LEARNING PORTFOLIO: RESEARCH METHODS IN FOREST SCIENCES

AMAITUM JOSHUA ELUKUT

2216274

MSc European Forestry

School of Forest Sciences

University of Eastern Finland

UNIVERSITY OF
EASTERN FINLAND

1. BASIC CONCEPTS IN STATISTICS

Statistics: The discipline that concerns the collection, organization, analysis, interpretation, and presentation of data.

Descriptive statistics: Summarize and describe the prominent features of data.

Inferential statistics: Methods that deal with evaluation of information present in data and the assessment of the new learning gained from this information.

Population: The vast collection of all potential observations that can be conceived in each context.

Sample: The pool of individuals within a population from which data is obtained.

Sampling unit: The distinct source from which each measurement in a data set originates.

Experimental group: The group of individuals to which a treatment is applied, according to the specification of the variable of particular interest.

Control: The group from the sample to which no treatment is applied.

Mean: The average value of a dataset.

Variance: A measure of how far a set of number is spread from the average value.

Standard deviation: The square root of variance.

Deviation score: Measure of by how much each point in a set of data lies above or below the mean for the entire data set.

Range: The difference between the largest and smallest value in a data set.

Degrees of freedom: The maximum number of logically independent values, that is the values which have the freedom to vary in a data sample. (df = 1 – item in the samples)

Statistical hypothesis: A statement made about the nature of a population.

Testing hypothesis: Process of determining whether a claim about some feature of the population, a parameter, is strongly supported by information obtained from the sample data.

Null hypothesis (H_o): The opposite statement, one which nullifies the alternative hypothesis.

Alternative hypothesis (H_a): The claim or research hypothesis that we seek to establish.

Test statistic: A number calculated by a statistical test whose value serves to determine the action, that is either accepting or rejecting the null hypothesis.

Type I error: Rejection of null hypothesis that is true in the population.

Type II error: Failing to reject a null hypothesis that is false in the population.

p-value: A value that serves as a measure of the strength of evidence against the null hypothesis. A small p-value, below the threshold, means that the null hypothesis is strongly rejected or that the result is highly statistically significant. This value also informs the probability that the results occurred by chance.

Normal distribution: A bell-shaped distribution curve whose density function is symmetrical about its mean value.

Properties of a normal distribution

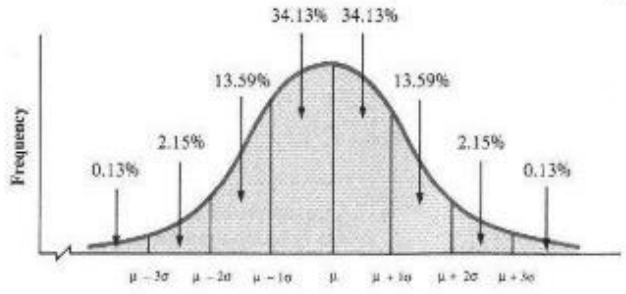
- The mean, mode and median are all equal.
- The curve is symmetric at the center (i.e., around the mean, μ).
- Exactly half of the values are to the left of center and exactly half the values are to the right.
- The total area under the curve is 1.

Descriptive Statistics Formula Sheet

	Sample	Population
Characteristic	statistic	Parameter
raw scores	$x, y, \ldots$	$X, Y, \ldots$
mean (central tendency)	$M = \frac{\sum x}{n}$	$\mu = \frac{\sum X}{N}$
range (interval/ratio data)	highest minus lowest value	highest minus lowest value
deviation (distance from mean)	Deviation = $(x - M)$	Deviation = $(X - \mu)$
average deviation (average distance from mean)	$\frac{\sum(x - M)}{n} = 0$	$\frac{\sum(X - \mu)}{N}$
sum of the squares (SS) *(computational formula)*	$SS = \sum x^2 - \frac{(\sum x)^2}{n}$	$SS = \sum X^2 - \frac{(\sum X)^2}{N}$
variance (average deviation^2 or standard deviation^2) *(computational formula)*	$s^2 = \frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1} = \frac{SS}{df}$	$\sigma^2 = \frac{\sum X^2 - \frac{(\sum X)^2}{N}}{N}$
standard deviation (average deviation or distance from mean) *(computational formula)*	$s = \sqrt{\frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}}$	$\sigma = \sqrt{\frac{\sum X^2 - \frac{(\sum X)^2}{N}}{N}}$
Z scores (standard scores) mean = 0 standard deviation = ± 1.0	$Z = \frac{x - M}{s} = \frac{\text{deviation}}{\text{stand. dev.}}$ $X = M + Zs$	$Z = \frac{X - \mu}{\sigma}$ $X = \mu + Z\sigma$

Area Under the Normal Curve

-1s to +1s = 68.3%
-2s to +2s = 95.4%
-3s to +3s = 99.7%



2. t-test and ANOVA

t-test and Analysis of variance (ANOVA) are examples of statistical tests, mathematical functions which provide a mechanism for making quantitative decisions about a process or processes.

t-test

t-test: A statistical test used to compare the means of two groups.

One sample t-test: A statistical test used to determine whether an unknown population mean is different from a specific value e.g., checking whether the mean weight of cereal boxes differs from 1.5kg.

Independent sample t-test: Compares the means of two independent groups to determine whether there is statistical evidence that the associated population means are statistically different e.g., Comparing earnings of randomly assigned individuals in two groups.

Paired samples t-test: A statistical procedure used to determine whether the mean difference between two sets of observations is zero e.g., comparing the blood pressure measurements of individuals before and after giving them a treatment. This would involve two measurements from a single sampling unit i.e., before and after.

Assumptions of t-test

- Independence: The observations in one sample are independent of the observations in another.
- Normality: Both samples are approximately normally distributed.
- Homogeneity of variances: Both samples have approximately equal variances.
- Random sampling method was used to obtain both samples.
- A reasonably large sample size is used.

ANOVA

ANOVA: Used to determine whether there are any statistical differences between three or more independent groups. There are two types of ANOVA tests, one-way and two-way and their characteristics are summarized in the table below.

	One-way ANOVA	Two-Way ANOVA
Definition	A test that allows one to make comparisons between the means of three or more groups of data	A test that allows one to make comparisons between the means of three or more groups of data, where two independent variables are considered
Number of Independent Variables	One	Two
What is Being Compared?	The means of three or more groups of an independent variable on a dependent variable	The effect of multiple groups of two independent variables on a dependent variable and on each other
Number of Groups of Samples	Three or more	Each variable should have multiple samples

Assumptions of ANOVA

- Homogeneity of variances: The variances across all the groups of between-subject effects are the same.
- Data is selected randomly.
- Normal population distribution.
- Samples are independent.

For this report, I examined one-way ANOVA concept practically. It is used when there is data on one quantitative dependent variable and one categorical independent variable which has at least three levels for example the type of fertilizer used in a crop field at different concentrations.

Example of ANOVA

Using data on different fertilizer concentration to examine impact on yield of a field.

H_0: There is no difference among group means

H_a: At least one group mean is different

Fertilizer Conc.	5%	10%	15%	20%
Field 1	7	12	14	19
Field 2	8	17	18	25
Field 3	15	13	19	22
Field 4	11	18	17	23
Field 5	9	19	16	18
Field 6	10	15	18	20

Anova: Single Factor

SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
5%	6	60	10	8
10%	6	94	15.66667	7.866667
15%	6	102	17	3.2
20%	6	127	21.16667	6.966667

ANOVA

Source of Varia	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between G	382.7917	3	127.5972	19.60521	3.59E-06	3.098391
Within Gro	130.1667	20	6.508333			
Total	512.9583	23				

From the analysis output, I observe that the p-value is less than 0.05, thus reject the null hypothesis and conclude that fertilizer concentration affects yield from the field. F (19.60521) is greater than F crit, so again we reject the null hypothesis.

Post-hoc testing

ANOVA tells us there are differences among the levels of the independent variable (in our case fertilizer), but not which differences are significant. To determine how treatment levels differ from one another, we perform a post-hoc test. I employed the Tukey-Kramer test, which compares the mean between each pairwise combination of groups.

First step is to determine the absolute mean difference between each group using averages listed in the ANOVA Output shared above. The pooled variance is calculated from the average of the variance of the groups, which is 6.508333

Next we find the Q critical value using the formula: Q critical value = $Q^*\sqrt{(s^2 pooled / n.)}$ where: Q = Value from Studentized Range Q Table; $s^2 pooled$ = Pooled variance across all groups; n. = Sample size for a given group. The Q value is obtained from the Studentized Range Q Table (k=4, and df=21). We obtained a Q value of 3.96.

Lastly, with a group size of 6, we calculate our Q Critical Value:

Q critical value = $3.96\sqrt{(6.508333/6)} = 4.124$

To conclude on which group means are statistically significant, we compare the absolute mean difference between each group to the Q critical value. The pairwise comparisons show that 20% fertilizer concentration has a significantly higher mean yield than the others, 15% has a significantly higher yield than 5%, and 10% has a significantly higher yield than 5% while the difference between the yield of 5% and 10% fertilizer concertation is not statistically significant.

Anova: Single Factor

SUMMARY

Groups	Count	Sum	Average	Variance		Pooled variance	
5%	6	60	10	8		6.508333	6.508333
10%	6	94	15.666667	7.866667			
15%	6	102	17	3.2			
20%	6	127	21.166667	6.966667			

ANOVA

Source of Variati	SS	df	MS	F	P-value	F crit
Between Gr	382.7916667	3	127.59722	19.60521	3.59E-06	3.098391
Within Grou	130.1666667	20	6.5083333			
Total	512.9583333	23				

Comparison	Abs. Mean Diff.	Q Critical Value	Significant	
5% vs. 10%	5.666666667	4.124	Yes	=IF(H31>I31,"Yes","No")
5% vs. 15%	7	4.124	Yes	
5% vs. 20%	11.16666667	4.124	Yes	
10% vs. 15%	1.333333333	4.124	No	
10% vs. 20%	5.5	4.124	Yes	
15% vs. 20%	4.166666667	4.124	Yes	

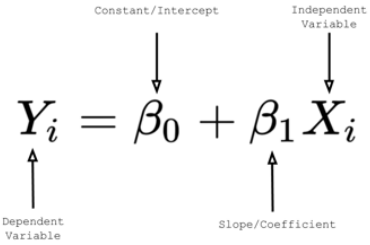
if **p>0.1**	supports $\mathbf{H_0}$
if **p>0.05**, but **p≤0.1**	Still supports $\mathbf{H_0}$
if **p≥0.01**, but **p≤0.05**	Not supporting enough $\mathbf{H_0}$: is dismissed and $\mathbf{H_1}$ accepted results is statistically almost significant: *
if **p≥0.001**, but **p≤0.01**	Not supporting enough $\mathbf{H_0}$: is dismissed and $\mathbf{H_1}$ accepted results is statistically significant: **
if **p≤0.001**	Not supporting enough $\mathbf{H_0}$: is dismissed and $\mathbf{H_1}$ accepted results is statistically very significant: ***

Interpretation of p-value

3. BASICS OF MODELING: SIMPLE REGRESSION

Regression is a statistical method used to predict a dependent output variable based on the values of independent input variables.

Linear regression is that which assumes a linear of straight-line relationship between the dependent variable and each predictor. Simple linear regression employs a single explanatory variable while multiple linear regression employs several explanatory variables.



Assumptions

- Linearity: The relationship between X and the mean of Y is linear.
- Homoscedasticity: The variance of residual is the same for any value of X.
- Independence: Observations are independent of each other.
- Normality: For any fixed value of X, Y is normally distributed.

When performing a linear regression analysis, the results yield an R^2 value which explains the strength of the relationship between the model and the dependent variable (the proportion of variance of the dependent variable that is explained by the independent variable). An ideal R^2 value is around 95% which means that 95% of the variability in the dependent variable id explained by the dependent variable and the other is explained by another factor.

Linear regression also primarily involves finding the best-fitting line, which can adequately explain how y varies with changes in x. It is defined by the intercept (where it crosses the y-axis) and slope values (and with the x-axis).

The ANOVA table tests the model's ability to explain any variation on the dependent variable but does not directly address the strength of the relationship. Furthermore, if ANOVA table Sig<0.05, we can state that it is a valid model, and the variation is accounted for by the model.

4. ADVANCED MODELS: ALTERNATIVES TO SIMPLE REGRESSION

Multiple linear regression is a model used to estimate the relationship between two or more explanatory variables and one response variable. Many relationships may require more than one variable for example when determining tree volume, it may require using parameters such as tree height and DBH.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_i X_i$$

Y : Dependent variable
β_0 : Intercept
β_i : Slope for X_i
X = Independent variable

The intercept B_0 is the value of the dependent variable when all the independent variables are zero. B_1 and B_2 are the slope coefficients for the independent variables X_1 and X_2. These represent the magnitude by which the dependent by which the dependent variable changes when they increase by one unit, while all other independent variables are constant. The intercepts and coefficients can be either negative or positive.

Five main assumptions underlying multiple regression models must be satisfied: (1) linearity, (2) homoskedasticity, (3) independence of errors, (4) normality, and (5) independence of independent variables. Diagnostic plots to determine whether these assumptions are satisfied. Scatterplots of dependent versus and independent variables are useful for detecting nonlinear relationships, while residual plots are useful for detecting violations of homoskedasticity and independence of errors.

Steps taken when checking adequacy of a model

- Makes sense.
- Visual exam.
- Fit model.
- ANOVA significance.
- Examine B_0, B_1, and standard error.
- B_0 and B_1 significance.
- R^2
- Errors
- Multicollinearity.

5. VALIDATING OF MODELS

A model is a simplified representation of some aspect of reality. Models aim to explain the behavior of systems and the right one is that which is most useful for the application needed and the choice must be based on the application and the resources available. Now we start with the basics.

The right model is one that is most useful for the application needed and its choice is based on the available parameters and resources. When a model is created, it is important to test its viability, and this can be done either through comparing its predictions with comparable values from which it was developed or from a completely new dataset altogether.

This can involve visual analysis of the plot of the modelled versus reference values to see how well they fit together. This involves analysis of the plot of residuals against fitted values as well as the histogram of the residuals. Residuals are defined as the difference between observed and model predicted values of the dependent variable. Usually, the first model developed may not be the best option and may display heteroscedastic patterns in the plot. This would require modifications like adding more variables to the model or transforming it using quadratic or exponential functions. In addition, a good model should have a normally distributed histogram of the residuals

Characteristics of residual plots which imply there is a good linear relationship in the model

- No odd patterns or curved trends in the plot
- The sum of the residuals is zero.
- The points are equally represented above and below the X-axis

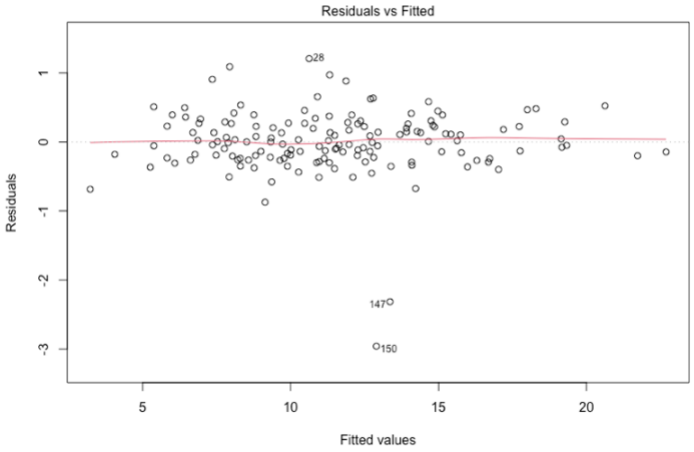
Furthermore, inferential concepts such as Root Mean Square Error (RMSE) and Bias can be used to validate the model. RMSE is used to check the precision of the model. It depicts how much scatter exists between the modeled and validated values. On the other hand, BIAS checks the accuracy of the model. It shows how much the average level of the modeled values differs from the validated values.

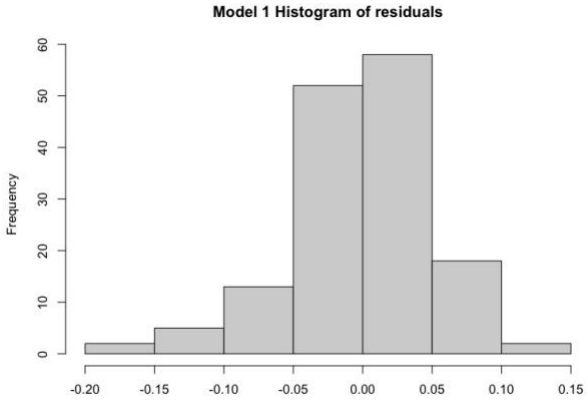
The following are key when steps when validating a model

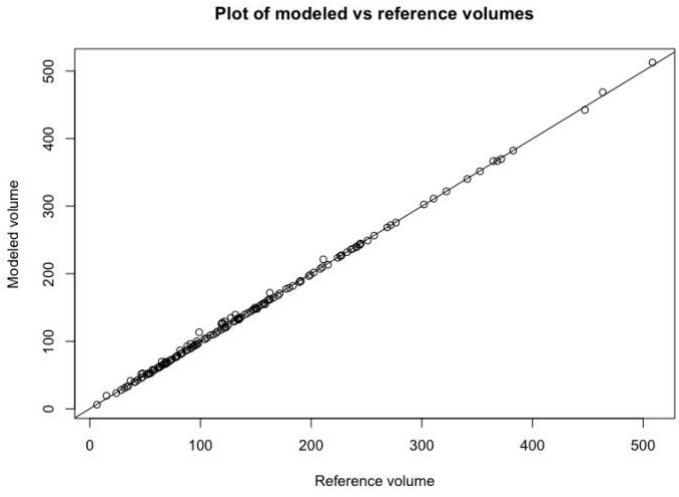
- Check whether B_0 and B_1 are statistically significant in the ANOVA and t-test.
- Residuals should be random.
- No heteroscedasticity.
- Independent variables are not multicollinear.
- Data follows a normal distribution.
- Check for outliers.
- Check the model's performance in a different dataset.

Ways of improving the model

- Transforming the variables e.g., by using the log function, squaring a parameter, or developing a reciprocal.
- Adding more variables e.g., when estimating volume, if diameter is not sufficient, consider adding height and basal area.
- Fixing non-linearity







6 . GIS TOOLS

Geographic Information System (GIS) is a computer-based program that can create, manage, analyze, and visualize different types of geographic data. It connects data on a map integrating location with descriptive information.

There are two types of coordinate systems namely: Geographic coordinate system (location measured from curved surface of the earth with units in in latitude and longitude) and Projected coordinate system (location based on flat surface with measurements units in feet, inches, and meters). When moving between them, distortions may occur except for very fine scale maps.

Ways of representing GIS data

Vector data: This GIS data type consists of points, lines, or polygons.

Advantages

- Efficient representation of topology.
- Adapts well to scale changes.
- Allows representing networks.
- Allows easy association with attribute data.

Disadvantages

- Complex data structure.
- Overlay more difficult to implement.
- Inefficient for image processing.
- More update intensive.

Raster data: Consists of a matrix of cells or pixels, each cell containing a value. Includes aerial photos, satellite imagery and scanned maps.

Advantages

- Simple data structure.
- Simple implementation of overlays.
- Efficient for image processing.

Disadvantages

- Less compact data structure.
- Difficulties in representing topology.
- Cell boundaries independent of feature boundaries.

Key GIS terminologies

Map projection: Mathematical formula for representing the curved surface of the earth on a flat map. Universal Transverse Mercator (UTM) is one of the most used map projections and it divides the earth into 60 zones, 6 degrees wide of longitude. Other common ones include Lambert azimuthal equal-area projection.

Topology: Expresses the spatial relationship of adjacent connecting features (point, lines, and polygons) and how they share the geometry.

Spatial analysis: Process where several thematic map layers are combined and overlaid.

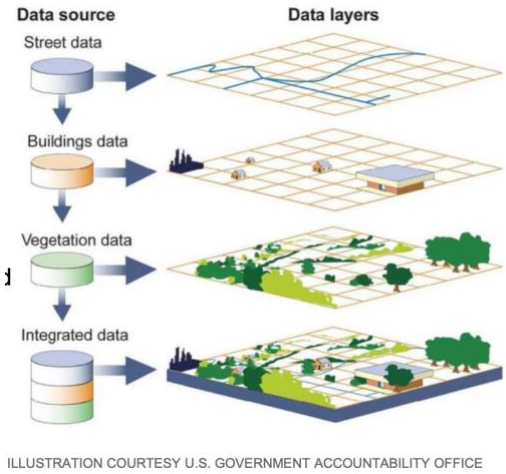
Spatial data: Any type of data which directly or indirectly references a specific geographic area, region, or location (describes the absolute and relative locations of geographic features).

Non-spatial data/Attribute data: That which is independent of geographic location (describes the characteristics of a geographic feature.).

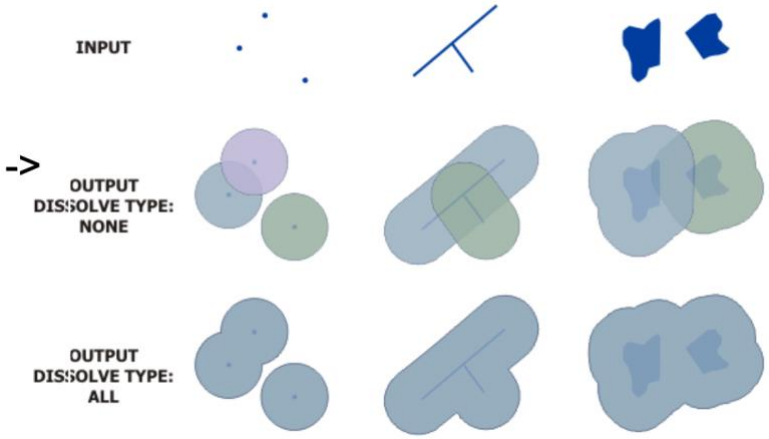
Database queries: A request using attribute table variables.

Spatial queries: Those which use spatial location e.g., intersection.

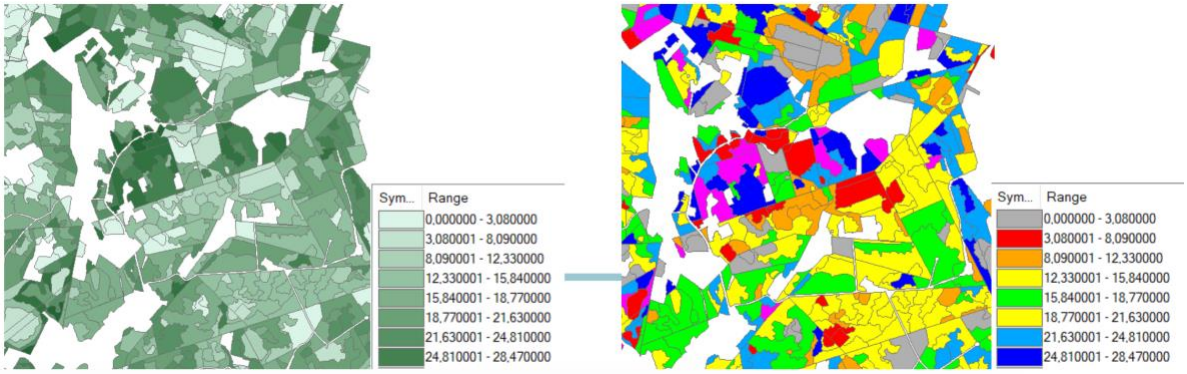
Buffers: Layers where specific distances from objects are defines



An illustration of several layers combined to develop a map



Buffers created around different map features



Different ways of presenting data on a map

7. REMOTE SENSING APPLICATIONS

Remote sensing refers to the process of obtaining information about objects or areas by using electromagnetic radiation without being in direct contact with them. In the field of forestry, it can be used for monitoring biomass as well as estimating forest structural attributes such as tree density to inform forest management operations.

It can be either passive (energy detected by sensor is coming from somewhere besides the sensor itself e.g., from the sun) or active (not dependent on an external source of illumination or radiation, instead sending energy itself at different wave lengths).

Different platforms are used in remote sensing, and these include satellites, airborne, terrestrial or mobile remote sensing tools. Each of these has different spatial, spectral, and temporal scales which determine their suitability for different purposes.

Predictions that can be made with remote sensing

- Above-ground vegetation water content.
- Measuring forest biomass to predict forest carbon stocks.
- Development of three-dimensional coordinates of targets based on images (through photogrammetry)
- Natural hazards study for example water pollution.
- Tracking changes in landscapes such as forests and glaciers over time.
- Mapping of forest fires to properly prepare their control.

Useful remote sensing tools in forestry

- Aerial photography.
- Multispectral scanners.
- Radio detecting and ranging.
- Light detection and ranging (Lidar).
- Videography data.
- Hyperspectral imaging.
- Airborne laser scanning.
- Thermal remote sensing.

REFERENCES

Johnson, R., & Bhattacharyya, G. (2019). Statistics : principles and methods (8th Edition). New York: Wiley.

https://www.statisticshowto.com/probability-and-statistics/normal-distributions/#:~:text=Properties%20of%20a%20normal%20distribution,under%20the%20curve%20is%201.

https://www.yumpu.com/en/document/view/33885377/descriptive-statistics-formula-sheet-sample-population-