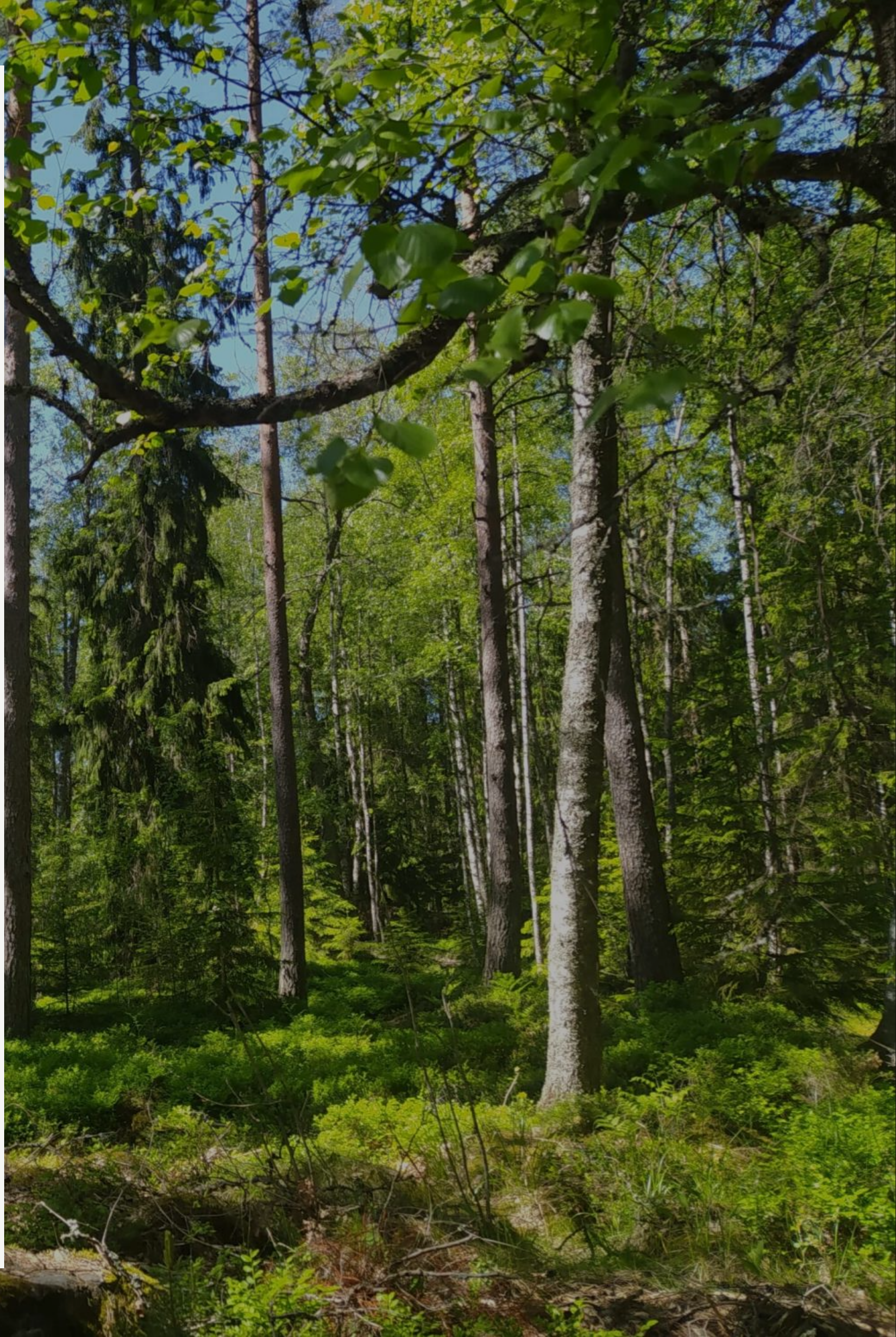


Transformation: Fuzzy Coding

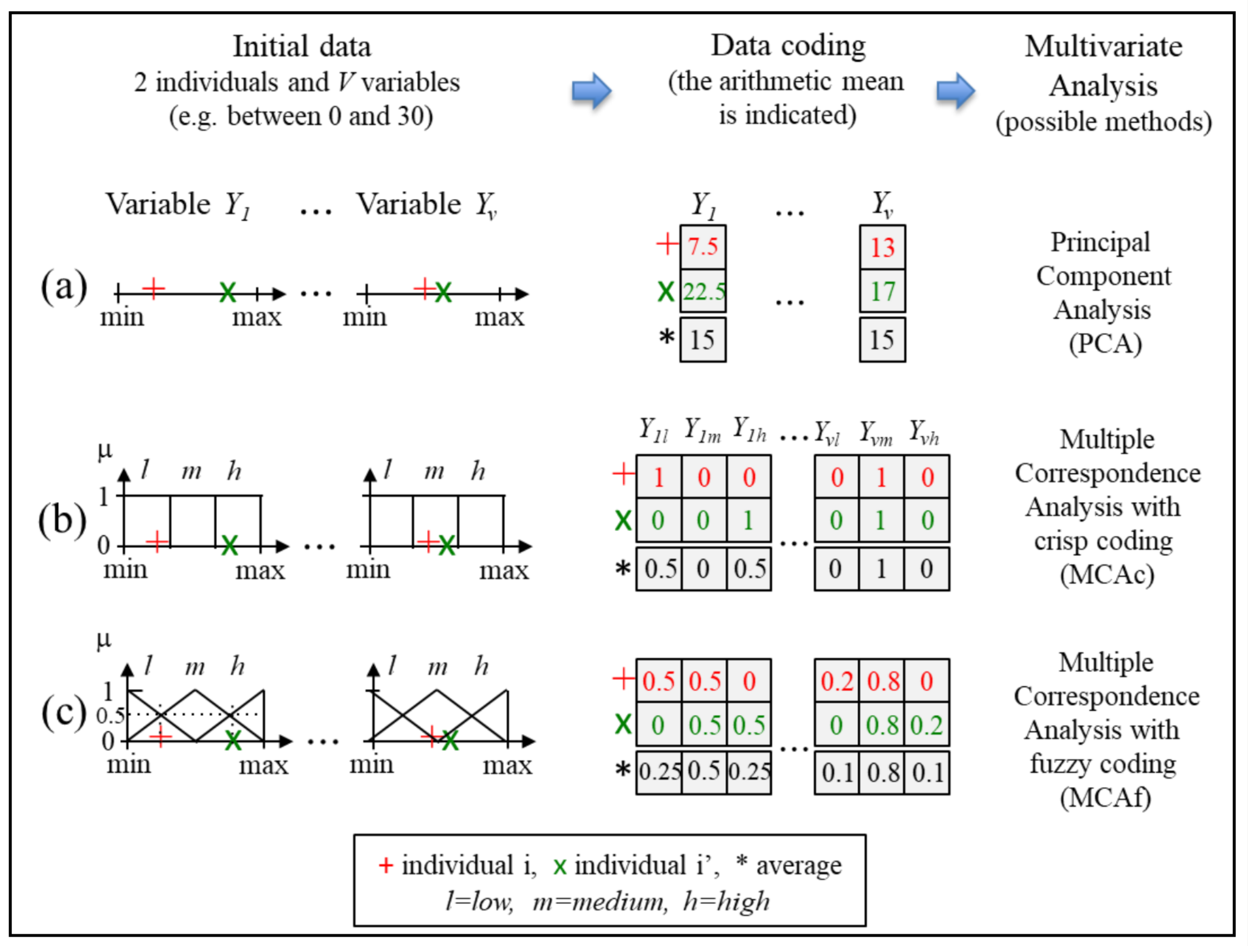
e.g., “low,” “medium,” and “high”



Greenacre and Primicerio



Transformation: Fuzzy Coding

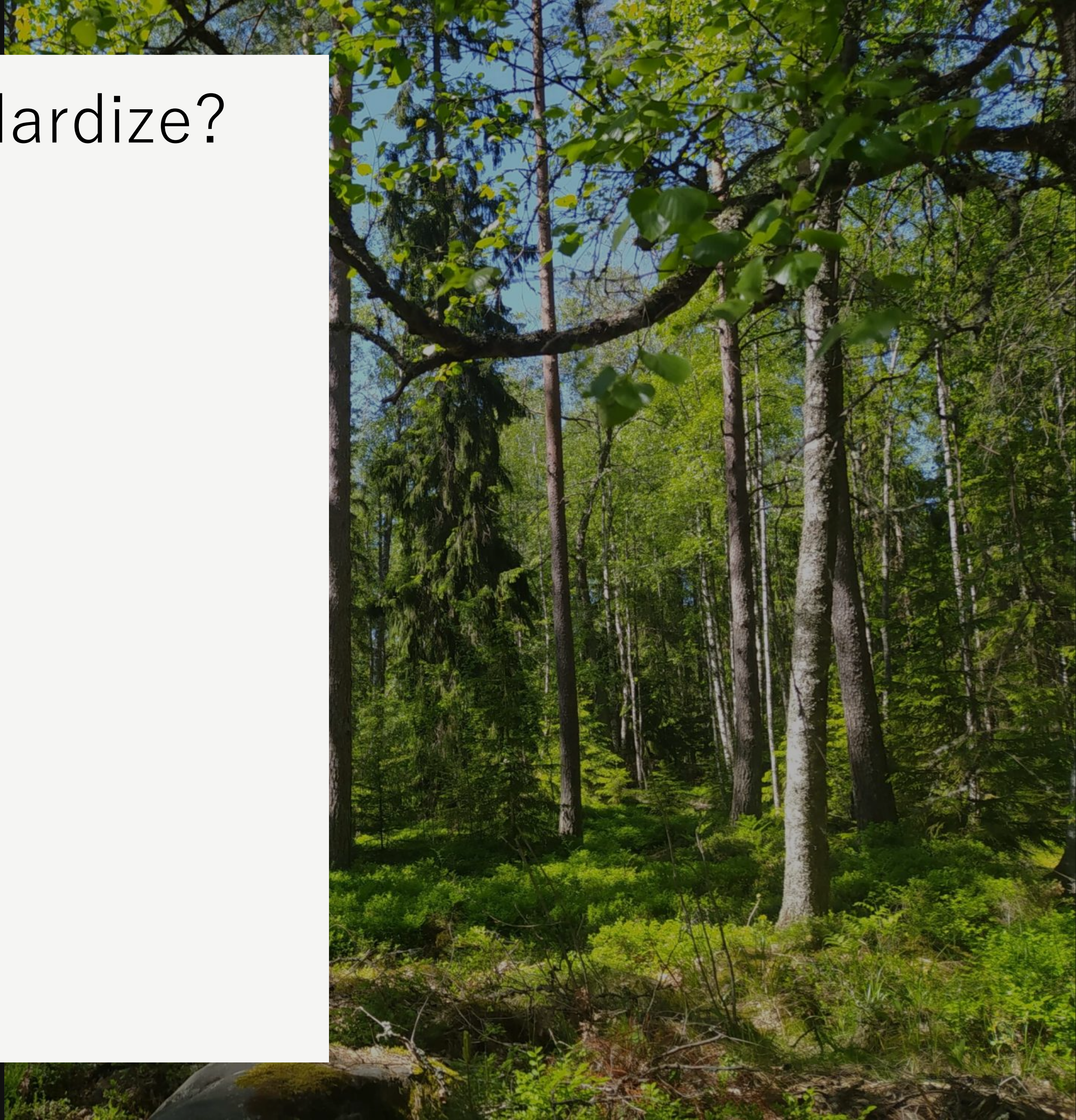


Loslever et al. 2019

Standardization

Standardization: Why Standardize?

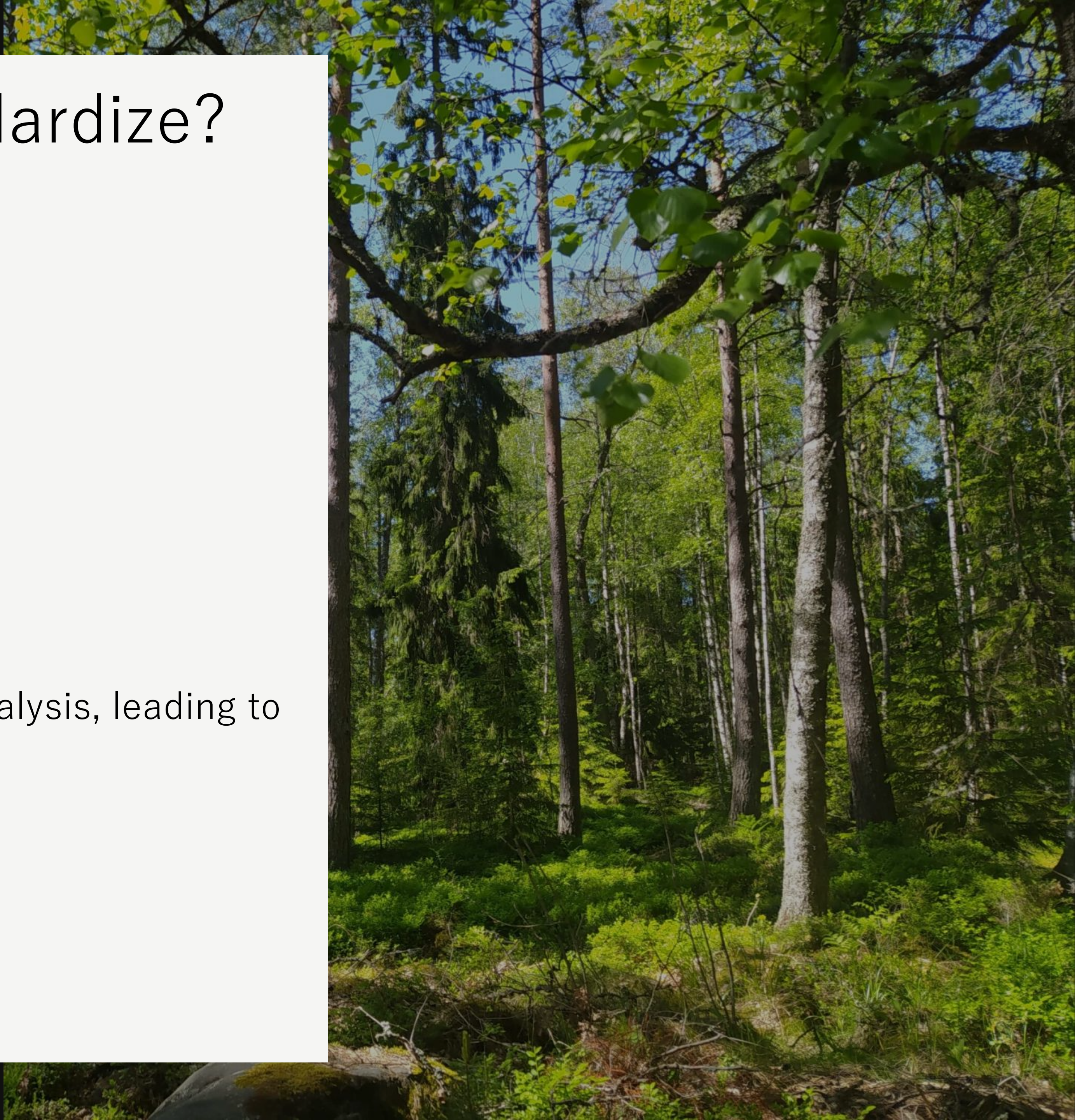
- Ensure comparability
- Equalize variable weighting
- Reduce impact of collinearity
- Improve distance calculations



Standardization: Why Standardize?

- **Ensure comparability**
- Equalize variable weighting
- Reduce impact of collinearity
- Improve distance calculations

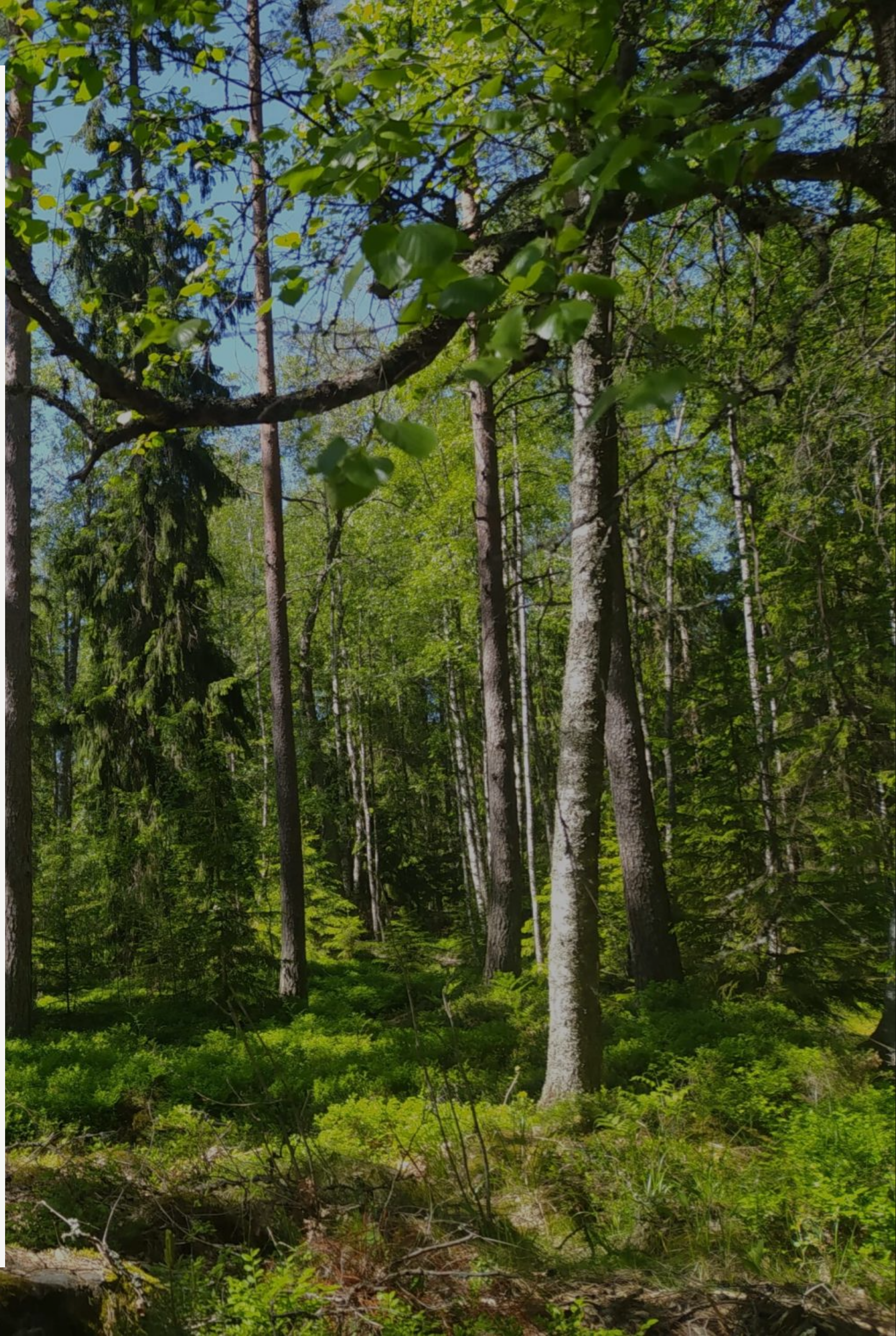
Variables with larger scales can dominate the analysis, leading to biased results



Standardization: Why Standardize?

- Ensure comparability
- **Equalize variable weighting**
- Reduce impact of collinearity
- Improve distance calculations

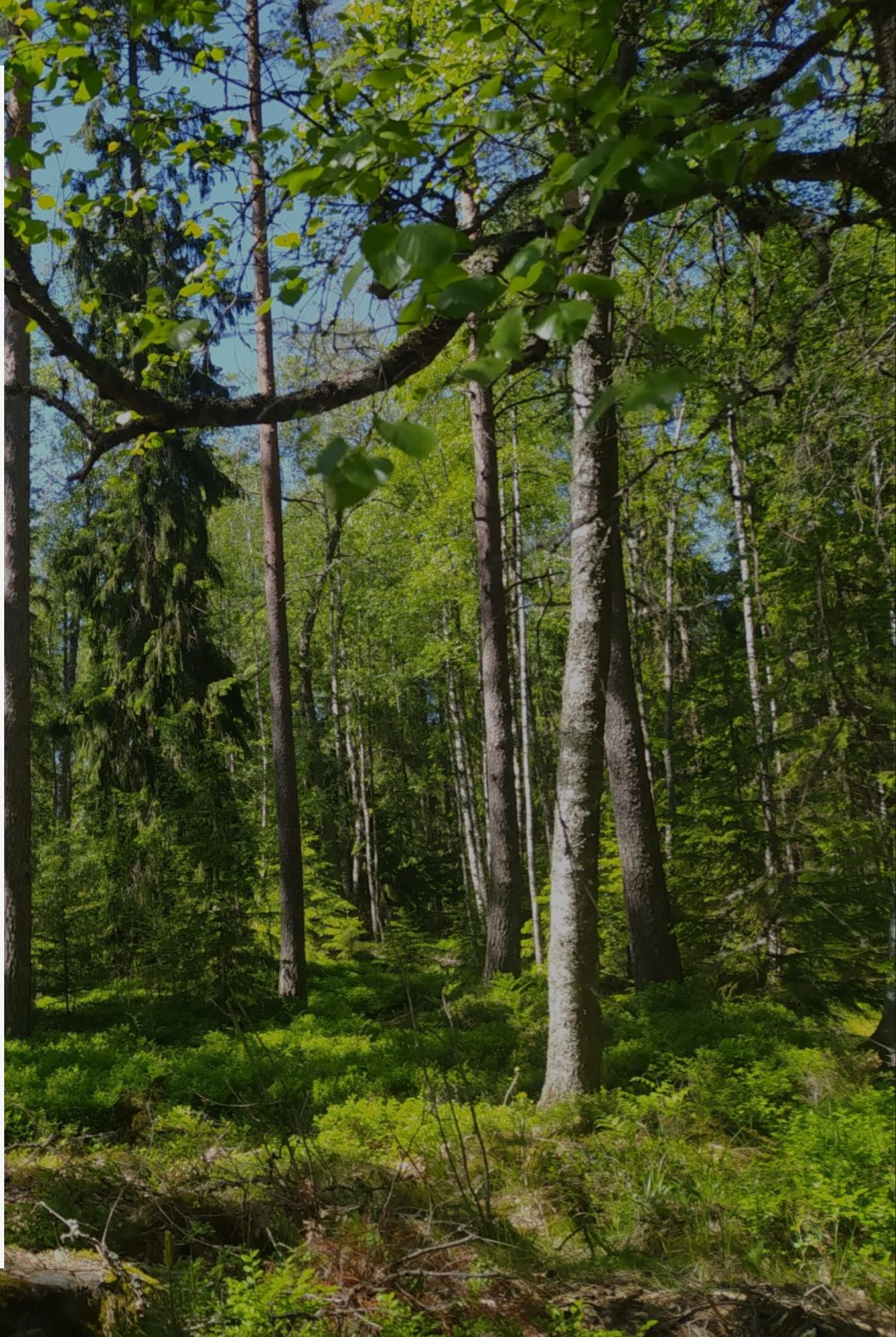
Standardization ensures that each variable is given equal importance in the analysis



Standardization: Why Standardize?

- Ensure comparability
- Equalize variable weighting
- **Reduce impact of collinearity**
- Improve distance calculations

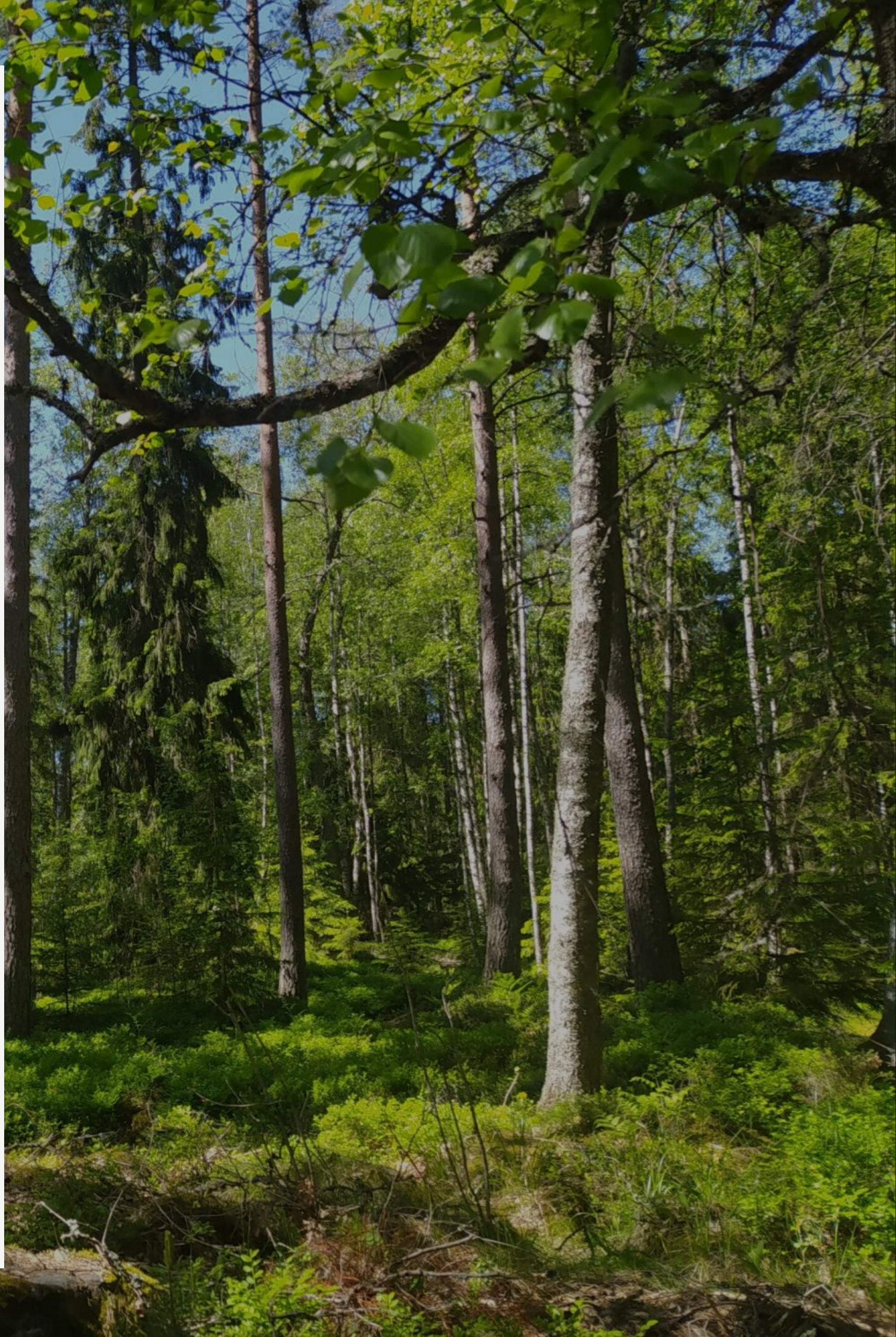
Collinearity between variables can inflate standard errors and lead to unstable estimates



Standardization: Why Standardize?

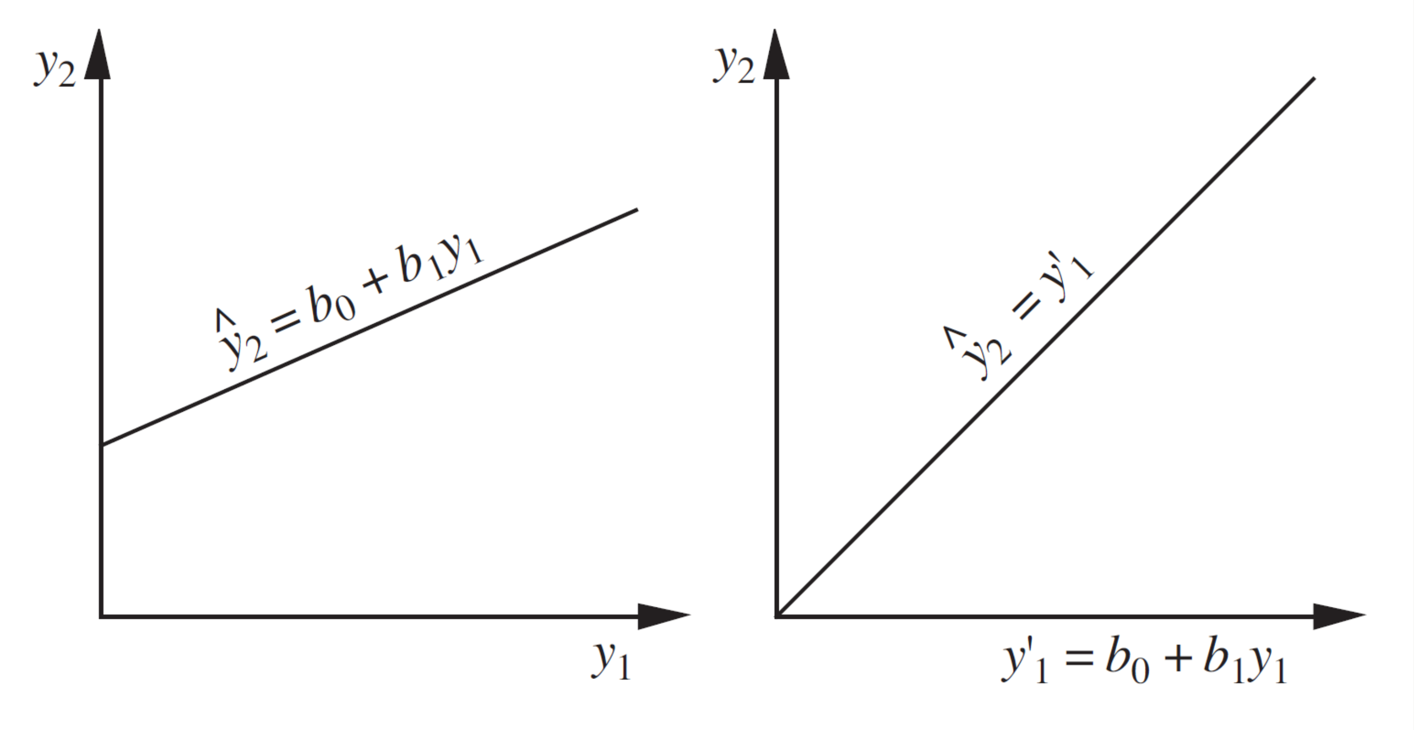
- Ensure comparability
- Equalize variable weighting
- Reduce impact of collinearity
- **Improve distance calculations**

Differences in scales can distort the distance metrics for methods like PCA and NMDS

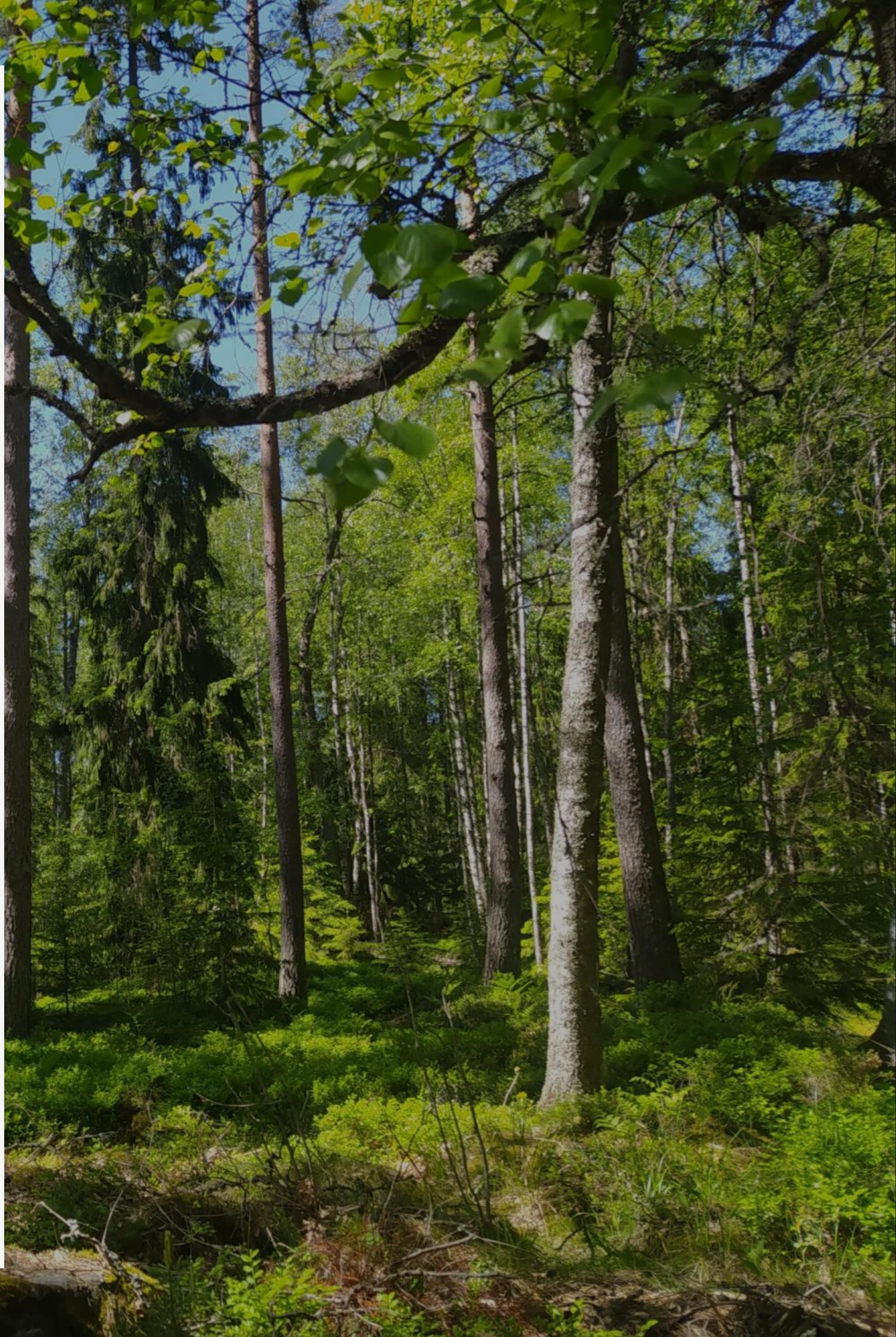


Standardization: Linear Transformation

Used to put quantitative descriptors on the same scale



Legendre and Legendre



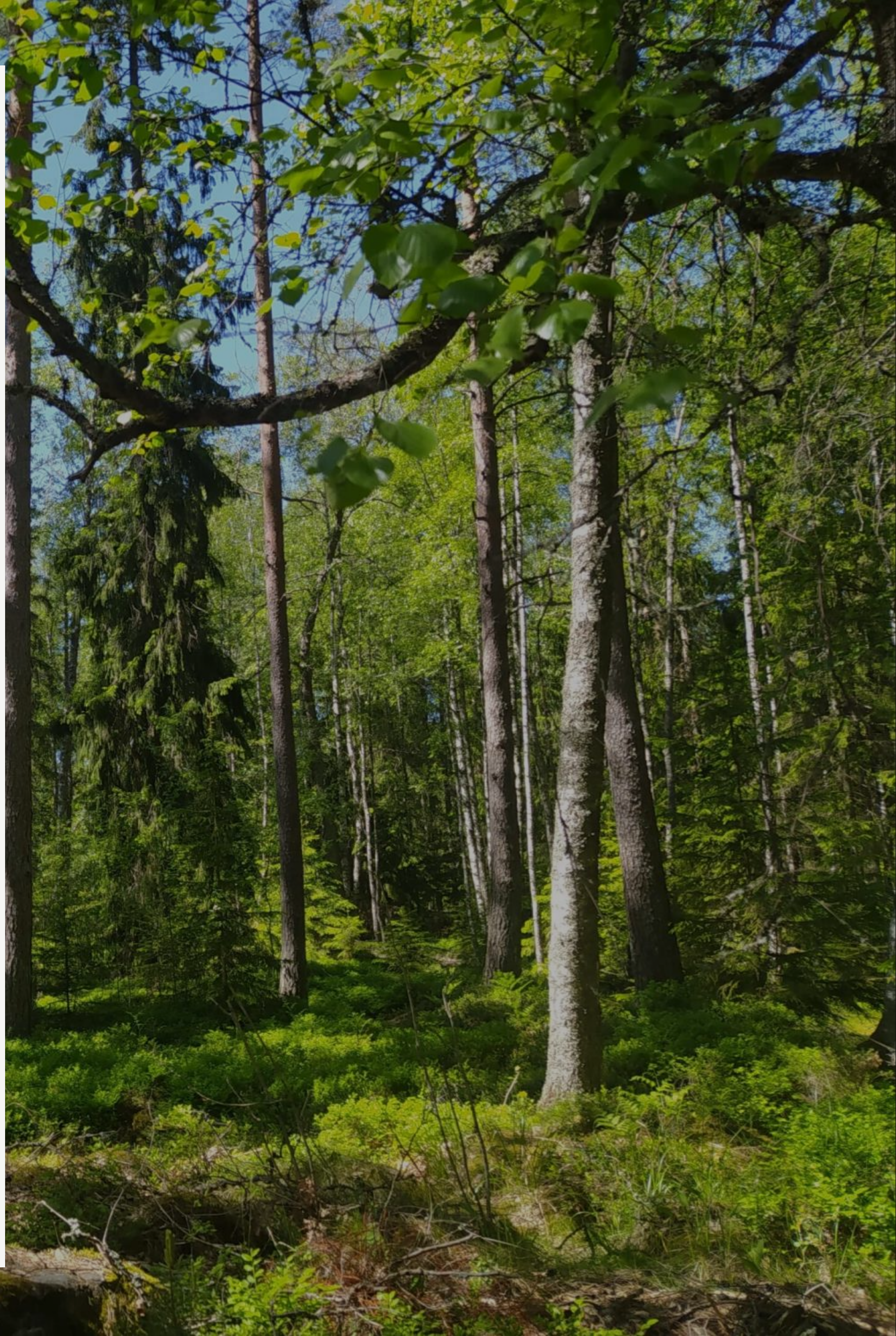
Standardization: Centering

$$\mathbf{X'} = \mathbf{X} - \bar{\mathbf{X}}$$

Where $\bar{\mathrm{X}}$ is the mean

- Centers data around zero but maintains shape and spread of data
- Centering can occur by row or by column

Use for: simplifying interpretation of covariance and correlation matrices



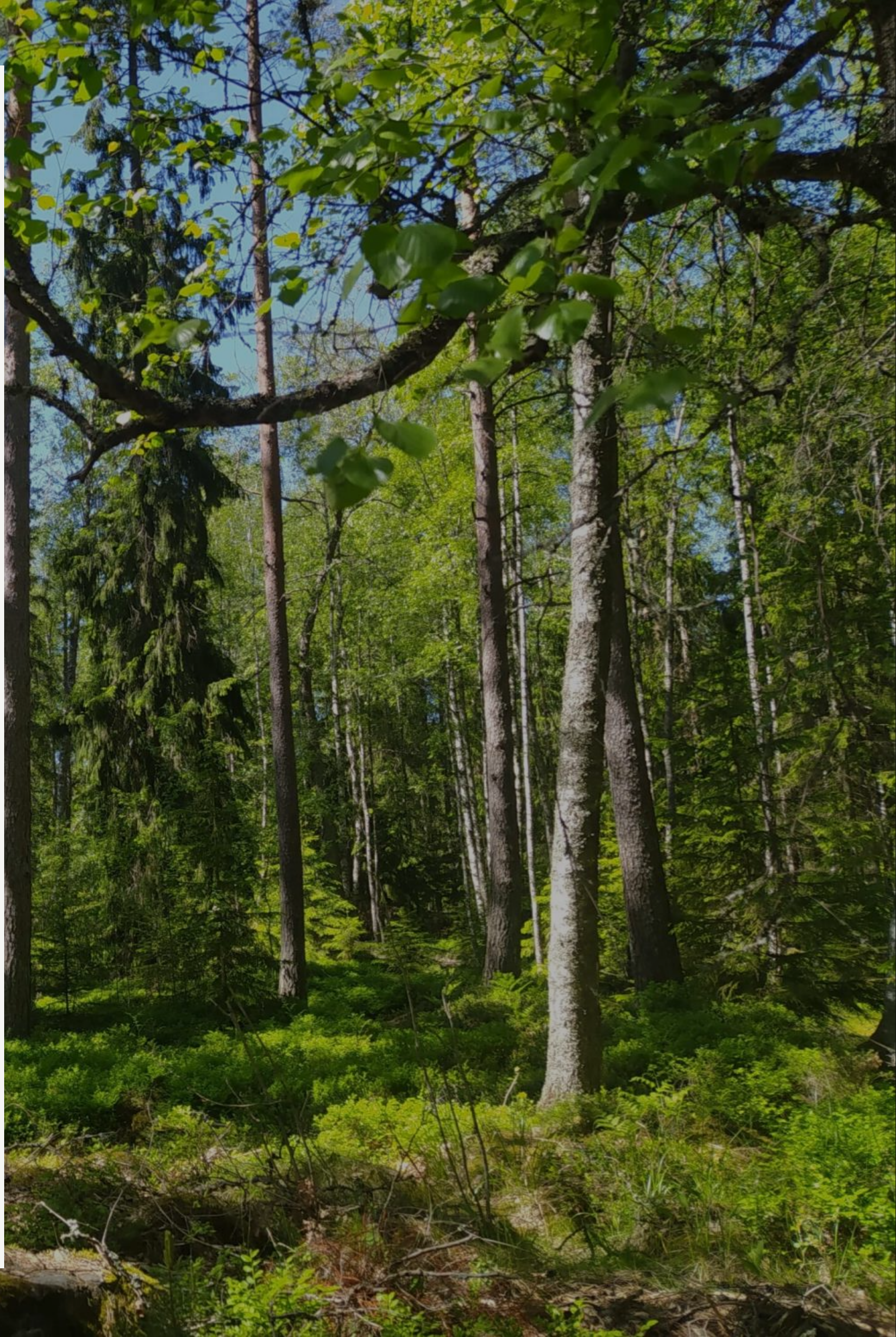
Standardization: Ranging (i.e., Min-Max Normalization)

$$\mathbf{X'} = \frac{\mathbf{X} - \mathbf{X}_{\mathbf{min}}}{\mathbf{X}_{\mathbf{max}} - \mathbf{X}_{\mathbf{min}}}$$

Where $\mathbf{X}_{min}$ and $\mathbf{X}_{max}$ are the minimum and maximum values of the original data

- Scales data to a specified range (0 – 1)
- Preserves relationships among data points but adjusts their scale

Use for: normalizing scores to a specific range



Standardization: Z-Scores

$$\mathbf{Z} = \frac{\mathbf{X} - \boldsymbol{\mu}}{\boldsymbol{\sigma}}$$

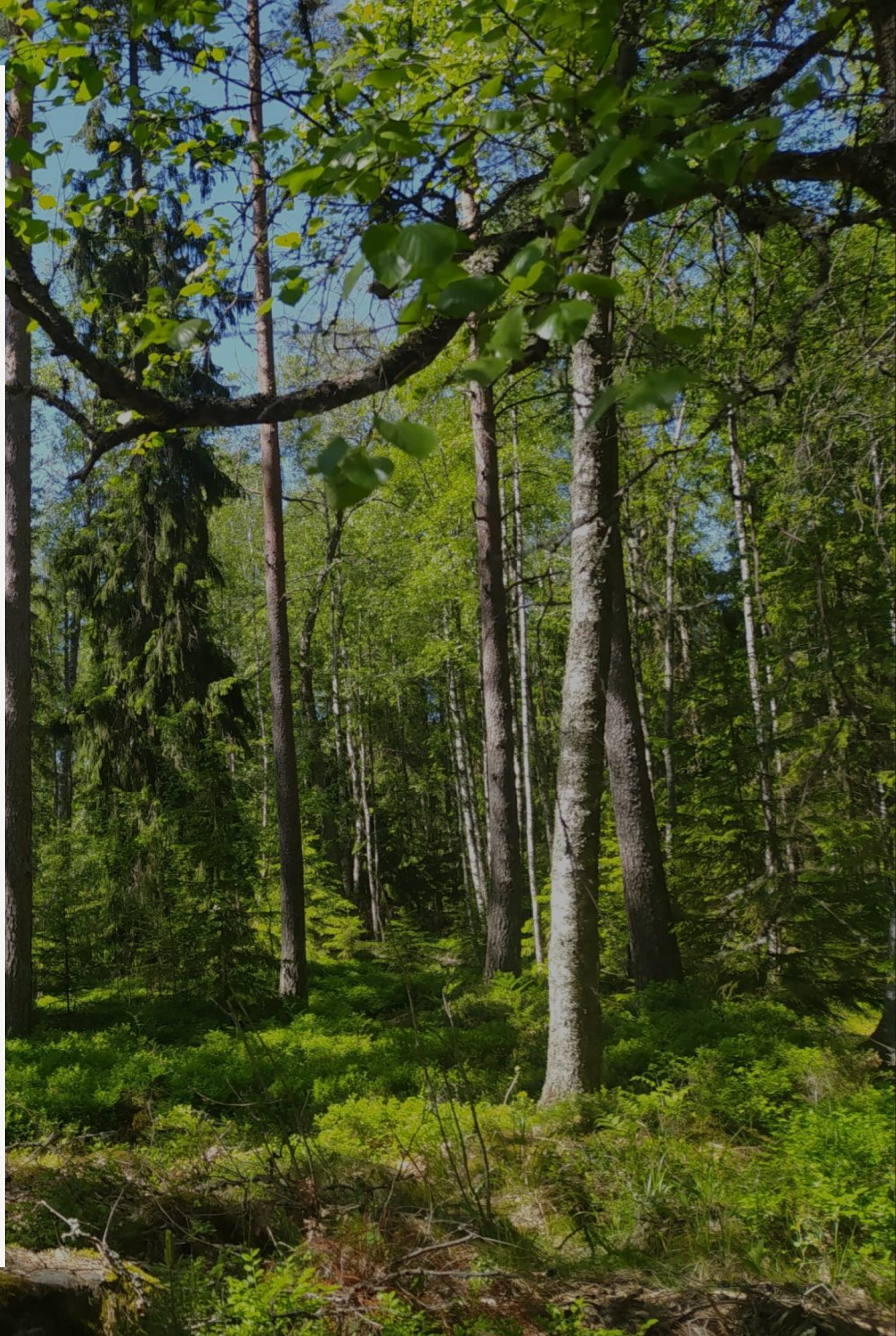
Where X = original data, μ = mean, σ = standard deviation

- Centers data around mean 0 and standard deviation of 1
- Removes units, makes data dimensionless and comparable

Use for: variables with different measurement units

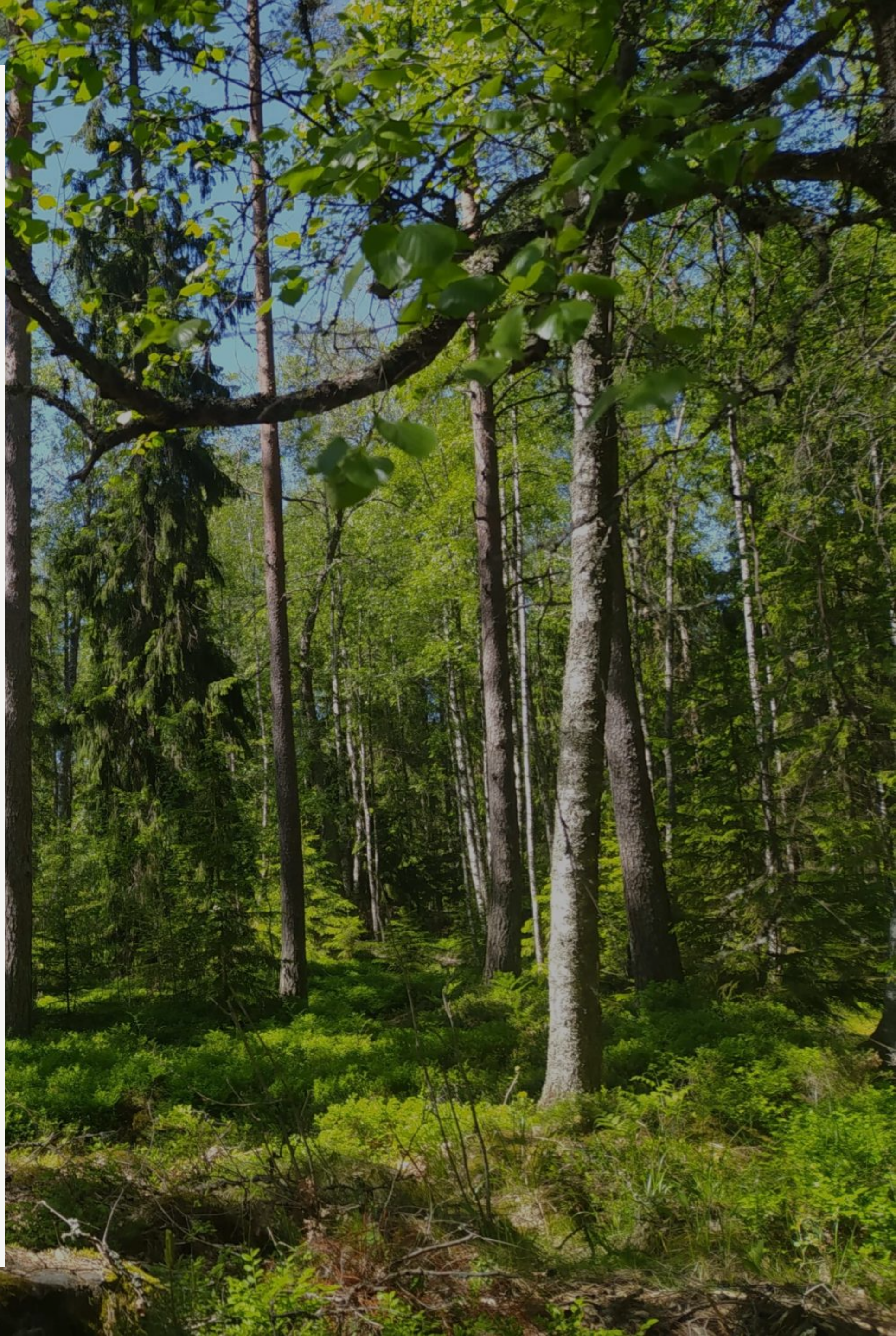
Standardization: Double Standardization

- Adjust each row and column by its sum or mean
- Ensures equal weight for objects (sites) *and* descriptors (species)



Standardization: Chord Transformation

- Reduces impact of varying magnitudes of species abundances among samples
- Emphasizes relative importances of species within a sample
- Normalizes each sample vector to have a “length” of 1



Standardization: Chord Transformation

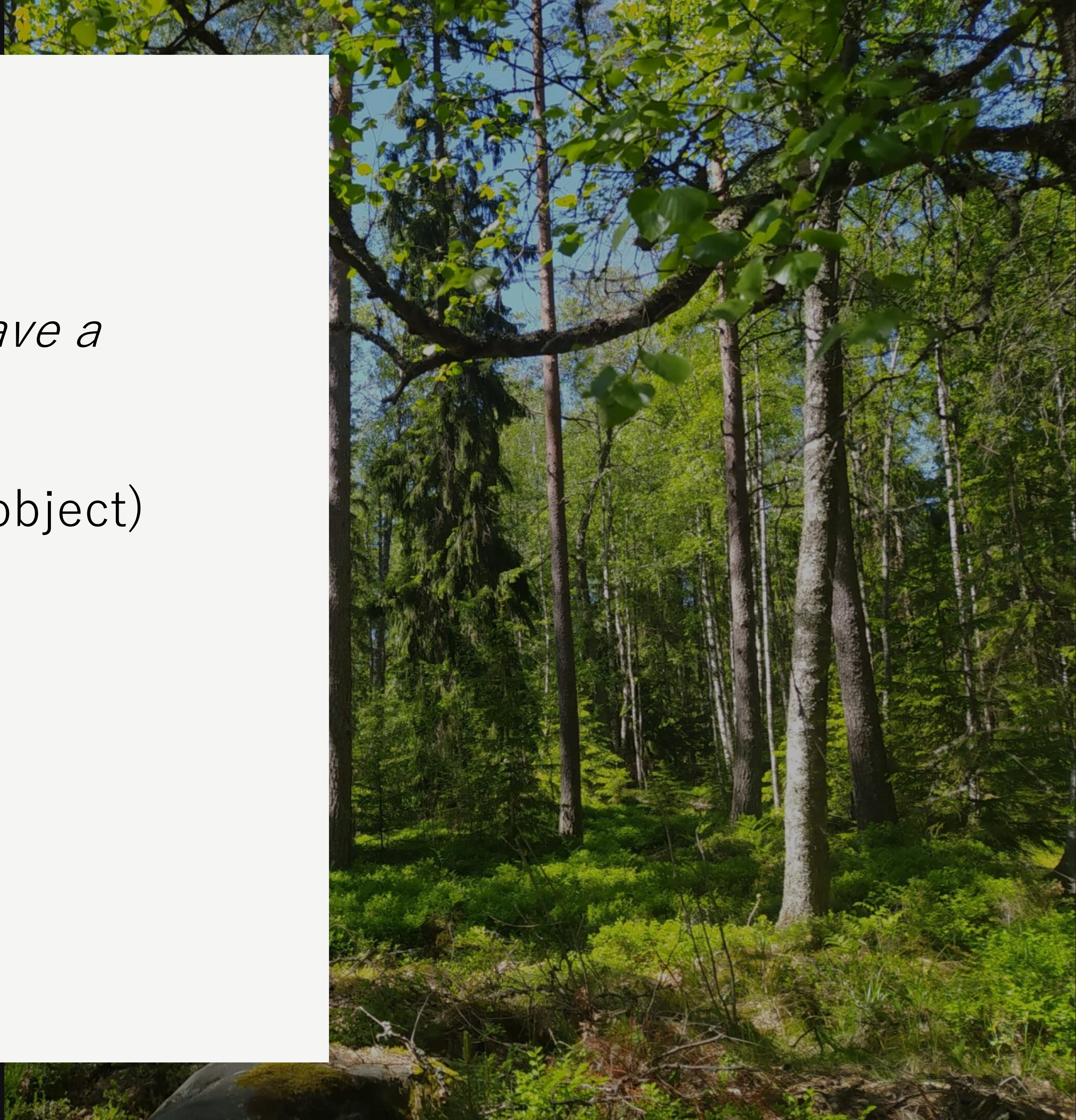
Normalizes each sample vector to have a "length" of 1

- Calculate length of each sample (object) vector:

$$||\mathbf{x}|| = \sqrt{\mathbf{x}_1^2 + \mathbf{x}_2^2 + \ldots \mathbf{x}_n^2}$$

- Normalize each sample vector

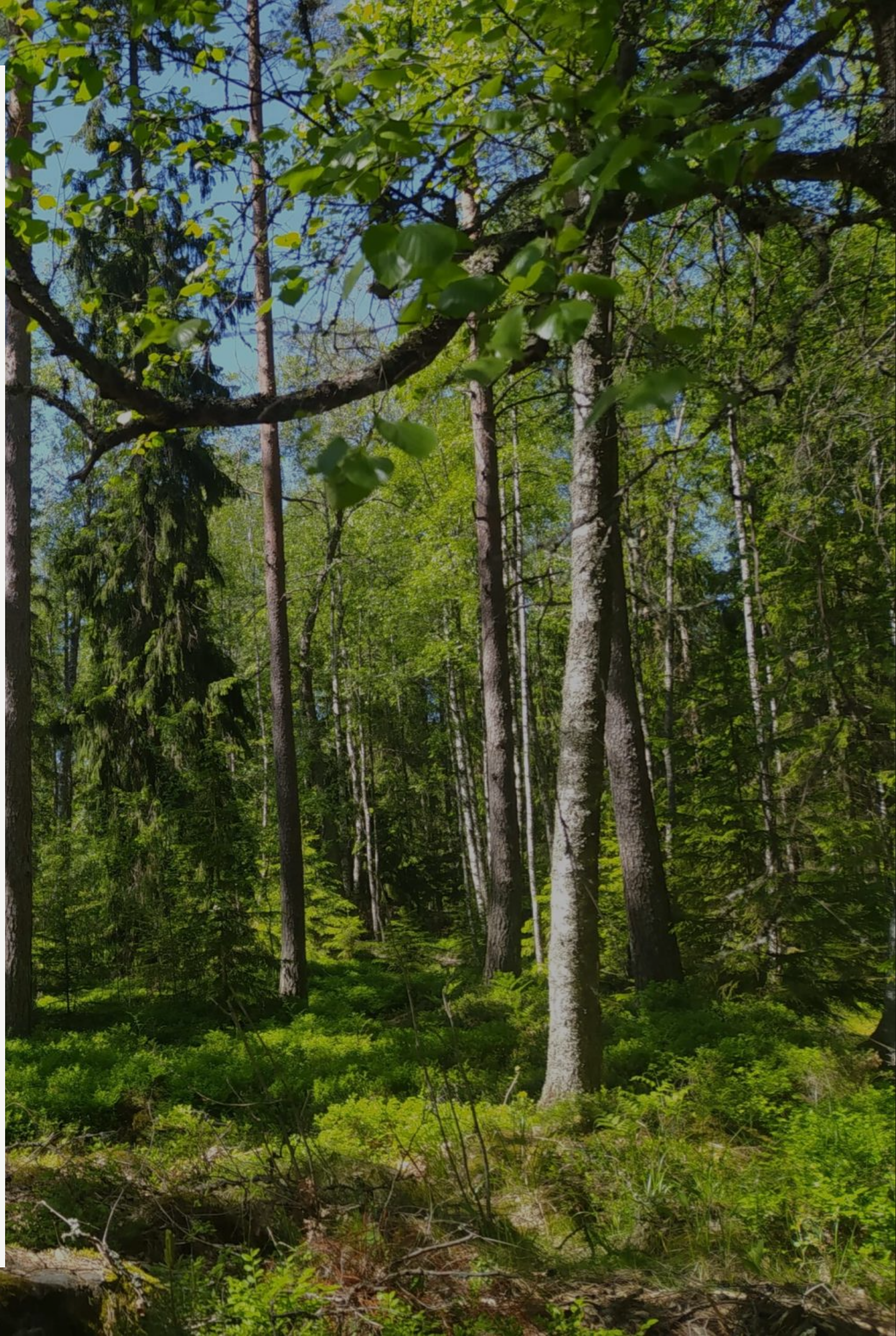
$$\mathbf{x}' = \mathbf{x}/||\mathbf{x}||$$



Standardization: Hellinger Transformation

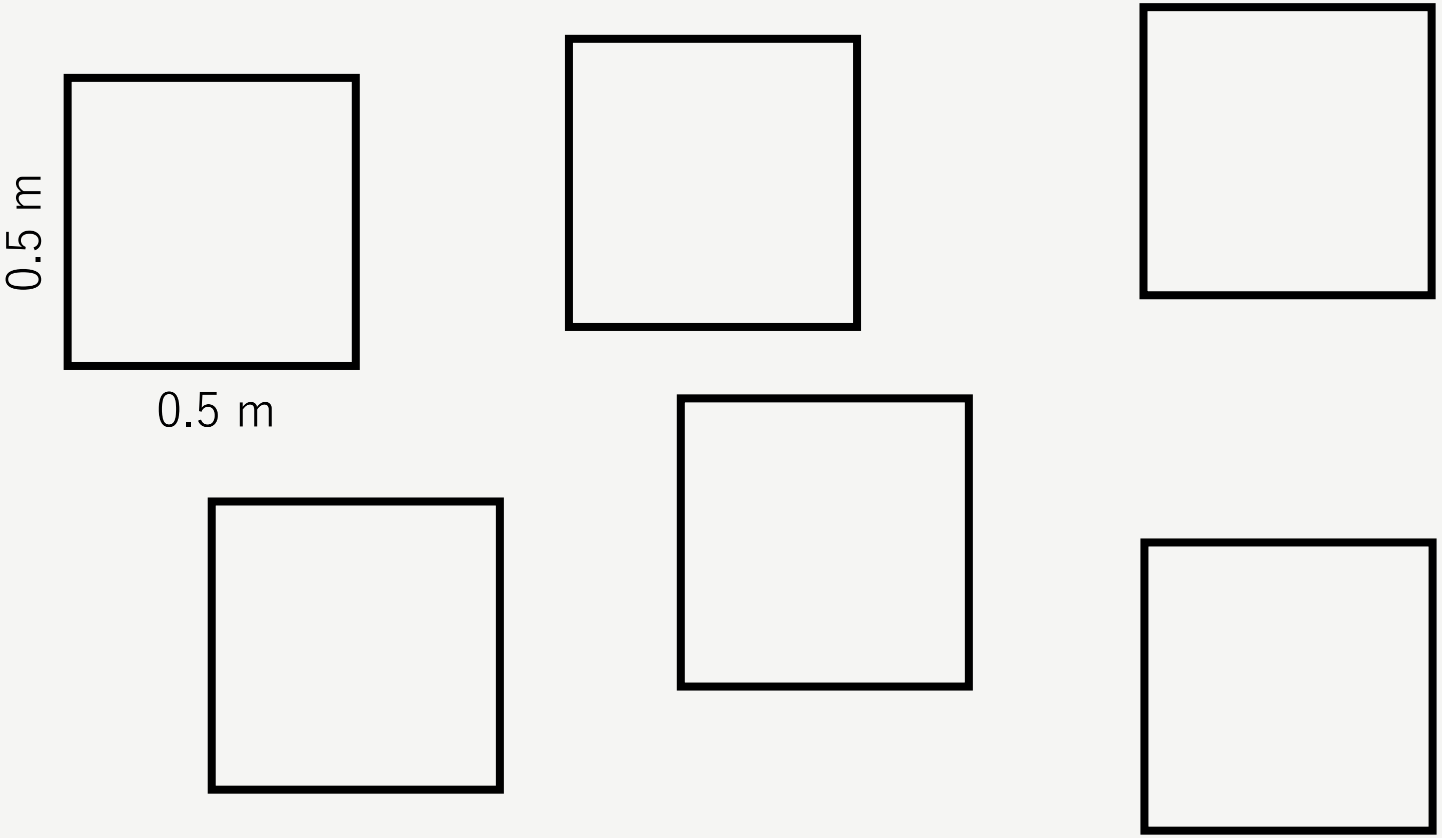
- Downweighs the influence of dominant species
- **Square root** each species' abundance
- **Normalize:** divide each species' square rooted abundance by the square root of the sum of squared-square root abundances

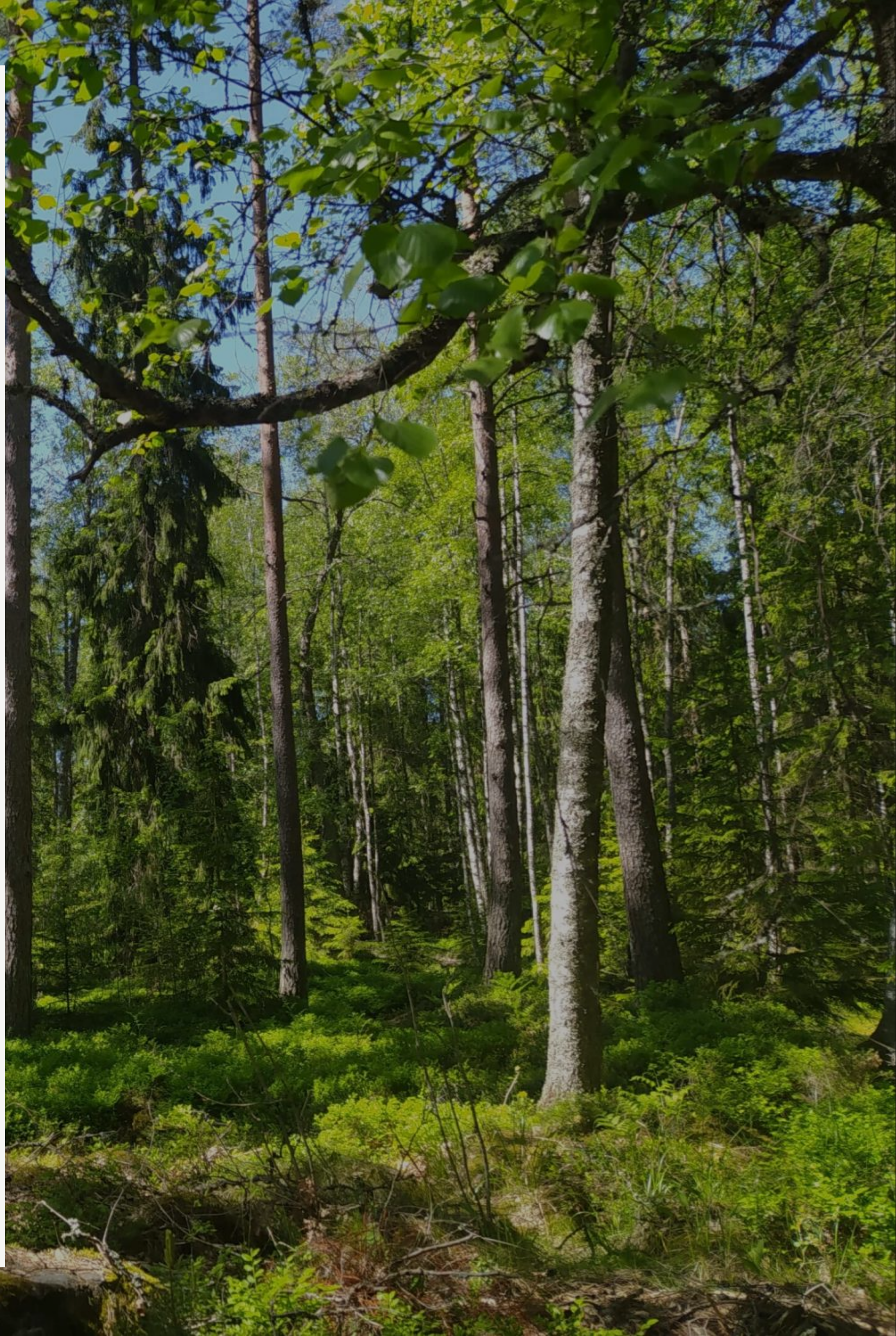
$$\mathbf{x}^{`}_{ij} = \frac{\sqrt{\mathbf{x}_{ij}}}{\sqrt{\mathbf{x}^2_{1j} + \mathbf{x}^2_{2j} + \ldots\, \mathbf{x}^2_{nj}}}$$



Standardization: By Sampling Effort

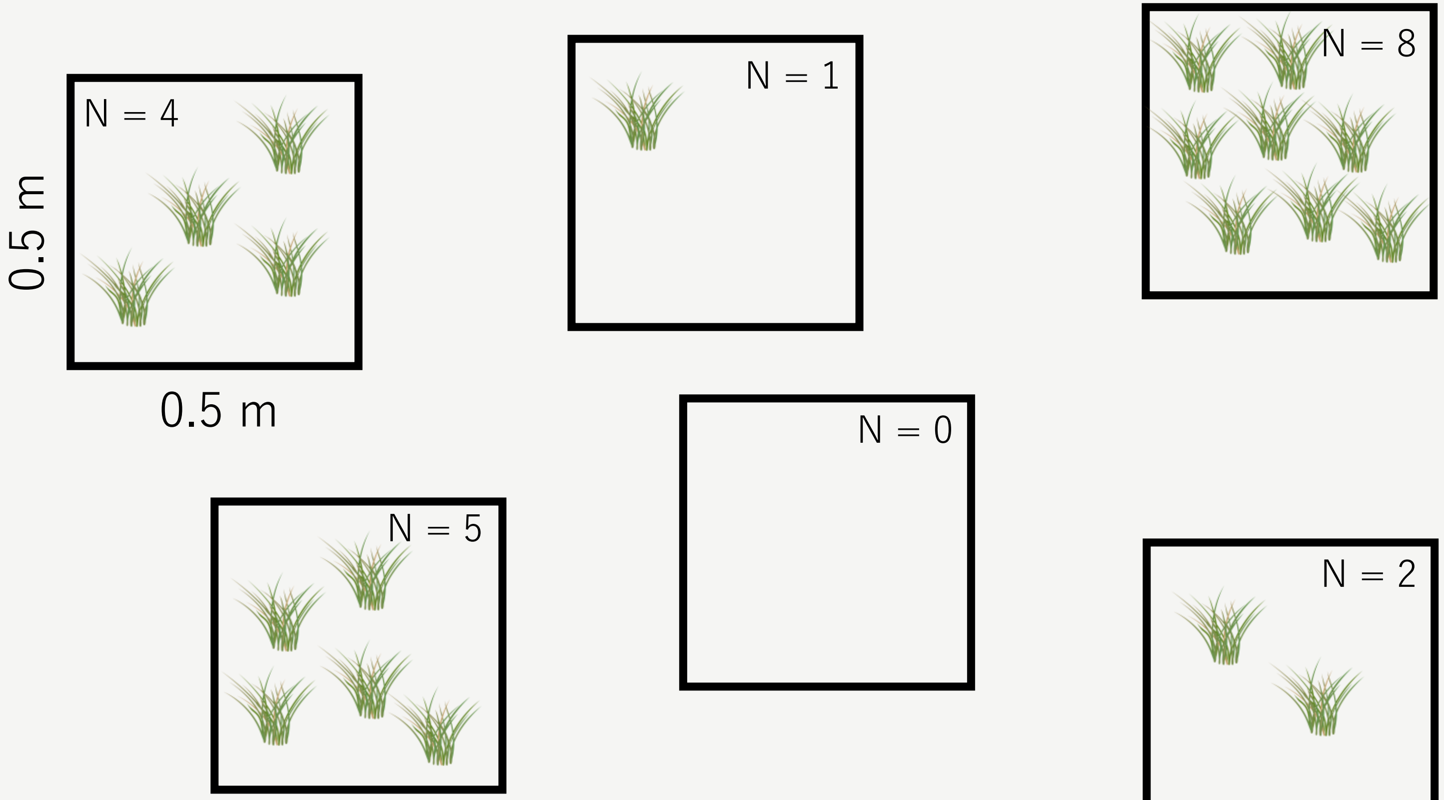
What units should you use for your response variable?

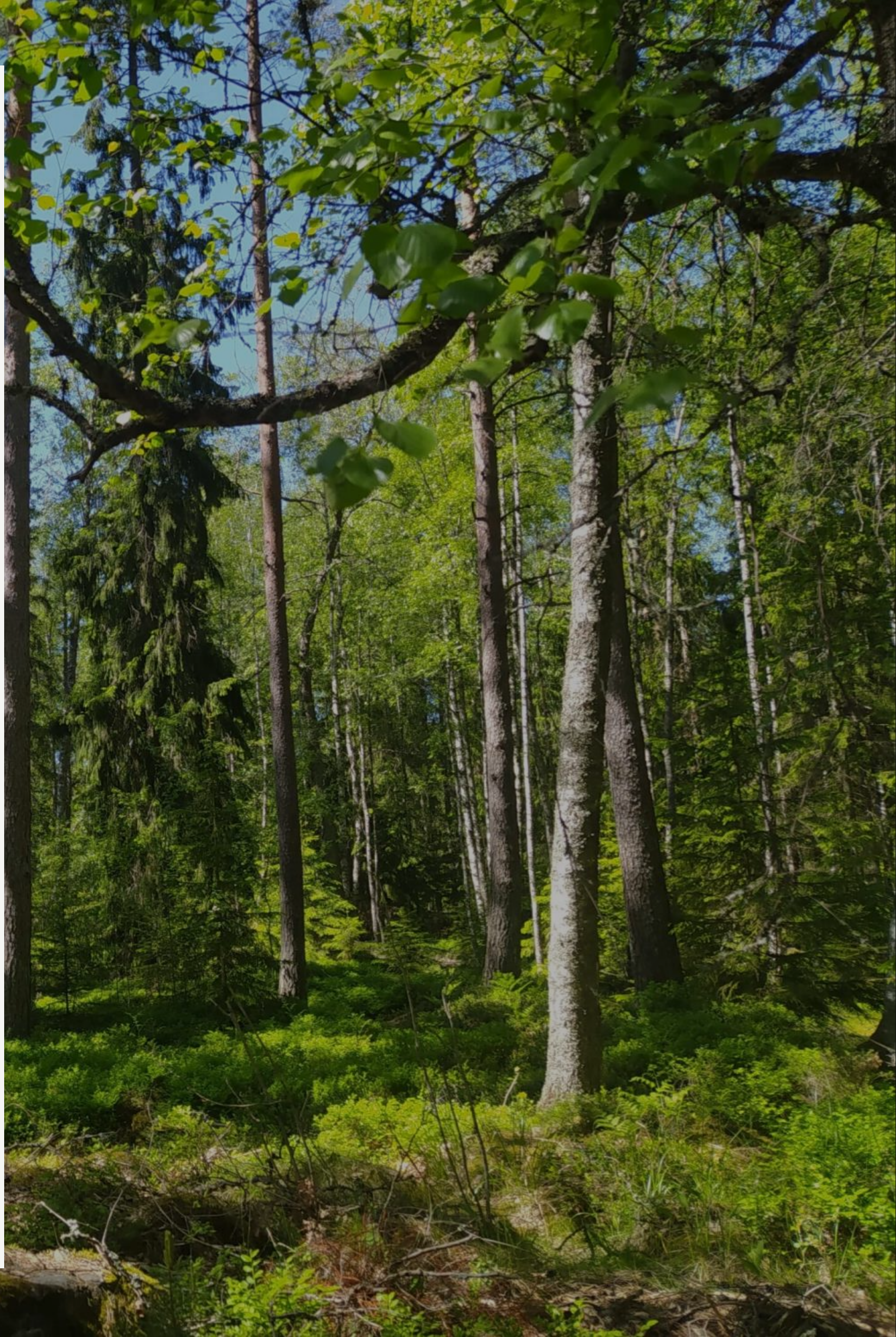




Standardization: By Sampling Effort

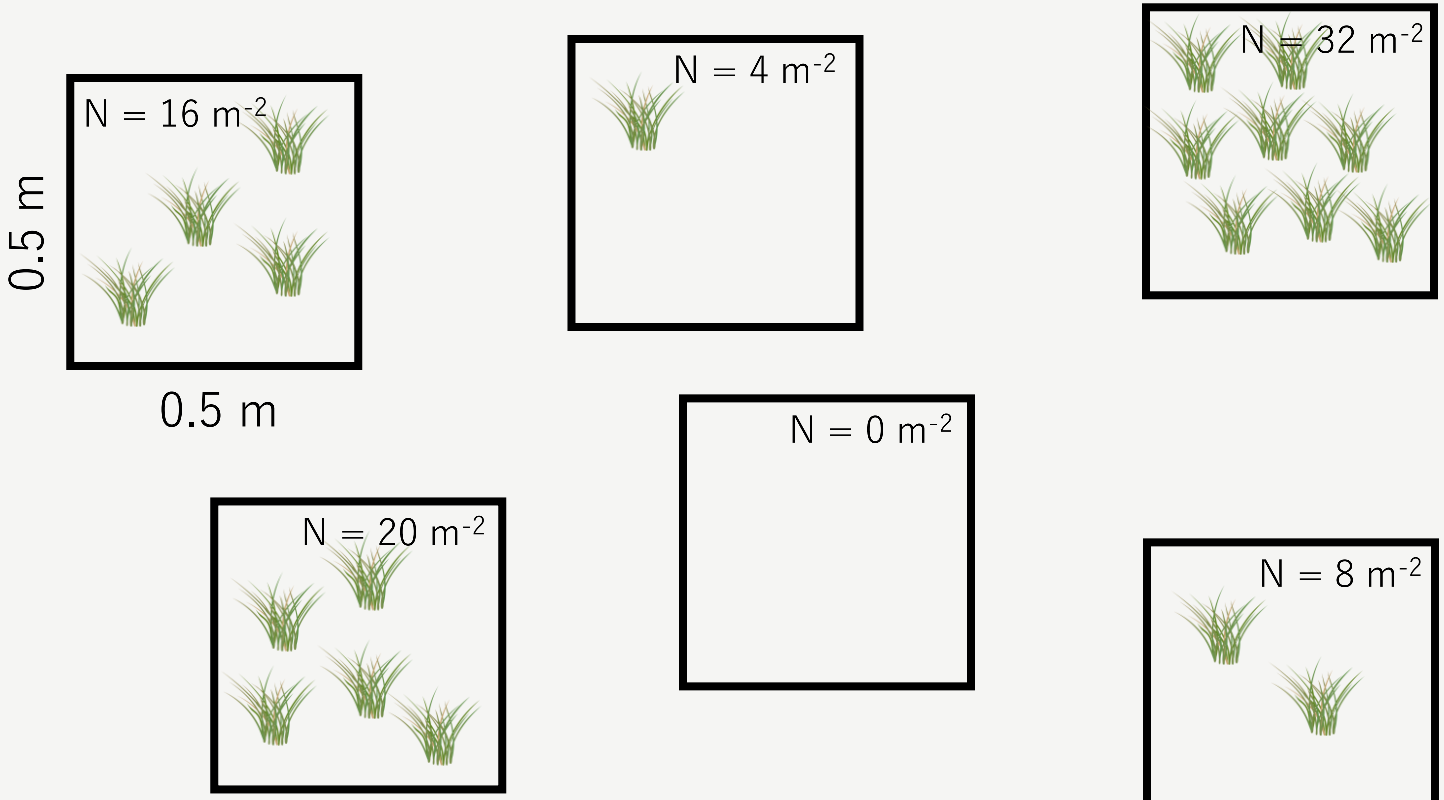
What units should you use for your response variable?

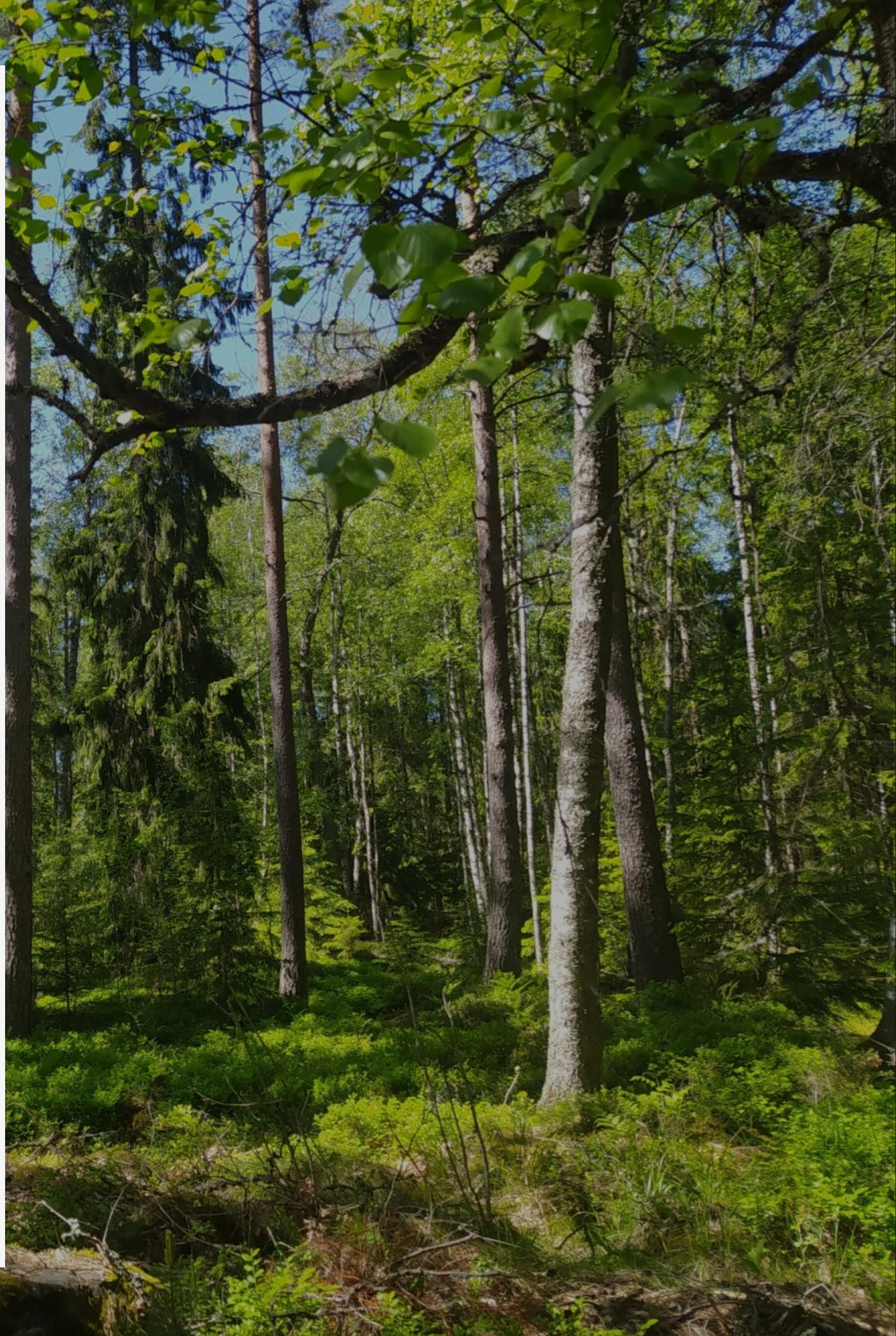




Standardization: By Sampling Effort

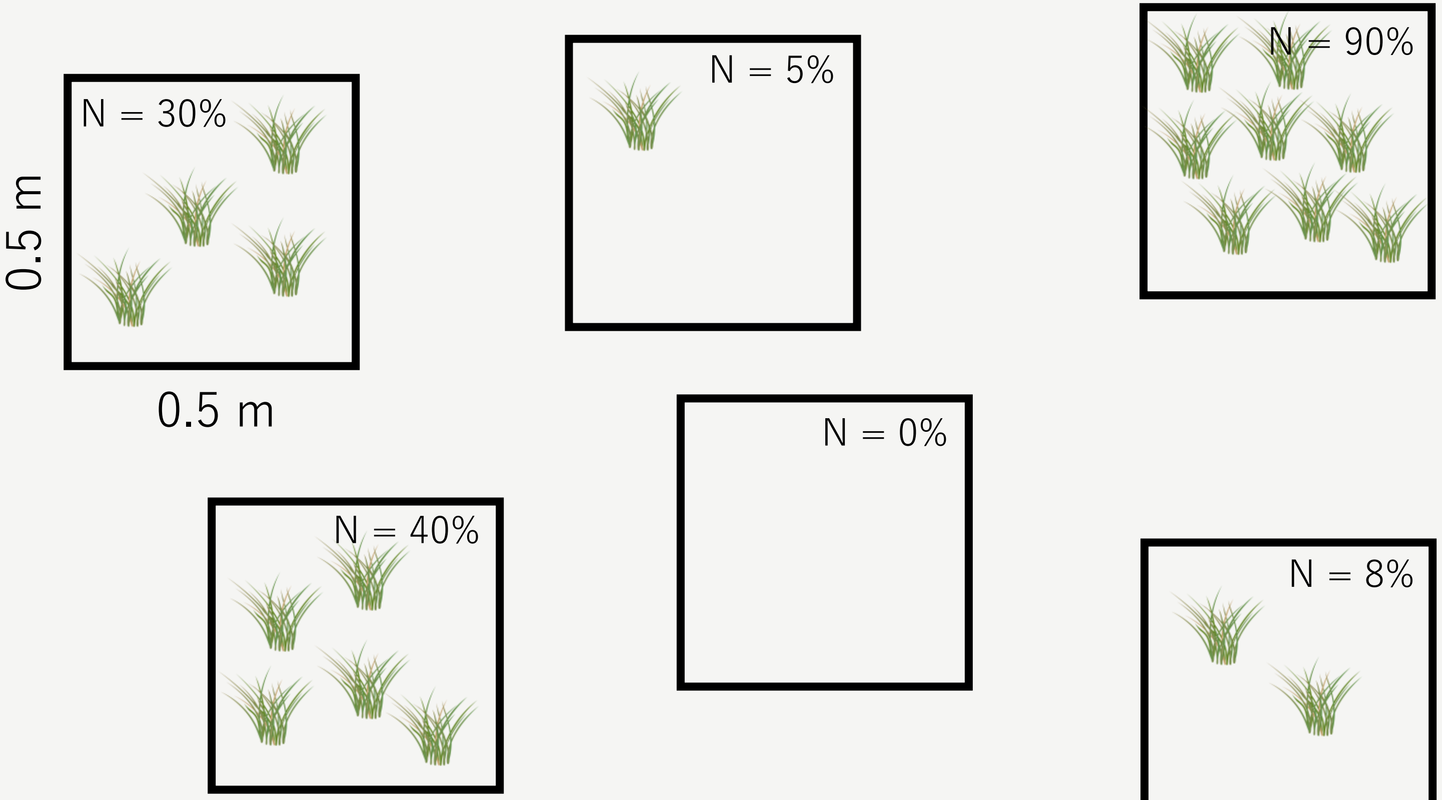
What units should you use for your response variable?

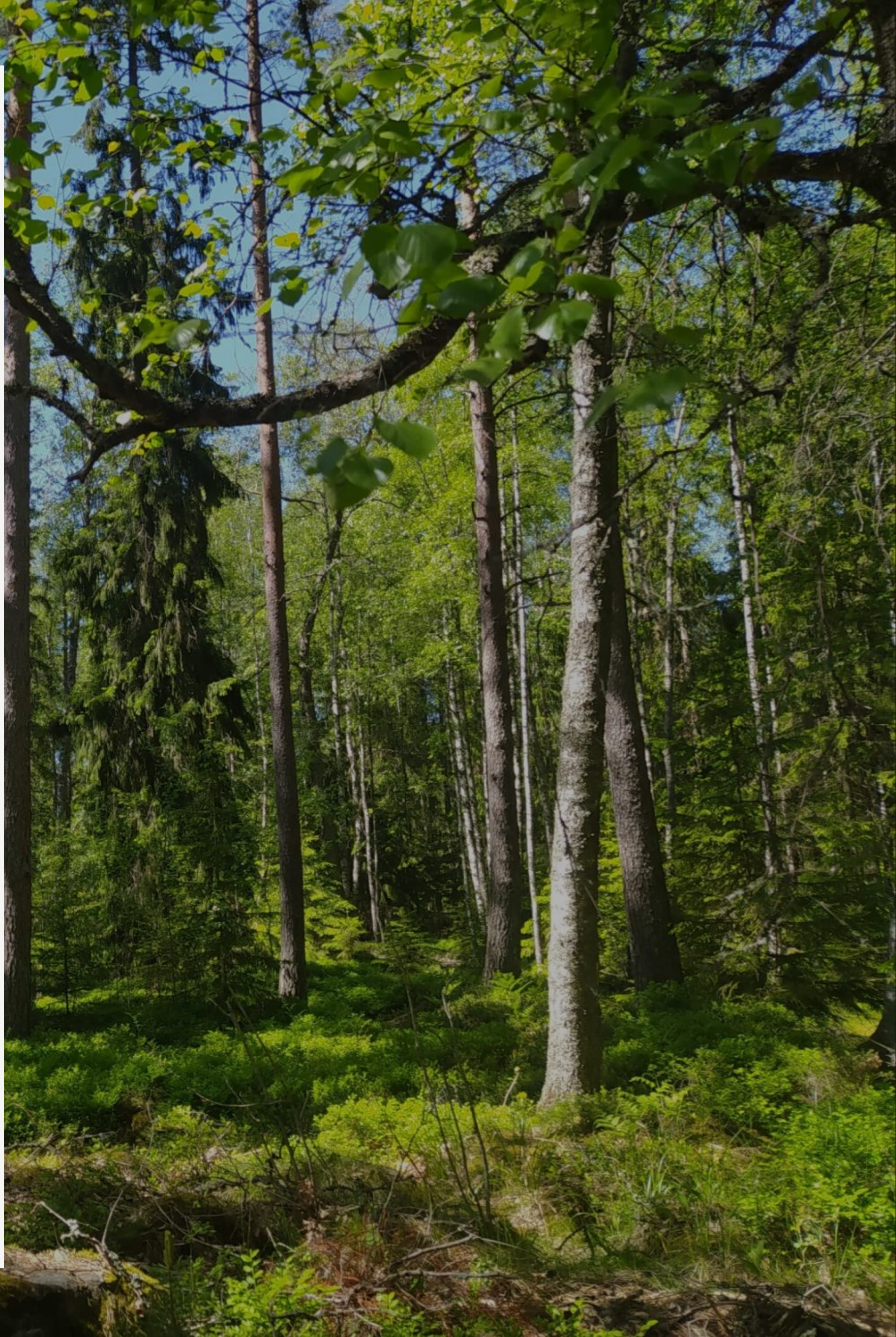




Standardization: By Sampling Effort

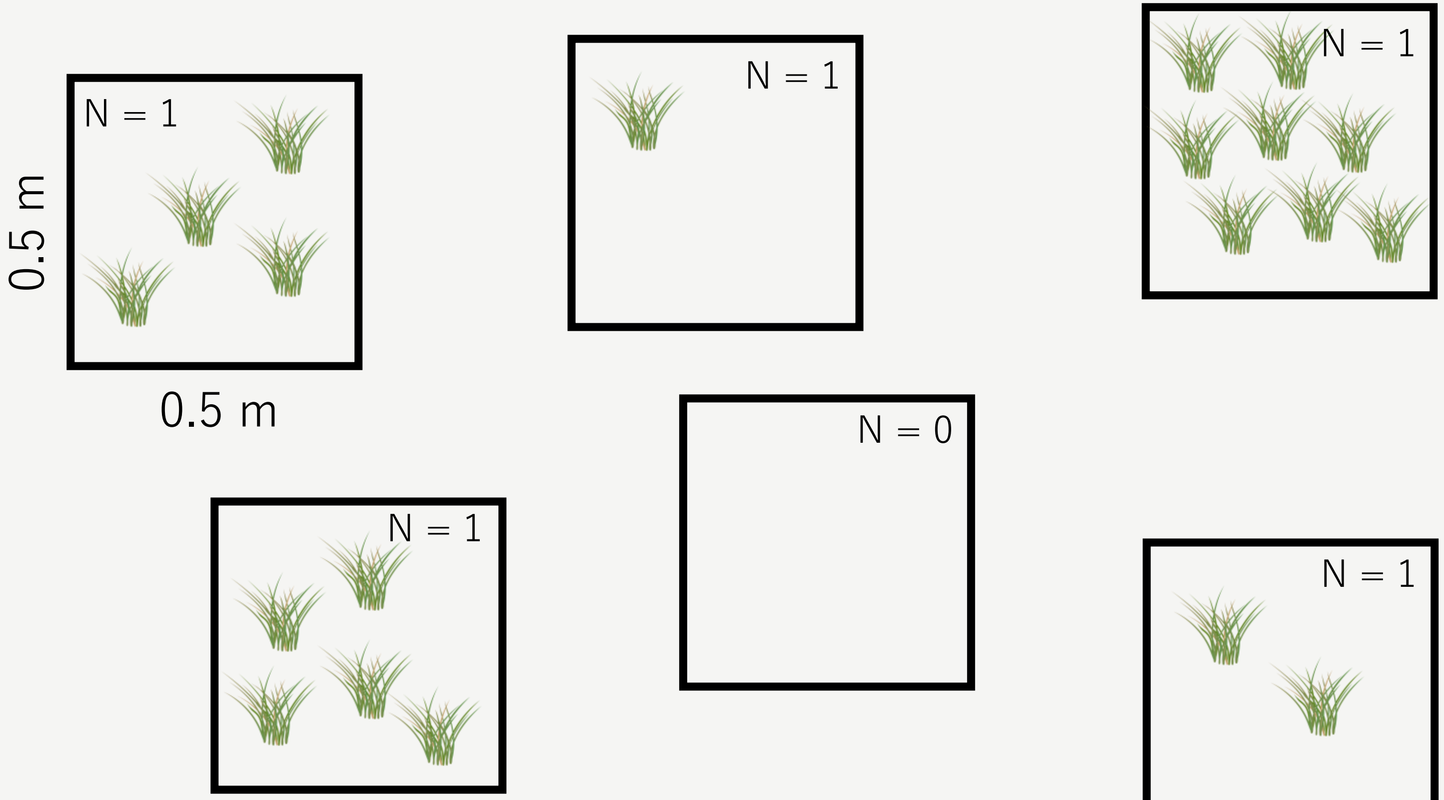
What units should you use for your response variable?

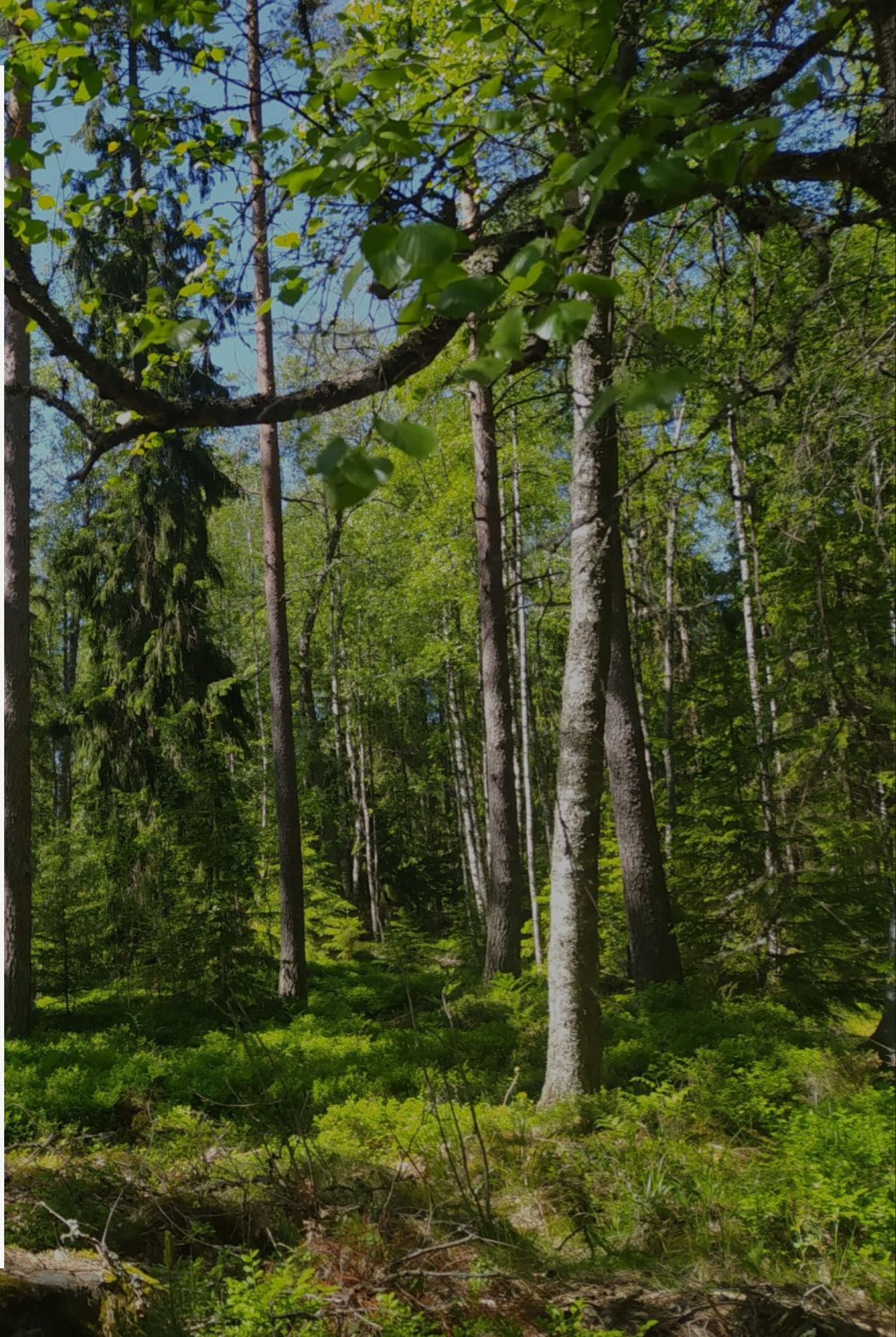




Standardization: By Sampling Effort

What units should you use for your response variable?





Standardization: By Sampling Effort

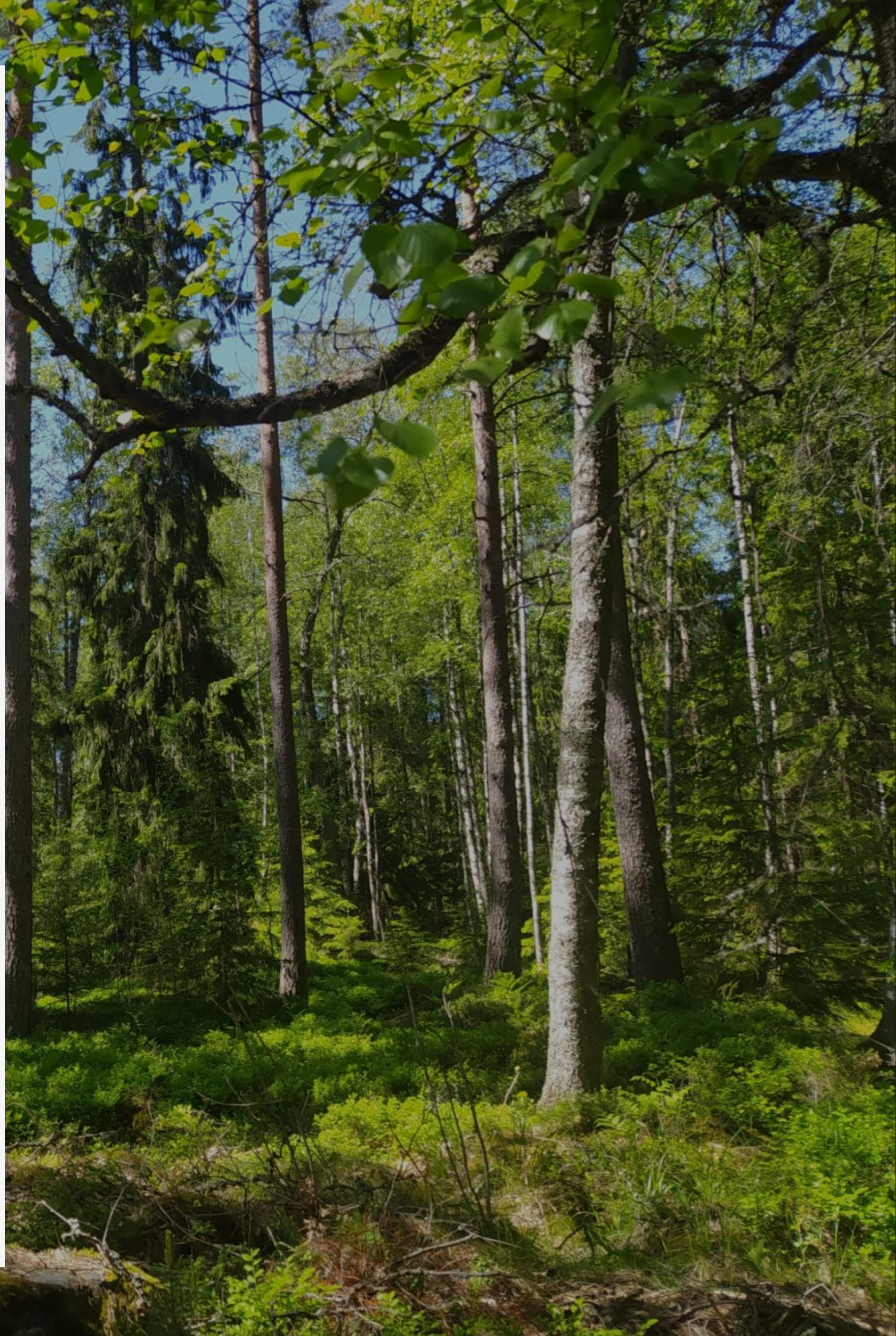
What units should you use for your response variable?

100 m

200 m

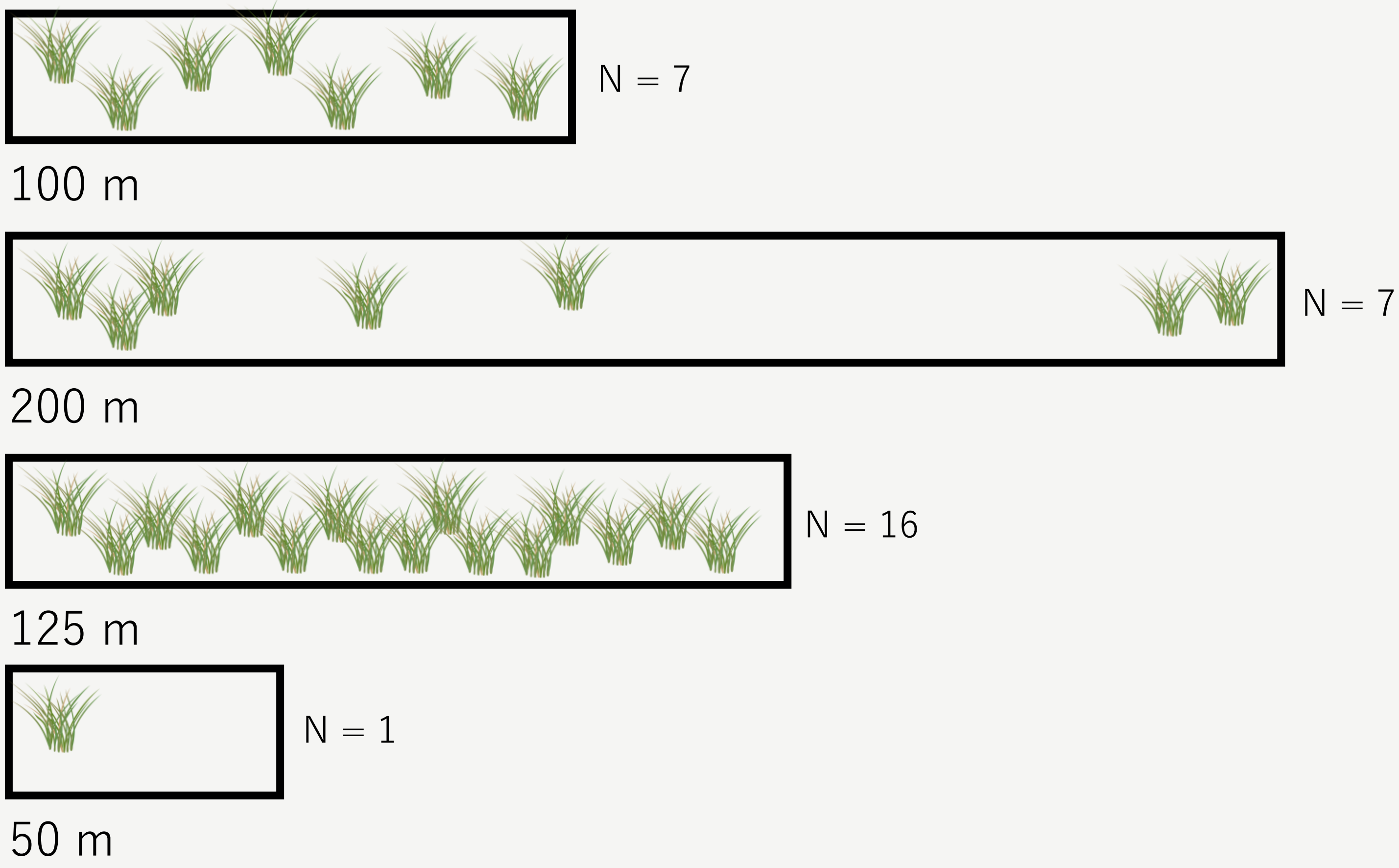
125 m

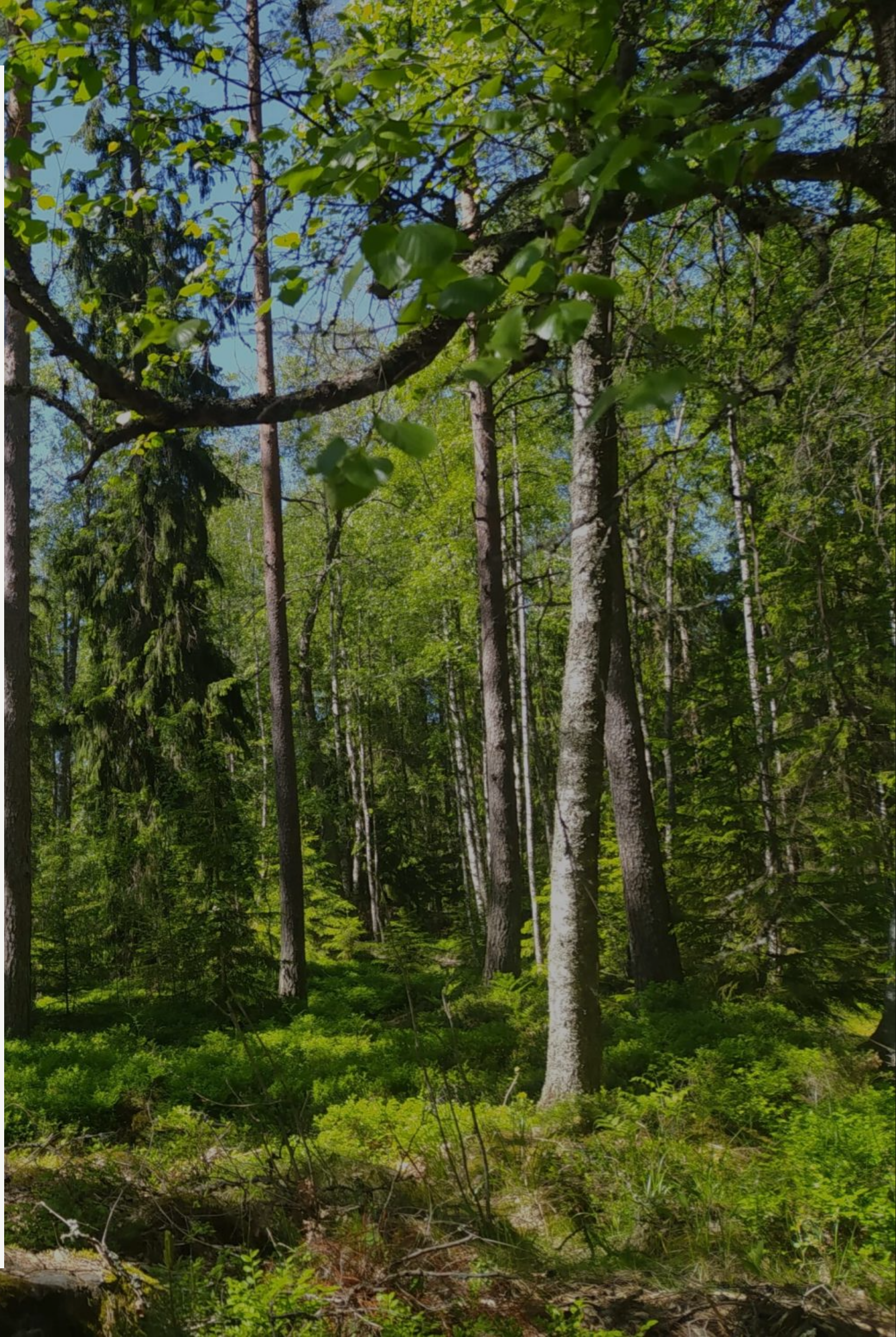
50 m



Standardization: By Sampling Effort

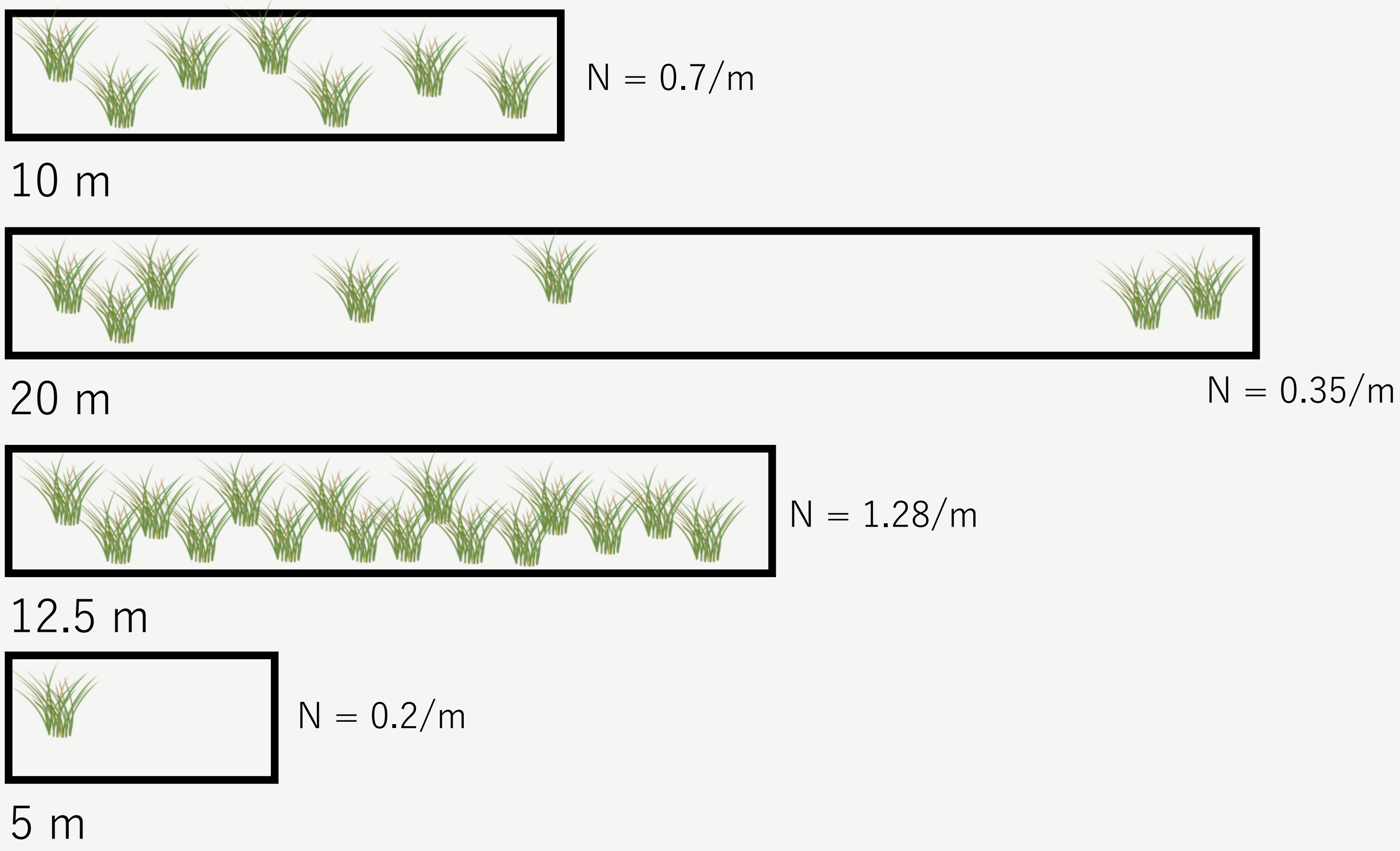
What units should you use for your response variable?

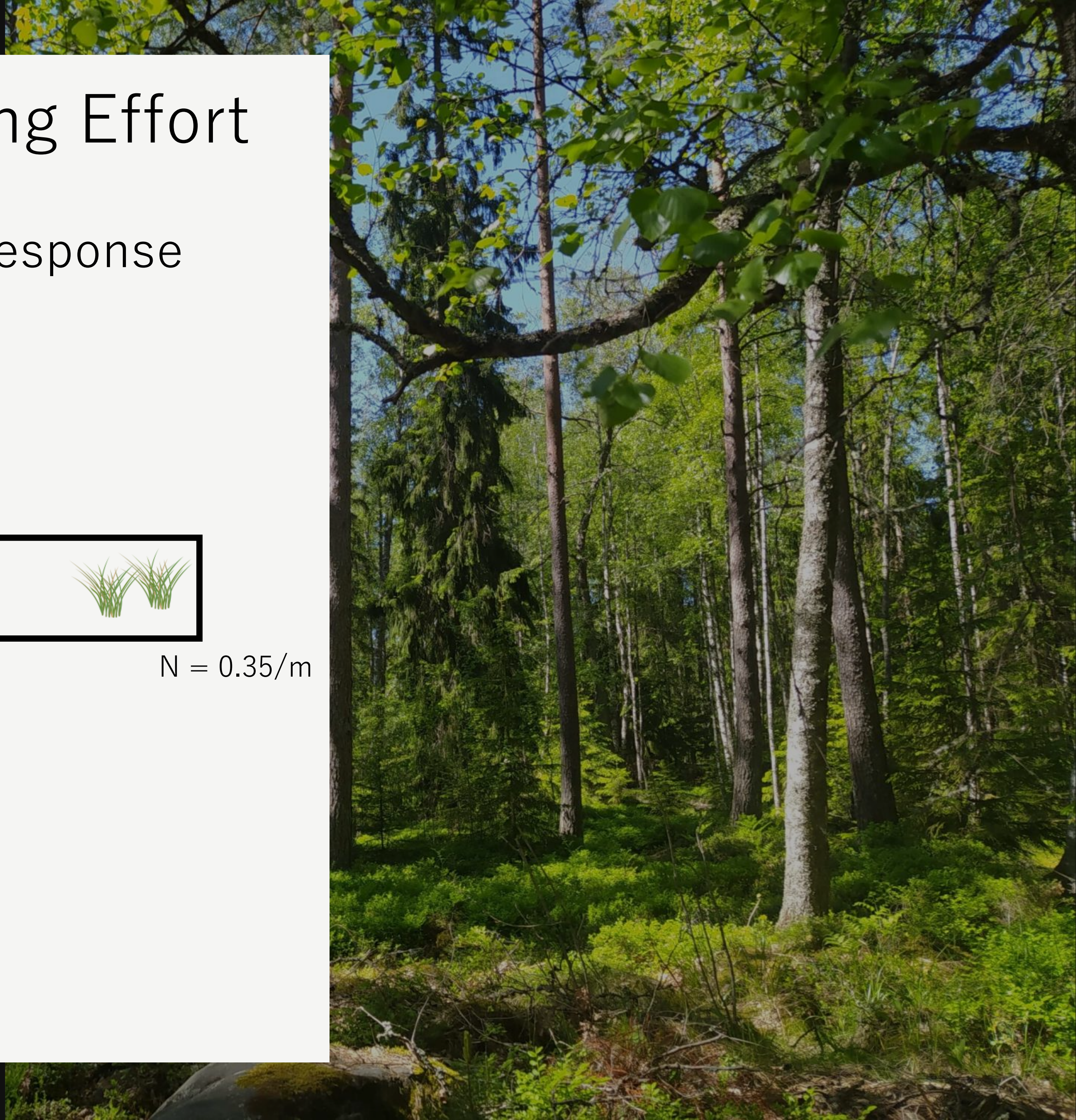




Standardization: By Sampling Effort

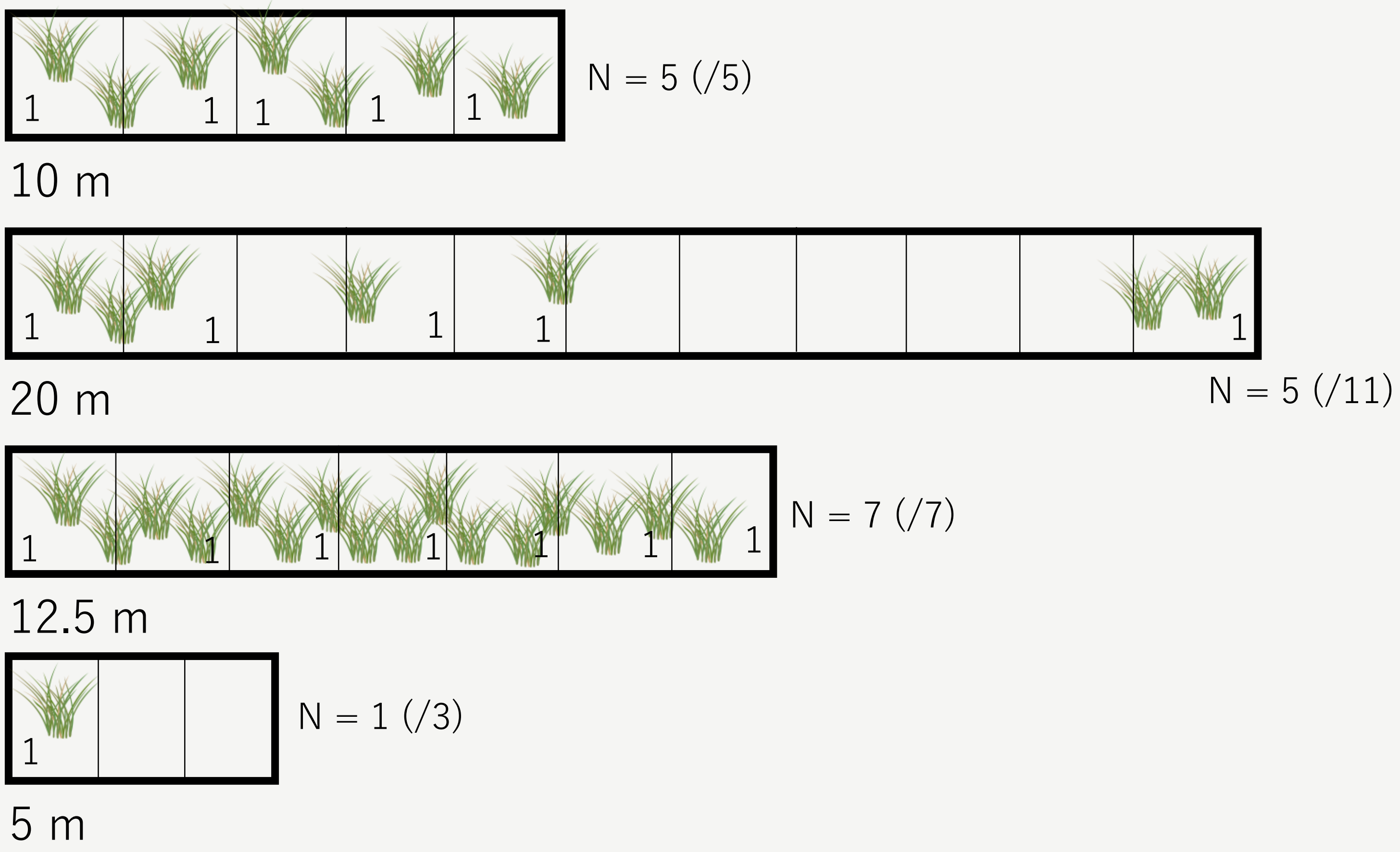
What units should you use for your response variable?

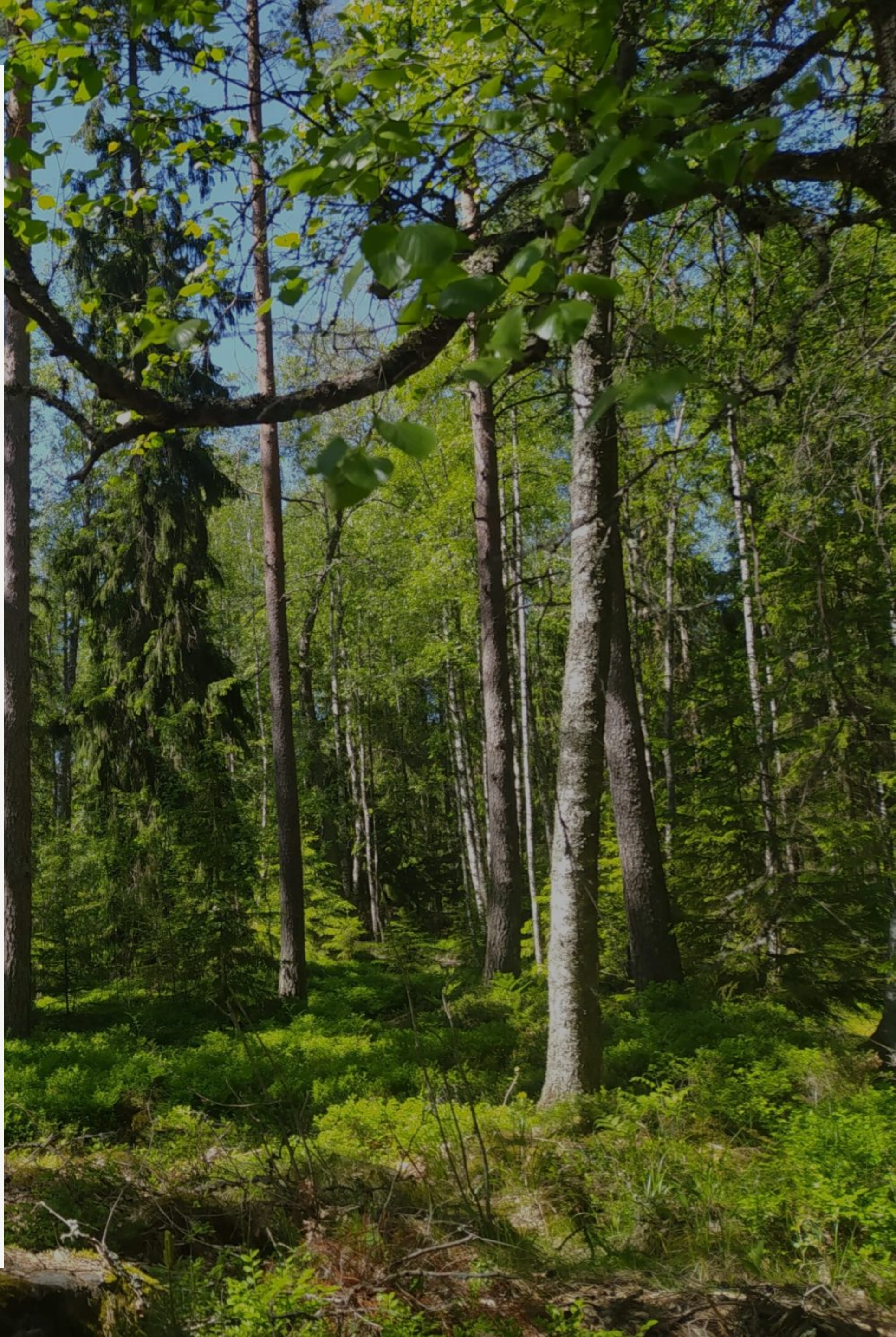




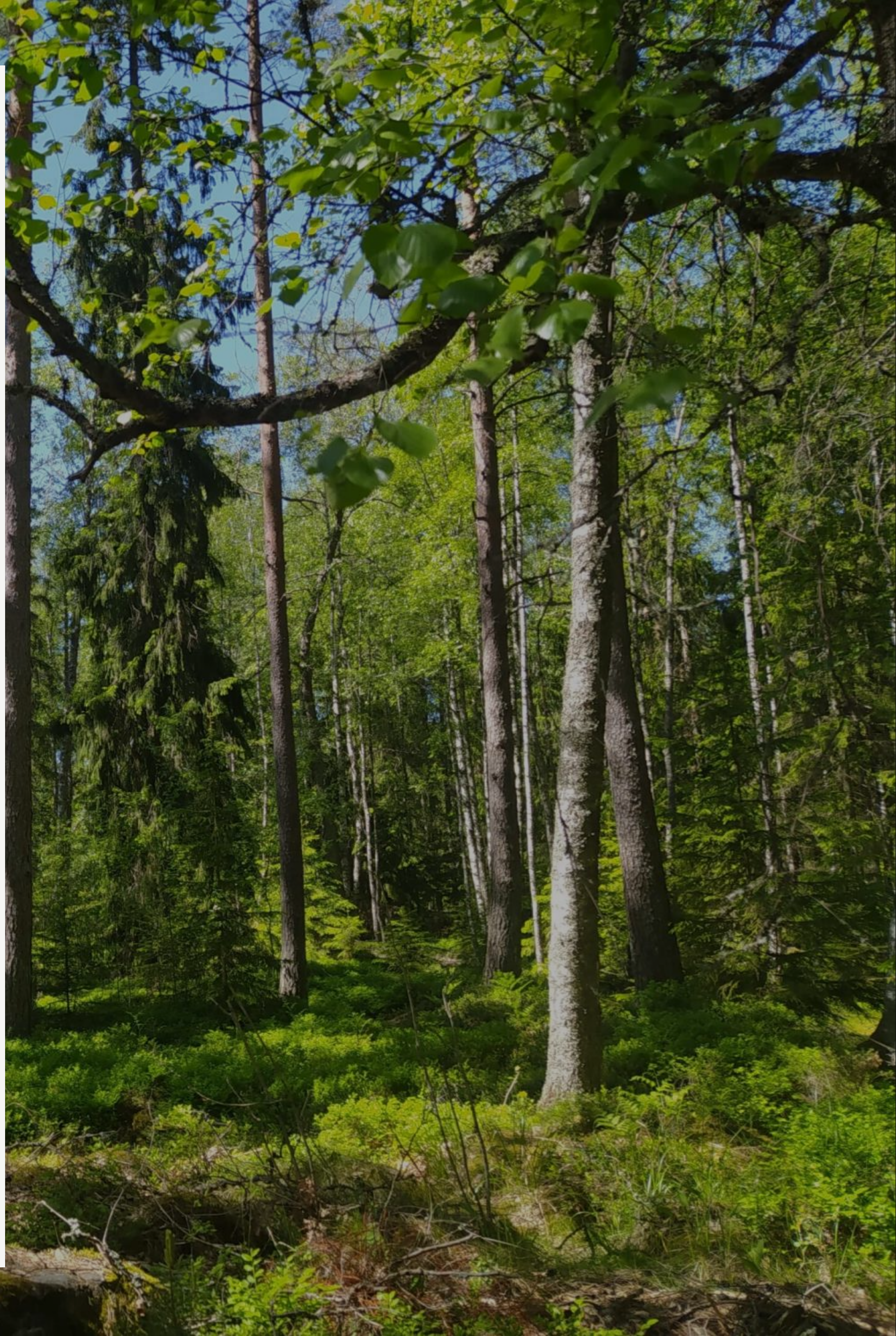
Standardization: By Sampling Effort

What units should you use for your response variable?



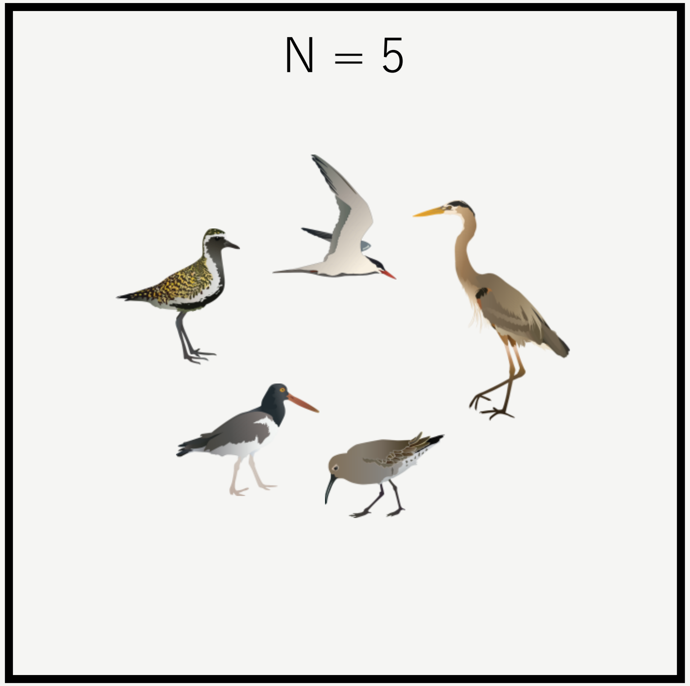


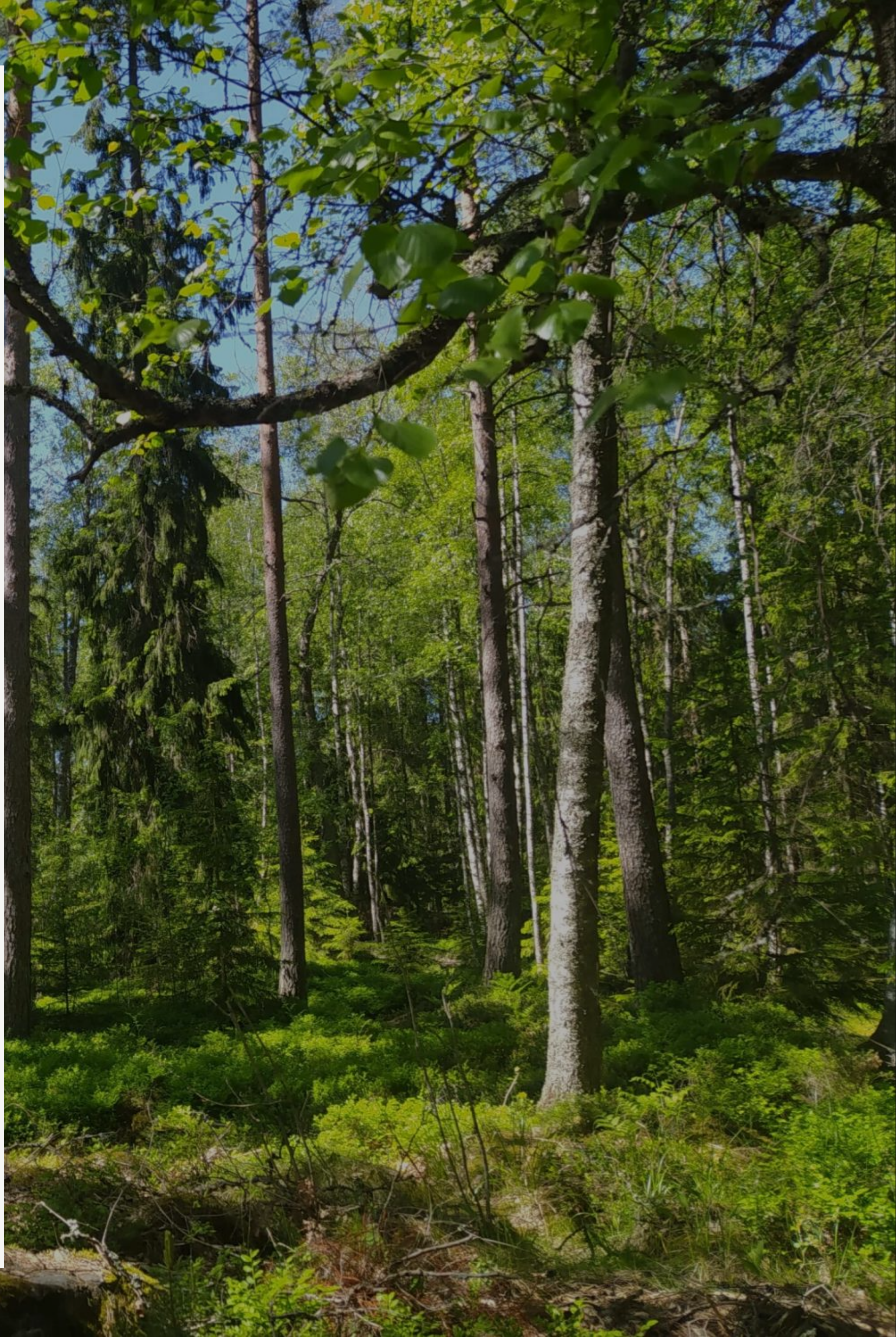
Univariate Metrics of Ecological Diversity



Univariate Metrics of Ecological Diversity: Species Richness

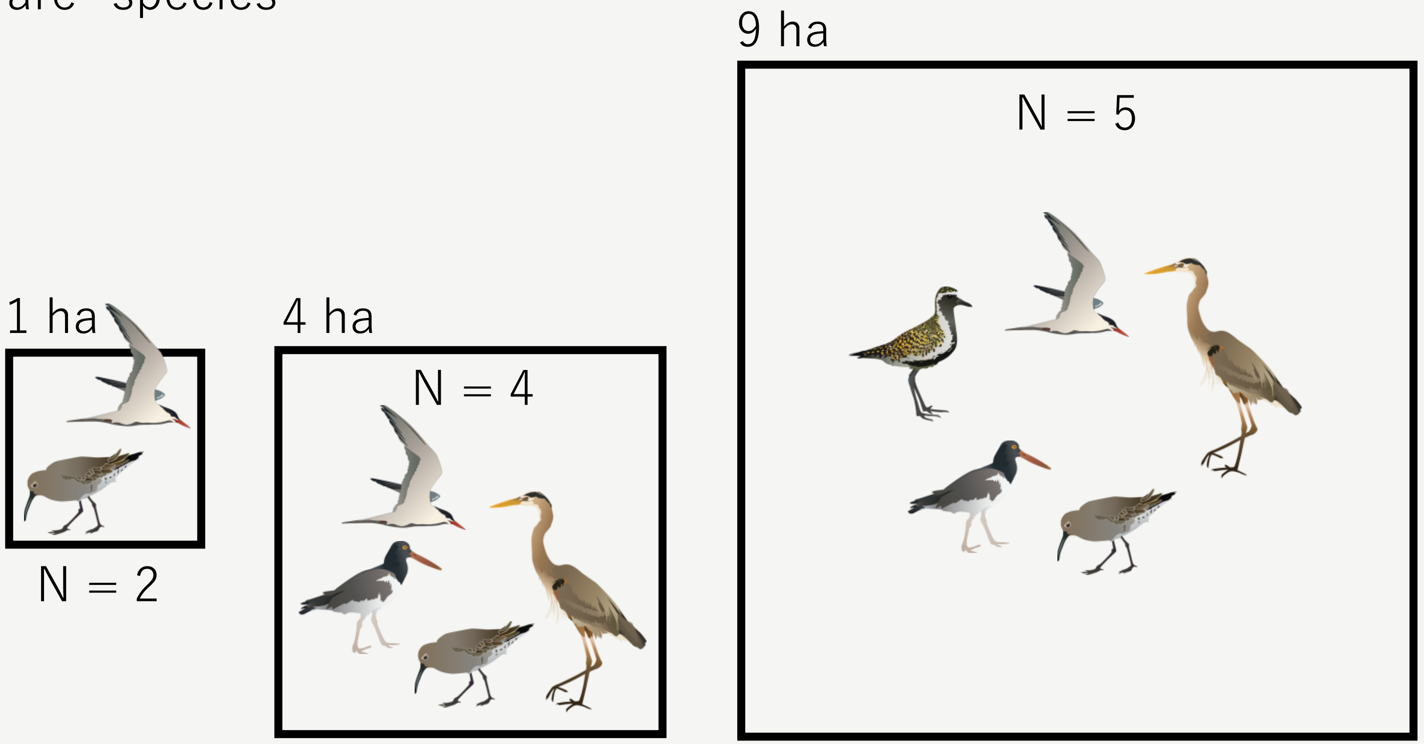
Species Richness: the number of species present in a given area

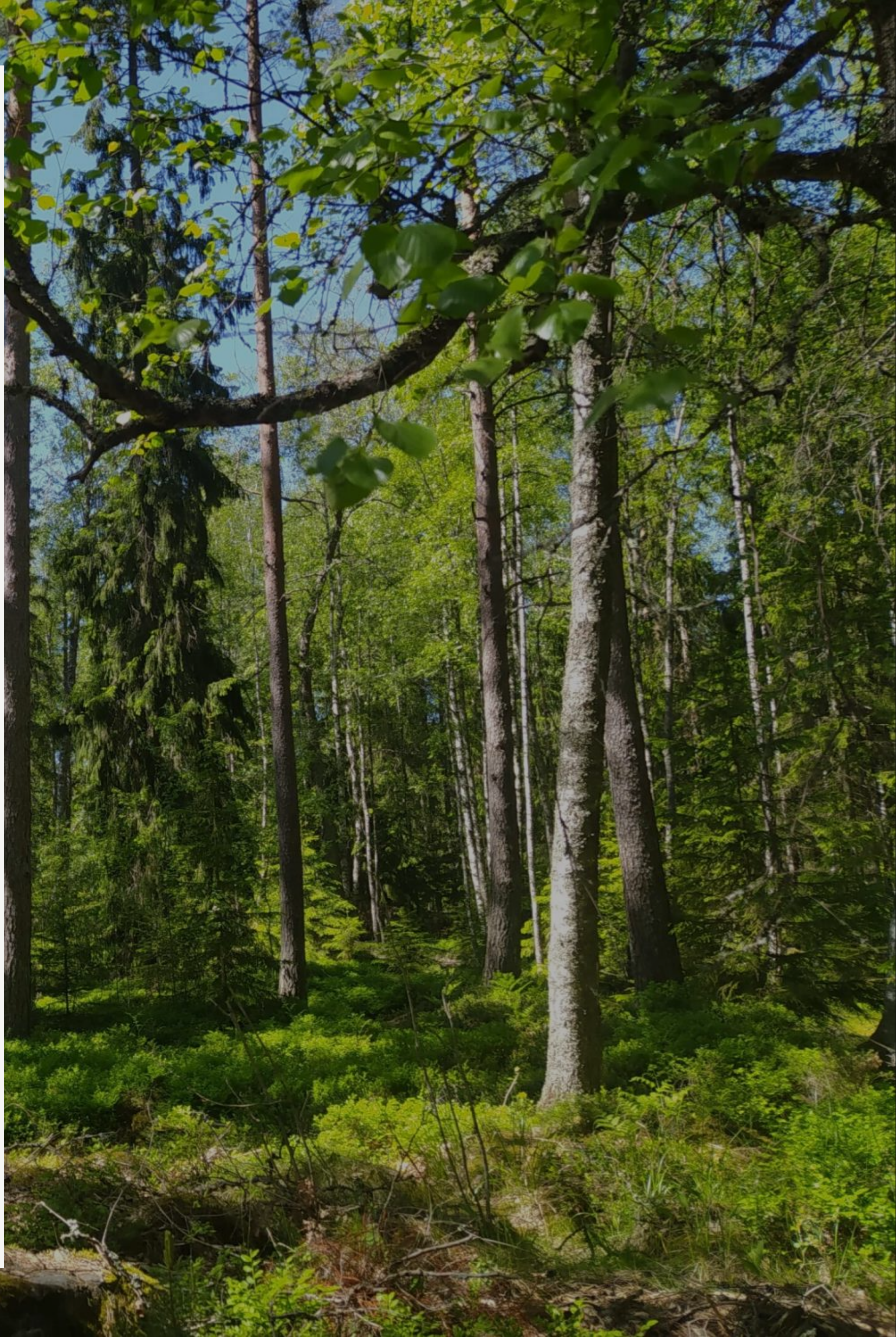




Univariate Metrics of Ecological Diversity: Species-Area Relationships

The larger the sampling area, the more likely one is to encounter "rare" species



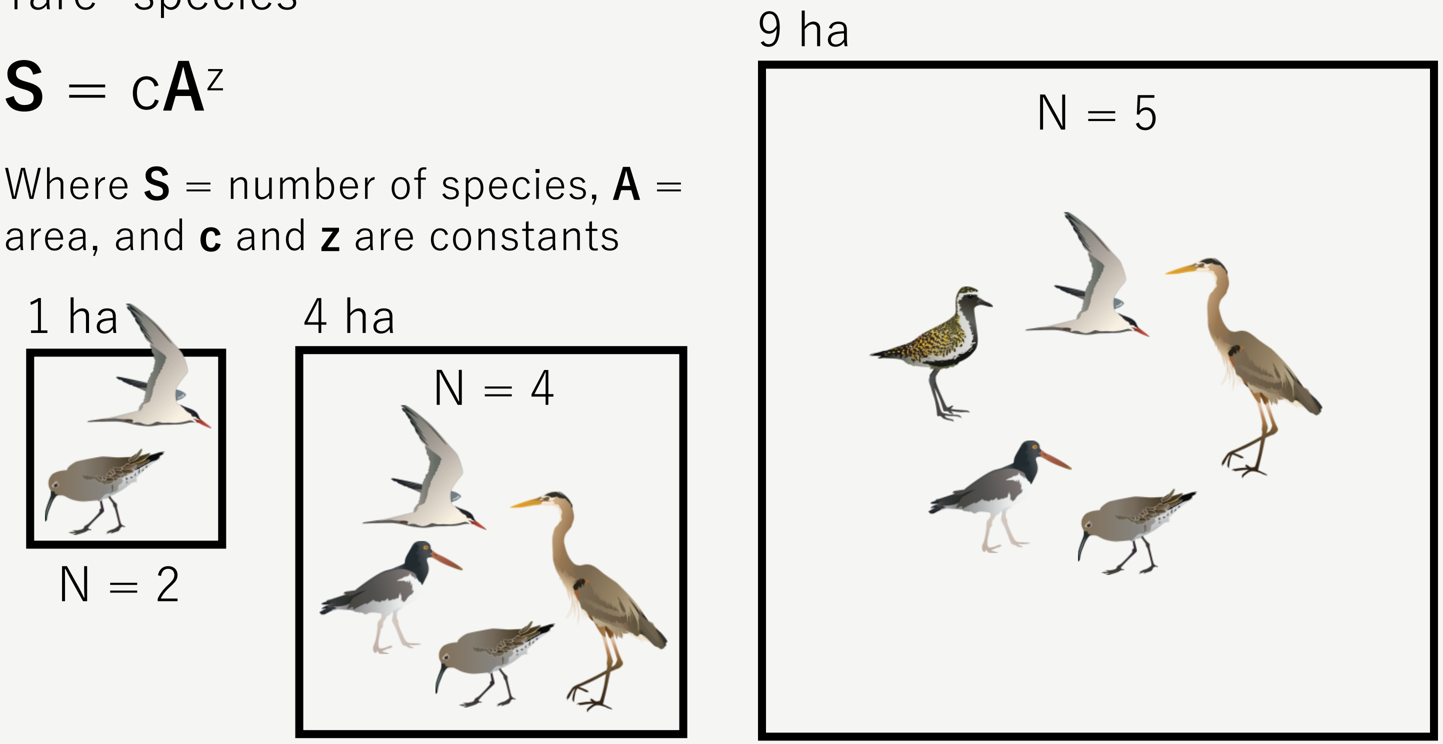


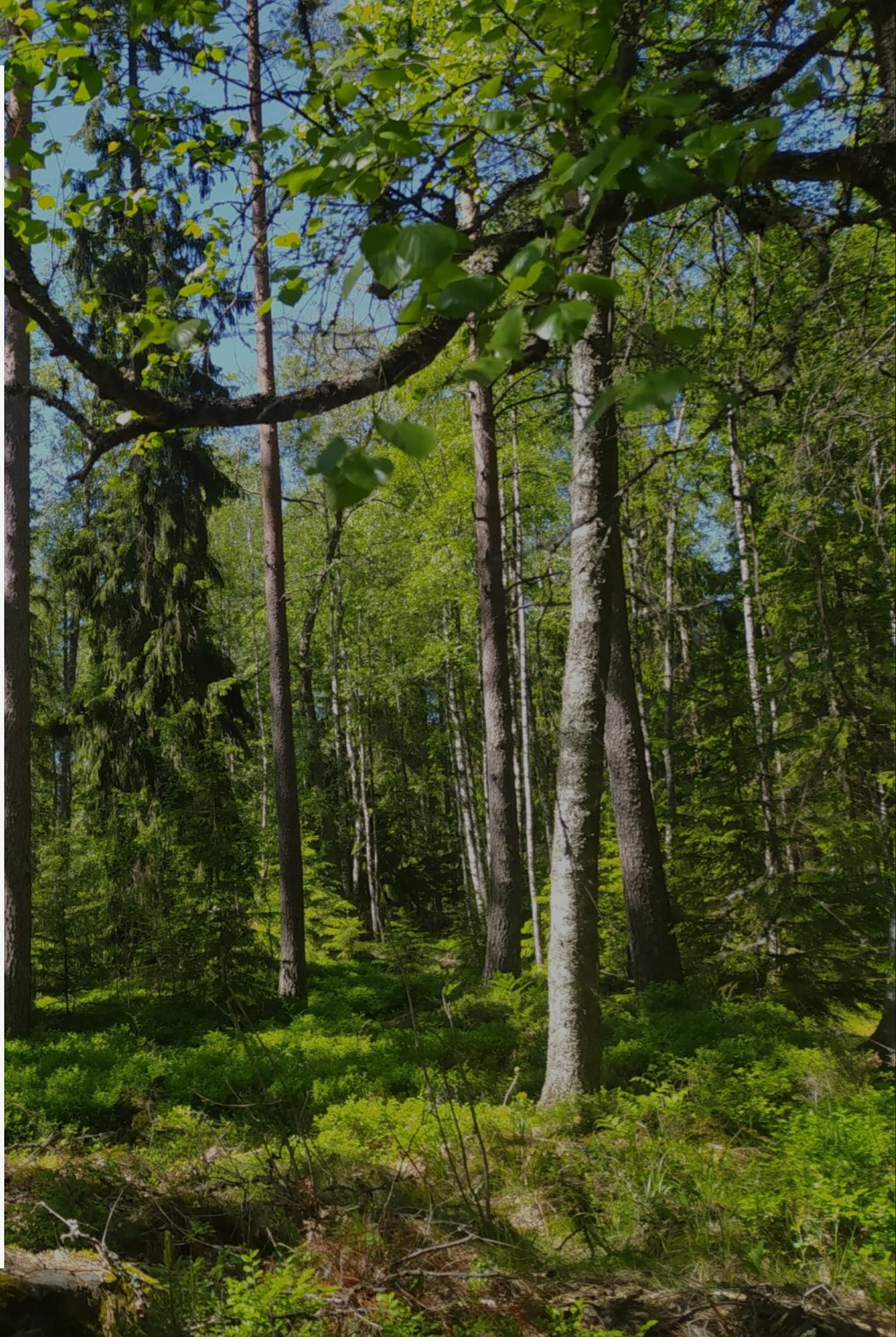
Univariate Metrics of Ecological Diversity: Species-Area Relationships

The larger the sampling area, the more likely one is to encounter "rare" species

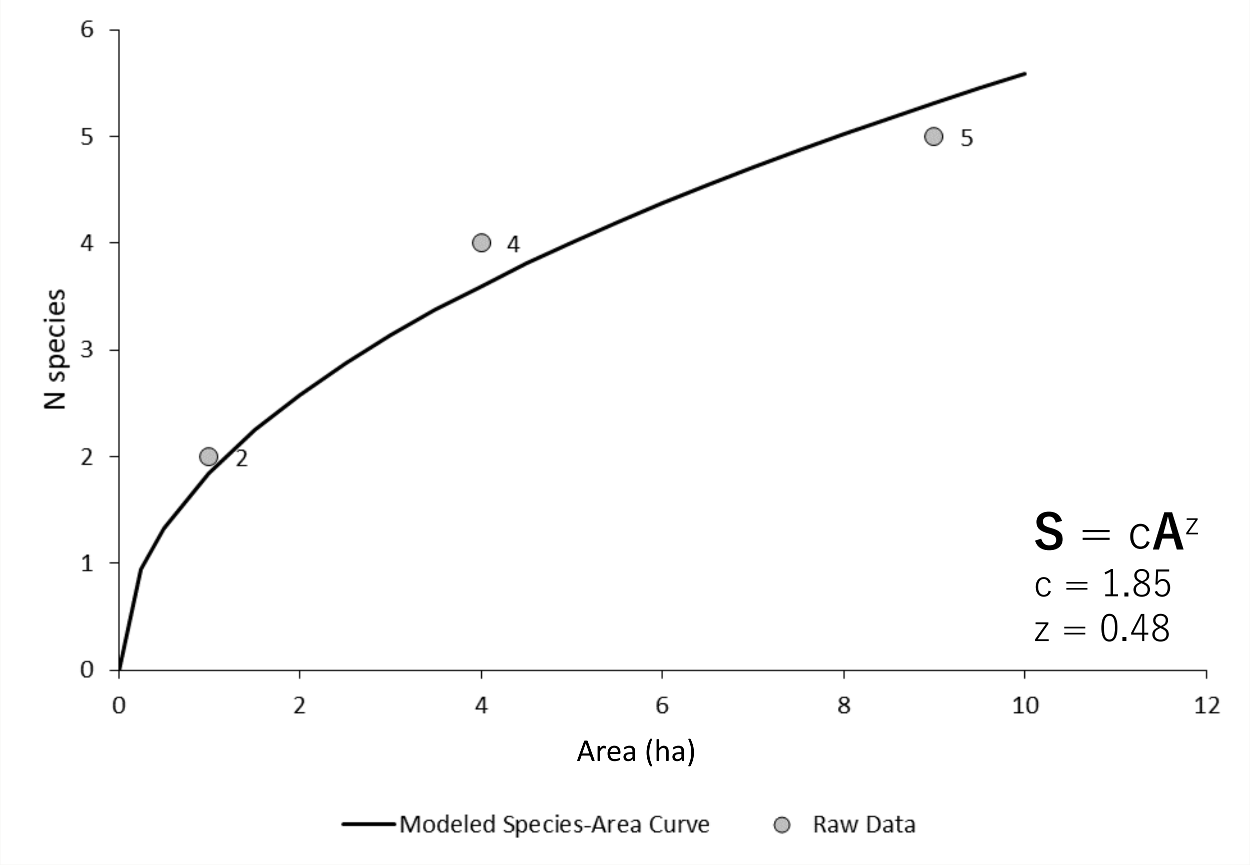
$$\mathbf{S} = c\mathbf{A}^{z}$$

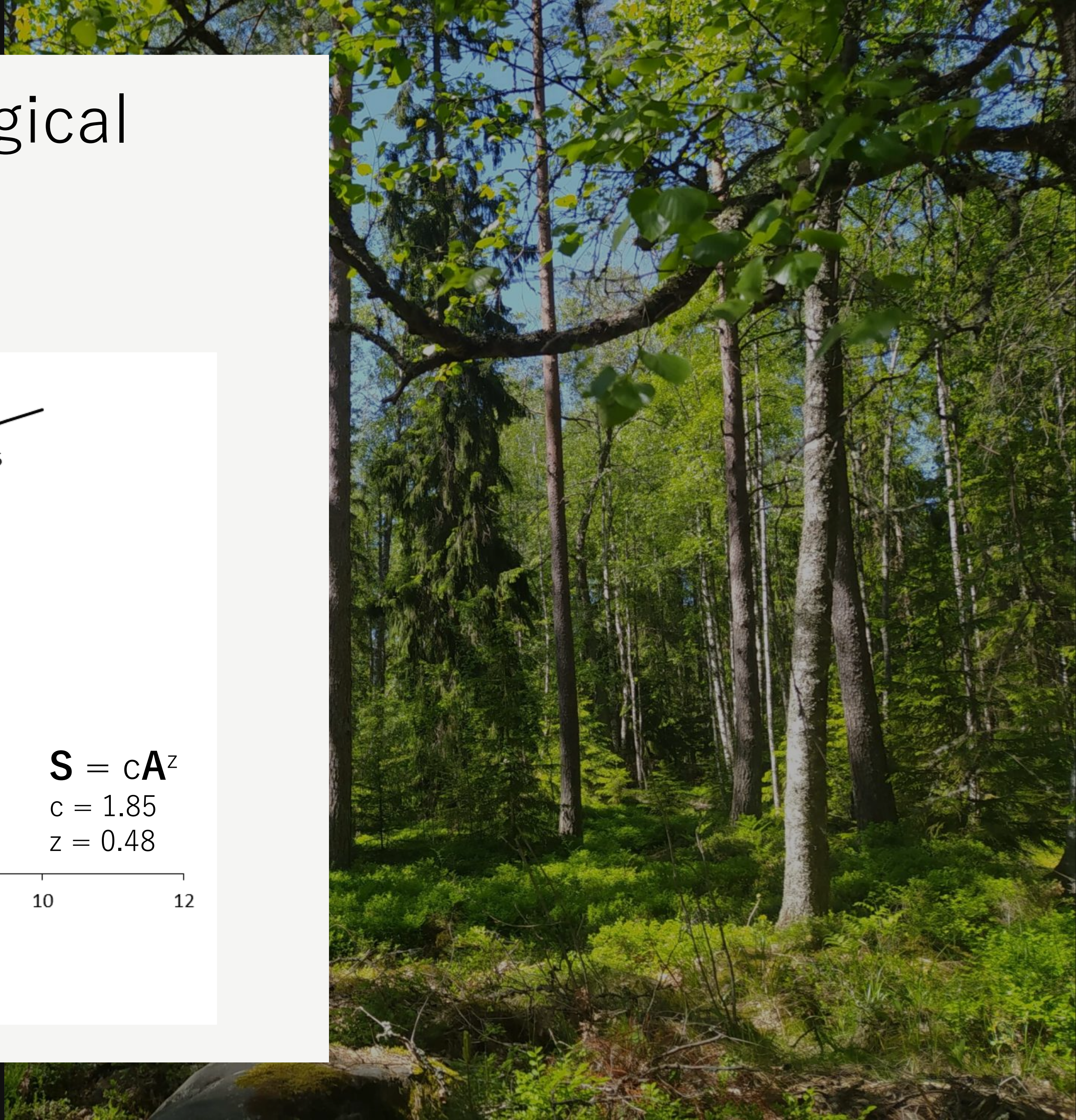
Where **S** = number of species, **A** = area, and **c** and **z** are constants





Univariate Metrics of Ecological Diversity: Species-Area Relationships





Univariate Metrics of Ecological Diversity: Species Diversity

Shannon-Weaver Diversity Index

H` $= -\Sigma\,(p_i \ln p_i)$

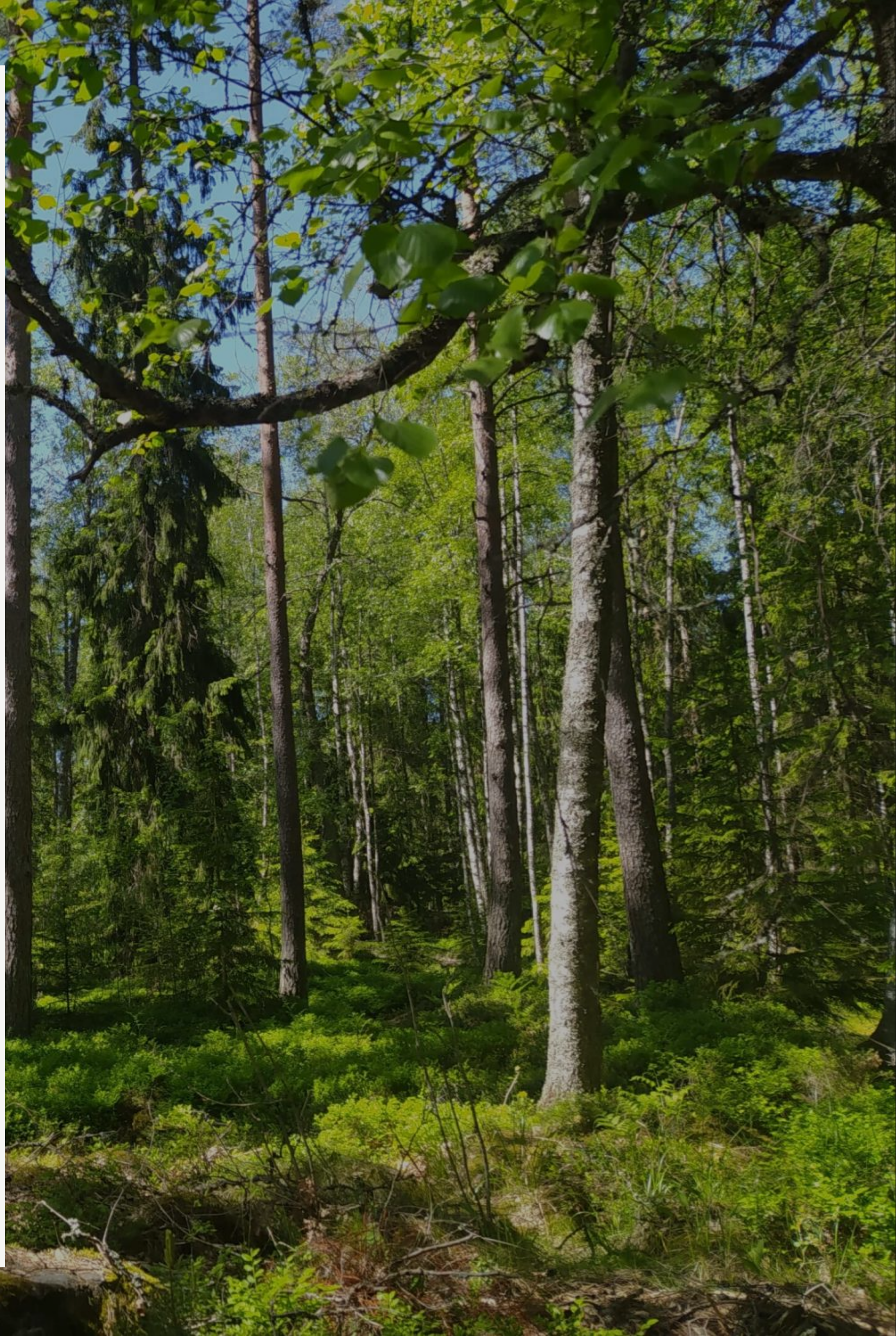
where p_i is the proportion of individuals of species i

Simpson’s Diversity Index

D $= 1 - \Sigma\, p_i^2$

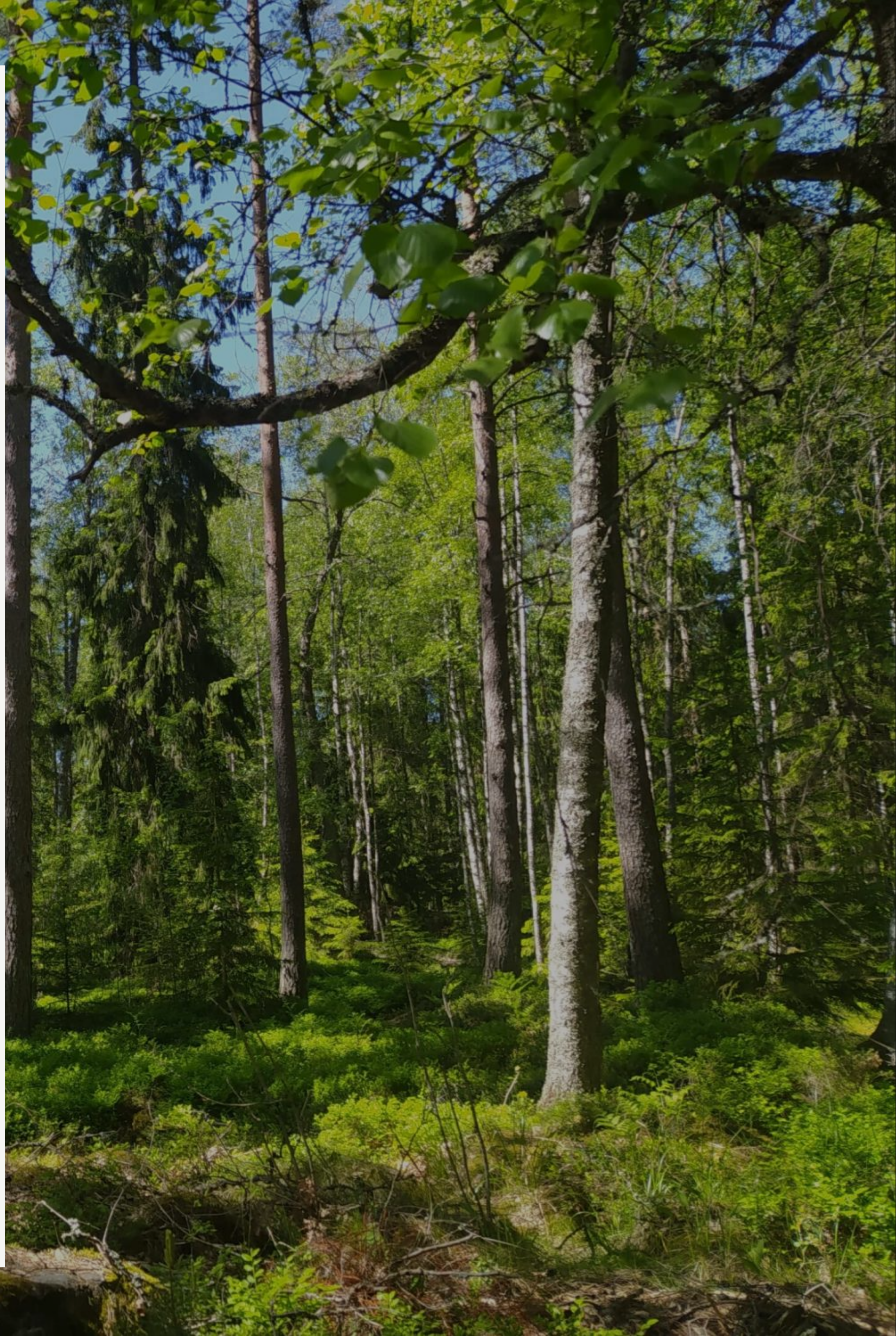
where p_i is the proportion of individuals of species i

Probability that two randomly selected individuals belong to a different species



Univariate Metrics of Ecological Diversity: Species Diversity

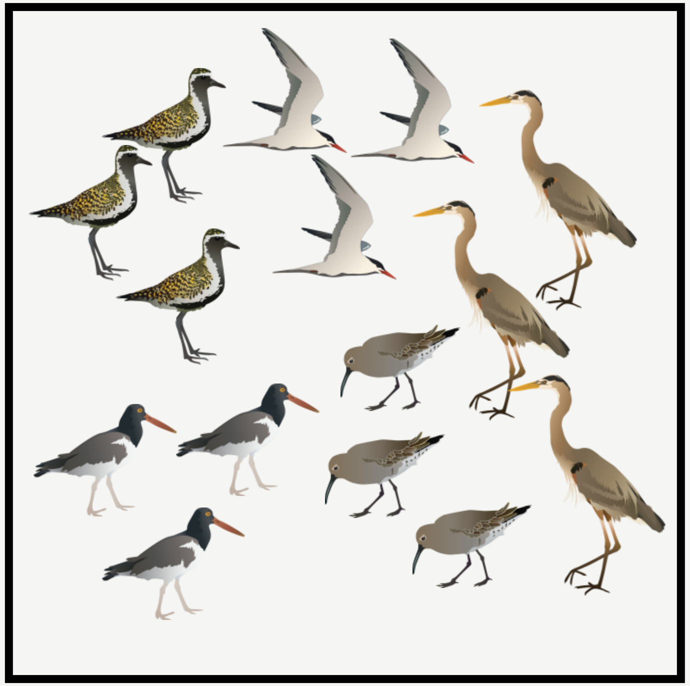
Criteria	Shannon-Weaver (H`)	Simpson’s (D)
Sensitivity to evenness	High	Moderate
Sensitivity to rare species	High	Low
Dominance weighting	Low	High
Ideal for high species richness	Yes	No
Ideal for low species richness	No	Yes
Community stability and dominance	Less suited	Well suited
Comparative studies of evenness	More informative	Less informative
Long-term monitoring	Suitable	Less sensitive to rare species changes
Appropriate for dominant species focus	Less appropriate	More appropriate



Univariate Metrics of Ecological Diversity: Species Evenness

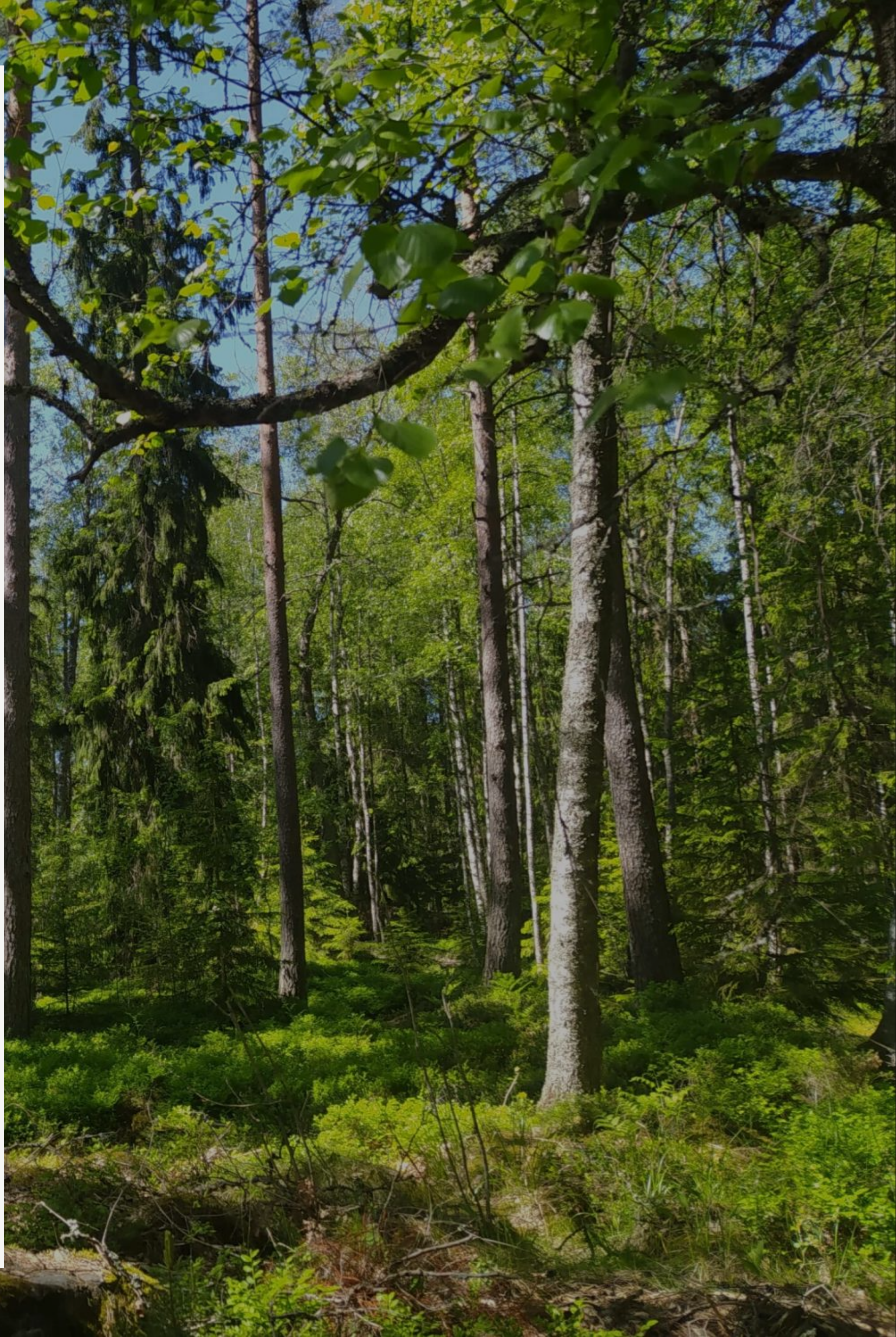
$\mathbf{E} = \frac{\mathbf{H`}}{\mathbf{\ln(S)}}$ where **H`** is Shannon-Weaver Diversity

$\mathbf{E} = \frac{\mathbf{D}}{\mathbf{S}}$ where **D** is Simpson's Diversity and **S** is species richness



H` = 1.61
E = 1.00

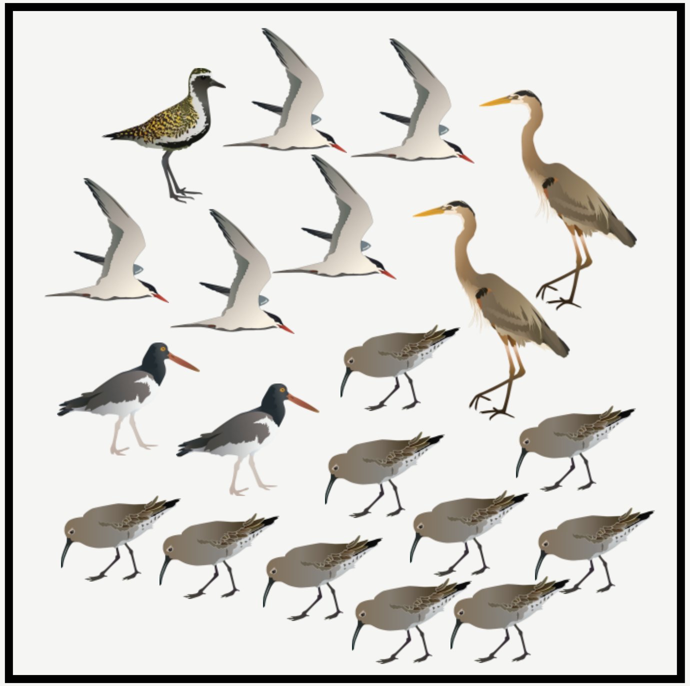
D = 0.80
E = 0.16



Univariate Metrics of Ecological Diversity: Species Evenness

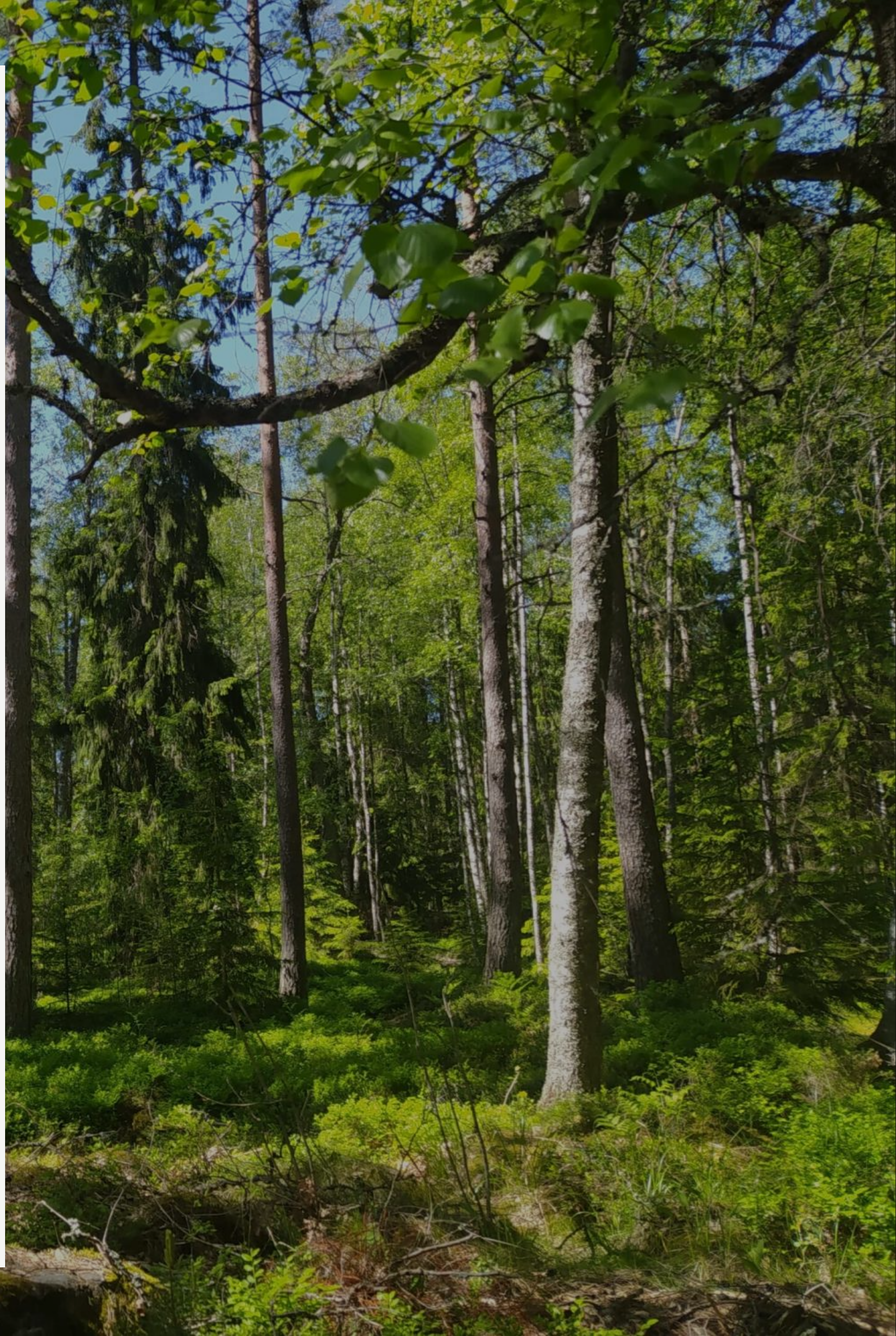
$\mathbf{E}=\frac{\mathbf{H`}}{\mathbf{\ln(S)}}$ where **H`** is Shannon-Weaver Diversity

$\mathbf{E}=\frac{\mathbf{D}}{\mathbf{S}}$ where **D** is Simpson's Diversity and **S** is species richness



H` = 1.30
E = 0.81

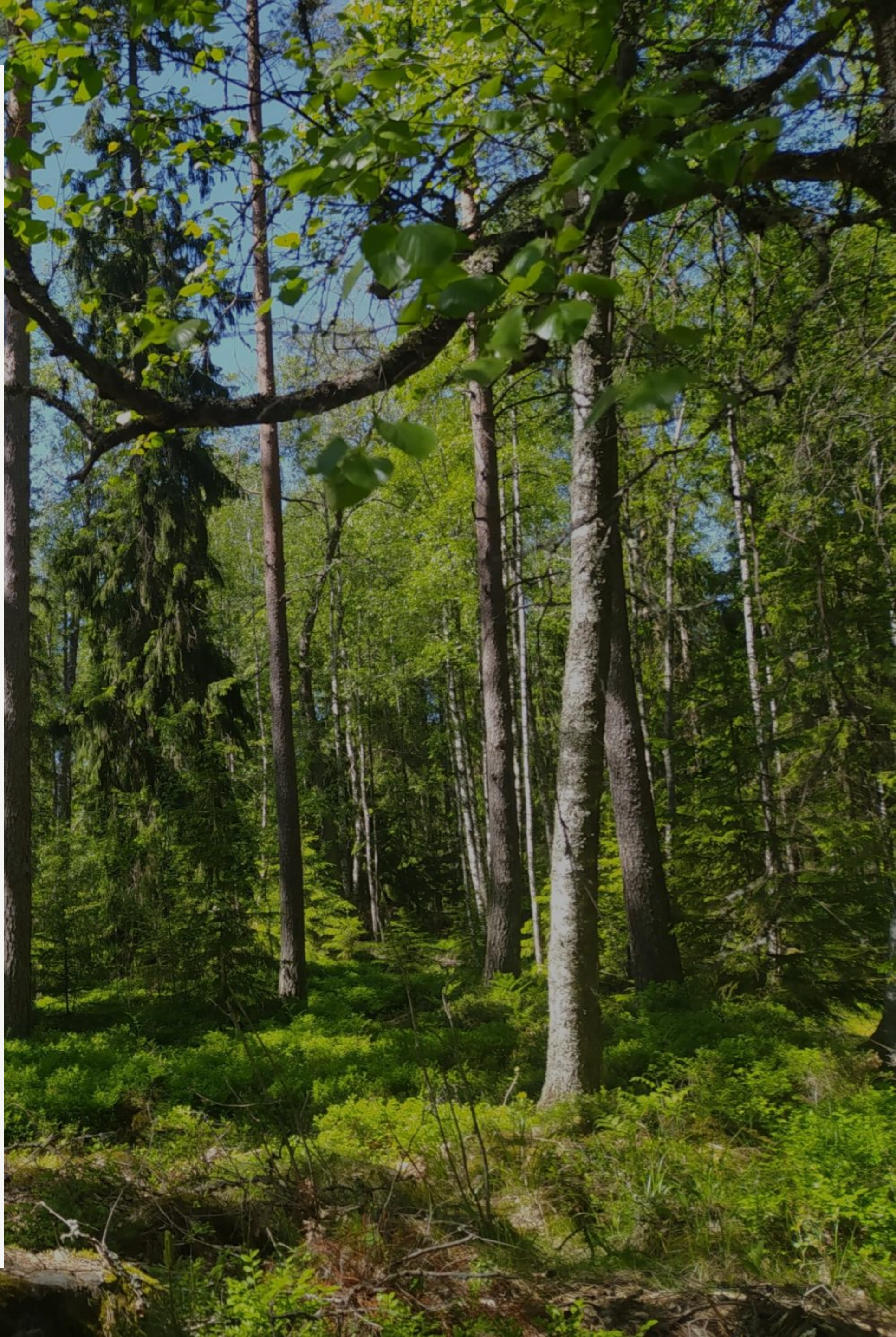
D = 0.67
E = 0.13



Univariate Metrics of Ecological Diversity: Functional Diversity

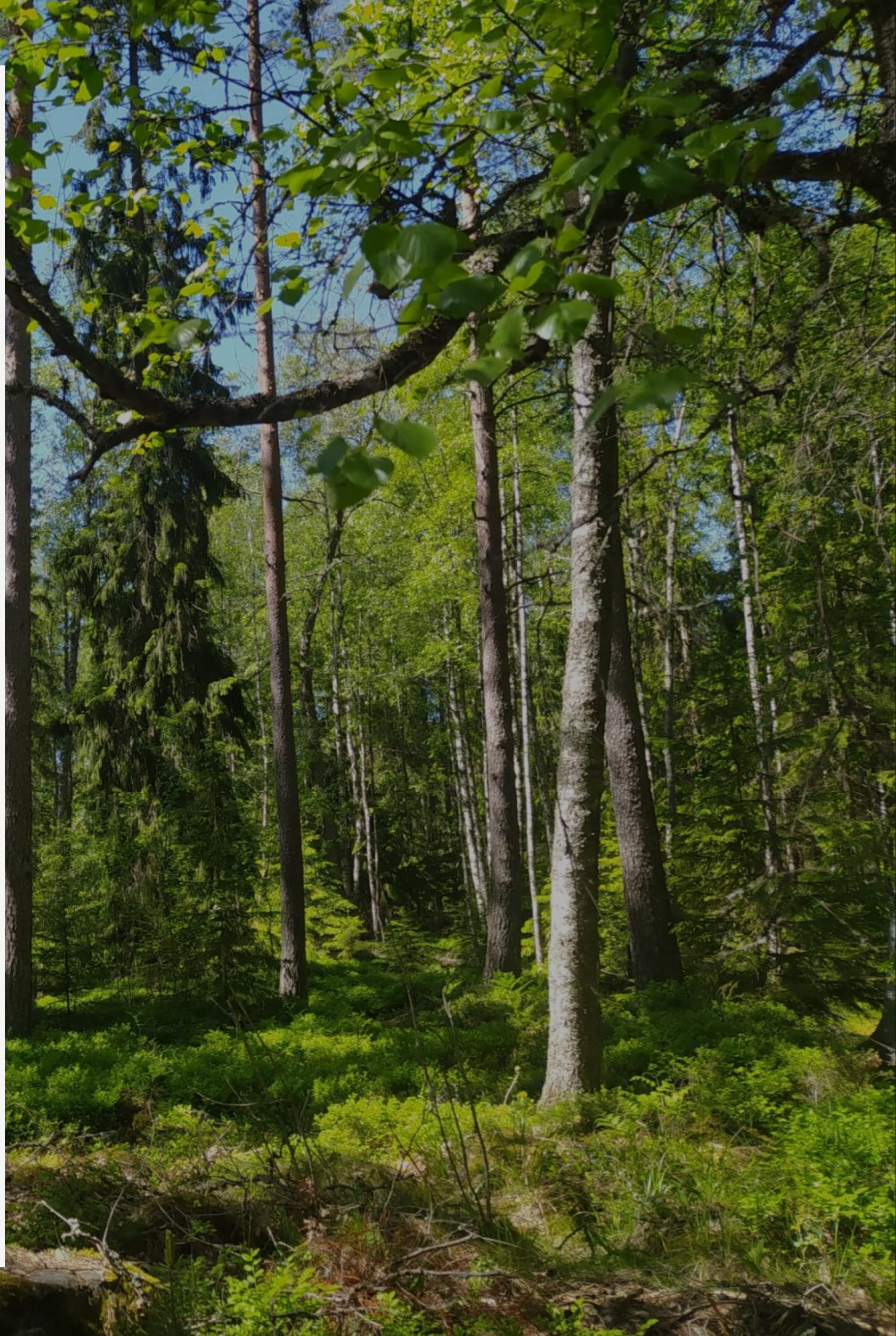
Functional Diversity: the range of different biological functions or roles that species within a community perform

- Feeding habits
- Reproductive strategies
- Other ecological roles



Conclusion: Summary of Key Points

- Is **transformation** necessary?
  - More likely for species data
- Is **standardization** necessary?
  - More likely for environmental data
- Have you appropriately accounted for differences in **sampling effort**?



Questions?