Alma Mater Studiorum · University of Bologna

Department of Physics and Astronomy
Department of Biological, Geological and Environmental Sciences

Master Degree in Science of Climate

PRELIMINARY EVALUATION OF FOREST BIODIVERSITY AND LANDSCAPE CLASSIFICATION USING ENMAP DATA

Supervisor:
Prof. Duccio Rocchini

Co-supervisor:
Dr. Michele Torresani

Submitted by:
Carlotta Zanetti

Academic Year 2023/2024

Abstract

Forest biodiversity and landscape classification are key issues for environmental management, particularly in the context of climate change. This thesis analyses the use of multispectral and hyperspectral remote sensing data for landscape classification and forest biodiversity assessment in the South Tyrol region.
This study is divided into two main parts: the classification of land cover using the Random Forest algorithm with multispectral (Sentinel-2, Landsat-8) and hyperspectral (EnMAP) satellite imagery, and the assessment of forest biodiversity using the Spectral Variation Hypothesis (SVH). The first part focuses on evaluating the accuracy of land cover classification by comparing the performance of different satellite data, while the second part explores biodiversity estimation by relating field data to the spectral heterogeneity of images from Sentinel-2 and EnMAP.
The Random Forest algorithm proved effective in identifying land cover types, including areas devastated by Storm Vaia, for all three satellites, demonstrating the value of remote sensing for monitoring environmental changes.
However, the application of the SVH for biodiversity assessment has shown mixed results: while the multispectral data from Sentinel-2 have provided good results in estimating biodiversity, the hyperspectral data of EnMAP did not produce any significant correlations with field data. Despite the high spectral resolution of EnMAP, its application to the SVH has not met expectations in terms of biodiversity assessment. Rao's Q index, used to quantify functional diversity, demonstrated its usefulness when combined with spectral data, although there were limitations in EnMAP data.
This study represents the first attempt to test the SVH using EnMAP images, highlighting both the strengths and weaknesses of remote sensing technologies, with a particular focus on EnMAP's hyperspectral data, for monitoring forest ecosystems.

Contents

Chapter 1

Introduction

1.1 Objectives

The main objective of this study is to evaluate the effectiveness of both multispectral and hyperspectral remote sensing data in monitoring biodiversity and in classifying forest landscapes. In particular, the study aims to:

- Assess the accuracy of soil classification using multispectral (Sentinel-2 and Landsat-8) and hyperspectral (EnMAP) data, through the application of the Random Forest algorithm, to identify areas of vegetation cover in the South Tyrol region.
- Apply the SVH to remote-sensing data to estimate forest biodiversity, comparing the performance of the multispectral satellite Sentinel-2 with those of the hyperspectral satellite EnMAP.
- Identify the limits and advantages of using hyperspectral data compared to multispectral data for the study of biodiversity and ecological complexity, with a particular focus on the potential of EnMAP to detect fine spectral variations and its applicability in biodiversity estimation.

1.2 Thesis Structure

This thesis is organized into six main chapters, which describe the theoretical background, the methods used, the results obtained and the final conclusions. The structure of the thesis is as follows:

- **Chapter 1 - Introduction**: It presents the objectives of the study and provides an overview on the importance of remote sensing for biodiversity monitoring and

landscape classification, with a focus on multispectral and hyperspectral technologies.

- **Chapter 2 - Literature Overview**: It analyses the state of the art on remote sensing technologies and their use in ecology, focusing on methodologies for land cover classification and biodiversity assessment. Machine learning algorithms, such as the Random Forest, and the SVH are described.

- **Chapter 3 - Data and Methods**: It describes the data used in the study, including the Sentinel-2, Landsat-8 and EnMAP satellites, and field data collected. The pre-processing process of satellite images and the application of the Random Forest algorithm for the classification and the SVH for biodiversity estimation are detailed.

- **Chapter 4 - Results**: It presents the results of the landcover classification for each satellite and the results of the application of the SVH. The classification accuracy and the capacity of remote sensing data to estimate biodiversity are evaluated

- **Chapter 5 - Discussion**: It interprets the results, comparing the performance of different satellites and discussing the limitations and advantages of the techniques used. It particularly focuses on the results of the classification and the applicability of the SVH.

- **Chapter 6 - Conclusion**: It summarises the main findings of the study and suggests potential future developments, such as improved remote sensing techniques for biodiversity and the expansion of EnMAP applications in complex ecological contexts.

Chapter 2

Literature Overview

Remote sensing has become an indispensable tool in ecology, particularly in assessing biodiversity and classifying landscapes at large scales. The integration of advanced imaging technologies, such as multispectral and hyperspectral sensors, with machine learning algorithms has opened up new possibilities for monitoring environmental changes and improving conservation efforts. This chapter explores the development of remote sensing technologies, their applications in biodiversity assessment, and the key methodologies—such as Random Forest algorithms and the SVH—used for landscape classification and biodiversity analysis.

2.1 Remote Sensing

Remote sensing is pivotal for biodiversity assessment and landscape classification, as it enables the collection of large-scale, repeated data providing insights into ecological structure, composition, and function. The technology allows for the identification and monitoring of different land cover types, essential for understanding habitat diversity and ecosystem dynamics. By leveraging multispectral and hyperspectral data from satellites like EnMAP, researchers can classify land cover with high precision, detecting subtle variations in vegetation and other landscape features. This approach enhances our ability to assess and manage biodiversity, offering a powerful tool for tackling global environmental challenges.

2.1.1 History and Development

Remote sensing, as a modern field, has evolved significantly over the last few decades due to technological advances and the growing need to monitor and understand our planet from a global perspective. Defined as the acquisition of information on an object or phenomenon without physical contact, remote sensing revolutionized the way we study

the Earth's surface and environment. Although aerial photography and other remote observation techniques date back more than a century, remote sensing as we know it today took shape mainly from the mid-20th century with the introduction of the first artificial satellites [Cracknell, 2018].

The earliest forms of remote sensing date back to the 19th century, with the use of hot air balloons and passenger pigeons equipped with small cameras for taking aerial photos. These experiments, although rudimentary, represented the first attempts to capture images of the Earth from above, foreshadowing future remote sensing techniques [Cracknell, 2018]. The first aerial photographs are attributed to Gaspard-Félix Tournachon, who conducted an experiment with a balloon in 1858. These early efforts marked the beginning of a new era in Earth observation, which would eventually find applications in cartography, geology, and urban planning [Dovgyi et al., 2019].

The launch of Sputnik in 1957 marked the beginning of the space age and, with it, the start of the first remote sensing applications from space. This satellite, launched by the Soviet Union, was the first artificial object to orbit the Earth, paving the way for a new era of global observation.
In the following years, the USA and the USSR launched a series of meteorological and Earth observation satellites, such as TIROS (Television Infrared Observation Satellite) and NIMBUS, which allowed large-scale monitoring of atmospheric phenomena and collected data crucial for meteorology and environmental science [Cracknell, 2018].

In 1972, the launch of Landsat-1 (initially called the Earth Resources Technology Satellite) marked the beginning of NASA's first satellite remote sensing program dedicated to observing the Earth's surface. This program was a turning point, providing multispectral data used for mapping, agriculture, natural resource management, and many other fields [Cracknell, 2018].

The year 1978 is considered crucial for remote sensing, with the launch of three key satellites: TIROS-N, SEASAT, and NIMBUS-7. These satellites introduced significant innovations, such as the AVHRR (Advanced Very High Resolution Radiometer) on TIROS-N and the SAR (Synthetic Aperture Radar) on SEASAT, which enabled the first radar images of the Earth's surface to be captured through clouds and darkness [Cracknell, 2018].

Since 1978, remote sensing has expanded in unprecedented ways, with the introduction of increasingly advanced satellites with multispectral and hyperspectral acquisition capabilities. The Landsat, SPOT, and Sentinel satellites have provided high spatial and temporal resolution images that are crucial for a variety of applications, ranging from natural resource management to urban planning. The combination of remote sensing and

geographic information systems (GIS) has opened up new possibilities for studying our planet, allowing us to explore inaccessible areas and monitor large-scale environmental changes. Advances in technology have led to the miniaturization of satellites, with the introduction of nanosatellites and CubeSats making remote sensing more accessible to universities, research institutions, and private companies. Additionally, the use of UAVs (Unmanned Aerial Vehicles) or drones for remote sensing has made it possible to acquire very high-resolution data at relatively low costs [Cracknell, 2018].

The Environmental Mapping and Analysis Program (EnMAP), launched by Germany in 2022, represents one of the most advanced developments in hyperspectral remote sensing [Earth Observation Portal]. Unlike multispectral sensors, which typically capture data across a limited number of broad spectral bands, hyperspectral sensors like those on EnMAP capture data in hundreds of narrow, contiguous bands across the electromagnetic spectrum. While multispectral sensors are effective for distinguishing broad land cover types—such as water, vegetation, and soil—they lack the spectral precision to detect subtle differences between materials with similar characteristics [Dovgyi et al., 2019]. EnMAP, with its ability to capture data across more than 200 spectral bands, allows researchers to detect these subtle variations in vegetation, soil, water bodies, and urban areas that are not visible with multispectral sensors.

Today, remote sensing is a key tool for environmental monitoring, resource management, and spatial planning. With the continued evolution of data acquisition technologies and the expansion of applications, remote sensing will continue to play a vital role in understanding and managing global environmental challenges.

2.1.2 Remote Sensing for Classification

Remote sensing is invaluable for classifying ecosystems and landscapes due to its ability to collect large-scale, repetitive data, enabling continuous monitoring of environmental changes. Multispectral data are ideal for distinguishing broad categories such as vegetation, water, and bare soil. On the other hand, hyperspectral data, which gather information in hundreds of narrow bands, allow for the differentiation of materials with very similar spectral signatures, such as different plant species or health conditions. In complex ecological studies, the higher spectral resolution offered by hyperspectral sensors is essential for the accurate classification of habitats. For example, hyperspectral data can detect plant stress, leaf chemistry, or soil conditions that would be difficult to identify using traditional multispectral sensors [Mehmood et al., 2022].

Spatial resolution is also a critical feature in land cover classification using remote sensing images. A particularly effective technique is the fusion of data from optical

sensors with different spatial resolutions. This method significantly improves the classification of soil and forests by combining the spectral richness of multispectral bands with higher spatial resolution. For example, the fusion of ZY-3 multispectral data with 10-band Sentinel-2 images resulted in a 14.2% increase in classification accuracy [Yu et al., 2020]. This approach is particularly useful in environments with high ecological heterogeneity, where greater resolution is essential for more detailed and accurate classification.

In this context, the integration of spectral, spatial, and topographic features has also demonstrated to significantly improve land cover classification in a subtropical ecosystem in China [Yu et al., 2020]. The study showed that the fusion of spectral data from ZiYuan-3 (ZY-3) and Sentinel-2, combined with topographic factors such as elevation and slope, achieved an overall classification accuracy of 83.5% across 11 different land cover classes.

The multispectral Sentinel-2 satellite mentioned in the study [Yu et al., 2020] is particularly powerful for land cover classification, especially in forest areas. Equipped with a wide range of spectral bands, including visible, near-infrared (NIR), short-wave infrared (SWIR), and red-edge bands, Sentinel-2 is highly sensitive to vegetation changes. These bands enable detailed and large-scale analysis of plant health, species identification, and monitoring of land cover changes. In a study conducted in the temperate forests of northern Iran [Nasiri et al., 2022], Sentinel-2 data were combined with aerial photogrammetry and machine learning algorithms to model forest canopy cover (FCC). Vegetation indices derived from Sentinel-2, such as the NDVI (Normalized Difference Vegetation Index) and the NDRE (Normalized Difference Red Edge Index), proved to be among the most significant predictors for estimating canopy coverage. Furthermore, combining Sentinel-2 data with high-resolution aerial imagery allowed FCC modeling to be extended over a larger spatial scale, demonstrating Sentinel-2's capability to provide accurate and up-to-date information on forest structure.

Another example of multispectral imaging is represented by the Landsat program. Operational for over 40 years, it's one of the longest-running and most consistent sources of data for monitoring land cover and landscapes. With a spatial resolution of 30 meters and a revisiting frequency of 16 days, Landsat data are widely used for forest classification, agricultural mapping, and coastal dynamics monitoring. The free availability of these data since 2008 has further expanded remote sensing applications, facilitating research in areas such as natural resource management and climate change assessment [Banskota et al., 2014].

On the other hand, the EnMAP satellite is one of the most advanced tools for acquiring high-resolution hyperspectral data. With its ability to collect data in hundreds of narrow spectral bands, EnMAP enables highly precise classification of materials with

similar spectral signatures, such as different plant species or health conditions. This makes EnMAP particularly useful in environments with high spectral diversity, such as tropical forests, grasslands, and other areas with complex biodiversity. This satellite has been successfully used to monitor ecological gradients and vegetation transitions, detecting changes in species composition and plant physiological states across spatial and temporal scales. For instance, a study in southern Portugal [Leitão et al., 2015] highlighted EnMAP's effectiveness in describing the gradual transition of shrub vegetation along an invasion gradient, confirming the essential role of hyperspectral data in capturing complex landscape details.

Applications of remote sensing for soil classification range from ecology to water management, urban planning to natural disaster mitigation. However, accurate classification may be hindered by factors such as atmospheric variability, landscape heterogeneity, and difficulties in distinguishing spectrally similar classes. Future challenges include developing more advanced data fusion techniques, integrating data from multiple sources (e.g., UAVs, aircraft, and satellites), and automating analytical processes for managing large data volumes.

2.1.3 Remote Sensing for Biodiversity Assessment

Remote sensing has become an essential tool for the assessment and monitoring of biodiversity on a global scale. Earth observation technologies provide fundamental data for the study and conservation of biodiversity, by providing the ability to observe large areas of the Earth's surface continuously and non-intrusively. This technology allows information to be gathered on various aspects of ecosystems, such as ecological structure, composition and functions, making it an ideal tool for tackling global challenges such as biodiversity loss.

Remote sensing for biodiversity monitoring can be applied at different levels, from genetic diversity to ecosystem diversity. At each level, remote sensing provides unique information useful for biodiversity assessment and management. The main parameters monitored include plant cover, ecosystem structure, vegetation health, ecosystem services and biogeochemical heterogeneity, which are crucial to understanding how different species respond to environmental changes [Reddy, 2021].

The use of multispectral and hyperspectral sensors has opened up new possibilities for biodiversity monitoring, offering substantial advantages in terms of spectral resolution and applicability. Hyperspectral sensors, like those on EnMAP allow for more precise identification of plant species and mapping of complex habitats with high biodiversity. These sensors provide detailed information on the structure and chemistry of vegetation, which is particularly useful in diverse environments like grasslands and peat bogs. In

contrast, multispectral sensors are suitable for large-scale applications or contexts where detailed species distinction is unnecessary. However, their limited spectral resolution may reduce accuracy in distinguishing species with similar spectral characteristics [Jarocińska et al., 2023, Reddy, 2021].

An example of multispectral sensors are the ones used in Sentinel-2. This satellite offers high spatial resolution and dense time series data due to a revisit time ranging from 2 to 5 days. Sentinel-2 images are particularly useful for monitoring multi-taxonomic biodiversity in forest environments. For example, in a study conducted in two national parks of the Apennines, Sentinel-2 images were used to derive metrics that showed a significant correlation with biodiversity indices, highlighting the potential of this satellite in identifying biodiversity hotspots[Parisi et al., 2023].

EnMAP, instead, uses imaging spectroscopy to provide high-resolution hyperspectral data, which is fundamental for biodiversity assessment. This program enables the characterization of ecosystem properties such as primary productivity, leaf water content, and vegetation chemistry, parameters fundamental for understanding ecosystem dynamics and ecological transitions[Leitão et al., 2015].

A key concept in the context of remote sensing for biodiversity is the Spectral Variation Hypothesis (SVH), which suggests that spectral variability observed in remote sensing images can serve as a proxy for environmental heterogeneity, which in turn correlates with biodiversity levels. The SVH is based on the ecological principle that greater habitat heterogeneity offers more ecological niches, supporting higher species diversity. This heterogeneity can be measured through various levels of spatial, spectral, and temporal resolution of remote sensing data[Torresani et al., 2024b, Rocchini et al., 2010].

Remote sensing technologies, including satellite platforms like Landsat and Sentinel, as well as airborne systems equipped with hyperspectral sensors, are fundamental to the SVH, providing large-scale data that enhance biodiversity assessments. [Torresani et al., 2024b, Rocchini and Neteler, 2012].

In particular, UAVs and drones are emerging as valuable tools for biodiversity monitoring due to their flexibility, high spatial resolution, and ability to operate in inaccessible areas. Although UAVs currently account for only 8% of the studies related to the SVH, their use is increasing rapidly as they become more accessible and equipped with specialized sensors for very high spatial resolution image analysis. UAVs are particularly useful for capturing detailed information in small areas, which can complement satellite data by providing finer-scale insights into habitat structure and species distribution[Rossi et al., 2022, Malavasi et al., 2021, Jackson et al., 2022].

The high spatial resolution provided by UAVs is critical in studies where precise mapping of species and habitats is required. In addition, UAVs can be deployed quickly and at relatively low cost, making them ideal for monitoring dynamic or hard-to-reach environments, such as wetlands, coastal areas, or dense forests. They are especially effective in regions where cloud cover often limits the utility of satellite imagery. The increasing use of UAVs equipped with multispectral and hyperspectral sensors provides new opportunities for monitoring plant diversity, ecological transitions, and habitat heterogeneity, where their high temporal, spectral and spatial resolution can be a valuable tool for biodiversity assessment. [Torresani et al., 2024b, Rossi et al., 2022].

The choice of spatial resolution is, in fact, crucial for biodiversity monitoring, affecting the accuracy and effectiveness of the analysis. The ideal spatial resolution is not an absolute value, but varies according to several factors, including the specific objectives of the study and the characteristics of the ecosystem under consideration. For example, studies focusing on the distinction of individual plant species require much higher resolution than those analysing forest cover on a large scale. Similarly, complex ecosystems with high spatial heterogeneity require higher resolution than more homogeneous ones. Pixel size, in particular, must be adapted to the crown size of plants to allow a reliable estimation of optical diversity, a parameter closely related to biodiversity. [Gamon et al., 2020]

Spectral resolution, which refers to the number and width of spectral bands acquired by a sensor, is equally important. Hyperspectral sensors, capable of capturing hundreds of narrow spectral bands, allow detailed characterisation of the properties of the earth's surface. This wealth of spectral information is valuable in distinguishing different plant species, even with similar morphological characteristics. For example, tree species can be identified based on their specific reflectance properties in the visible red and near infrared region [Torresani et al., 2024b]. In contrast, multispectral sensors with fewer larger bands may not be sufficient to discriminate plant species with similar spectral signatures, especially in complex ecosystems. It is important to note that the choice between hyperspectral and multispectral sensors depends not only on spectral resolution, but also on other factors such as study objectives, scale of analysis and ecosystem characteristics. For example, for large-scale monitoring of forest cover, multispectral sensors such as Landsat and Sentinel-2 can provide valuable information at low cost [Torresani et al., 2024b]. However, for studies of plant diversity at the species level, especially in heterogeneous ecosystems, hyperspectral sensors offer a significant advantage.

Temporal resolution, or the frequency of data collection, is vital for capturing dynamic changes in ecosystems and understanding the temporal patterns of species and communities. For instance, using complete time series, such as those derived from Landsat,

leverages temporal differences between species and communities, enhancing the ability to monitor biodiversity changes in highly dynamic environments [Rossi et al., 2021, Banskota et al., 2014].

Despite its many advantages, the use of remote sensing in biodiversity assessment also has limitations. The spatial and spectral resolution of sensors may not always be sufficient to detect species-level biodiversity, especially in complex environments like tropical forests. Additionally, satellite data must be complemented by field observations to obtain accurate estimates, particularly when monitoring specific species or rare habitats. The combined use of ecological models and remote sensing data, such as species distribution models, can help overcome these limitations, but requires a deep understanding of ecological dynamics and remote sensing technologies[Reddy, 2021].

To conclude, remote sensing is a powerful and versatile technology for monitoring biodiversity, offering a unique perspective on how ecosystems and species respond to global environmental changes. The integration of multispectral and hyperspectral data with field observations continues to enhance our capacity to monitor and manage biodiversity, contributing to global efforts for the conservation of natural resources and achieving sustainable development goals. However, challenges related to spatial resolution, ecological complexity, and data integration must be addressed to maximize the effectiveness of these technologies.

2.2 Random Forest Algorithms

Random Forests are a versatile and powerful machine learning technique used for classification and regression. They operate by constructing a multitude of decision trees during training and outputting the mode of the classes (for classification) or the mean prediction (for regression) of the individual trees. This ensemble method is known for its robustness, ability to handle large datasets with many variables, and resistance to overfitting. By leveraging randomness in both data selection and feature selection, Random Forests improve model generalization and provide valuable insights into variable importance, making them a popular choice in various fields of research and industry.

2.2.1 History and Development

The concept of ensemble learning, which underlies the Random Forest method, has deep roots in the machine learning community. One of the first similar techniques was "bagging" (Bootstrap Aggregating), developed by Breiman in 1996. Bagging involves generating multiple versions of a predictor by training on different bootstrap samples

and using the average of these versions to improve robustness and accuracy of the model [Breiman, 1996] [Biau and Scornet, 2016].

Later, in 1998, Dietterich proposed the "random split selection", a method where at each node of the tree a random split is chosen from the best K-split. So instead of always choosing the optimal subdivision according to a specific criterion, such as the Gini index or the entropy impurity, the "random split selection" introduces an element of randomness by considering a random subset of the best K subdivisions [Breiman, 2001]. Another significant technique is the "random subspace method" by Ho, which randomly selects a subset of features to construct each tree [Ho, 1998].

The central idea of Random Forest is to combine the advantages of these approaches, adding further randomness in the construction of trees [Biau and Scornet, 2016]. In particular, each tree in the Random Forest is built using a random sample of the training dataset (with repetition) and, for each split in the tree nodes, a random subset of the available characteristics is considered. This not only makes each tree different from the others, but also less sensitive to the predominant characteristics, thus contributing to a better overall performance of the model [Breiman, 2001].

Breiman demonstrated that as the number of trees in a Random Forest grows, it is highly likely to reach a limit in generalization error, effectively reducing the risk of overfitting, an issue where a model performs very well on the training data but poorly on unseen data [Breiman, 2001].
One of the most significant contributions of Random Forest is its ability to provide estimates of the importance of variables, offering valuable insights into data and facilitating model interpretation. Breiman developed methods to assess the importance of variables based on reduction in purity in tree nodes and decrease in classification accuracy when the values of the variables are permuted [Breiman, 2001].

Additionally, the Random Forest is robust to noise in data. Breiman observed that the use of a random selection of features for each split makes the model less sensitive to disturbances in input data, a significant advantage over other algorithms such as Adaboost, which may be more susceptible to noise [Breiman, 1996] [Breiman, 2001].
Boosting is another ensemble technique that sequentially applies weak classifiers to reweighted versions of the data, focusing on previously misclassified instances. This approach, developed by Yoav Freund and Robert Schapire, aims to convert weak learners into strong ones through iterative refinement [Freund and Schapire, 1997].

Since their creation, Random Forests have been perfected and extended. Variants such as extremely random trees (ExtraTrees) further randomize the tree building process by selecting splits randomly, which can sometimes lead to improved performance

and efficiency [Geurts et al., 2006].

In recent years, the Random Forest has continued to evolve. The introduction of advanced data pre-processing techniques, such as missing data management and class balancing, has further improved the performance of the algorithm. In addition, integration with other machine learning techniques has further expanded the capabilities of Random Forest. For example, Random Forest-based kernels can be used in algorithms such as the Kernel Principal Component Analysis and Support Vector Machines [Biau and Scornet, 2016].

2.2.2 Theory and Mechanism of Random Forest

The fundamental idea behind Random Forest is straightforward: construct an ensemble of independent decision trees and employ the average or most common outcome from these trees to make predictions for new data [Breiman, 2001]. Each decision tree starts from a root node that poses a question about the data, typically involving one or more features (such as radiance values in specific spectral bands). The tree branches are based on the answers to these questions, leading to further decision nodes. These nodes split the data until reaching terminal leaf nodes, where a final decision or classification is made. The final output for a given input is determined by aggregating the outputs of all individual trees, often using a majority voting scheme for classification tasks.

The main mechanisms of Random Forest include:

Bootstrap Sampling
For each tree in the Random Forest, a random sample with replacement (bootstrap sample) is selected from the training data set. This means that some samples may be selected multiple times, while others may be excluded. This process creates different variations of the original dataset, allowing each tree to train on a different subset of available data [Breiman, 2001]. On average, each bootstrap sample contains about 63% of the original data, with some data repeated several times and some excluded, the latter called Out-Of-Bag data (OOB). The Random Forest uses these OOB data to estimate the classification or regression error, providing an internal estimation of the model's performance without the need for a separate validation set [Cutler et al., 2007].
Bootstrap sampling introduces diversity between trees, helping to reduce the variance of the overall model. Diversity among trees is crucial to the success of the method, as it ensures that trees are not too closely correlated [Breiman, 2001] [Louppe, 2014].

Random Feature Selection
For each node in each tree, instead of considering all possible variables, the Random

Forest selects a random subset of features. This subset is generally of size $\sqrt{p}$ for classification problems and $p/3$ for regression problems, where p is the total number of variables [Breiman, 2001]. This mechanism increases diversity among trees, since each tree can make different decisions even if trained on the same data set. The random selection of features then prevents some trees from becoming dominant or making the same choices, thus improving the ability of the Random Forest to generalize on unseen data [Ho, 1995].

Tree Growth
Each tree is built to the maximum possible depth without pruning. This means that each tree is trained to fit the training data as much as possible, creating a series of complex decision trees.
Although each individual tree can be highly complex and susceptible to overfitting, the mean of many unrelated trees tends to balance out individual errors, improving the robustness and generalizability of the overall model [Breiman, 2001].

Prediction
After all trees have been built, the Random Forest uses an aggregation process to make its final predictions.
For classification problems, each tree votes for a class. The class with the highest number of votes among all trees is chosen as the final prediction. This process is known as majority voting and helps to mitigate the effect of any inaccurate trees, since the wrong predictions of some trees are offset by the correct predictions of the majority [Breiman, 2001].
For regression problems, each tree provides a numerical prediction. The final prediction of the Random Forest is the average of all the tree predictions. This approach helps to smooth predictions and reduce the impact of any outliers in the training data [Breiman, 2001, Liaw et al., 2002].

Variable Importance
One of the significant benefits of Random Forests is their ability to provide insights into the importance of different features in making predictions. This is achieved by measuring the decrease in accuracy when a particular feature's values are permuted. Features that, when altered, lead to a significant drop in model accuracy are deemed important. This information is useful not only for interpreting the model, but also for selecting characteristics, allowing to reduce the dimensionality of the dataset without losing predictive precision [Breiman, 2001].

2.2.3 Applications of Random Forest in Various Fields

Random Forests have gained extensive acceptance across multiple disciplines due to their adaptable nature and proficiency in managing intricate datasets.

Ecology, in particular, has seen increasing application of this technique, given the often complex and non-linear nature of ecological data.

A study by Castaldi et al. [Castaldi et al., 2019] analyses the effectiveness of multispectral (Sentinel-2) and hyperspectral (EnMAP) satellite data for mapping soil organic carbon (SOC) content using Random Forest models. The authors evaluated different sampling strategies for mapping soil organic carbon content. Among the various regression algorithms tested, Random Forest proved to deliver the best results. To further optimize the Random Forest model, the Puchwein sample selection algorithm was implemented. This method is based on the Mahalanobis distance (a measure that takes into account the correlation between variables to assess the distance between points in a multi-dimensional space) between spectra to iteratively select the most dissimilar samples, thus ensuring a calibration dataset representative of SOC variability in the study area. The results show that Sentinel-2 data, due to their 10-meter spatial resolution, were more suitable for SOC mapping compared to EnMAP data (30 meters). Using a Random Forest model calibrated with Sentinel-2 data, the authors achieved high accuracy in SOC prediction, with an nRMSE of 8.7% (normalized Root Mean Square Error, a measure of model prediction error relative to the range of observed values) and an RPD of 2.5 (Ratio of Performance to Deviation, an indicator of model reliability in predicting the target variable).

One of the most prominent uses of Random Forests is in modelling the distribution of species. This approach has been used to predict the presence of invasive plant species, rare lichen species and cavity-nesting birds. For example, Random Forest was able to predict the presence of the invasive plant Verbascum thapsus in the Lava Beds National Monument with a specificity (percentage of correctly classified absences) of 84.5% using 10-fold cross-validation, significantly outperforming other methods such as logistic regression (51.4% specificity) and linear discriminant analysis (48.6% specificity). The ability of Random Forests to manage complex interactions between predictive variables and provide accurate classifications makes this technique extremely useful in predicting species distribution in various ecological studies [Cutler et al., 2007] .

In addition to ecology and remote sensing, Random Forest has found application in many fields, including genomics. Here, its robustness in handling large-scale datasets with numerous variables makes them an excellent tool. For instance, the references by Díaz-Uriarte and De Andres [Díaz-Uriarte and Alvarez de Andrés, 2006] provide detailed insights into the methodological considerations and findings in the application of Random Forests for microarray data analysis and gene selection. Their research highlights the ability of Random Forests to achieve accurate classification even when the number of variables (e.g., genes) far exceeds the number of samples. Specifically, the study showed that Random Forests, applied to nine different microarray datasets, achieved error rates

comparable or lower than other classification methods such as DLDA, KNN and SVM. For example, for the "Leukemia" dataset, Random Forests obtained an error rate of 0.087 with variable selection and 0.075 without selection, compared to higher errors for the other methods. This is possible thanks to the Random Forest strategy of creating multiple independent decision trees, each based on a random sample of data and a subset of features. The aggregation of the results from these trees significantly reduces the risk of overfitting, thus improving the robustness and accuracy of the model. In addition, Random Forests provides estimates of the importance of variables, allowing researchers to identify which genes have the most significant impact on classification. This feature is particularly useful in contexts where it is crucial to understand which genes are involved in a disease, facilitating the discovery of potential biomarkers for diagnosis or treatment.

2.2.4 Benefits and Limitations

One of the main advantages of Random Forest is its ability to handle data with a large number of variables, such as those derived from satellite images. The algorithms for random selection of variables during the tree construction phase allow to deal effectively with problems of high dimensionality, reducing the risk of overfitting and improving the generalization capacity of the model [Ho, 1995, Louppe, 2014].

In addition to this, this Random Forests are able to reduce the overall model error by using substitution sampling (bootstrap) to generate multiple trees and aggregating results through a majority vote. This approach allows for better management of the inherent variability in ecological and remote sensing data, reducing the possibility that a single tree will dominate the final forecast and improving the model's ability to generalize to new data [Breiman, 2001].
Random Forests are therefore robust to noisy data and outliers. This is due to the fact that each tree is built on a different sample of data, and only a subset of the characteristics are considered for each node, reducing the impact of outliers on the final result [Breiman, 2001]. The algorithm is extremely flexible, allowing it to be used not only for classification but also for regression and other applications such as missing value imputation and survival analysis [Cutler et al., 2007].

The ability of the algorithm to estimate the importance of variables is another crucial aspect in the context of satellite image classification. This tool allows to identify which spectral bands or vegetation indices influence the classification of land cover the most, improving the interpretability of the model and enabling ecologists and remote sensing analysts to better understand the environmental dynamics underlying their observations [Cutler et al., 2007, Louppe, 2014].

Despite their advantages, Random Forests can be computationally intensive, especially with very large datasets. Building hundreds or thousands of trees takes time and significant computational resources. This may limit their applicability in contexts where the speed of computation is crucial [Breiman, 2001, Cutler et al., 2007].

Also, although the Random Forest provides an estimate of the importance of variables, model's interpretability can be more difficult than in other simpler models. Each decision tree in the Random Forest is relatively easy to interpret, but the combination of hundreds of trees makes it difficult to get a clear view of relationships in data [Cutler et al., 2007]. This is a limitation, especially in areas where model interpretability is as important as accuracy.
Moreover, even though Random Forest is generally resistant to overfitting, using too many trees may cause over-adapting to extremely noisy data, reducing the model's ability to generalize to new data [Cutler et al., 2007].

In conclusion, Random Forest is a powerful and flexible tool for many applications, but its effectiveness depends on careful consideration of its limits and optimization of parameters for the specific context of use.

2.3 Spectral Variation Hypothesis

The Spectral Variation Hypothesis (SVH) posits that the spectral heterogeneity observed in remote sensing imagery is correlated with the biodiversity of a given area. This relationship is grounded in the concept that different species or functional groups within an ecosystem exhibit unique spectral signatures due to variations in their biochemical, structural, and phenological properties. Thus, areas with higher species diversity are expected to exhibit greater spectral variability. The diversity of these spectral signals, therefore, can serve as a proxy for biodiversity, offering a non-invasive means to assess and monitor ecological variation over large spatial scales.

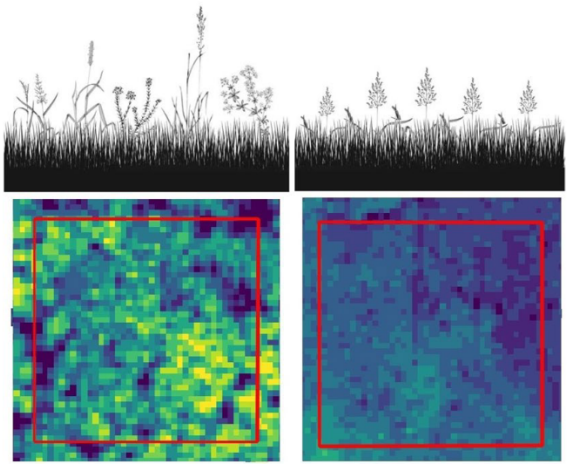


Figure 2.1: Visual representation of species diversity and spectral heterogeneity in a grassland ecosystem [Torresani et al., 2024a]

2.3.1 History and Development

Biodiversity, as defined by the Convention on Biological Diversity, encompasses the variety and variability of life forms at ecosystem, species, and genetic levels. This broad concept includes three main components: alpha diversity (species diversity within a specific area or ecosystem), beta diversity (differences in species composition between ecosystems), and gamma diversity (overall diversity across a larger region). Understanding and monitoring these components is crucial in the context of climate change, as biodiversity underpins ecosystem resilience and functions, providing essential services like carbon sequestration, climate regulation, and maintaining water cycles [Opoku, 2019, Torresani et al., 2019].

Historically, biodiversity studies relied heavily on field-based observations by botanists and ecologists, who cataloged species within specific plots. This method, while detailed, was constrained by the significant time, cost, and labor required, particularly in large or remote areas. The advent of remote sensing technologies revolutionized this field, enabling comprehensive, repeatable, and non-destructive sampling over extensive areas and varied temporal scales. Satellite imagery and aerial sensors have democratized access to environmental data, providing new tools for detailed monitoring that were previously unavailable [Palmer et al., 2002, Torresani et al., 2024b].

The conceptual foundations of the SVH were laid by Palmer in the late 1990s and early 2000s. Palmer et al. [Palmer et al., 2002] formally articulated the hypothesis that variability in the spectral signal captured by optical images could serve as an indicator of biodiversity. They initially tested this hypothesis in the Tallgrass Prairie Preserve (Oklahoma, USA), demonstrating that spectral heterogeneity metrics correlated with

various biodiversity indices, including species richness, rarity, and the number of infrequent species.

Earlier, Gould [Gould, 2000] had explored similar ideas, showing that the variability of the NDVI derived from Landsat images could indicate landscape heterogeneity, correlating well with plant species richness in the Arctic ecosystem of the Hood River Region, Canada. Gould's work highlighted the potential of integrating vegetation type information with NDVI variability to enhance biodiversity estimation accuracy. .

The SVH has since been explored in a wide range of ecosystems, from forests and grasslands to wetlands, coastal regions, savannahs, and even urban areas. This growing body of research underscores the increasing recognition of spectral heterogeneity as a valuable tool in biodiversity monitoring [Torresani et al., 2024b].

2.3.2 Application of Spectral Variation Hypothesis in Various Fields

The SVH is mainly used in ecology to estimate biodiversity through remote sensing data. It has been widely applied to assess vegetation diversity in various ecosystems, including forests, grasslands, wetlands, and agroforestry systems [Torresani et al., 2024b]. The hypothesis is often tested using different remote sensing data, such as multispectral and hyperspectral images, to measure spectral indices and their correlation with species diversity metrics such as species richness, the Shannon index, or the Simpson index.

The hypothesis is commonly used in forest ecosystems to monitor plant species diversity, and has been applied in various types of forests, including tropical, Mediterranean, temperate, and alpine forests. In this context, the SVH helps to assess plant species diversity by analysing the spectral variability captured in optical images. In addition, the SVH has been tested in grasslands and wetlands to estimate species diversity, using spectral data from various sensors to identify and monitor plant communities and their diversity, taking into account seasonal and spatial variations [Torresani et al., 2024b]. The study by Torresani et al. (2019) [Torresani et al., 2019] tested the SVH in an alpine coniferous forest to estimate tree species diversity. Using Sentinel-2 and Landsat-8 satellite images, the study analysed the relationship between spectral heterogeneity, calculated with Rao's Q diversity index, and tree diversity measured in the field by Shannon index. The results showed a strong correlation between Rao's Q index, applied to Sentinel-2 data, and tree species diversity, particularly during the peak period of the NDVI (between June and July). During this period, the coefficient of determination (R^2) reached values of 0,48 for 2016 and 0,70 for 2017. The temporal analysis of the data showed that the relationship between spectral heterogeneity and species diversity

is seasonal, with lower R^2 values in winter and spring. This result suggests that the ability of NDVI to capture small variations in leaf reflectance, typical of different tree species, is greater when the vegetation index reaches its maximum values. The study also highlighted the importance of spatial resolution of data, with Sentinel-2 (10m) providing better results than Landsat-8 (30m) in estimating tree species diversity.

The study by Chraibi et al. [Chraibi et al., 2021] applied the SVH to assess tree diversity during secondary succession phases in abandoned cocoa forests in Trinidad and Tobago. The aim was to see whether remote sensing, based on SVH, could be used to monitor forest biodiversity on a large scale. The researchers compared field data on tree diversity with remote sensing data from Sentinel-2 satellite images. Although they did not find a direct correlation between the diversity indices based on field data and the pixel heterogeneity of the images, they observed that the beta-diversity derived from remote sensing was able to identify a regeneration gradient in forests. In other words, remote sensing allowed to distinguish active cocoa agroforestry from secondary forests at different stages of succession, suggesting that SVH, applied to remote sensing, can provide information useful for monitoring biodiversity on a large scale, especially in areas that are difficult to access. However, the results also indicate that remote sensing may not be sensitive enough to capture tree diversity locally with the same precision as field data.

Although originally developed for plant diversity studies, the SVH has found applications in other fields, exploiting its ability to use spectral data to infer diversity in different contexts. Recent studies have extended the use of the SVH to assess animal diversity, such as that of mammals, birds, and benthic invertebrates. The study by Oindo et al. [Oindo and Skidmore, 2002] explored the correlation between spectral variability, measured by the NDVI, and the richness of mammalian species in different regions of Kenya. The analysis revealed that the spatial heterogeneity of NDVI, indicative of environmental variability, is positively correlated with the richness of mammalian species on a small spatial scale (10 x 10 km). This means that areas with higher spectral variability tend to host a higher diversity of mammalian species, confirming the effectiveness of SVH as a tool for assessing mammalian biodiversity in diverse ecological contexts.

In marine ecosystems, the SVH has been tested to study the diversity of benthic invertebrates. By examining the spectral reflectance data of marine environments, researchers can estimate the diversity and distribution of different species in underwater habitats. A study by Herkül et al. [Herkül et al., 2013] was the first attempt to apply the SVH in the marine environment, using hyperspectral images recorded by the "Compact Airborne Spectrographic Imager" (CASI) to correlate spectral variability with the biodiversity of benthic macro-organisms, including species richness and Shannon index. The results show that all diversity measures derived from coverage data had significant positive correlations with SV at all spatial scales. For the species richness and Shannon

index, the strongest correlation was found at the scale of 10 metres, with r=0.24 for the Shannon index, and r=0.32 for the species richness. This result suggests that, despite the influence of the water column on signal absorption, there is a positive relationship between spectral variability and benthic diversity, measured in terms of both Shannon Index and species richness.

The SVH continues to evolve, finding new applications in various fields thanks to the increasing availability and sophistication of remote sensing data. As technology advances, the field of application is expected to expand further, offering new insights into different disciplines.

2.3.3 Benefits and Limitations

One of the main advantages of the SVH is its ability to use spectral data to efficiently estimate biodiversity on a large scale. This approach overcomes the logistical and time constraints associated with traditional field surveys, which often require considerable resources and time [Torresani et al., 2024b]. This capability makes the SVH a valuable tool for ecosystem management and biodiversity conservation, enabling a quick and cost-effective identification of areas of high biodiversity.

In addition, the use of the SVH can contribute to dynamic monitoring of biodiversity, particularly in complex ecosystems where species can respond rapidly to environmental changes. Therefore, it is particularly useful in highly variable environments, where spectral heterogeneity can reflect significant variations in species composition, allowing the spatial and temporal dynamics of biodiversity to be monitored and adapting to different ecological scales, from alpha diversity to gamma diversity. This flexibility makes the SVH a versatile tool for environmental monitoring and long-term conservation planning [Torresani et al., 2024b].

Despite the many advantages, the SVH also has some limitations. One of the main limitations is its non-universal applicability, which varies greatly depending on the ecological context, the characteristics of the sensor used, and the spatial scale of the analysis. For example, the SVH has shown mixed results in grasslands where the relationship between spectral heterogeneity and biodiversity can be influenced by factors such as the spatial resolution of data and the complexity of the ecosystem [Torresani et al., 2024b].

Another issue concerns the spectral and spatial resolution of remote sensing data. Although a higher spatial resolution may improve the ability to detect changes in biodiversity, it can also introduce noise and complications, due to spectral redundancy and the difficulties in managing the variability introduced by pixel-level detail. Furthermore,

the effectiveness of the SVH is strongly dependent on the quality and timing of data collected. Seasonal changes and phenological variations may affect spectral measurements, making the use of time series data necessary to obtain more accurate estimates [Torresani et al., 2024b].

In conclusion, the SVH represents an innovative and powerful approach for biodiversity analysis, with significant advantages in terms of efficiency, scale, and applicability. However, the effectiveness of the SVH is closely linked to the quality of spectral data and the scale of analysis, requiring a rigorous methodological approach to overcome its limitations. Despite these challenges, the SVH remains an important tool in ecological research and ecosystem management, with significant potential for future applications.

2.3.4 Rao's Q index

Rao's Q index, also known as Rao's Quadratic Entropy, is a measure of diversity that considers not only the relative frequency or abundance of elements within an ensemble, but also the differences between these elements.
In the SVH, Rao's Q index allows to move from a purely qualitative analysis of spectral diversity to a quantitative measure that can be directly correlated with ecological data from the field. In fact, this index has proven to be effective in assessing and monitoring biodiversity at landscape level, when used in combination with spectral data [Rocchini et al., 2017].

Before the introduction of Rao's Q index in 1982, there was already a large literature on measures of diversity and on dissimilarity or similarity between populations. These measurements have been used in a wide range of studies in different fields, including anthropology [Rao, 1948, 1971a,b, 1977], genetics [Karlin et al., 1979, Morton and Lalouel, 1973, Nei, 1978], economics [Gini, 1912, Sen, 1997], sociology [Agresti and Agresti, 1978] and biology [Sokal and Sneath, 1963]. These measurements were often based on heuristic considerations, some derived from mathematically well-postulated axioms, while others were constructed using hypothetical models for the genetic and environmental mechanisms that cause differences between individuals and populations.

By introducing Rao's Q index in 1982 [Rao, 1982], C. R. Rao contributed to a significant advance in measuring diversity by providing a tool that incorporates dissimilarity between elements, providing a more complete view of diversity than traditional measures such as the Shannon index or the Simpson index, which do not take into account distance between species or, in the context of remote sensing, pixels.
The index is calculated by considering all pairs of elements in the set and summing the products of their dissimilarity and their relative frequencies. The general formula for

Rao's Q is:

$$Q = \sum_{i=1}^{S} \sum_{j=1}^{S} d_{ij} p_i p_j$$

where d_{ij} represents the dissimilarity between species i and j, and p_i and p_j are the relative frequencies of species i e j.
This formulation allows to capture the complexity of diversity in an ecological system, reflecting not only the variety (how many different types of elements are present) but also how different these elements are [Rao, 1982].
In the context of remote sensing, Rao's Q index can be applied to quantify the spectral diversity of an area, using pixel spectral values as system elements. In this scenario, the dissimilarity d_{ij} can be defined as the spectral distance between the reflectance values of two pixels, while p_i and p_j can represent the relative frequencies of these spectral values [Rocchini et al., 2017].

To conclude, Rao's Q index is a powerful and flexible tool for diversity analysis, both in traditional ecological contexts and in the field of remote sensing. Its ability to integrate dissimilarity between elements makes it particularly suitable for Spectral Variation Analysis, providing a quantitative measure that can be correlated with actual biodiversity. The growing application of Rao's Q in ecological and remote sensing models marks a significant advance in our ability to understand and manage biological diversity on a global scale.

Chapter 3

Data and Methods

This chapter describes the data sources and methodological approaches used to classify landscapes and assess forest biodiversity in the South Tyrol region using remote sensing data. It details the characteristics and preprocessing of satellite imagery from Sentinel-2, Landsat-8, and EnMAP, as well as the collection and processing of field data. The chapter further explains the use of the Random Forest algorithm for land cover classification and the application of the SVH for biodiversity assessment, integrating both remote sensing and field data to evaluate the effectiveness of different sensors in ecological monitoring.

3.1 Satellite Imagery

Satellite imagery is a crucial component of remote sensing and plays a vital role in monitoring and understanding environmental and climatic changes. This section provides an overview of the main features of the satellites used in this study: Sentinel-2, Landsat-8, and EnMAP.

3.1.1 Sentinel-2

Sentinel-2 is part of the Copernicus Programme, a joint initiative by the European Space Agency (ESA) and the European Commission [European Space Agency]. It consisted of two satellites, Sentinel-2A and Sentinel-2B, launched in June 2015 and March 2017, respectively. In September 2024, Sentinel-2C was successfully launched, further enhancing the mission's capacity for Earth observation. The primary features of Sentinel-2 are:

- **Spatial Resolution:** Sentinel-2 provides imagery at three different spatial resolutions: 10 meters, 20 meters, and 60 meters, depending on the spectral band.
- **Spectral Bands:** It has 13 spectral bands, ranging from the visible and near-infrared to the shortwave infrared.

- **Revisit Time:** The revisit time is approximately 5 days at the equator with both satellites (Sentinel-2A and Sentinel-2B) in operation.
- **Applications:** It is widely used for land monitoring, including agriculture, forestry, land cover classification, and natural disaster management.

3.1.2 Landsat-8

Landsat-8, launched in February 2013, is part of the Landsat program managed by NASA and the U.S. Geological Survey (USGS) [United States Geological Survey (USGS)]. It continues the mission of providing high-quality, long-term data on Earth's surface. The main features of Landsat-8 include:

- **Spatial Resolution:** Landsat-8 offers a spatial resolution of 30 meters for multispectral bands and 15 meters for the panchromatic band.
- **Spectral Bands:** It includes 11 spectral bands, spanning the visible, near-infrared, shortwave infrared, and thermal infrared regions.
- **Revisit Time:** The revisit time is 16 days.
- **Applications:** It is used for various applications, including agriculture, forestry, geology, land cover change, and water resources management.

3.1.3 EnMAP

EnMAP (Environmental Mapping and Analysis Program) is a German hyperspectral satellite mission to provide detailed spectral information for environmental monitoring and analysis [Earth Observation Portal]. The satellite, launched in 2022, carries the Hyperspectral Imager (HSI). Key features of EnMAP are:

- **Spatial Resolution:** EnMAP provides a spatial resolution of 30 meters.
- **Spectral Bands:** The HSI onboard EnMAP covers 224 spectral bands, ranging from the visible (420 nm) to the shortwave infrared (2450 nm).
- **Revisit Time:** EnMAP has a standard revisit time of 27 days at nadir but can revisit a target much more quickly—within 4 days—by tilting its sensor 30° off-nadir.
- **Applications:** It is designed for a wide range of applications, including agriculture, forestry, soil and geology, coastal and inland waters, and urban areas. EnMAP's hyperspectral capability allows for detailed analysis of material composition and biochemical properties.

3.1.4 Comparison of Satellite Features

Table 3.1 provides a comparative summary of the main features of Sentinel-2, Landsat-8, and EnMAP.

Feature	Sentinel-2	Landsat-8	EnMAP
Spatial Resolution	10m, 20m, 60m	30m (MS), 15m (PAN)	30m
Spectral Bands	13	11	224
Revisit Time	5 days	16 days	27 days

Table 3.1: Comparison of Satellite Features

This study focuses on the South Tyrol region and utilizes satellite imagery taken on September 10, 2023 for Sentinel-2 A and Lansat-8, and on September 9, 2023, for EnMAP. The satellites provided comprehensive data for monitoring vegetation, land use, and environmental characteristics. These images were crucial for developing machine learning models to classify land cover and study spectral variation.

Figure 3.1: Satellite images used in the study: EnMAP, Sentinel-2 and Landsat-8

3.2 Field Data

Field data were used to calculate key biodiversity metrics, such as species richness and the Shannon index, in the South Tyrol region. The field data, which include information on tree species and their abundance at various locations, were used to compute these metrics and compare them with Rao's Q index derived from remote sensing imagery. This comparison was crucial to assess the effectiveness of different indices in capturing forest biodiversity using the SVH.

Location	Species
GS001	LD (1), PA (5), PC (27)
GS002	AA (7), LD (11), PA (34), PC (27)
GS003	PA (4), PC (30)
GS004	Alnus Sup (4), LD (1), PA (4), PC (11), SA (15)
GS007	PA (2), PC (30)
GS008	PA (1), PC (45), SA (17)
GS010	AA (14), LD (8), PA (9), PC (4)
GS011	AA (1), PA (19), PS (1)
GS012	Betulus spp. (1), LD (1), PA (27)
GS013	LD (4), PA (23), PC (2)
GS014	PA (13), PC (11), Ulmus spp. (1)
GS015	LD (4), PA (25)
GS019	PA (29)
GS020	PA (75)
GS022	PC (12)
GS024	AA (1), LD (2), PA (15), PC (1)
GS025	AA (19), LD (5), PA. (25)
GS026	AA (11), LD (1), PA (10), PS (1)
GS027	LD (1), PA (28)
GS028	PA (6), PC (23)
GS029	LD (5), PA (24)
Chiusa10GS001	AA (3), PA (26)
Chiusa10GS002	PA (4), PC (17)

Table 3.2: LD: Larix decidua, PA: Picea abies, PC: Pinus cembra, AA: Acer campestre, Alnus sup.: Alnus species, SA: Sorbus aucuparia, PS: Pinus sylvestris, Betulus spp.: Betula species, Ulmus spp.: Ulmus species

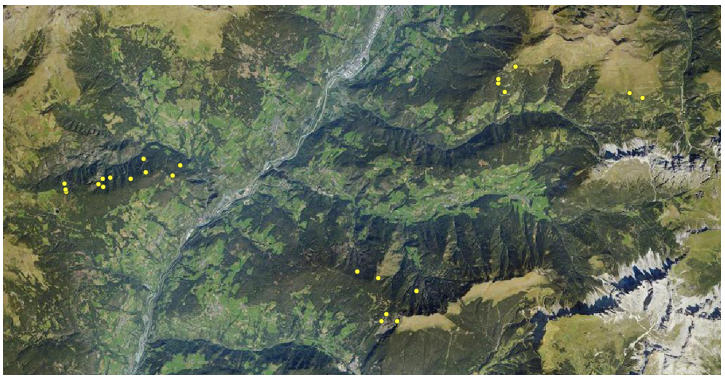


Figure 3.2: Locations where the field data were collected

3.3 Preprocessing

3.3.1 Satellite Imagery Processing for Classification

Preprocessing of satellite imagery is a critical step in ensuring accurate classification results. In this study, specific bands were selected and processed for each satellite to optimize the classification tasks. The preprocessing steps varied slightly between Sentinel-2, Landsat-8, and EnMAP due to their different spectral and spatial resolutions.

3.3.2 Sentinel-2 Preprocessing

For the Sentinel-2 satellite, three different classifications were performed, each utilizing a different combination of spectral bands and spatial resolutions:

10m Resolution Classification

The first classification was carried out using bands available at a 10m spatial resolution:

- B2 (Blue - 492.7 nm)
- B3 (Green - 559.8 nm)
- B4 (Red - 664.6 nm)
- B8 (NIR - 832.8 nm)

These bands were used directly for the classification process without any resampling.

20m Resolution Classification

For the second classification at a 20m resolution, the following bands were used:

- B1 (Coastal - 442.7 nm)
- B5 (Red Edge 1 - 704.1 nm)
- B6 (Red Edge 2 - 740.5 nm)
- B7 (Red Edge 3 - 782.8 nm)
- B8A (NIR - 864.7 nm)
- B11 (SWIR 1 - 1613.7 nm)
- B12 (SWIR 2 - 2202.4 nm)

Additionally, bands B2, B3, B4 and B8 were used. While bands B2, B3, and B4 were provided in Sentinel's data at both 10m and 20m resolutions, band B8 was only available at 10m resolution. To ensure consistency in spatial resolution across all bands, B8 was resampled to 20m resolution.

Listing 3.1: Resampling bands to 10m resolution

```
library(raster)

rst_lst[["B08"]] <- raster::resample(x = rst_lst[["B08"]],
                                                    y = rst_lst$B05)
```

Combined 10m and 20m Resolution Classification

The third classification involved resampling all bands to a 10m resolution. This process included:

- Using the original 10m resolution bands (B2, B3, B4, B8).
- Resampling the 20m resolution bands (B1, B5, B6, B7, B8A, B11, B12) to 10m resolution.

In order to make the resampling more computationally efficient, the raster image was cropped on each area of interest.

Listing 3.2: Resampling bands to 10m

```
library(raster)

#Initializing the vector
rst_for_prediction <- vector(mode = "list", length = length(rst_lst))
names(rst_for_prediction) <- names(rst_lst)

#Cropping the vector
rst_for_prediction[['B08']] <- crop(rst_lst[['B08']],area_A)

for (x in bands_names) {
     if (x == 'B08'){
          print(paste0(x, 'is already at 10m of resolution '))
}
     else{
          print(paste0('resampling ',x))
          rst_for_prediction[[x]] <- crop(rst_lst[[x]],area_A)
          rst_for_prediction[[x]] <- raster::resample(x = rst_for_
             ↪ prediction[[x]],  y = rst_for_prediction$B08)
} }
```

These combined bands were then used to perform the classification at a 10m resolution.

3.3.3 Landsat-8 Preprocessing

For Landsat-8, all available bands were used, each having a spatial resolution of 30m:

- B1 (Coastal - 440 nm)
- B2 (Blue - 480 nm)
- B3 (Green - 560 nm)
- B4 (Red - 650 nm)
- B5 (NIR - 870 nm)
- B6 (SWIR 1 - 1600 nm)
- B7 (SWIR 2 - 2200 nm)
- B10 (Thermal - 10900 nm)

These bands were directly used for the classification process without any resampling.

3.3.4 EnMAP Preprocessing

EnMAP provides hyperspectral imagery with a large number of spectral bands. However, some bands needed to be excluded from the analysis due to the presence of null values:

- Bands from 131 to 135 were removed from the dataset.

After the removal of these bands, the remaining spectral bands were used for the classification tasks.

In summary, the preprocessing involved selecting appropriate bands for each satellite, resampling bands to ensure consistent spatial resolution where necessary, and excluding problematic bands. These steps were essential to prepare the data for accurate and effective land cover classification of the South Tyrol area.

3.3.5 Field Data Processing for Spectral Variation Hypothesis

For each location where field data were collected, a circular area with a radius of 15 meters was considered. This radius was chosen to ensure that the biodiversity metrics would be calculated over a consistent spatial scale, big enough to incorporate more than one pixel. Satellite images were cropped using the circles defined around each data collection point. This step ensured that the satellite data used in the analysis corresponded

to the exact locations where field data were collected.
Any circles that extended beyond forested areas into other land covers (such as pastures) were excluded from further analysis to maintain the integrity of the forest biodiversity assessment.

Two key biodiversity metrics were calculated for each data collection point: species richness and the Shannon index.
Species richness is a simple count of the number of different species present in a given area. It is calculated as:

$$S = \sum_{i=1}^{n} 1$$

where n is the total number of species observed in the area.
The Shannon index (also known as Shannon-Wiener index) is a measure of species diversity that takes into account both the number of species and the evenness of their abundances. It is calculated as:

$$H' = -\sum_{i=1}^{n} p_i \ln(p_i)$$

where n is the total number of species and p_i is the proportion of individuals of species i relative to the total number of individuals of all species.

These biodiversity metrics provided a quantitative basis for assessing the variation in species composition across the study area (Table 3.3).
The Shannon index and the species richness were then related to Rao's Q index, allowing for an analysis of the relationship between biodiversity and spectral variation.

In summary, the field data processing involved selecting appropriate circles around data collection points, cropping satellite images to match these circles, and calculating key biodiversity metrics. These steps ensured that the field data were accurately represented and ready for further analysis.

Location	Shannon Index	Species Richness
GS001	0.556	3
GS002	1.22	4
GS003	0.362	2
GS008	0.66	3
GS009	0.358	2
GS010	1.35	4
GS011	0.381	3
GS012	0.238	4
GS013	1.15	5
GS014	0.83	3
GS015	0.377	2
GS024	0.734	4
GS025	0.958	3
GS026	0.988	4
GS027	0.271	3
GS028	0.219	2
Chiusa10GS001	0.181	2
Chiusa10GS001	0.487	2

Table 3.3: Shannon Index and Species Richness calculated at each location

3.4 Methodology for Land Cover Classification

This section outlines the approach used to classify land cover types in the South Tyrol region using remote sensing data. It details the selection of areas of interest, the generation of random points for training data, and the methods used to split datasets and validate the classification model. The Random Forest algorithm was applied to the preprocessed satellite imagery, enabling accurate classification of different land cover types

3.4.1 Areas of Interest and Classes

Four areas of interest (A, B, C and D) were selected in South Tyrol to be representative of the local land cover. The land cover classes considered for the classification were five: forests, urban areas, mountains, pastures, and areas devastated by the Vaia storm.

Storm Adrian (also known as Vaia), which occurred in late October 2018, was an extreme weather event that caused widespread damage throughout northern Italy, particularly in the Alpine regions. The storm, which was marked by high winds of over 200 km/h and heavy rain, uprooted millions of trees and devastated vast portions of forests. In South Tyrol, the effects of the Vaia have been particularly pronounced, with

significant parts of the forest area permanently altered. These areas have become of great ecological interest, as they represent altered landscapes in the regeneration phase and pose challenges for land management and biodiversity conservation.

To accurately capture the range of land cover in the region, ten polygons were drawn for each class in every area of interest using the geographic information system QGIS. Each polygon was then used for the creation of training and testing dataset for its specific class. In the picture (Figure 3.3) the polygons drawn for area A are shown.

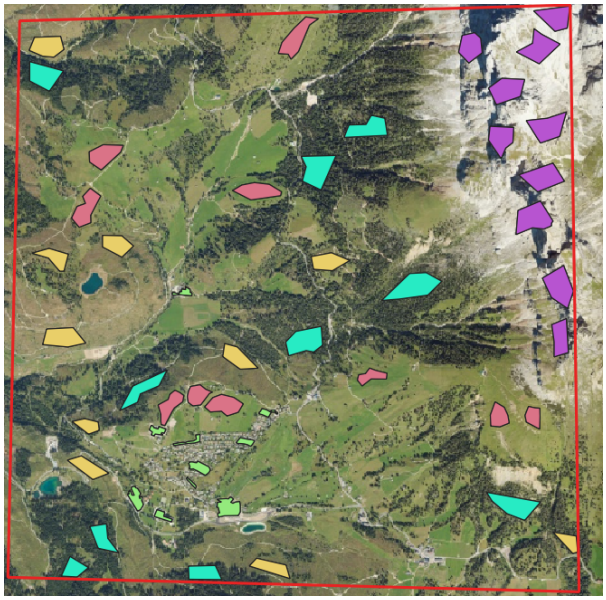


Figure 3.3: Area A with its polygons (cyan = forests, purple = mountains, green = urban, pink = pastures, yellow = Vaia).

3.4.2 Generating Random Points for Training Data

To train the Random Forest model, 750 random points were generated for each land cover class within the polygons. The **spsample** function from the **sp** package was used to generate these points, ensuring a regular distribution. Each point was associated with its respective class ID using the **over** function.

Listing 3.3: Generating Random Points for Training Data

```
# Libraries required
library(sp)          # Spatial data handling
library(rgdal)       # Reading and writing spatial data
library(raster)      # Raster data manipulation
library(data.table) # Fast data manipulation
library(dplyr)       # Data manipulation

```

```r


# Creating 750 random points for each class
# Here, id==1 corresponds to forests

# Selecting the plygons with id=1
ptsamp1 <- subset(poly_area_A, id == "1")

ptsamp1_1 <- spsample(ptsamp1, 750, type='regular')
# spsample: Generates 750 random points within the polygons

ptsamp1_1$class <- over(ptsamp1_1, ptsamp1)$id
# over: Associates each random point with the class ID

# Saving the results
saveRDS(ptsamp1_1, file=paste0 ("path_to_file", file="_ptsamp1_A.rds"))


# Taking the infromation of the pixel where the random point
# landed and saving them in a dataframe
dt1 <- brick_for_prediction %>%
  raster::extract(y = ptsamp1_1) %>%
  as.data.table %>%
  .[, id_cls := ptsamp1_1@data] # add the class names to each row

#After doing the same thing for all 5 classes,
#the dataframes are merged into a single dataframe
dt <- rbind(dt1, dt2, dt3, dt4, dt5)

names(dt)[names(dt) == 'id_cls'] <- 'class'
dt <- dt %>% drop_na() #deletes the rows with null values
dt$class <- factor(dt$class, labels=c('forest','urban', 'mountain', '
   ↪ vaia', 'pasture'))
#factor: Converts the class column to a factor with meaningful labels
```

3.4.3 Splitting the Dataset

The generated points were split into training and test datasets. A stratified random split ensured that 70% of the data was used for training and 30% for testing, preserving the class distribution. The **caret** package's **createDataPartition** function was used for this purpose.

Listing 3.4: Splitting the Dataset

```r
# Libraries required
library(caret)  # For creating data partitions


```

```
set.seed(321)
#set.seed: Sets the seed for reproducibility.

# A stratified random split of the data
idx_train <- createDataPartition(dt$class, p = 0.7, list = FALSE)


dt_train <- dt[idx_train] # Training set
dt_test <- dt[-idx_train] # Test set
```

3.4.4 Cross-Validation

The training dataset was used for performing cross-validation and grid search to tune the model. Cross-validation helps assess how the model will generalize to an independent dataset by providing a more reliable estimate of model performance. This is achieved by reducing the bias that can occur with a single train-test split.
In this process, the dataset is divided into a number of subsets, or "folds" (in this case, 10). During each iteration of cross-validation, one fold is used as the validation set, while the remaining folds are used for training the model. This procedure is repeated multiple times so that each fold serves as the validation set once. The performance metrics are averaged over all iterations to obtain a robust estimate of the model's performance.

Grid search is used alongside cross-validation to find the optimal hyperparameters for the model. The **tuneGrid** function was used to specify a grid of **mtry** values to be tested. The best **mtry** value (which represents the number of features randomly sampled as candidates at each split) was selected based on the cross-validation results.

Once the optimal parameters were determined, a final model was trained on the entire training dataset using these optimal hyperparameters. This final model was then used for making predictions on new data.

Listing 3.5: Cross-Validation for Sentinel 10m image of area A

```
# Libraries required
library(caret)      # For createFolds, trainControl, and train
library(MLmetrics)  # For multiClassSummary
library(randomForest) # For random forest method
library(dplyr)      # For data manipulation
library(data.table) # For data handling


# The dataset is divided into 10 subsets
n_folds <- 10
set.seed(321)
folds <- createFolds(1:nrow(dt_train), k = n_folds)
```

```
13
14 # Set the seed at each resampling iteration.
15 seeds <- vector(mode = "list", length = n_folds + 1)
16
17 # For each of the 10 iterations, one fold is used as the validation set
   ↪ and the remaining 9 folds are used for training.
18 for(i in 1:n_folds) seeds[[i]] <- sample.int(1000, n_folds)
19 seeds[n_folds + 1] <- sample.int(1000, 1)
20
21
22 # Specifying how the training should be controlled and validated
23 ctrl <- trainControl(summaryFunction = multiClassSummary,
24                      method = "cv",
25                      number = n_folds,
26                      search = "grid",
27                      classProbs = TRUE,
28                      savePredictions = TRUE,
29                      index = folds,
30                      seeds = seeds)
31
32
33 # This function sets up a grid of tuning parameters
34 # for a number of classification routines
35 model_rf <- caret::train(class ~ . ,
36                          method = "rf",
37                          data = dt_train,
38                          importance = TRUE,
39                          tuneGrid = data.frame(mtry = c(2, 3, 4, 5, 8))
                               ↪ ,
40                          trControl = ctrl)
```

3.4.5 Saving and Applying the Random Forest Model

After the model has been tuned, it is saved for future use. This trained model is then applied to make predictions on the entire area of interest (in this case, area A). The predicted classifications for each pixel in the raster data are saved as a new raster file.

Listing 3.6: Saving and Applying the Random Forest Model

```
1 # Libraries required
2 library(caret)  # For saving the trained model
3 library(raster) # For applying the model to raster data and writing the
   ↪ output
4
5
6 #saving the model
7 saveRDS(model_rf, file = paste0("Path_to_file","model_rf_10m","area_A",
   ↪ ".rds"))
```

```
8
9  # Using the trained model to predict classes in the raster data
10 predict_rf <- raster::predict(object = brick_for_prediction,
11                                model = model_rf, type = 'raw')
12 writeRaster(predict_rf, paste0("Path_to_file", "sentinel_10m_area_A_
   ↪ classification",".tiff"),overwrite=T )
```

3.5 Methodology for Spectral Variation Hypothesis

This section describes the methods used to evaluate biodiversity in the study area by integrating remote sensing data with field observations through the Spectral Variation Hypothesis. It explains the calculation of Rao's Q index from Sentinel-2 and EnMAP imagery, which was then compared with biodiversity metrics like species richness and the Shannon index derived from field data. This comparative analysis aimed to determine the effectiveness of remote sensing-based indices in assessing forest biodiversity.

3.5.1 Calculating Rao's Q index for Sentinel images

First, Sentinel-2 10m resolution images were taken and the Normalized Differnce Vegetation Index was calculated.

Normalized Difference Vegetation Index (NDVI) is a widely used remote sensing index that measures the health and density of vegetation in a given area. It is calculated using the difference between the near-infrared (NIR) light, which vegetation strongly reflects, and the red light, which vegetation absorbs. The formula for NDVI is:

$$\text{NDVI} = \frac{\text{NIR} - \text{Red}}{\text{NIR} + \text{Red}}$$

In this case, bands B08 (NIR) and B04 (Red) were used.
NDVI values range from -1 to 1, where values closer to 1 indicate healthy, dense vegetation, and values near 0 or negative suggest little to no vegetation, bare soil, or non-vegetated surfaces like water or urban areas. NDVI is a critical tool in agriculture, forestry, and environmental monitoring, allowing researchers and land managers to assess plant growth, detect changes in vegetation over time, and make informed decisions about land use and conservation.

Rao's Q index was then calculated

Listing 3.7: Rao's Q Index Calculation on Sentinel NDVI

```
1  # Libraries required
2  library(terra)          # For raster and vector data manipulation
```

```
library(exactextractr) # For exact extraction of raster values
library(stats)

# Iterate over each circle
for (i in seq_along(circles_filtered)) {
 # Extract the polygon representing the current circle
  circle_polygon <- circles_filtered[i, ]

  # Crop and mask the raster with the polygon
  extracted_values <- exact_extract(ndvi, circle_polygon, fun = NULL)

  #Extracting the value from each list element and converting them into
     ↪  a signle vector
  mat_s <- unlist(lapply(extracted_values, function(x) x$value))
  #Removing NA values
  mat_s <- mat_s[!is.na(mat_s)]

  # Rao's Q index calculation
  n_s <- length(mat_s)

  # Squaring the number of values. This will be used to normalize the
     ↪ Rao's Q index by the number of pairwise comparisons.
  n2_s <- n_s^2

  #Calculating the pairwise Euclidean distances between all NDVI values
     ↪  in mat_s
  distm_s <- as.matrix(dist(mat_s))

  #Computing the Rao's Q index by summing all the pairwise distances
     ↪ and dividing by n2_s
  rao_index <- sum(distm_s) / n2_s

  #Storing the Rao's Q index in the vector
  sentinel_rao_indices[i] <- rao_index
}
```

3.5.2 Calculating Rao's Q index for EnMAP images

For EnMAP images, the Principal Component was calculated. Principal Component Analysis (PCA) is a statistical technique used to simplify a dataset by reducing its dimensionality. It does this by transforming the original variables into a new set of variables called Principal Components (PCs). These PCs are uncorrelated and are ordered such that the first few retain most of the variation present in the original dataset. In this case, the PC1 was considered. The first principal component (PC1) captures the maximum variance from the dataset, so it's the direction in the data that has the most information or variability.

PCA is particularly useful in remote sensing and image analysis, where it helps to compress data from multiple bands or layers into fewer components that still retain most of the essential information.

Listing 3.8: PC1 Calculation on EnMAP images

```
# Libraries required
library(terra)          # For raster and vector data manipulation
library(exactextractr) # For exact extraction of raster values
library(stats)

# Define a function to standardize a single layer
standardize_layer <- function(layer) {
  values <- getValues(layer)
  #Subtracting the mean and divideing by the standard deviation,
     ↪ resulting in a distribution with a mean of 0 and a standard
     ↪ deviation of 1
  standardized_values <- scale(values, center = TRUE, scale = TRUE)
  #Updating the original raster layer with the standardized values.
  setValues(layer, standardized_values)
}

# Standardize each layer in the raster stack
standardized_layers <- stack(lapply(1:nlayers(enmap), function(i)
   ↪ standardize_layer(enmap[[i]])))

#performing the PCA
enmap_pca <- rasterPCA(standardized_layers)
```

Rao's Q index was then calculated on the first PC, using the same procedure as was used for the Sentinel images.

Since the results were not satisfactory, an alternative approach was employed by calculating Rao's Q index on specific optical traits. These traits were derived using a look-up table (LUT) within the EnMAP Box plugin in QGIS. A LUT is a data structure that maps input values to desired output values, facilitating the classification of data based on predefined criteria. In this context, the LUT was used to relate spectral data from EnMAP images to specific biophysical and biochemical properties of vegetation, referred to as optical traits. The use of optical traits provides a more direct link to the vegetation's physiological characteristics, potentially leading to better estimation of biodiversity.

The optical traits considered in this study include:

- **Structure Parameter (N)**: Represents the structural properties of vegetation, influencing light scattering within the canopy.

- **Chlorophyll A + B (Cab)**: Indicates the chlorophyll content, which is crucial for photosynthetic activity.
- **Water Content (Cw)**: Reflects the water content within the vegetation, affecting the spectral reflectance in specific wavelengths.
- **Dry Matter Content (Cm)**: Relates to the amount of dry biomass in the leaves, influencing their reflectance properties.
- **Carotenoids (Ccx)**: Pigments in the leaves that protect against photooxidative damage, affecting the spectral response.
- **Brown Pigments (Cbrown)**: Represent non-photosynthetic pigments, which are indicative of senescence or stress conditions.
- **Anthocyanins (Canth)**: Pigments that provide protective functions in plants, contributing to the red coloration in leaves.
- **Proteins (Cp)**: Related to the nitrogen content, which is essential for various physiological processes.
- **Carbon-Based Constituents (CBC)**: Represents carbon compounds like lignin and cellulose, influencing the structural integrity of leaves.

Additional canopy model parameters used include:

- **Leaf Area Index (LAI)**: Measures the total leaf area per unit ground area, affecting the interception of light.
- **Leaf Angle (ALIA)**: Describes the angular distribution of leaves, which influences light absorption and scattering.
- **Hot Spot Size Parameter**: Affects the reflectance when the observation and illumination directions align.

By incorporating these optical traits, the study aims to leverage specific spectral properties that are directly linked to the physiological and structural characteristics of vegetation, potentially enhancing the accuracy of biodiversity estimations through Rao's Q index.

Chapter 4

Results

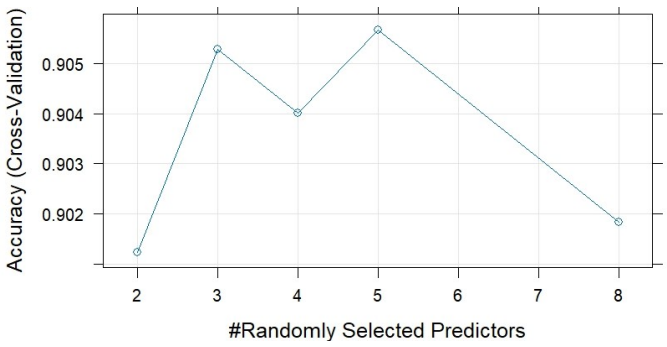
This chapter presents the outcomes of the land cover classification and SVH in the South Tyrol region. After calculating the accuracy, variable importance and confusion matrices for each satellite's classification and area of interest, satellite statistics were calculated, including overall accuracy, kappa statistics, and variable importance. The chapter also compares the Rao's Q index derived from remote sensing imagery with field-based biodiversity metrics, such as species richness and the Shannon index, to evaluate the effectiveness of various approaches in capturing forest biodiversity.

4.1 Land Cover Classification

4.1.1 Tuning Results

The **plot(model_rf)** function is used to visualize the performance of the model with different hyperparameters. The plot displays the accuracy against different values of the tuning parameters used in the model. In the case of a Random Forest model, the tuning parameter is **mtry**, which represents the number of variables randomly sampled as candidates at each split.

The accuracy of a model is defined as the proportion of correct predictions (the count of data points for which the predicted class label matches the actual class label) out of the total number of predictions.
In the graph, an example of the tuning results for the classification of Landsat-8 image of area A are shown.



The **caret::train** function automatically selects the best model based on the highest performance metric. This model is saved and can be used for making predictions.
The value of mtry balances the trade-off between bias and variance: a smaller mtry introduces more randomness, leading to more diverse trees and potentially reducing overfitting. However, too small a value can increase bias. Conversely, a larger mtry allows more features to be considered, which can reduce bias but increase the risk of overfitting.
In this case, the cross-validation process determined that mtry = 5 provides the best performance for the model, indicating that considering 5 bands at a time for splits results in the highest accuracy.

4.1.2 Variable Importance

The function **randomForest::varImpPlot(model_rf$finalModel)** generates a variable importance plot (Figure 4.1), which is a useful tool for evaluating the significance of each feature in the Random Forest model. the plot consists of two metrics: Mean Decreased Accuracy and Mean Decreased Gini.

Mean Decreased Accuracy

- This metric measures how much the model's accuracy decreases when a particular feature (in this case the satellite bands) is excluded.
- A higher value indicates that the feature is more important because removing it causes a significant drop in model accuracy.

Mean Decreased Gini

- This metric measures the total decrease in node impurity (measured by the Gini index) that a feature achieves across all trees in the forest.
- A higher value indicates that the feature is important in improving the purity of the nodes and thus the model's classification performance.

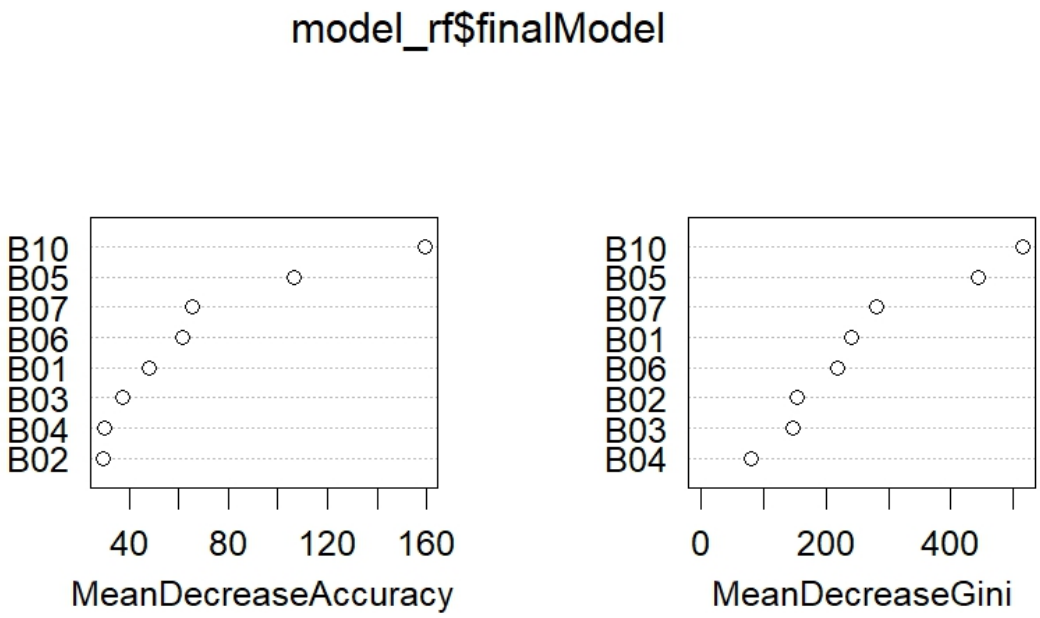


Figure 4.1: Mean Decreased Accuracy and Mean Decreased Gini of Landsat-8 classification of area A

The features at the top of the plots are the most important for the model. These are the bands that the model relies on the most for making accurate predictions. If some features have very low importance scores, one might consider removing them to simplify the model. This can make the model more interpretable and reduce computational costs without significantly affecting performance.

By using the **caret::varImp(model_rf)** function one can also generate a heatmap visualization of feature importance scores (Figure 4.2), where the color intensity represents the magnitude of importance assigned to each feature.

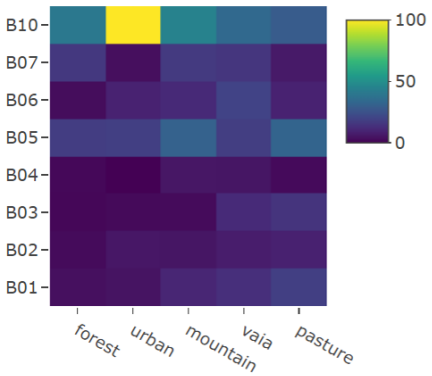


Figure 4.2: Predictor importance of Landsat-8 classification of area A

As expected from the mean decreased accuracy plot, bands 10 and 5 seem to be more influential in determining the class labels in this example. In particular, band 10 plays a crucial role in distinguishing urban areas.

4.1.3 Confusion Matrix

By using the function **confusionMatrix(data , reference)** the confusion matrix was computed, to investigate predictions versus actual class labels, allowing for a thorough evaluation of the model's performance (Figure 4.3).

```
          Reference
Prediction forest urban mountain vaia pasture
  forest      226     0        0    5       0
  urban         0   223        0    0       0
  mountain      0     0      212    0       0
  vaia          0     0        0  219       0
  pasture       0     0        1    0     221

Overall Statistics

               Accuracy : 0.9946
                 95% CI : (0.9882, 0.998)
    No Information Rate : 0.2042
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9932
```

Figure 4.3: Confusion Matrix for the classification of area A using Landsat images

To obtain this, the function **predict(model_rf, newdata = dt_test)** was used to use the model to make predictions on the unseen test dataset.
In the table, each row represents the instances in an actual class, whereas each column represents the instances in a predicted class. The main diagonal (from top left to bottom right) indicates the number of correct predictions for each class, while the off-diagonal elements show the misclassifications.

Parameters

- **Accuracy**: The ratio of correctly predicted instances to the total instances. In a confusion matrix, it's the sum of the diagonal elements divided by the total number of instances.
- **95% CI (Confidence Interval)**: This is the range within which the true accuracy of the model is expected to lie with 95% confidence. It gives an idea of the reliability of the accuracy measurement. In this case, one can be 95% confident that the accuracy of this model lies between 0.9882 and 0.9880.
- **No Information Rate (NIR)**: It's a baseline measure that represents the accuracy of a simple classification model that always predicts the most frequent class

in the dataset.This rate helps to understand how well a more complex model (like this Random Forest model) is performing compared to this simple baseline.

- **P-value (acc > NIR)**: This p-value tests the null hypothesis that the model's accuracy is no better than the NIR. Essentially, it helps determining if the observed accuracy improvement of the model over the NIR is statistically significant or if it could have happened by chance. A low p-value (typically < 0.05) indicates that the model's accuracy is significantly better than just predicting the most frequent class.

- **Kappa**: It is a measure of agreement between predicted and observed classifications, correcting for the possibility of agreement occurring by chance

4.1.4 Satellites Statistics

To retrieve the accuracy and kappa statistics of each satellite, the test datasets of each of the four areas of interest were joint and the confusion matrix was calculated. This was done for every satellite.

Listing 4.1: Confusion Matrix and Statistics of Sentinel-2 10m

```
# Libraries required
library(caret)

#SENTINEL 10m CONFUSION MATRIX
test_s10 <- c(test_s10A, test_s10B, test_s10C, test_s10D)
class_s10 <- c(class_s10A, class_s10B, class_s10C, class_s10D)
cm_s10 <- confusionMatrix(data = test_s10, class_s10)
cm_s10
```

The statistics obtained from the models for each satellite are shown in Table 4.1. The classification results from different satellite data highlight the trade-offs between resolution, the number of spectral bands, and classification accuracy.

Satellite	Resolution	Bands	Accuracy	Kappa	Most Important Predictors
Sentinel-2	10m	4	0.885	0.856	B08 (832.8 nm) and B02 (492.4 nm)
Sentinel-2 resampled	10m	11	0.972	0.965	B01 (442.7 nm) and B11 (1613.7 nm)
Sentinel-2	20m	7	0.985	0.981	B01 (442.7 nm) and B11 (1613.7 nm)
Landsat-8	30m	11	0.998	0.998	B10 (10600 – 11190 nm) and B05 (850 - 880 nm)
EnMAP	30m	224	0.99	0.987	It varies depending on the area

Table 4.1: Model statistics for each satellite.

Prediction	Reference: forest	urban	mountain	vaia	pasture
forest	782	4	8	116	5
urban	3	802	82	17	1
mountain	23	49	789	19	4
vaia	83	33	11	724	24
pasture	5	6	1	22	857

Figure 4.4: Sentinel-2 at 10m Confusion Matrix

Prediction	Reference: forest	urban	mountain	vaia	pasture
forest	865	0	0	42	3
urban	1	880	6	4	0
mountain	2	7	894	0	0
vaia	28	9	1	850	11
pasture	3	2	0	6	887

Figure 4.5: Sentinel-2 at 10m Resampled Confusion Matrix

Prediction	Reference: forest	urban	mountain	vaia	pasture
forest	895	0	0	29	1
urban	0	888	8	0	0
mountain	0	2	888	0	0
vaia	6	4	0	866	13
pasture	1	2	0	1	878

Figure 4.6: Sentinel-2 at 20m Confusion Matrix

Prediction	Reference: forest	urban	mountain	vaia	pasture
forest	902	0	0	5	0
urban	0	897	0	0	1
mountain	0	0	866	0	0
vaia	0	0	2	892	0
pasture	0	0	1	0	891

Figure 4.7: Landsat Confusion Matrix

Prediction	Reference: forest	urban	mountain	vaia	pasture
forest	895	2	1	9	1
urban	1	890	1	4	4
mountain	1	0	895	1	1
vaia	3	4	2	883	6
pasture	0	1	0	3	889

Figure 4.8: EnMAP Confusion Matrix

4.2 Spectral Variation Hypothesis

4.2.1 Rao's Q index for Sentinel-2 images

This section presents the results of the analysis of the calculated Rao's Q index on the Sentinel-2 images and their comparison with the field data related to the Shannon index and species richness. Two graphs were generated showing the relationships between Rao's Q index and the Shannon index, and between Rao's index and species richness.

In the first graph (Figure 4.9), Rao's Q index was related to the Shannon index, obtaining a value of R^2 equal to 0.547. The R^2 value, or coefficient of determination, indicates the proportion of variance in the Shannon index which is explained by variance in Rao's Q index. In this case, a value of 0.547 suggests that about 54.7% of the variation in the Shannon index can be explained by the variation in Rao's Q index, indicating a moderate correlation between the two diversity metrics.

In the second graph (Figure 4.10), the relationship between Rao's Q index and species richness is shown, with a value of R^2 equal to 0.267. This value indicates that only 26.7% of the variance in the species species richness is explained by Rao's Q index, suggesting a weaker correlation than that observed with the Shannon index.

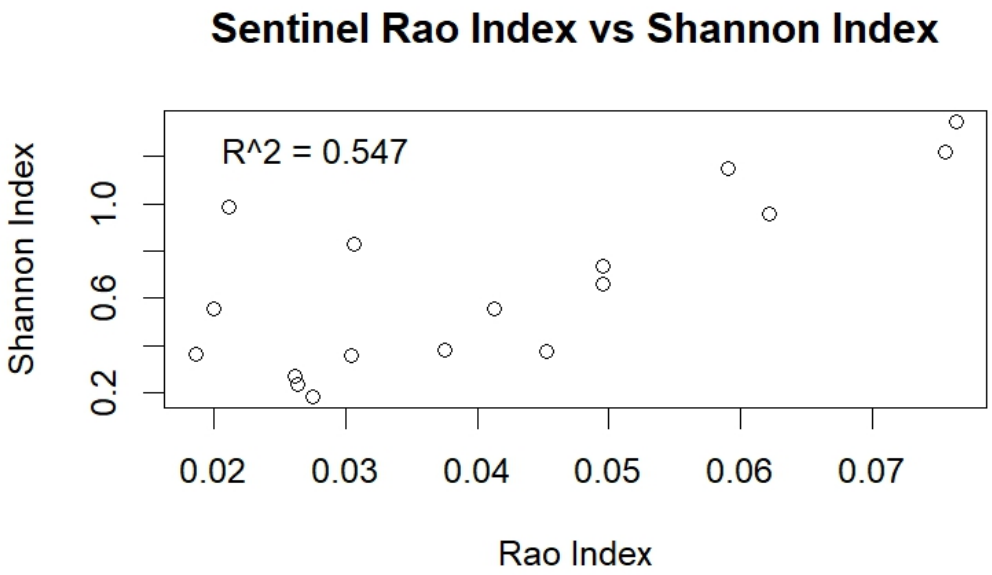


Figure 4.9: Relationship between Rao's Q index and the Shannon index.

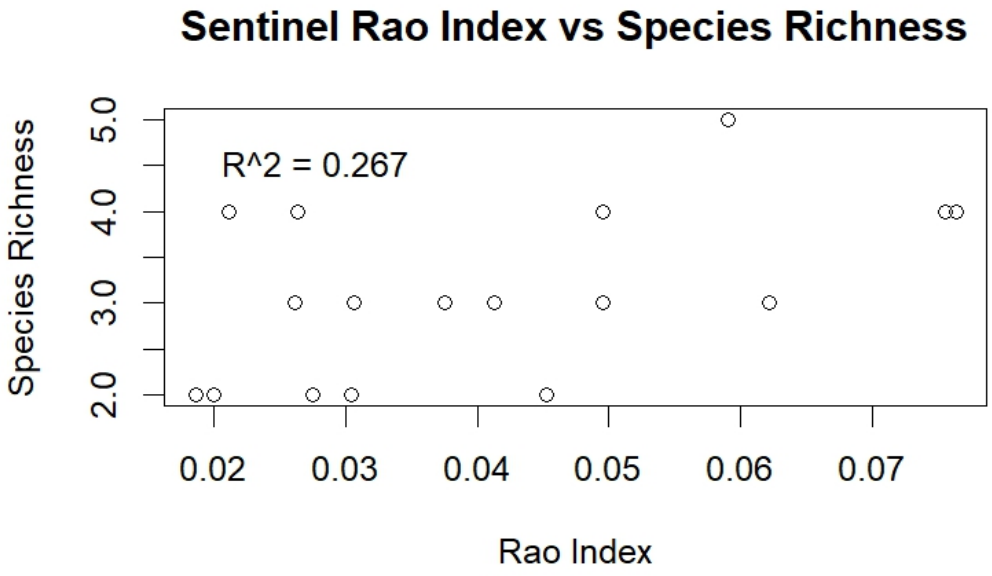


Figure 4.10: Relationship between Rao's Q index and the species richness.

4.2.2 Rao's Q index for EnMAP images

For the EnMAP images, Rao's Q index was initially calculated on the principal component (PC1) and subsequently compared with the Shannon index and species richness. The results of these correlations were unsatisfactory, with very low R^2 values: 1e-04 for the relationship between Rao's Q index (calculated on the PCA) and the Shannon index, and 0.012 for the relationship between Rao's Q index (PCA) and species richness (Figures 4.11 and 4.12).

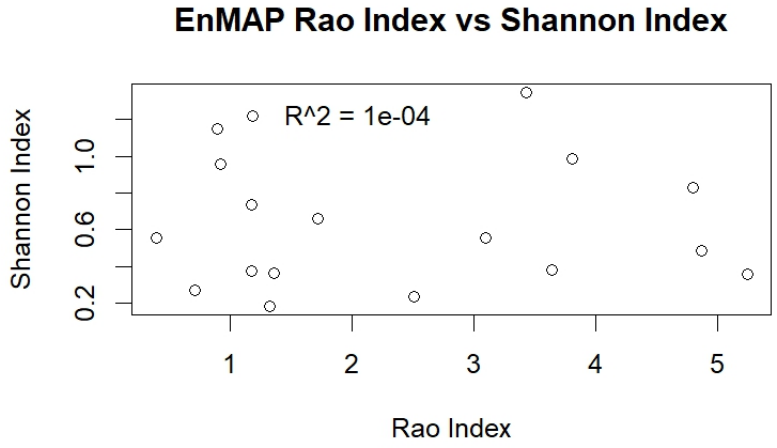


Figure 4.11: Relationship between Rao's Q index calculated on the PCA and the Shannon index.

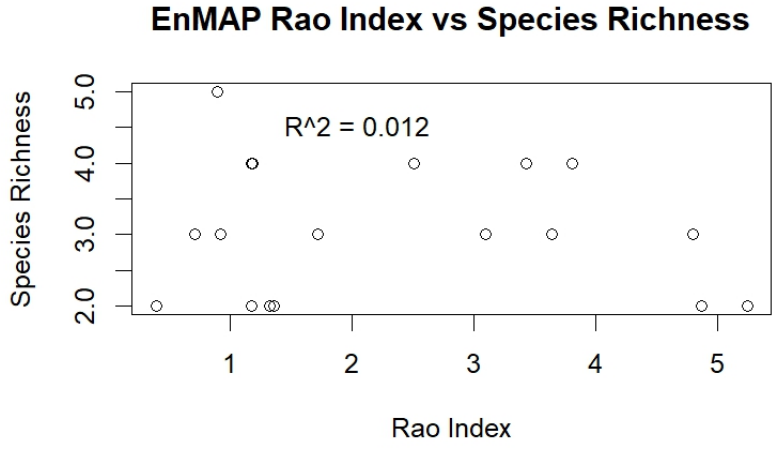


Figure 4.12: Relationship between Rao's Q index calculated on the PCA and species richness.

Subsequently, Rao's Q index was calculated on various optical traits to improve the correlation. The results show the following relationships:

- **Anthocyanins (Anth)**: $R^2 = 0.003$ with the Shannon index and $R^2 = 0.006$ with species richness.
- **Chlorophyll A + B (Cab)**: $R^2 = 0.138$ with the Shannon index and $R^2 = 0.126$ with species richness.
- **Brown Pigments (Cbrown)**: $R^2 = 0.073$ with the Shannon index and $R^2 = 0.004$ with species richness.
- **Dry Matter Content (Cm)**: $R^2 = 0.003$ with the Shannon index and $R^2 = 0.001$ with species richness.
- **Water Content (Cw)**: $R^2 = 0.185$ with the Shannon index and $R^2 = 0.052$ with species richness.
- **Hot-Spot Size Parameter (Hspot)**: $R^2 = 0.013$ with the Shannon index and $R^2 = 0.081$ with species richness.
- **Leaf Area Index (LAI)**: $R^2 = 0.013$ with the Shannon index and $R^2 = 0.01$ with species richness.
- **Leaf Inclination Distribution Function (LIDF)**: $R^2 = 1e\text{-}04$ with the Shannon index and $R^2 = 0.016$ with species richness.
- **Structure Parameter (N)**: $R^2 = 0.01$ with the Shannon index and $R^2 = 0.006$ with species richness.

The corresponding graphs (Figures 4.13 and 4.14) show how the variability of Rao's Q index calculated on optical traits relates to the Shannon index and species richness, generally highlighting weak correlations.

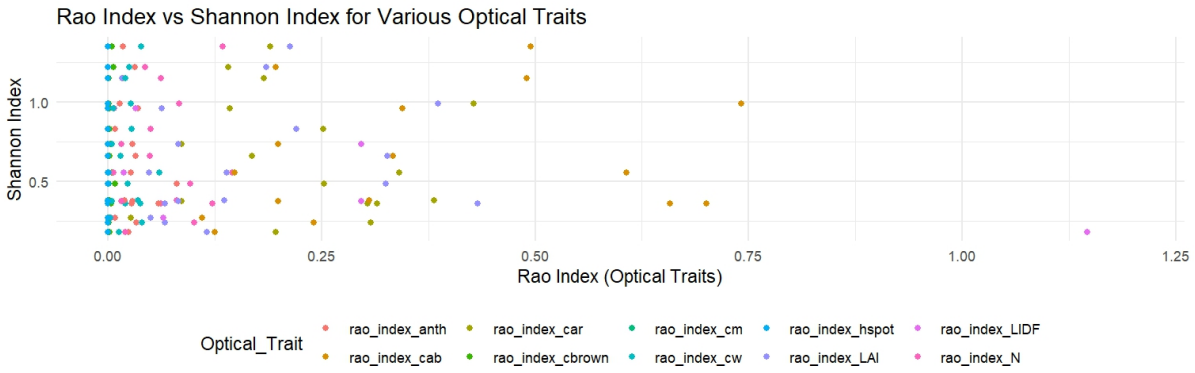


Figure 4.13: Relationship between Rao's Q index calculated on optical traits and the Shannon index.

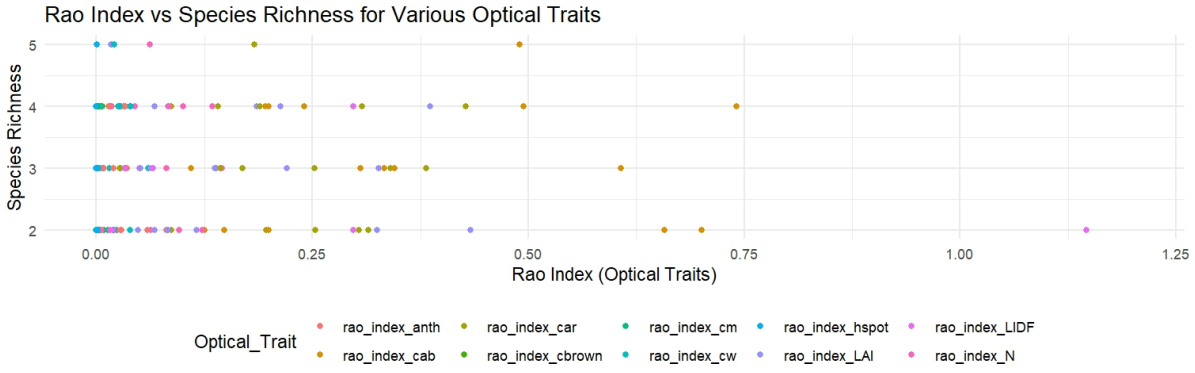


Figure 4.14: Relationship between Rao's Q index calculated on optical traits and species richness.

Chapter 5

Discussion

In this chapter, the main results obtained by the application of the land classification methods and the SVH using remote sensing data are analysed and discussed. The analysis focuses on the benefits and limitations of each methodology, with particular attention to the performance of the EnMAP satellite compared to multispectral sensors such as Sentinel-2. The ecological implications of these results are also explored, providing a more complete picture of the effectiveness of remote sensing for biodiversity management and monitoring environmental changes.

5.1 Land Cover Classification

5.1.1 Summary of Findings

Accuracy and Kappa coefficient

- **Landsat-8** shows the highest accuracy (99.8%) and Kappa coefficient (0.998), indicating it provides the most reliable classification results among the satellites tested. This high performance can be attributed to the inclusion of the thermal infrared band (B10), which is crucial for distinguishing between different land cover types. The confusion matrix for Landsat-8 further supports this high accuracy, with minimal misclassifications across different land cover types.

- **EnMAP** also demonstrates high accuracy (99%) and Kappa (0.987). Its extensive spectral coverage with 224 bands contributes to its robust performance, allowing for finer distinctions between land cover classes, even at a 30m resolution.

- **Sentinel-2** at 20m resolution achieves excellent accuracy (98.5%) and Kappa (0.981). Despite a coarser spatial resolution than the 10m counterpart, the 20m bands provide valuable spectral information for classification, especially the coastal (B01) and shortwave infrared (B11) bands, which are critical for soil and vegetation

moisture assessment. The confusion matrix for this dataset reveals a high level of precision, with only a few errors between urban and forest classes.

- **Sentinel-2** (10m resampled) showed significantly improved performance than the original 10m bands with an accuracy of 97.2% and a Kappa coefficient of 0.965. This resampling to 10m from 20m resolution provided finer spatial details while still leveraging the richer spectral information from the 11 bands, such as the B01 and B11 bands, which are critical for distinguishing vegetation and soil characteristics. The confusion matrix for Sentinel-2 10m resampled data supports this finding, with reduced misclassification compared to the original 10m data.

- **Sentinel-2** (10m original) exhibited the lowest accuracy of 88.5% and a Kappa coefficient of 0.856. While the high spatial resolution (10m) is beneficial for detailed mapping, the limitation of only four bands restricts the model's ability to capture subtle spectral differences. The confusion matrix reflects this, showing higher misclassification rates, especially between forest and mountain categories.

Predictor Importance

- For **Landsat-8**, bands B10 (thermal infrared) and B05 (near-infrared) are key.
  The thermal infrated band (10900 nm) measures the thermal radiation emitted from the earth's surface, which is very effective in differentiating surfaces that emit heat differently. Urban areas, with concrete and asphalt, retain more heat than vegetation and water bodies, which cool faster. Therefore, this band is particularly useful for classifying urban areas from natural ones. In addition, thermal bands are sensitive to soil moisture content and vegetation transpiration, helping to distinguish water content in soil and plants.
  The near-infrared (NIR) band (870 nm) is crucial for vegetation analysis because of its interaction with plant cell structure. Healthy vegetation strongly reflects NIR radiation, making this band useful for distinguishing vegetated from unvegetated areas. This band is particularly effective for assessing the health of vegetation, its density and for distinguishing between different types of vegetation cover (e.g., forests versus grasslands).

- For **EnMAP**, the importance of bands varies by area due to its extensive spectral range, allowing for flexible and detailed spectral analysis tailored to specific applications.

- For **Sentinel-2** (20 m) the most important predictors were B01 and B11.
  B01 (442.7 nm) is particularly sensitive to atmospheric particles such as aerosols and fine dust, as well as coastal water conditions. In addition, B01 is useful in coastal studies to detect water quality, sediment concentrations and to differentiate land-water boundaries, especially in wetlands or coastal environments.

Band B11 (1613.7 nm) is sensitive to moisture content in vegetation and soils. Water strongly absorbs radiation in this region, making B11 essential for detecting soil and plant water stress, monitoring plant transpiration and distinguishing between wet and dry surfaces. It is particularly useful in agricultural and forestry areas, where water content directly affects plant health and soil classification.

- For **Sentinel-2** (10m resampled) bands B01 (442.7 nm) and B11 (1613.7 nm) were similarly crucial. Their importance lies in the ability to capture detailed spectral information, allowing the model to perform more accurate classifications compared to Sentinel-2 (10m original).

- For **Sentinel-2** (10m original), bands B08 and B02 were the most influential.
  B08 (832.8 nm) is essential for assessing vegetation health. Near infrared (NIR) is reflected by the leaves due to their internal structure, and variations in NIR reflectance help distinguish between healthy and stressed vegetation. This band is particularly useful for distinguishing between different types of vegetation and is a key component of vegetation indices such as the NDVI.
  B02 (492.4 nm) is particularly useful for monitoring surface water, as pure water reflects blue light significantly, allowing sediment, algae and pollutants to be detected. It is also sensitive to atmospheric diffusion, so it can be used to study aerosols and correct for atmospheric effects. B02 is also effective in mapping coastal zones and glaciers, and helps to distinguish reflective surfaces (such as snow and ice) from vegetation, which absorbs blue for photosynthesis.

The results of the land cover classification showed that Landsat-8 achieved the highest accuracy (99.8%) and Kappa coefficient (0.998), mainly due to the inclusion of the infrared thermal band (B10). EnMAP showed high accuracy (99%) and a Kappa coefficient of 0.987, with solid performance due to its wide spectral coverage of 224 bands, although it presented some difficulties in distinguishing between areas affected by the Storm Vaia and grasslands. Sentinel-2 at 10m resolution showed the lowest accuracy despite having the highest spatial resolution, which improved drastically with the inclusion of the resampled bands. This highlighted the importance of the SWIR (B11) band, which was also crucial for the classification of the sentinel 20m resolution images.

5.1.2 Benefits and Limitations of the Methodology

The use of multispectral and hyperspectral data for landscape classification offered significant advantages, especially in the context of monitoring and classifying land cover on a large scale.

One of the main advantages of this methodology was its ability to handle large amounts of spatial and spectral data, allowing detailed classification of land cover. In-

tegration of data from satellites such as Sentinel-2 and EnMAP provided a wide and detailed spatial coverage, making it possible to monitor variations in difficult-to-access environments, such as dense forests or areas devastated by the Adrian Storm.

One of the main benefits of EnMAP, in particular, was its ability to detect subtle spectral variations, making it particularly effective in monitoring habitats with high biological diversity. Unlike other satellites with fewer bands, EnMAP's hyperspectral images provided a vast range of features to select from when building the Random Forest model. This flexibility allowed for more nuanced tuning of hyperparameters. Interestingly, the feature importance analysis did not highlight specific bands consistently across the entire classification area. Instead, different bands emerged as important depending on the specific habitat or classification region, underscoring the adaptability of EnMAP's hyperspectral data to a variety of ecological contexts.

An additional advantage came from the use of the robust and flexible Random Forest algorithm, which is particularly effective in handling high-dimensional data, like those of satellite images. The algorithm reduced the risk of overfitting, common in remote sensing data, thus improving the model's ability to generalize to new data.

However, even if the results were satisfactory for all the considered satellites, the methodology had some limitations. The accuracy of classification depended heavily on the spatial and spectral resolution of the satellite data used. Spectral resolution appeared to be more crucial than spatial resoltuion in the accuracy of the classification, but while the use of hyperspectral data offered greater detail, it also required more demanding computational handling and in-depth data preparation, such as the removal of problematic bands.

Moreover, EnMAP's spatial resolution (30 metres), although sufficient for this particular study, limits the ability to detect fine details in heterogeneous landscapes, which could represent a limitation in more ecologically complex areas. To overcome these limitations, it would be appropriate to include auxiliary data such as LiDAR images or UAVs (drones), which would add structural and altitude information, improving the distinction between different land cover and enhancing the accuracy of classification.

5.1.3 Ecological Implications

The classification of land use using hyperspectral and multispectral satellite images has fundamental implications for ecosystem management, biodiversity conservation and landscape planning. By carefully distinguishing the different types of land cover, especially in diverse ecosystems such as that of South Tyrol, this study provides valuable information on forest health, species distribution and ecosystem dynamics.

Monitoring and Management of Ecosystem Health
One significant ecological implication of the classification results lies in the potential for improved ecosystem monitoring. Remote sensing techniques, such as those used in this study, allow for a continuous and large-scale assessment of forest ecosystems. This capability is particularly relevant for fragmented or difficult to access environments, such as the mountainous areas of South Tyrol. The high precision achieved by classification models, in particular with Landsat-8 and EnMAP, supports the use of satellite images to detect changes in vegetation composition and ecosystem degradation due to natural or anthropogenic factors.

For example, the ability to differentiate areas affected by events such as Storm Vaia from other types of land cover provides an early warning system for disturbances to ecosystems. This facilitates more responsive and adaptive management strategies, allowing conservationists to intervene before irreversible losses of biodiversity occur. In addition, the hyperspectral resolution of EnMAP, although not always easy to apply, provides a better understanding of vegetation health indicators, It can be used to guide efforts to reforest or regenerate forests in degraded areas.

Improving Biodiversity Conservation
The results of the classification also have direct implications for biodiversity conservation. The ability to accurately classify and monitor different types of land cover allows for the identification of critical habitats and areas of high biodiversity. For example, the precise delineation of forest boundaries and the distinction between grasslands and degraded areas support priority conservation efforts. Therefore, satellite-based classification methods provide a non-invasive means of continuously monitoring habitat changes, facilitating the implementation of conservation strategies aligned with dynamic environmental conditions.

In addition, the ecological richness detected through multispectral and hyperspectral imaging provides information on the spatial distribution of species, thus supporting conservation planning. Protected areas can be adapted based on real-time data to maximise species conservation and minimise the impact of land use changes, climate change and human invasion.

Adaptation to Climate Change
In light of the growing impacts of climate change, the classification results offer a valuable tool for understanding the responses of ecosystems to climate change. The ability to map changes in vegetation patterns over time provides crucial data on how ecosystems like the forests of South Tyrol are adapting to changes in temperature, precipitation and the frequency of extreme weather events. This information can inform climate resilience

strategies by identifying areas of greatest ecological vulnerability and suggesting regions that may require targeted interventions to maintain biodiversity and ecosystem functionality.

Support for a Sustainable Management of the Territory
Finally, classification data support land use planning and sustainable forest management by providing a detailed picture of the dynamics of soil cover. The ability to assess large areas with high precision ensures that decisions regarding land use-whether for conservation, agriculture or development-are based on solid ecological data. For example, distinguishing between agricultural fields, grasslands and natural forests can help in zonal regulations that promote sustainable development by minimizing habitat destruction and biodiversity loss.

5.2 Spectral Variation Hypothesis

5.2.1 Summary of Findings

The results obtained by applying the SVH in the study area of South Tyrol have shown some interesting correlations, but also limitations that deserve reflection.

The analysis of Rao's Q Index on the images of Sentinel-2 showed a moderate correlation with the Shannon index, with a coefficient of determination (R^2) of 0.547, suggesting that about 54,7% of the variation in Shannon's index can be explained by the variation in Rao's Q index. This correlation, although not perfect, indicates that spectral heterogeneity can be a reasonable indicator of species diversity in a forest context. However, the relationship between Rao's Q index and species richness was weaker, with an R^2 value of 0.267, suggesting that only 26.7% of the variation in species richness can be explained by spectral variation.

The results for EnMAP images were less satisfactory. Initial analysis of the calculated Rao's Q index on the first principal component (PC1) showed very low correlations with the Shannon index (R^2 = 1e-04) and species richness (R^2 = 0.012), indicating that the spectral information captured was not sufficient to accurately represent biodiversity. Subsequently, the calculation of the Rao's Q index on individual optical traits slightly improved the correlation with the Shannon index, in particular for water content (R^2 = 0,185) and chlorophyll A + B (R^2 = 0,138). However, the overall values of correlations remained weak, suggesting that the optical traits considered are not fully representative of the ecological diversity in this area.

An assumption about the unsatisfactory results of EnMAP may be related to its

spatial resolution of 30 metres, which may not be fine enough to capture the ecological complexity and variability of species in a fragmented forest area. In addition, the hyperspectrality of EnMAP, while offering high spectral resolution, may have introduced spectral noise due to overlapping information not relevant to the context studied. This may have compromised the ability to accurately record species diversity.
Another factor that deserves attention is the seasonal growth cycles of plants. In September, vegetation may be transitioning into senescence (beginning to slow down before winter), which could reduce the distinctiveness of spectral signals. This change could make it harder to detect differences in species and reduce the effectiveness of the SVH. Therefore, using imagery from summer might provide stronger signals related to biodiversity, especially in forested ecosystems where phenological changes can significantly affect spectral reflectance.

In summary, the results confirm that spectral heterogeneity can be used as a proxy for biodiversity, but with some limitations related to the spectral and spatial resolution of the data, as well as the complexity of the ecosystem. Better results have been obtained with the Sentinel-2 images, while the use of EnMAP data requires further exploration to improve the accuracy of biodiversity estimates.

5.2.2 Benefits and Limitations of the Methodology

The Spectral Variation Hypothesis has proven to be a promising tool for non-invasive large-scale biodiversity estimation using remote sensing data. However, its effectiveness varied according to different technical and environmental factors, which influenced both the benefits and limitations of the method.

One of the main advantages of SVH was the ability to provide biodiversity estimates over large geographical areas without the need for complex and costly field campaigns. This efficiency is particularly useful in environments that are difficult to reach or where continuous monitoring is required, such as in forest or mountain environments. The possibility of using multispectral or hyperspectral images, such as those from Sentinel-2 and EnMAP, allowed detailed information on spectral variations related to the structure and composition of vegetation.

The use of indices such as Rao's Q index has proved to be an effective method for quantifying spectral heterogeneity, providing a proxy indicator of ecological diversity. In particular, the ability of Rao's Q index to consider both the composition and the functional dissimilarity of species makes this measure particularly suitable for capturing the complexity of plant communities.

Despite its benefits, the application of SVH had several limitations, mainly related to spatial resolution and spectral data quality. In the case of EnMAP images, for example, the spatial resolution of 30 metres may have not be fine enough to accurately distinguish biodiversity in environments with high fragmentation or small variations in plant composition.

Another limitation concerned the hyperspectrality itself. Although hyperspectral data provided a great abundance of information due to the large number of spectral bands, it can introduce spectral noise and information redundancy. Overlapping bands that are too similar may have confused the algorithm, reducing its ability to distinguish accurately between species or functional groups. This problem was evident in the results obtained with EnMAP images, where the correlation between Rao's Q index and biodiversity metrics was weak.

In addition, SVH can be sensitive to environmental conditions and seasonal variations. Phenomena such as plant phenology or the influence of climatic factors, such as drought or soil moisture, can affect spectral reflections, leading to temporary variations not necessarily related to biodiversity, but rather to contingent environmental conditions.

5.2.3 Ecological Implications

The results obtained by applying the SVH in this study offer several important ecological implications, especially for the management and conservation of forest ecosystems. The ability to estimate large-scale biodiversity through remote sensing represents a unique opportunity to address the challenges of biodiversity loss, climate change and sustainable management of natural resources.

Monitoring Biodiversity
The correlation between spectral heterogeneity, measured by Rao's Q index, and species diversity in forest contexts shows that remote sensing can be a powerful tool for monitoring biodiversity. In particular, the results obtained from Sentinel-2 data, which show a significant correlation between spectral variability and Shannon's index, highlight the potential of this approach to identify areas with high species diversity. This is crucial to identify and protect "biodiversity hotspots", which are areas of high biodiversity that may be threatened by human activities or environmental changes.

However, the use of hyperspectral data such as EnMAP has not yielded equally promising results, raising questions about the effectiveness of hyperspectrality for monitoring biodiversity in complex environments. This suggests that, in some contexts, a combination of multispectral and hyperspectral data may be necessary to accurately

capture ecological diversity.

Forest management and conservation
Remote sensing based on SVH offers new perspectives for forest management, as it allows the health and composition of forests to be assessed continuously and on a large scale. This is particularly relevant in mountain areas such as South Tyrol, where the forest structure is often fragmented and difficult to monitor with traditional methods. The application of SVH could therefore improve the ability to detect changes in species composition by providing early warning of forest degradation, invasion of non-native species, or reduction of biodiversity due to human activities.

The possibility of monitoring biodiversity continuously over time also allows to assess the effectiveness of conservation measures already in place and adapt them dynamically. For example, information from satellite imagery could be used to test the effect of reforestation policies, land management or deforestation protection, allowing more targeted management of natural resources.

Response to climate change
Loss of biodiversity is closely linked to climate change, and the ability to monitor changes in ecosystems in real time could be a key tool for understanding how plant communities respond to phenomena such as rising temperatures, the variation in precipitation or the frequency of extreme events such as storms and droughts. Spectral variability analysis could be used to study how forests change their functional composition and resilience in response to these environmental stresses.

In addition, the SVH approach could help to understand how climate change affects the distribution of species at landscape level. For example, identifying species sensitive to climate change could help predict future shifts in species distribution areas and develop adaptation strategies that minimise habitat and biodiversity loss.

Implications for ecological research
Finally, the integration of remote sensing methods with classic ecological approaches, such as the use of biodiversity metrics in the field, opens up new possibilities for ecological research. This study demonstrates the importance of combining different data sources to obtain a comprehensive view of ecosystems. The results indicate that, although spectral analysis can provide valuable insights into biodiversity, it should always be accompanied by field observations to ensure a correct interpretation of ecological dynamics.

Chapter 6

Conclusion

This study aimed to evaluate the effectiveness of remote sensing technologies, specifically multispectral data from Sentinel-2 and Landsat-8, and hyperspectral data from EnMAP, in the classification of forest landscapes and the assessment of biodiversity in the South Tyrol region. Using a combination of the Random Forest algorithm and the Spectral Variation Hypothesis, this research represents the first attempt to test the SVH using EnMAP data, offering insights into both the strengths and limitations of these methodologies.

The findings confirm the high potential of multispectral data for land cover classification. The Random Forest algorithm demonstrated high classification accuracy across all datasets, with Sentinel-2 and Landsat-8 performing well in distinguishing different land cover types. However, it is worth noting that despite its higher spectral resolution, EnMAP's performance in land cover classification did not significantly surpass that of the multispectral sensors. This suggests that while hyperspectral data offers greater spectral detail, the added complexity may not always translate into improved classification accuracy for certain landscape types.

The application of the SVH for biodiversity estimation yielded mixed results. Sentinel-2 multispectral data provided relatively robust correlations between spectral variation and biodiversity indices, particularly Rao's Q index, highlighting the potential of multispectral data to serve as proxies for biodiversity in forested ecosystems. By contrast, EnMAP data, despite its high spectral resolution, did not produce significant correlations with field-measured biodiversity indices. One possible explanation for this outcome could be the timing of the EnMAP image acquisition in September, when vegetation activity may have been lower, reducing the detectable spectral variability associated with biodiversity.

These results underscore the importance of temporal factors when using remote sens-

ing data for biodiversity monitoring. The seasonal dynamics of vegetation, which can strongly influence spectral signatures, suggest that satellite imagery captured during peak growing seasons may provide more reliable estimates of biodiversity, particularly when using hyperspectral data. Future research should explore the seasonal variability of spectral data to better understand the optimal conditions for biodiversity assessment.

Moreover, this research highlights some limitations in the hyperspectrality itself for biodiversity monitoring. While the fine spectral resolution of EnMAP theoretically allows for more precise detection of subtle vegetation traits, this advantage may be outweighed by challenges in data processing, as well as the potential for oversaturation of spectral information that complicates the analysis. Multispectral data, in contrast, offers a more accessible and computationally efficient alternative for large-scale biodiversity monitoring, with results that are still highly relevant for ecological applications.

In conclusion, this study contributes to the growing body of knowledge on the use of remote sensing for biodiversity assessment, providing valuable insights into the performance of both multispectral and hyperspectral sensors. The results suggest that while hyperspectral data holds promise for detecting subtle ecological variations, multispectral data currently offers a more practical solution for large-scale biodiversity monitoring. Future advancements in hyperspectral technology, particularly in terms of processing algorithms and temporal resolution, may further enhance its utility in this field. Additionally, the application of the SVH in diverse ecological contexts and the integration of time-series analysis represent promising avenues for future research, with the potential to improve the precision and applicability of remote sensing for biodiversity conservation efforts.

6.1 Appendix A: Additional R Code for EnMAP Classification

Listing 6.1: Complete code for EnMAP classification

```
#importing packages
library(raster)
library(rgdal)
library(sf)
library(sp)
library(RStoolbox)
library(rasterVis)
library(mapview)
library(data.table)
library(RColorBrewer)
library(plotly)
```

```
library(grDevices)
library(caret)
library(randomForest)#mis
library(ranger)
library(MLmetrics)
library(nnet)
library(NeuralNetTools)
library(LiblineaR)
library(data.table)
library(dplyr)
library(stringr)
library(doParallel)
library(snow)
library(parallel)
library(tidyr)
library(maptools)

############### WORKING WITH ENMAP DATASET ##########################################
#224 bands with 30 m resolution
#bands 131, 132, 133, 134, 135 have missing values

#####WORKING ON AREA A
#Importing only the area of interest (A)
rst_lst <- stack('path_to_file.tif')
rst_lst <- as.list(rst_lst)   #transforming rasterstack into list
names(rst_lst) <- 1:224

#dropping the columns with missing values
rst_lst <- rst_lst[-c(131:135)]

#Visualize the image in Natural Color (R = Red, G = Green, B = Blue).
suppressWarnings({viewRGB(brick(rst_lst[1:44]), r = 44, g = 21, b = 5) })

brick_for_prediction <- brick(rst_lst)

#importing the shp file of area A
poly_area_A <-shapefile('path_to_file.shp')
poly_area_A@data$id <- as.integer(factor(poly_area_A@data$id))
setDT(poly_area_A@data)


ptsamp1<-subset(poly_area_A, id == "1") #selecting polygons with id=1
ptsamp1_1 <- spsample(ptsamp1, 750, type='regular') #selecting 750 random points in the polygons with id=1
ptsamp1_1$class <- over(ptsamp1_1, ptsamp1)$id #assigning the id to the  random points
saveRDS(ptsamp1_1, file=paste0 ("path_to_file", file="_ptsamp1_A.rds"))
```

```
57
58 ptsamp2<-subset(poly_area_A, id == "2") #selecting polygons with id=2
59 ptsamp2_2 <- spsample(ptsamp2, 750, type='regular') #selecting 750
   ↪ random points in the polygons with id=2
60 ptsamp2_2$class <- over(ptsamp2_2, ptsamp2)$id #assigning the id to the
   ↪  random points
61 saveRDS(ptsamp2_2, file=paste0 ("path_to_file", file= "_ptsamp2_A.rds")
   ↪ )
62
63 ptsamp3<-subset(poly_area_A, id == "3") #selecting polygons with id=3
64 ptsamp3_3 <- spsample(ptsamp3, 750, type='regular') #selecting 750
   ↪ random points in the polygons with id=3
65 ptsamp3_3$class <- over(ptsamp3_3, ptsamp3)$id #assigning the id to the
   ↪  random points
66 saveRDS(ptsamp3_3, file=paste0 ("path_to_file", file= "_ptsamp3_A.rds")
   ↪ )
67
68 ptsamp4<-subset(poly_area_A, id == "4") #selecting polygons with id=4
69 ptsamp4_4 <- spsample(ptsamp4, 750, type='regular')  #selecting 750
   ↪ random points in the polygons with id=4
70 ptsamp4_4$class <- over(ptsamp4_4, ptsamp4)$id #assigning the id to the
   ↪  random points
71 saveRDS(ptsamp4_4, file=paste0 ("path_to_file", file= "_ptsamp4_A.rds")
   ↪ )
72
73 ptsamp5<-subset(poly_area_A, id == "5") #selecting polygons with id=5
74 ptsamp5_5 <- spsample(ptsamp5, 750, type='regular')   #selecting 750
   ↪ random points in the polygons with id=5
75 ptsamp5_5$class <- over(ptsamp5_5, ptsamp5)$id  #assigning the id to
   ↪ the random points
76 saveRDS(ptsamp5_5, file=paste0 ("path", file="_ptsamp5_A.rds"))
77
78
79 #saving the information of the random points into a data frame
80 dt1 <- brick_for_prediction %>%
81   raster::extract(y = ptsamp1_1) %>%
82   as.data.table %>%
83   .[, id_cls := ptsamp1_1@data] # add the class names to each row
84
85 dt2 <- brick_for_prediction %>%
86   raster::extract(y = ptsamp2_2) %>%
87   as.data.table %>%
88   .[, id_cls := ptsamp2_2@data] # add the class names to each row
89
90 dt3 <- brick_for_prediction %>%
91   raster::extract(y = ptsamp3_3) %>%
92   as.data.table %>%
93   .[, id_cls := ptsamp3_3@data] # add the class names to each row
94
```

```
dt4 <- brick_for_prediction %>%
  raster::extract(y = ptsamp4_4) %>%
  as.data.table %>%
  .[, id_cls := ptsamp4_4@data] # add the class names to each row

dt5 <- brick_for_prediction %>%
  raster::extract(y = ptsamp5_5) %>%
  as.data.table %>%
  .[, id_cls := ptsamp5_5@data] # add the class names to each row


#merging the dataframes into a single dataframe
dt<-rbind(dt1, dt2, dt3, dt4, dt5)
names(dt)[names(dt) == 'id_cls'] <- 'class'
dt<-dt %>% drop_na()
dt$class <- factor(dt$class, labels=c('forest','urban', 'mountain', '
    ↪ vaia', 'pasture'))



#Random Forest algorith
set.seed(321)
# A stratified random split of the data
idx_train <- createDataPartition(dt$class,
                                 p = 0.7, # percentage of data as
                                     ↪ training
                                 list = FALSE)


dt_train <- dt[idx_train]
dt_test <- dt[-idx_train]


# create cross-validation folds (splits the data into n random groups)
n_folds <- 10
set.seed(321)
folds <- createFolds(1:nrow(dt_train), k = n_folds)

#Explanding mtry interval to consider more bands
tuneGrid <- expand.grid(mtry = seq(2, 215, by = 10)) # from 2 to 244,
    ↪ step=10

# Set the seed at each resampling iteration. Useful when running CV in
    ↪ parallel.
seeds <- vector(mode = "list", length = n_folds + 1)
for(i in 1:n_folds) seeds[[i]] <- sample.int(1000, nrow(tuneGrid))
seeds[n_folds + 1] <- sample.int(1000, 1)

ctrl <- trainControl(summaryFunction = multiClassSummary,
```

```
 method = "cv", number = n_folds,  search = "grid", classProbs = TRUE,
    ↪  savePredictions = TRUE, index = folds, seeds = seeds)

model_rf <- caret::train(class ~ . , method = "rf", data = dt_train,
   ↪ importance = TRUE, tuneGrid = tuneGrid, trControl = ctrl)

#saving the model
saveRDS(model_rf, file = paste0("path_to_file","model_rf_","area_A",".
   ↪ rds"))

predict_rf <- raster::predict(object = brick_for_prediction,
                              model = model_rf, type = 'raw')
writeRaster(predict_rf, paste0("path_to_file","enmap_area_A_
   ↪ classification",".tiff"),overwrite=T )


####EVALUATION OF THE MODEL
plot(model_rf) # tuning results

#confusion matrix and statistics
test_e30A <- predict(model_rf, newdata = dt_test)
class_e30A <- dt_test$class
cm <- confusionMatrix(data = test_e30A, class_e30A)
cm


#ordering them by predictor importance across the classes
vi <- varImp(model_rf)$importance
vi$max <- apply(vi, 1, max)
vi <- vi[order(-vi$max),]
#selecting only the 20 most important bands
vi20 <- head(vi, 20)

vi20%>%
     as.matrix %>%
     plot_ly(x = colnames(.)[1:5], y = rownames(.), z = ., type = "
        ↪ heatmap",
             width = 350, height = 300)

#mean decrease accuracy and mean decrease gini
randomForest::varImpPlot(model_rf$finalModel)




####WORKING ON AREA B
#Importing only the area of interest (B)
rst_lst <- stack('path_to_file.tif')
rst_lst <- as.list(rst_lst)   #transforming rasterstack into list
```

```
names(rst_lst) <- 1:224

#dropping the columns with missing values
rst_lst <- rst_lst[-c(131:135)]

#Visualize the image in Natural Color (R = Red, G = Green, B = Blue).
suppressWarnings({viewRGB(brick(rst_lst[1:44]), r = 44, g = 21, b = 5) })

brick_for_prediction <- brick(rst_lst)

#importing the shp file of area B
poly_area_B <-shapefile('path_to_file.shp')
poly_area_B@data$id <- as.integer(factor(poly_area_B@data$id))
setDT(poly_area_B@data)

ptsamp1<-subset(poly_area_B, id == "1") #selecting the polygons with id =1
ptsamp1_1 <- spsample(ptsamp1, 750, type='regular') #selecting 750 random points within the selected polygons
ptsamp1_1$class <- over(ptsamp1_1, ptsamp1)$id #assigning the value id =1 to the random points
saveRDS(ptsamp1_1, file=paste0 ("path_to_file", file="_ptsamp1_B.rds"))

ptsamp2<-subset(poly_area_B, id == "2") #selecting the polygons with id =2
ptsamp2_2 <- spsample(ptsamp2, 750, type='regular') #selecting 750 random points within the selected polygons
ptsamp2_2$class <- over(ptsamp2_2, ptsamp2)$id #assigning the value id =2 to the random points
saveRDS(ptsamp2_2, file=paste0 ("path_to_file", file= "_ptsamp2_B.rds") )

ptsamp3<-subset(poly_area_B, id == "3") #selecting the polygons with id =3
ptsamp3_3 <- spsample(ptsamp3, 750, type='regular') #selecting 750 random points within the selected polygons
ptsamp3_3$class <- over(ptsamp3_3, ptsamp3)$id #assigning the value id =3 to the random points
saveRDS(ptsamp3_3, file=paste0 ("path_to_file", file= "_ptsamp3_B.rds") )

ptsamp4<-subset(poly_area_B, id == "4") #selecting the polygons with id =4
ptsamp4_4 <- spsample(ptsamp4, 750, type='regular') #selecting 750 random points within the selected polygons
ptsamp4_4$class <- over(ptsamp4_4, ptsamp4)$id #assigning the value id =4 to the random points
```

```
saveRDS(ptsamp4_4, file=paste0 ("path_to_file", file= "_ptsamp4_B.rds") )

ptsamp5<-subset(poly_area_B, id == "5") #selecting the polygons with id =5
ptsamp5_5 <- spsample(ptsamp5, 750, type='regular') #selecting 750 random points within the selected polygons
ptsamp5_5$class <- over(ptsamp5_5, ptsamp5)$id #assigning the value id =5 to the random points
saveRDS(ptsamp5_5, file=paste0 ("path_to_file", file="_ptsamp5_B.rds"))


#saving the information of the random points into a dataframe
dt1 <- brick_for_prediction %>%
  raster::extract(y = ptsamp1_1) %>%
  as.data.table %>%
  .[, id_cls := ptsamp1_1@data] # add the class names to each row

dt2 <- brick_for_prediction %>%
  raster::extract(y = ptsamp2_2) %>%
  as.data.table %>%
  .[, id_cls := ptsamp2_2@data] # add the class names to each row

dt3 <- brick_for_prediction %>%
  raster::extract(y = ptsamp3_3) %>%
  as.data.table %>%
  .[, id_cls := ptsamp3_3@data] # add the class names to each row

dt4 <- brick_for_prediction %>%
  raster::extract(y = ptsamp4_4) %>%
  as.data.table %>%
  .[, id_cls := ptsamp4_4@data] # add the class names to each row

dt5 <- brick_for_prediction %>%
  raster::extract(y = ptsamp5_5) %>%
  as.data.table %>%
  .[, id_cls := ptsamp5_5@data] # add the class names to each row


#merging the dataframes into one
dt<-rbind(dt1, dt2, dt3, dt4, dt5)
names(dt)[names(dt) == 'id_cls'] <- 'class'
dt<-dt %>% drop_na()
dt$class <- factor(dt$class, labels=c('forest','urban', 'mountain', 'vaia', 'pasture'))



#Random Forest algorithm
```

```
set.seed(321)
# A stratified random split of the data
idx_train <- createDataPartition(dt$class,
                                 p = 0.7, # percentage of data as
                                    ↪ training
                                 list = FALSE)


dt_train <- dt[idx_train]
dt_test <- dt[-idx_train]


# create cross-validation folds (splits the data into n random groups)
n_folds <- 10
set.seed(321)
folds <- createFolds(1:nrow(dt_train), k = n_folds)

#Explanding mtry interval to consider more bands
tuneGrid <- expand.grid(mtry = seq(2, 215, by = 10))

# Set the seed at each resampling iteration. Useful when running CV in
   ↪ parallel.
seeds <- vector(mode = "list", length = n_folds + 1)
for(i in 1:n_folds) seeds[[i]] <- sample.int(1000, nrow(tuneGrid))
seeds[n_folds + 1] <- sample.int(1000, 1)

ctrl <- trainControl(summaryFunction = multiClassSummary, method = "cv"
   ↪ , number = n_folds, search = "grid", classProbs = TRUE,
   ↪ savePredictions = TRUE, index = folds, seeds = seeds)

model_rf <- caret::train(class ~ . , method = "rf", data = dt_train,
   ↪ importance = TRUE,tuneGrid = tuneGrid, trControl = ctrl)

#saving the model
saveRDS(model_rf, file = paste0("path_to_file","model_rf_","area_B",".
   ↪ rds"))

predict_rf <- raster::predict(object = brick_for_prediction,
                              model = model_rf, type = 'raw')
writeRaster(predict_rf, paste0("path_to_file",".tiff"),overwrite=T )

####EVALUATION OF THE MODEL
plot(model_rf) # tuning results

#confusion matrix and statistics
test_e30B <- predict(model_rf, newdata = dt_test)
class_e30B <- dt_test$class
cm <- confusionMatrix(data = test_e30B, class_e30B)
cm
```

```
304
305
306 #ordering them by predictor importance across the classes
307 vi <- varImp(model_rf)$importance
308 vi$max <- apply(vi, 1, max)
309 vi <- vi[order(-vi$max),]
310 #selecting only the 20 most important bands
311 vi20 <- head(vi, 20)
312
313 vi20%>%
314      as.matrix %>%
315      plot_ly(x = colnames(.)[1:5], y = rownames(.), z = ., type = "
         ↪ heatmap",
316              width = 350, height = 300)
317
318 #mean decrease accuracy and mean decrease gini
319 randomForest::varImpPlot(model_rf$finalModel)
320
321
322
323
324 ####WORKING ON AREA C
325 rst_lst <- stack('path_to_file.tif')
326 rst_lst <- as.list(rst_lst)   #transforming rasterstack into list
327 names(rst_lst) <- 1:224
328
329 #dropping the columns with missing values
330 rst_lst <- rst_lst[-c(131:135)]
331
332 #Visualize the image in Natural Color (R = Red, G = Green, B = Blue).
333 suppressWarnings({viewRGB(brick(rst_lst[1:44]), r = 44, g = 21, b = 5)
    ↪ })
334
335 brick_for_prediction <- brick(rst_lst)
336
337 #importing the shp file of area C
338 poly_area_C <-shapefile('path_to_file.shp')
339 poly_area_C@data$id <- as.integer(factor(poly_area_C@data$id))
340 setDT(poly_area_C@data)
341
342 ptsamp1<-subset(poly_area_C, id == "1") #selecting the polygons with id
    ↪ =1
343 ptsamp1_1 <- spsample(ptsamp1, 750, type='regular') #selecting 750
    ↪ random points within the chosen polygons
344 ptsamp1_1$class <- over(ptsamp1_1, ptsamp1)$id #assigning the id to the
    ↪  random points
345 saveRDS(ptsamp1_1, file=paste0 ("path_to_file", file="_ptsamp1_C.rds"))
346
```

```
ptsamp2<-subset(poly_area_C, id == "2") #selecting the polygons with id =2
ptsamp2_2 <- spsample(ptsamp2, 750, type='regular')  #selecting 750 random points within the chosen polygons
ptsamp2_2$class <- over(ptsamp2_2, ptsamp2)$id #assigning the id to the random points
saveRDS(ptsamp2_2, file=paste0 ("path_to_file", file= "_ptsamp2_C.rds") )

ptsamp3<-subset(poly_area_C, id == "3") #selecting the polygons with id =3
ptsamp3_3 <- spsample(ptsamp3, 750, type='regular') #selecting 750 random points within the chosen polygons
ptsamp3_3$class <- over(ptsamp3_3, ptsamp3)$id #assigning the id to the random points
saveRDS(ptsamp3_3, file=paste0 ("path_to_file", file= "_ptsamp3_C.rds") )

ptsamp4<-subset(poly_area_C, id == "4") #selecting the polygons with id =4
ptsamp4_4 <- spsample(ptsamp4, 750, type='regular')  #selecting 750 random points within the chosen polygons
ptsamp4_4$class <- over(ptsamp4_4, ptsamp4)$id #assigning the id to the random points
saveRDS(ptsamp4_4, file=paste0 ("path_to_file", file= "_ptsamp4_C.rds") )

ptsamp5<-subset(poly_area_C, id == "5") #selecting the polygons with id =5
ptsamp5_5 <- spsample(ptsamp5, 750, type='regular') #selecting 750 random points within the chosen polygons
ptsamp5_5$class <- over(ptsamp5_5, ptsamp5)$id  #assigning the id to the random points
saveRDS(ptsamp5_5, file=paste0 ("path_to_file", file="_ptsamp5_C.rds"))


#saving the information of the random points into a dataframe
dt1 <- brick_for_prediction %>%
  raster::extract(y = ptsamp1_1) %>%
  as.data.table %>%
  .[, id_cls := ptsamp1_1@data] # add the class names to each row

dt2 <- brick_for_prediction %>%
  raster::extract(y = ptsamp2_2) %>%
  as.data.table %>%
  .[, id_cls := ptsamp2_2@data] # add the class names to each row

dt3 <- brick_for_prediction %>%
  raster::extract(y = ptsamp3_3) %>%
```

```
  as.data.table %>%
  .[, id_cls := ptsamp3_3@data] # add the class names to each row

dt4 <- brick_for_prediction %>%
  raster::extract(y = ptsamp4_4) %>%
  as.data.table %>%
  .[, id_cls := ptsamp4_4@data] # add the class names to each row

dt5 <- brick_for_prediction %>%
  raster::extract(y = ptsamp5_5) %>%
  as.data.table %>%
  .[, id_cls := ptsamp5_5@data] # add the class names to each row


#merging the dataframes into a single dataframe
dt<-rbind(dt1, dt2, dt3, dt4, dt5)
names(dt)[names(dt) == 'id_cls'] <- 'class'
dt<-dt %>% drop_na()
dt$class <- factor(dt$class, labels=c('forest','urban', 'mountain', '
   ↪ vaia', 'pasture'))



#random forest algorithm
set.seed(321)
# A stratified random split of the data
idx_train <- createDataPartition(dt$class,
                                 p = 0.7, # percentage of data as
                                    ↪ training
                                 list = FALSE)


dt_train <- dt[idx_train]
dt_test <- dt[-idx_train]


# create cross-validation folds (splits the data into n random groups)
n_folds <- 10
set.seed(321)
folds <- createFolds(1:nrow(dt_train), k = n_folds)

#Explanding mtry interval to consider more bands
tuneGrid <- expand.grid(mtry = seq(2, 215, by = 10))

# Set the seed at each resampling iteration. Useful when running CV in
   ↪ parallel.
seeds <- vector(mode = "list", length = n_folds + 1)
for(i in 1:n_folds) seeds[[i]] <- sample.int(1000, nrow(tuneGrid))
seeds[n_folds + 1] <- sample.int(1000, 1)
```

```
427
428 ctrl <- trainControl(summaryFunction = multiClassSummary, method = "cv"
    ↪ , number = n_folds, search = "grid", classProbs = TRUE,
    ↪ savePredictions = TRUE, index = folds, seeds = seeds)
429
430 model_rf <- caret::train(class ~ . , method = "rf", data = dt_train,
    ↪ importance = TRUE, tuneGrid = tuneGrid, trControl = ctrl)
431
432 #saving the model
433 saveRDS(model_rf, file = paste0("path_to_file","model_rf_","area_C",".
    ↪ rds"))
434
435 predict_rf <- raster::predict(object = brick_for_prediction,
436                               model = model_rf, type = 'raw')
437 writeRaster(predict_rf, paste0("path_to_file",".tiff"),overwrite=T )
438
439 ####EVALUATION OF THE MODEL
440 plot(model_rf) # tuning results
441
442 #confusion matrix and statistics
443 test_e30C <- predict(model_rf, newdata = dt_test)
444 class_e30C <- dt_test$class
445 cm <- confusionMatrix(data = test_e30C, class_e30C)
446 cm
447
448
449 #ordering them by predictor importance across the classes
450 vi <- varImp(model_rf)$importance
451 vi$max <- apply(vi, 1, max)
452 vi <- vi[order(-vi$max),]
453 #selecting only the 20 most important bands
454 vi20 <- head(vi, 20)
455
456 vi20%>%
457      as.matrix %>%
458      plot_ly(x = colnames(.)[1:5], y = rownames(.), z = ., type = "
        ↪ heatmap",
459              width = 350, height = 300)
460
461 #mean decrease accuracy and mean decrease gini
462 randomForest::varImpPlot(model_rf$finalModel)
463
464
465
466 ####WORKING ON AREA D
467 rst_lst <- stack('path_to_file.tif')
468 rst_lst <- as.list(rst_lst)   #transforming rasterstack into list
469 names(rst_lst) <- 1:224
470
```

```
#dropping the columns with missing values
rst_lst <- rst_lst[-c(131:135)]

#Visualize the image in Natural Color (R = Red, G = Green, B = Blue).
suppressWarnings({viewRGB(brick(rst_lst[1:44]), r = 44, g = 21, b = 5) })

brick_for_prediction <- brick(rst_lst)

#importing the shp file of area D
poly_area_D <-shapefile('path_to_file.shp')
poly_area_D@data$id <- as.integer(factor(poly_area_D@data$id))
setDT(poly_area_D@data)

ptsamp1<-subset(poly_area_D, id == "1") #selecting the polygons with id =1
ptsamp1_1 <- spsample(ptsamp1, 750, type='regular') # selecting 750 random points within the polygons with id=1
ptsamp1_1$class <- over(ptsamp1_1, ptsamp1)$id #giving the value id=1 to the random points
saveRDS(ptsamp1_1, file=paste0 ("path_to_file", file="_ptsamp1_D.rds"))

ptsamp2<-subset(poly_area_D, id == "2") #selecting polygons with id=2
ptsamp2_2 <- spsample(ptsamp2, 750, type='regular') # selecting 750 random points within the polygons with id=2
ptsamp2_2$class <- over(ptsamp2_2, ptsamp2)$id #giving the value id=2 to the random points
saveRDS(ptsamp2_2, file=paste0 ("path_to_file", file= "_ptsamp2_D.rds") )

ptsamp3<-subset(poly_area_D, id == "3") #selecting polygons with id=3
ptsamp3_3 <- spsample(ptsamp3, 750, type='regular') # selecting 750 random points within the polygons with id=3
ptsamp3_3$class <- over(ptsamp3_3, ptsamp3)$id #giving the value id=3 to the random points
saveRDS(ptsamp3_3, file=paste0 ("path_to_file", file= "_ptsamp3_D.rds") )

ptsamp4<-subset(poly_area_D, id == "4") #selecting polygons with id=4
ptsamp4_4 <- spsample(ptsamp4, 750, type='regular') # selecting 750 random points within the polygons with id==4
ptsamp4_4$class <- over(ptsamp4_4, ptsamp4)$id #giving the value id=4 to the random points
saveRDS(ptsamp4_4, file=paste0 ("path_to_file", file= "_ptsamp4_D.rds") )

ptsamp5<-subset(poly_area_D, id == "5") #selecting polygons with id=5
ptsamp5_5 <- spsample(ptsamp5, 750, type='regular') # selecting 750 random points within the polygons with id==5
```

```
ptsamp5_5$class <- over(ptsamp5_5, ptsamp5)$id #giving the value id=5 to the random points
saveRDS(ptsamp5_5, file=paste0 ("path_to_file", file="_ptsamp5_D.rds"))


dt1 <- brick_for_prediction %>%
  raster::extract(y = ptsamp1_1) %>%
  as.data.table %>%
  .[, id_cls := ptsamp1_1@data] # add the class names to each row

dt2 <- brick_for_prediction %>%
  raster::extract(y = ptsamp2_2) %>%
  as.data.table %>%
  .[, id_cls := ptsamp2_2@data] # add the class names to each row

dt3 <- brick_for_prediction %>%
  raster::extract(y = ptsamp3_3) %>%
  as.data.table %>%
  .[, id_cls := ptsamp3_3@data] # add the class names to each row

dt4 <- brick_for_prediction %>%
  raster::extract(y = ptsamp4_4) %>%
  as.data.table %>%
  .[, id_cls := ptsamp4_4@data] # add the class names to each row

dt5 <- brick_for_prediction %>%
  raster::extract(y = ptsamp5_5) %>%
  as.data.table %>%
  .[, id_cls := ptsamp5_5@data] # add the class names to each row


dt<-rbind(dt1, dt2, dt3, dt4, dt5)
names(dt)[names(dt) == 'id_cls'] <- 'class'
dt<-dt %>% drop_na()
dt$class <- factor(dt$class, labels=c('forest','urban', 'mountain', 'vaia', 'pasture'))



#random forest algorithm
set.seed(321)
# A stratified random split of the data
idx_train <- createDataPartition(dt$class,
                                 p = 0.7, # percentage of data as training
                                 list = FALSE)


dt_train <- dt[idx_train]
```

```r
dt_test <- dt[-idx_train]


# create cross-validation folds (splits the data into n random groups)
n_folds <- 10
set.seed(321)
folds <- createFolds(1:nrow(dt_train), k = n_folds)

#Explanding mtry interval to consider more bands
tuneGrid <- expand.grid(mtry = seq(2, 215, by = 10))

# Set the seed at each resampling iteration. Useful when running CV in parallel.
seeds <- vector(mode = "list", length = n_folds + 1)
for(i in 1:n_folds) seeds[[i]] <- sample.int(1000, nrow(tuneGrid))
seeds[n_folds + 1] <- sample.int(1000, 1)


ctrl <- trainControl(summaryFunction = multiClassSummary, method = "cv", number = n_folds, search = "grid", classProbs = TRUE, savePredictions = TRUE, index = folds, seeds = seeds)

model_rf <- caret::train(class ~ . , method = "rf", data = dt_train, importance = TRUE, tuneGrid =tuneGrid, trControl = ctrl)

#saving the model
saveRDS(model_rf, file = paste0("path_to_file","model_rf_","area_D",".rds"))

predict_rf <- raster::predict(object = brick_for_prediction,
                              model = model_rf, type = 'raw')
writeRaster(predict_rf, paste0("path_to_file",".tiff"),overwrite=T )

####EVALUATION OF THE MODEL
plot(model_rf) # tuning results


#confusion matrix and statistics
test_e30D <- predict(model_rf, newdata = dt_test)
class_e30D <- dt_test$class
cm <- confusionMatrix(data = test_e30D, class_e30D)
cm


#ordering them by predictor importance across the classes
vi <- varImp(model_rf)$importance
vi$max <- apply(vi, 1, max)
vi <- vi[order(-vi$max),]
#selecting only the 20 most important bands
```

```r
vi20 <- head(vi, 20)

vi20%>%
     as.matrix %>%
     plot_ly(x = colnames(.)[1:5], y = rownames(.), z = ., type = "heatmap",
             width = 350, height = 300)

#mean decrease accuracy and mean decrease gini
randomForest::varImpPlot(model_rf$finalModel)

#ENMAP 30m ACCURACY
test_e30 <- c(test_e30A, test_e30B, test_e30C, test_e30D)
class_e30 <- c(class_e30A, class_e30B, class_e30C, class_e30D)
cm_e30  <- confusionMatrix(data = test_e30, class_e30)
cm_e30
```

6.2 Appendix B: Additional R Code for SVH

Listing 6.2: Complete code for SVH on EnMAP images

```r
library(RStoolbox)
library(raster)
library(rasterdiv)
library(sp)
library(terra)
library(exactextractr)

#INITIALIZING CIRCLES
#Importing the areas where I want to calculate the pc
cricles <- shapefile('path_to_file.shp')

# Removing points that are not in forests or for which we don't have data
values_to_remove <- c("GS0004", "GS0005", "GS0006", "GS0007", "GS0008", "GS0016", "GS0017", "GS0018", "GS0019", "GS0020", "GS0021", "GS0022", "GS0029")

# Filter the shapefile to exclude these values
circles_filtered <- circles[!(circles@data[["PUNTO"]] %in% values_to_remove), ]

# Convert SpatialPolygonsDataFrame to SpatVector
circles_vect <- vect(circles_filtered)

# Initialize a list to store the centroids
centroids_list <- list()
```

```
23
24 species_richness <- c(2,2,3,4,2,3,2,4,3,4,5,3,2, 4,3,4,3,2)
25 shannon_indices <- c
   ↪ (0.487,0.181,0.556,1.22,0.362,0.66,0.358,1.35,0.381,
26 0.238,1.15,0.83,0.377,0.734,0.958,0.988,0.271,0.554)
27
28 #CALCULATING RAO INDEX ON ENMAP'S PC1
29 #importing Enmap image
30 suppressWarnings({
31 rst_lst <- stack('path_to_file')
32 })
33 cropped_area <-shapefile('path_to_file.shp')
34 enmap <-crop(rst_lst, extent(cropped_area))
35
36 names(enmap) <- as.character(1:224)
37
38 #dropping the layers with missing values and anomalous behaviour
39 enmap <- dropLayer(enmap, paste0("X", c(131:135, 80:102)))
40
41 #Transforming into a brick since we have many layers
42 enmap <- brick(enmap)
43
44 # Define a function to standardize a single layer
45 standardize_layer <- function(layer) {
46   values <- getValues(layer)
47   standardized_values <- scale(values, center = TRUE, scale = TRUE)
48   setValues(layer, standardized_values)
49 }
50
51 # Standardize each layer in the raster stack
52 standardized_layers <- stack(lapply(1:nlayers(enmap), function(i)
   ↪ standardize_layer(enmap[[i]])))
53
54 #performing the PCA
55 enmap_pca <- rasterPCA(standardized_layers)
56
57 #Plot the map of the first principal component (PC1)
58 plot(enmap_pca$map[[1]], main = "PC1")
59
60
61 #Calculating Rao index on the PC1
62 PC1_raster_layer <- enmap_pca$map[[1]]
63
64 # Initialize a vector to store Rao indices for each circle
65 enmap_rao_indices <- numeric(length(circles_vect))
66
67 # Iterate over each circle
68 for (i in seq_along(circles_filtered)) {
69   # Extract the polygon representing the current circle
```

```
  circle_polygon <- circles_filtered[i,]

  # Calculate centroid of the circle
  centroid <- centroids(circles_vect[i])
  centroids_list[[i]] <- centroid

  # Use exact_extract to get the values of PC1_raster_layer within the
    ↪ circle polygon
  extracted_values <- exact_extract(PC1_raster_layer, circle_polygon)

  mat_s <- unlist(lapply(extracted_values, function(x) x$value))
  mat_s <- mat_s[!is.na(mat_s)]

  # Rao index calculation
  n_s <- length(mat_s)
  n2_s <- n_s^2
  distm_s <- as.matrix(dist(mat_s))
  rao_index <- sum(distm_s) / n2_s

  # Store the Rao index in the vector
  enmap_rao_indices[i] <- rao_index
}

# Combine centroids into a data frame
centroids_df <- do.call(rbind, lapply(centroids_list, function(x) cbind
  ↪ (x[,1], x[,2])))
centroids_df <- as.data.frame(centroids_df)
colnames(centroids_df) <- c("Longitude", "Latitude")

# Create a combined data frame with centroids, Rao indices, Shannon
  ↪ indices, and number of trees
biodiversity_enmap <- data.frame(Longitude = centroids_df$Longitude,
  ↪ Latitude = centroids_df$Latitude, Rao_Index = enmap_rao_indices,
  ↪ Shannon_Index = shannon_indices, Species_Richness = species_
  ↪ richness)
# Save as CSV file
write.csv(biodiversity_enmap, file = "path_to_file.csv", row.names =
  ↪ FALSE)

#Calculating the R^2 value Shannon Index and Rao Index
lm_shannon <- lm(biodiversity_enmap$Shannon_Index ~ biodiversity_enmap$
  ↪ Rao_Index)
r2_shannon <- summary(lm_shannon)$r.squared

# Calculate R^2 Species Richness and Rao Index
lm_species <- lm(biodiversity_enmap$Species_Richness ~ biodiversity_
  ↪ enmap$Rao_Index)
r2_species <- summary(lm_species)$r.squared

```

```
#CALCULATING RAO INDEX ON ENMAP OPTICAL TRAITS
#importing Enmap image
rst_lst <- rast("path_to_file.bsq")
cropped_area <-shapefile('path_to_tile.shp')
opt_traits <-crop(rst_lst, extent(cropped_area))

#Transforming into a brick since we have many layers
opt_traits <- brick(opt_traits)


#Initialize the vector where I will store the Rao indices for each layer
rao_indices <- numeric(length(circles_vect))

for (k in opt_traits@data@names){
# Initialize a vector to store Rao indices for each optical trait
vector_rao_name <- paste0("rao_index_", k)

# Iterate over each circle
for (i in seq_along(circles_filtered)) {
  # Extract the polygon representing the current circle
  circle_polygon <- circles_filtered[i,]

  # Calculate centroid of the circle
  centroid <- centroids(circles_vect[i])
  centroids_list[[i]] <- centroid

  # Use exact_extract to get the values of the layer within the circle polygon
  extracted_values <- exact_extract(opt_traits[[k]], circle_polygon)

  mat_s <- unlist(lapply(extracted_values, function(x) x$value))
  mat_s <- mat_s[!is.na(mat_s)]

  # Rao index calculation
  n_s <- length(mat_s)
  n2_s <- n_s^2
  distm_s <- as.matrix(dist(mat_s))
  rao_index <- sum(distm_s) / n2_s

  # Store the Rao index in the vector
  enmap_rao_indices[i] <- rao_index
}
assign(vector_rao_name, enmap_rao_indices)

cat(vector_rao_name, "vector created\n")
}

# Combine centroids into a data frame
```

```
centroids_df <- do.call(rbind, lapply(centroids_list, function(x) cbind
    ↪ (x[,1], x[,2])))
centroids_df <- as.data.frame(centroids_df)
colnames(centroids_df) <- c("Longitude", "Latitude")

# Create a combined data frame with centroids, Rao indices (only
    ↪ keeping the ones that are statistically significant), Shannon
    ↪ indices, and number of trees
biodiversity_enmap_opt_traits <- data.frame(Longitude = centroids_df$
    ↪ Longitude,
Latitude = centroids_df$Latitude, rao_index_car=rao_index_car, rao_
    ↪ index_anth = rao_index_anth, rao_index_N = rao_index_N, rao_index
    ↪ _cab=rao_index_cab, rao_index_cbrown=rao_index_cbrown, rao_index_
    ↪ cm=rao_index_cm, rao_index_LAI=rao_index_LAI, rao_index_cw=rao_
    ↪ index_cw,  rao_index_LIDF=rao_index_LIDF, rao_index_hspot=rao_
    ↪ index_hspot,               Shannon_Index = shannon_indices,
    ↪ Species_Richness = species_richness)

# Save as CSV file
write.csv(biodiversity_enmap_opt_traits, file = "path_to_file.csv", row
    ↪ .names = FALSE)

#Calculating the R^2 value
for (k in 3:12){
# Calculate R^2 between Shannon Index and Rao Index
lm_shannon <- lm(biodiversity_enmap_opt_traits[[13]] ~ biodiversity_
    ↪ enmap_opt_traits[[k]])
r2_shannon <- summary(lm_shannon)$r.squared
R2_shannonVSrao_name <- paste0("R2_Shannon_VS_", colnames(biodiversity_
    ↪ enmap_opt_traits)[k])
assign(R2_shannonVSrao_name, r2_shannon)

# Calculate R^2 Species Richness and Rao Index
lm_species <- lm(biodiversity_enmap_opt_traits[[14]] ~ biodiversity_
    ↪ enmap_opt_traits[[k]])
r2_species <- summary(lm_species)$r.squared
R2_speciesVSrao_name <- paste0("R2_Species_VS_", colnames(biodiversity_
    ↪ enmap_opt_traits)[k])
assign(R2_speciesVSrao_name, r2_species)
}
```

Bibliography

A. Agresti and B. F. Agresti. Statistical analysis of qualitative variation. *Sociological methodology*, 9:204–237, 1978.

A. Banskota, N. Kayastha, M. J. Falkowski, M. A. Wulder, R. E. Froese, and J. C. White. Forest monitoring using landsat time series data: A review. *Canadian Journal of Remote Sensing*, 40(5):362–384, 2014.

G. Biau and E. Scornet. A random forest guided tour. *Test*, 25:197–227, 2016.

L. Breiman. Bagging predictors. *Machine learning*, 24:123–140, 1996.

L. Breiman. Random forests. *Machine learning*, 45:5–32, 2001.

F. Castaldi, S. Chabrillat, and B. van Wesemael. Sampling strategies for soil property mapping using multispectral sentinel-2 and hyperspectral enmap satellite data. *Remote Sensing*, 11(3):309, 2019.

E. Chraibi, H. Arnold, S. Luque, A. Deacon, A. E. Magurran, and J.-B. Féret. A remote sensing approach to understanding patterns of secondary succession in tropical forest. *Remote Sensing*, 13(11):2148, 2021.

A. P. Cracknell. The development of remote sensing in the last 40 years. *International Journal of Remote Sensing*, 39(23):8387–8427, 2018. doi: 10.1080/01431161.2018.1550919. URL `https://doi.org/10.1080/01431161.2018.1550919`.

D. R. Cutler, T. C. Edwards Jr, K. H. Beard, A. Cutler, K. T. Hess, J. Gibson, and J. J. Lawler. Random forests for classification in ecology. *Ecology*, 88(11):2783–2792, 2007.

R. Díaz-Uriarte and S. Alvarez de Andrés. Gene selection and classification of microarray data using random forest. *BMC bioinformatics*, 7:1–13, 2006.

S. Dovgyi, V. Lialko, S. Babiichuk, T. Kuchma, O. Tomchenko, and L. Iurkiv. Fundamentals of remote sensing: history and practice. 2019.

Earth Observation Portal. Enmap (environmental mapping and analysis program). Retrieved from https://www.eoportal.org/satellite-missions/enmap#hsi-hyperspectral-imager.

European Space Agency. Sentinel-2 overview. Retrieved from https://sentinel.esa.int/web/sentinel/missions/sentinel-2.

Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

J. A. Gamon, R. Wang, H. Gholizadeh, B. Zutta, P. A. Townsend, and J. Cavender-Bares. Consideration of scale in remote sensing of biodiversity. *Remote sensing of plant biodiversity*, pages 425–447, 2020.

P. Geurts, D. Ernst, and L. Wehenkel. Extremely randomized trees. *Machine learning*, 63:3–42, 2006.

C. Gini. Variabiliti e mutabiliti. studi economicoaguridici della facotta di giurisprudenza dell, 1912.

W. Gould. Remote sensing of vegetation, plant species richness, and regional biodiversity hotspots. *Ecological applications*, 10(6):1861–1870, 2000.

K. Herkül, J. Kotta, T. Kutser, and E. Vahtmäe. Relating remotely sensed optical variability to marine benthic biodiversity. *PLoS One*, 8(2):e55624, 2013.

T. K. Ho. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE, 1995.

T. K. Ho. The random subspace method for constructing decision forests. *IEEE transactions on pattern analysis and machine intelligence*, 20(8):832–844, 1998.

J. Jackson, C. S. Lawson, C. Adelmant, E. Huhtala, P. Fernandes, R. Hodgson, H. King, L. Williamson, K. Maseyk, N. Hawes, A. Hector, and R. Salguero-Gómez. Short-range multispectral imaging is an inexpensive, fast, and accurate approach to estimate biodiversity in a temperate calcareous grassland. *Ecology and Evolution*, 12(12):e9623, 2022. doi: https://doi.org/10.1002/ece3.9623. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/ece3.9623. e9623 ECE-2022-09-01372.

A. Jarocińska, D. Kopeć, J. Niedzielko, J. Wylazłowska, A. Halladin-Dabrowska, J. Charyton, A. Piernik, and D. Kamiński. The utility of airborne hyperspectral and satellite multispectral images in identifying natura 2000 non-forest habitats for conservation purposes. *Scientific Reports*, 13(1):4549, 2023.

S. Karlin, R. Kenett, and B. Bonné-Tamir. Analysis of biochemical genetic data on jewish populations: Ii. results and interpretations of heterogeneity indices and distance measures with respect to standards. *American journal of human genetics*, 31(3):341, 1979.

P. J. Leitão, M. Schwieder, S. Suess, A. Okujeni, L. S. Galvão, S. v. d. Linden, and P. Hostert. Monitoring natural ecosystem and ecological gradients: perspectives with enmap. *Remote Sensing*, 7(10):13098–13119, 2015.

A. Liaw, M. Wiener, et al. Classification and regression by randomforest. *R news*, 2(3): 18–22, 2002.

G. Louppe. Understanding random forests: From theory to practice. *arXiv preprint arXiv:1407.7502*, 2014.

M. Malavasi, M. Bazzichetto, J. Komárek, V. Moudrý, D. Rocchini, S. Bagella, A. T. R. Acosta, and M. L. Carranza. Unmanned aerial systems-based monitoring of the eco-geomorphology of coastal dunes through spectral rao's q. *Applied Vegetation Science*, 24(1):e12567, 2021. doi: https://doi.org/10.1111/avsc.12567. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/avsc.12567`.

M. Mehmood, A. Shahzad, B. Zafar, A. Shabbir, and N. Ali. Remote sensing image classification: A comprehensive review and applications. *Mathematical Problems in Engineering*, 2022(1):5880959, 2022.

N. Morton and J. Lalouel. Topology of kinship in micronesia. *American Journal of Human Genetics*, 25(4):422, 1973.

V. Nasiri, A. A. Darvishsefat, H. Arefi, V. C. Griess, S. M. M. Sadeghi, and S. A. Borz. Modeling forest canopy cover: a synergistic use of sentinel-2, aerial photogrammetry data, and machine learning. *Remote Sensing*, 14(6):1453, 2022.

M. Nei. The theory of genetic distance and evolution of human races. *Japanese Journal of Human Genetics*, 23(4):341–369, 1978.

B. Oindo and A. Skidmore. Interannual variability of ndvi and species richness in kenya. *International journal of remote sensing*, 23(2):285–298, 2002.

A. Opoku. Biodiversity and the built environment: Implications for the sustainable development goals (sdgs). *Resources, Conservation and Recycling*, 141:1–7, 2019. ISSN 0921-3449. doi: https://doi.org/10.1016/j.resconrec.2018.10.011. URL `https://www.sciencedirect.com/science/article/pii/S0921344918303768`.

M. W. Palmer, P. G. Earls, B. W. Hoagland, P. S. White, and T. Wohlgemuth. Quantitative tools for perfecting species lists. *Environmetrics*, 13(2):121–137, 2002.

F. Parisi, E. Vangi, S. Francini, G. D'Amico, G. Chirici, M. Marchetti, F. Lombardi, D. Travaglini, S. Ravera, E. De Santis, et al. Sentinel-2 time series analysis for monitoring multi-taxon biodiversity in mountain beech forests. *Frontiers in Forests and Global Change*, 6:1020477, 2023.

C. R. Rao. The utilization of multiple measurements in problems of biological classification. *Journal of the Royal Statistical Society. Series B (Methodological)*, 10(2): 159–203, 1948.

C. R. Rao. Advanced statistical methods in biometric research. 1971a.

C. R. Rao. Taxonomy in anthropology. *Mathematics in The Archeological and Historical Sciences*, 1971b.

C. R. Rao. Cluster analysis applied to a study of race mixture in human populations. In *Classification and clustering*, pages 175–197. Elsevier, 1977.

C. R. Rao. Diversity and dissimilarity coefficients: a unified approach. *Theoretical population biology*, 21(1):24–43, 1982.

C. S. Reddy. Remote sensing of biodiversity: what to measure and monitor from space to species? *Biodiversity and Conservation*, 30(10):2617–2631, 2021.

D. Rocchini and M. Neteler. Let the four freedoms paradigm apply to ecology. *Trends in Ecology Evolution*, 27(6):310–311, 2012. ISSN 0169-5347. doi: https://doi.org/10.1016/j.tree.2012.03.009. URL `https://www.sciencedirect.com/science/article/pii/S0169534712000742`.

D. Rocchini, N. Balkenhol, G. A. Carter, G. M. Foody, T. W. Gillespie, K. S. He, S. Kark, N. Levin, K. Lucas, M. Luoto, H. Nagendra, J. Oldeland, C. Ricotta, J. Southworth, and M. Neteler. Remotely sensed spectral heterogeneity as a proxy of species diversity: Recent advances and open challenges. *Ecological Informatics*, 5(5):318–329, 2010. ISSN 1574-9541. doi: https://doi.org/10.1016/j.ecoinf.2010.06.001. URL `https://www.sciencedirect.com/science/article/pii/S1574954110000646`. Special Issue on Advances of Ecological Remote Sensing Under Global Change.

D. Rocchini, M. Marcantonio, and C. Ricotta. Measuring rao's q diversity index from remote sensing: An open source solution. *Ecological indicators*, 72:234–238, 2017.

C. Rossi, M. Kneubühler, M. Schütz, M. E. Schaepman, R. M. Haller, and A. C. Risch. Remote sensing of spectral diversity: A new methodological approach to account for spatio-temporal dissimilarities between plant communities. *Ecological Indicators*, 130: 108106, 2021.

C. Rossi, M. Kneubühler, M. Schütz, M. E. Schaepman, R. M. Haller, and A. C. Risch. Spatial resolution, spectral metrics and biomass are key aspects in estimating plant species richness from spectral diversity in species-rich grasslands. *Remote Sensing in Ecology and Conservation*, 8(3):297–314, 2022. doi: https://doi.org/10.1002/rse2.244. URL `https://zslpublications.onlinelibrary.wiley.com/doi/abs/10.1002/rse2.244`.

A. Sen. *On economic inequality*. Oxford university press (first edition 1973), 1997.

R. R. Sokal and P. H. Sneath. Principles of numerical taxonomy. 1963.

M. Torresani, D. Rocchini, R. Sonnenschein, M. Zebisch, M. Marcantonio, C. Ricotta, and G. Tonon. Estimating tree species diversity from space in an alpine conifer forest: The rao's q diversity index meets the spectral variation hypothesis. *Ecological Informatics*, 52:26–34, 2019. ISSN 1574-9541. doi: https://doi.org/10.1016/j.ecoinf.2019.04.001. URL `https://www.sciencedirect.com/science/article/pii/S1574954119300561`.

M. Torresani, D. Rocchini, G. Ceola, J. P. R. de Vries, H. Feilhauer, V. Moudrỳ, H. Bartholomeus, M. Perrone, M. Anderle, H. A. Gamper, et al. Grassland vertical height heterogeneity predicts flower and bee diversity: an uav photogrammetric approach. *Scientific Reports*, 14(1):809, 2024a.

M. Torresani, C. Rossi, M. Perrone, L. T. Hauser, J.-B. Féret, V. Moudrý, P. Simova, C. Ricotta, G. M. Foody, P. Kacic, H. Feilhauer, M. Malavasi, R. Tognetti, and D. Rocchini. Reviewing the spectral variation hypothesis: Twenty years in the tumultuous sea of biodiversity estimation by remote sensing. *Ecological Informatics*, 82:102702, 2024b. ISSN 1574-9541. doi: https://doi.org/10.1016/j.ecoinf.2024.102702. URL `https://www.sciencedirect.com/science/article/pii/S1574954124002449`.

United States Geological Survey (USGS). Landsat 8. Retrieved from `https://landsat.gsfc.nasa.gov/landsat-8/`.

X. Yu, D. Lu, X. Jiang, G. Li, Y. Chen, D. Li, and E. Chen. Examining the roles of spectral, spatial, and topographic features in improving land-cover and forest classifications in a subtropical region. *Remote Sensing*, 12(18):2907, 2020.