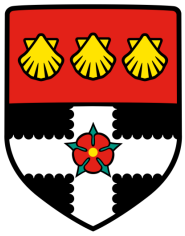

University of Reading

Department of Computer Science

Generating Tree Species Classification and Change Maps to Assist Mitigate Climate Change

Pedro Junio

Supervisor: Muhammad Shahzad

A report submitted in partial fulfilment of the requirements of
the University of Reading for the degree of
Master of Science in *Data Science and Advanced Computing*

August 19, 2024

Declaration

I, Pedro Junio, of the Department of Computer Science, University of Reading, confirm that this is my own work and figures, tables, equations, code snippets, artworks, and illustrations in this report are original and have not been taken from any other person's work, except where the works of others have been explicitly acknowledged, quoted, and referenced. I understand that if failing to do so will be considered a case of plagiarism. Plagiarism is a form of academic misconduct and will be penalised accordingly.

I give consent to a copy of my report being shared with future students as an exemplar.

I give consent for my work to be made available more widely to members of UoR and public with interest in teaching, learning and research.

Pedro Junio
August 19, 2024

Abstract

TODO update at the end

Variety of tree species are crucial in reducing the vulnerabilities and offering stable ecosystem functioning. The precise quantification and assessment of existing tree species on global-scale is therefore essential in filling the science-policy gaps by providing key insights essential in promoting the success of natural climate solutions and devising effective climate mitigation policies. To this end, this study aims to develop novel deep learning based algorithms by using the multi-temporal multi-spectral imagery to generate large-scale forest/tree species classification maps.

Keywords: Convolutional Neural Network, Sentinel-2, Copernicus, SoilGrid, EU-Forest, Forests, Tree Genera, Climate

Report's total word count: over 9000

Contents

List of Figures

List of Tables

Chapter 1

Data Exploration

1.1 EU-Forest Labels

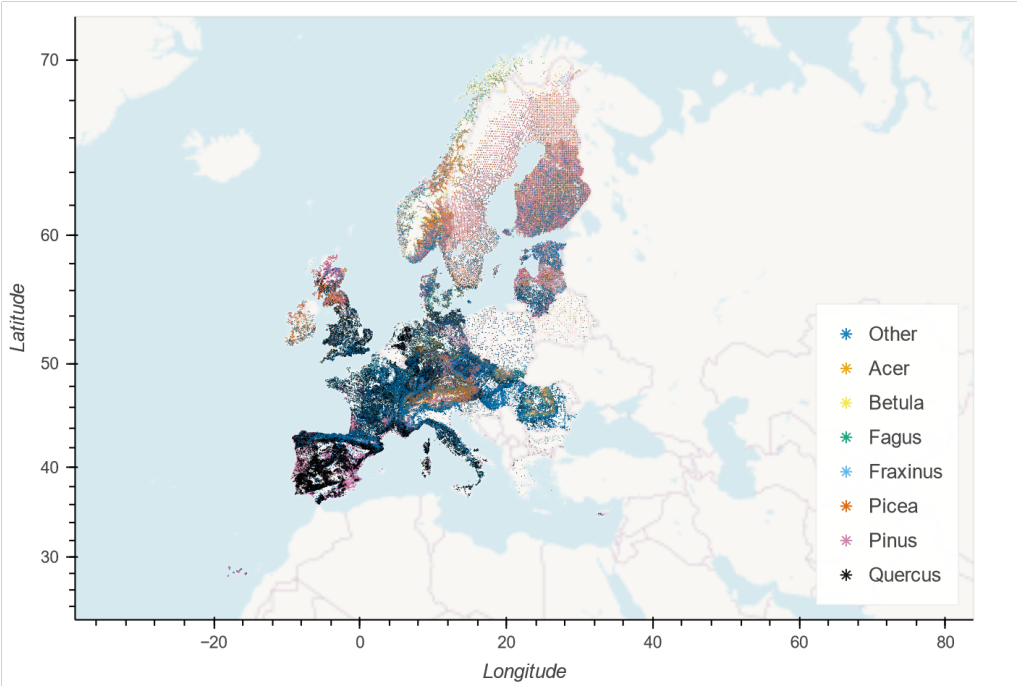

Figure 1.1: Map of the most common tree genera in EU-Forest.

EU-Forest is a dataset containing tree species and genera for nearly $250,000$ locations across Europe Mauri et al. (2017). Each plot is 1 km $\times$ 1 km and may contain multiple tree species and genera. Fig. 1.1 shows the distribution of tree genera in the EU-Forest dataset across 21 European countries. In this figure, the label 'Other' is an umbrella class for 70 tree genera with less than 20,000 occurrences each.

Using the EU-Forest dataset to train a CNN classifier of tree genera with Sentinel-2 data offers several significant advantages. Firstly, the dataset's high spatial resolution enables fine-grained analysis of tree species distribution, which enhances the accuracy of the classifier. Its comprehensive coverage across Europe, including diverse forest types and geographical areas, allows the model to learn from a wide variety of environments and tree genera. The rich occurrence data provides detailed information on tree species, aiding in precise identification and classification.

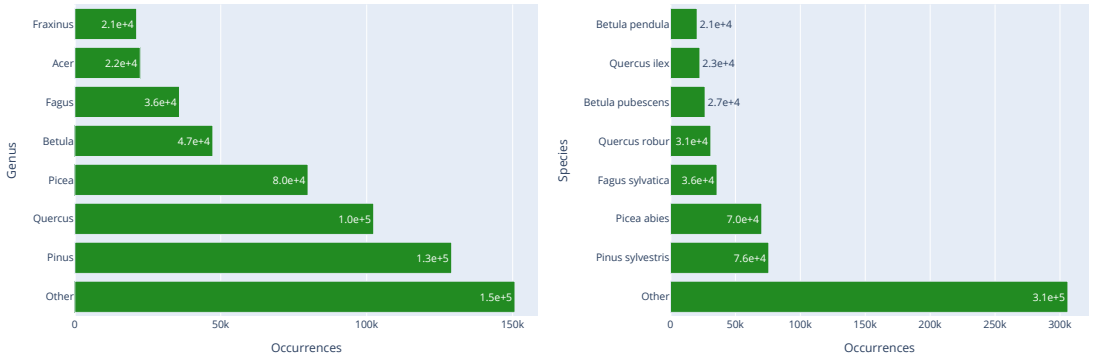

Figure 1.2: Distribution of genera (left) and species (right) in EU-Forest.

Integration with Sentinel-2 satellite data, which provides high-resolution multispectral images, allows for a robust model that leverages both ground-truth data and spectral information. The dataset, being relatively recent, offers a contemporary snapshot of forest conditions, ensuring that the trained model is relevant to current ecological and environmental conditions.

The plots in Fig. 1.2 underscore the prevalence of certain genera and species in European forests, providing valuable insights for training a CNN classifier. The dominance of specific genera and species in the dataset can enhance the classifier's ability to accurately identify and classify tree types when combined with Sentinel-2 data.

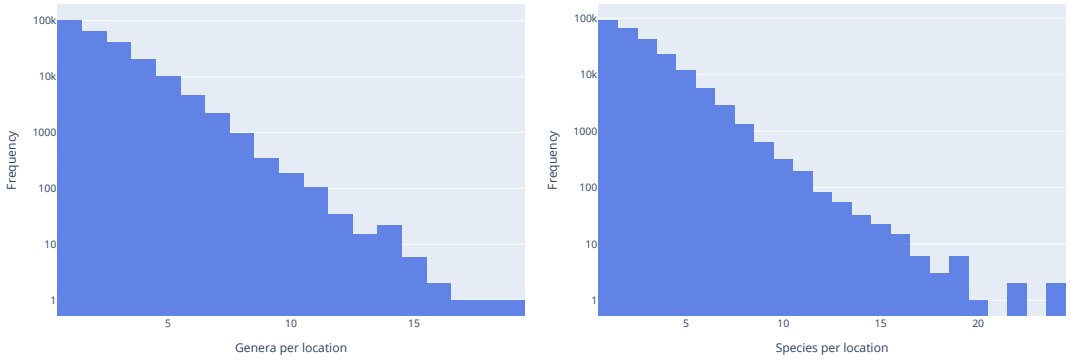

Figure 1.3: Distribution of genera (left) and species (right) per location.

The plots in Fig. 1.3 reveal a common pattern in biodiversity studies: most locations are characterized by a limited number of dominant genera and species, with a smaller number of locations exhibiting higher diversity. This pattern is important for training a CNN classifier, as it indicates that the classifier will often encounter locations with limited genera and species. However, it must also be capable of handling the less common, more diverse locations. The high-frequency, low-diversity areas will likely dominate the training process, influencing the classifier's ability to generalise across different forest types.

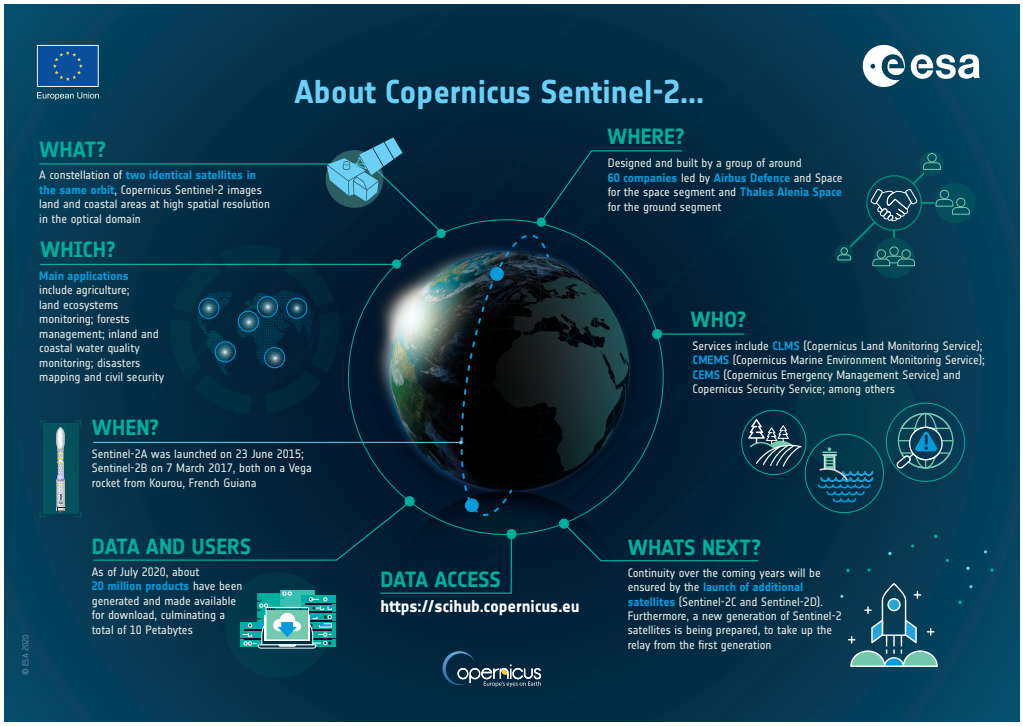

Figure 1.4: Sentinel-2 mission infographic. It highlights important facts and achievements of the mission. Courtesy of ESA.

1.2 Sentinel-2 Features

Using Sentinel-2, Fig. 1.4, specifically the 10-meter and 20-meter resolution bands, for training a CNN classifier of tree genera offers several significant advantages. Sentinel-2 provides high-resolution imagery with these bands capturing detailed spatial information essential for precise classification tasks.

The 10-meter resolution bands include visible red, green, and blue (RGB) wavelengths, as well as near-infrared (NIR) wavelengths, which are essential for assessing vegetation health and differentiating tree genera based on their reflectance properties. The 20-meter resolution bands encompass the red-edge, shortwave infrared (SWIR), and additional near-infrared regions, which enhance the classifier's ability to distinguish between tree genera by capturing subtle variations in spectral signatures. For subsequent analysis, the 20-meter bands were resampled to match the 10-meter resolution.

The multispectral imaging capability of Sentinel-2, with these selected bands, allows for detailed analysis and precise classification of tree genera. Each genus reflects and absorbs light differently across these wavelengths, providing rich data for the classifier to learn from and accurately identify tree types.

Moreover, Sentinel-2 has a frequent revisit time, with satellites passing over the same area every 5 days at the equator. This frequent update cycle is crucial for handling cloud cover, as it increases the likelihood of acquiring cloud-free images, ensuring that the classifier is trained on clear and usable data.

Sentinel-2 also offers extensive geographical coverage, capturing large areas in each image. This comprehensive coverage is essential for training classifiers intended for wide-ranging applications across different forest types and regions, and it supports the development of global models for tree genus classification.

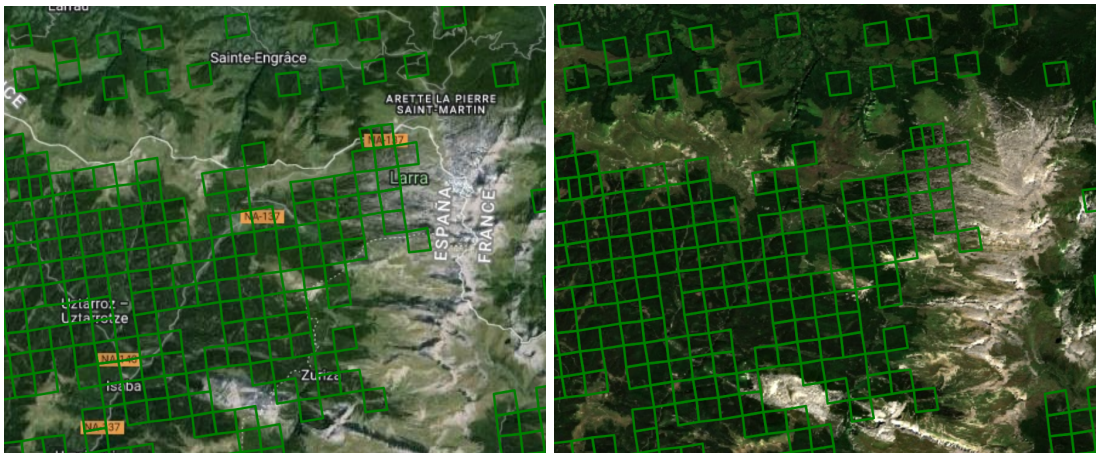

Figure 1.5: Multiple sample locations overlaid on Google Earth (left) and Sentinel-2 (right) images.

Additionally, Sentinel-2 data is freely available through the European Space Agency's Copernicus program and Google Earth Engine. This open access removes budget constraints, making high-quality satellite data accessible for various research and operational purposes.

Figure 1.6: Sample location overlaid on Google Earth (left) and Sentinel-2 (right) images.

Figs. 1.5 and 1.6 illustrate the integration of high-resolution geographical context with Sentinel-2 satellite data for detailed environmental analysis. The grid overlay in the images represents areas for data collection and analysis. The left images provide a more detailed geographical context, while the right images show how Sentinel-2 bands B2, B3, and B4 (blue, green, and red respectively) are structured and utilised for spectral analysis.

Figs. 1.7, 1.8, and 1.9 provide a detailed look at the distribution of surface reflectance values and their z-scores for various Sentinel-2 spectral bands, a normalisation method which is crucial for many classification tasks using CNNs. These figures use medians taken over the summer months (June, July, and August) between 2017, 2018, and 2019.

Fig. 1.7 shows that the reflectance values for the bands B2, B3, and B4 generally follow a log-normal distribution. The z-scores for these bands peak near zero, demonstrating that the reflectance values have been normalized effectively. This normalisation is essential for ensuring that the values from different bands are on the same scale, which is particularly important

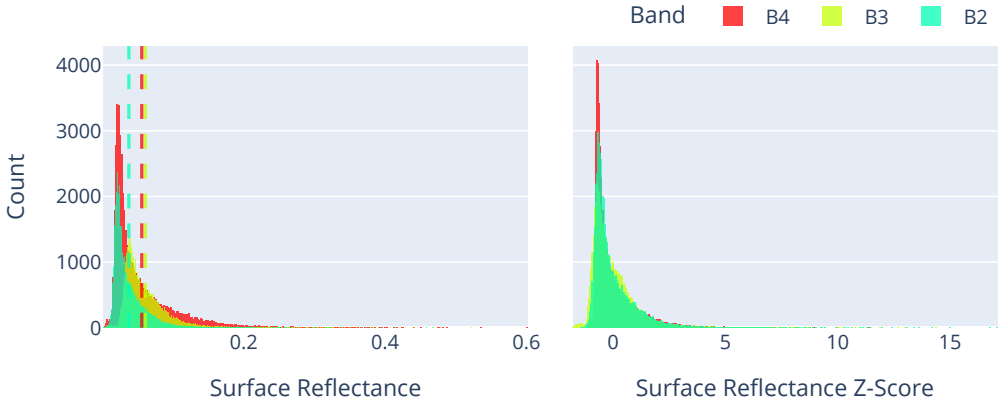

Figure 1.7: Distributions for surface reflectance (left), including a sample means as a dotted line, and z-score normalisation (right) for RGB.

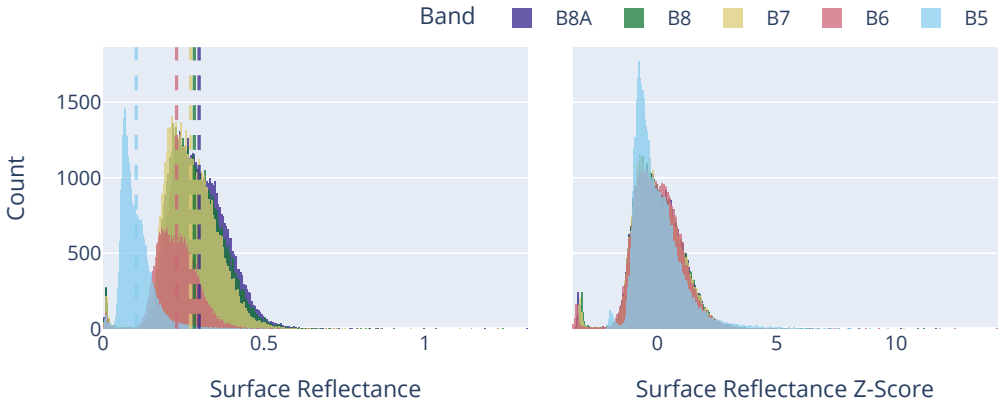

Figure 1.8: Distributions for surface reflectance (left), including a sample means as a dotted line, and z-score normalisation (right) for NIR.

when using CNN for classification.

Fig. 1.8 presents the distribution for the NIR bands B5, B6, B7, B8, and B8A. The reflectance values for these bands display a central tendency, with a significant number of pixels having values around the mean. The z-score distributions, peaking near zero, indicate successful normalisation of the reflectance values. Normalising the data in this way ensures that the CNN can process and compare these values more effectively, enhancing the model's ability to classify different types of tree genera accurately.

Fig. 1.9 shows histograms for the SWIR bands B11 and B12. The bands display a similar pattern, with the majority of reflectance values clustering on the left of the mean and tailing off to the right of the mean. The z-scores for these bands also peak at zero, confirming that the data has been normalised successfully.

The correlations among different spectral bands of Sentinel-2 data across seasons, as depicted in the correlation matrices in Fig. 1.10, highlight significant patterns that are useful for tree genus classification. Each figure shows the correlation coefficients between various bands, providing insights into how these relationships change with the seasons.

During winter, all bands in NIR and RGB groups show very strong correlations with each other. These high correlations, often close to 1, indicate that the spectral responses of these bands are highly similar during this season. This consistency is likely due to the uniform reflectance properties of vegetation and the ground cover in winter. This strong correlation

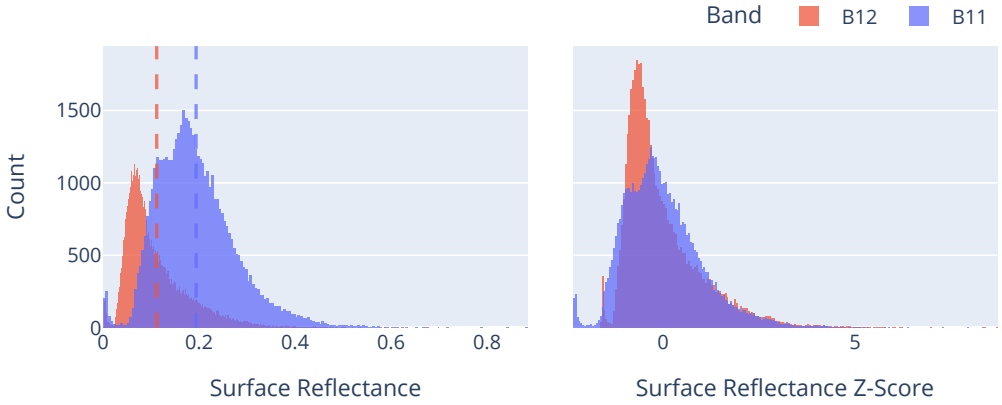

Figure 1.9: Distributions for surface reflectance (left), including a sample means as a dotted line, and z-score normalisation (right) for SWIR.

is beneficial for tree genus classification as it suggests that the data from these bands can be reliably used to distinguish tree genera based on their reflectance characteristics, or lack thereof, in winter.

In spring, the correlations remain high within the NIR and RGB groups but to a slightly lesser degree compared to winter. The correlation between these two groups starts to decrease, reflecting the changes in vegetation as new growth begins. This seasonal variation provides additional information that can be leveraged to improve the classification models, as the differences in reflectance between the bands become more pronounced.

During summer and autumn, while there are still strong correlations within the NIR and RGB groups, the correlation between these groups is notably lower. This reduced correlation can be attributed to the varying phenological stages of the trees, including differences in leaf development and moisture content. These seasonal changes affect the reflectance properties differently in the NIR and RGB bands. The distinct spectral responses during these seasons offer complementary information that can enhance the classification of tree genera by providing a more diverse set of data points that capture the variability in tree characteristics.

Overall, the varying correlations across seasons underscore the robustness of using Sentinel-2 data for tree genus classification. The normalisation of data using z-scores is crucial for bringing the values of different bands to the same scale, which is particularly important for CNNs. This normalisation ensures that the model can effectively process and compare the multi-band data, leading to more accurate classification results.

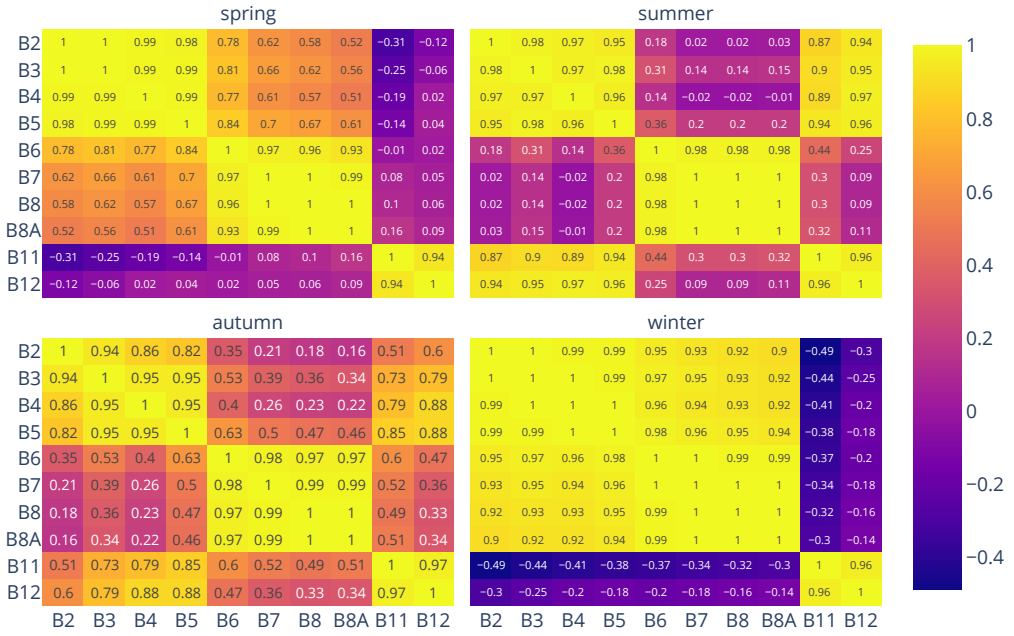

Figure 1.10: Seasonal Sentinel-2 band correlations.

1.3 Copernicus DEM GLO-30

Integrating the Copernicus Global 30m Digital Elevation Model (DEM) data into the classification model can significantly enhance its performance for tree genus classification. Elevation data provides critical information about the terrain, which influences various ecological factors such as temperature, moisture availability, and soil type. These factors, in turn, affect vegetation types and distribution. By incorporating elevation data, the model can better understand the environmental context of each location, leading to more accurate predictions. For example, certain tree genera may be more prevalent at specific elevation ranges due to their adaptation to particular climatic conditions or soil properties. Therefore, integrating elevation data can help in distinguishing between tree genera that occupy different ecological niches.

Fig. 1.11 presents the distribution of elevation values in meters, as well as the z-scores, which standardise these values. The elevation values range from 0 to around 2500 meters, with most data points clustered below 500 meters. This suggests that the majority of the study area is relatively low-lying. Standardising elevation values using z-scores is beneficial for integrating elevation data with spectral data, as it ensures that elevation values are on a comparable scale to the reflectance values from Sentinel-2 bands.

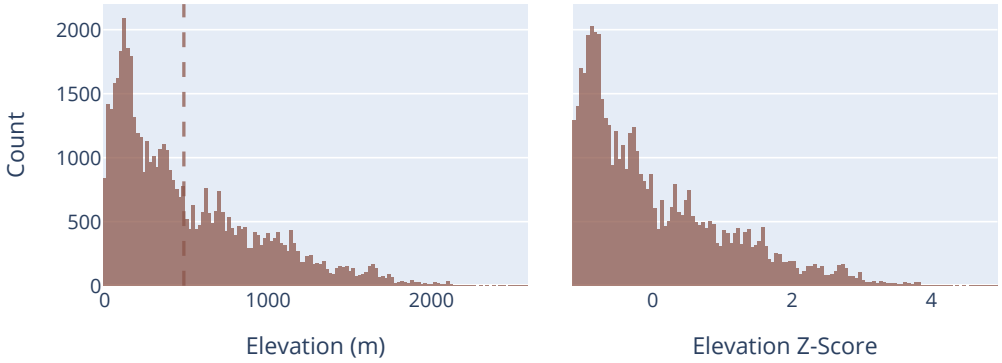

Figure 1.11: Distributions for elevation (left), including a sample means as a dotted line, and z-score normalisation (right) for the Copernicus DEM GLO-30 dataset.

1.4 SoilGrids

The SoilGrids dataset (Poggio et al. (2021)) provides global soil information at a spatial resolution of 250 meters and incorporates quantified spatial uncertainty. It includes key soil properties such as organic carbon content, pH, sand, silt, and clay percentages, bulk density, cation exchange capacity, and more. These data are derived from machine learning models trained on extensive soil sample databases and environmental covariates.

Integrating SoilGrids data into the tree genus classification model can significantly enhance its performance. Soil properties profoundly influence vegetation types and distribution, as different tree genera have specific soil requirements and preferences. For instance, soil pH, nutrient content, and texture can determine the suitability of a habitat for particular tree species. By incorporating detailed soil composition data, the model can better understand the environmental context, leading to more accurate predictions of tree genera. This integration helps to account for the ecological niche of each genus, improving the robustness and precision of the classification.

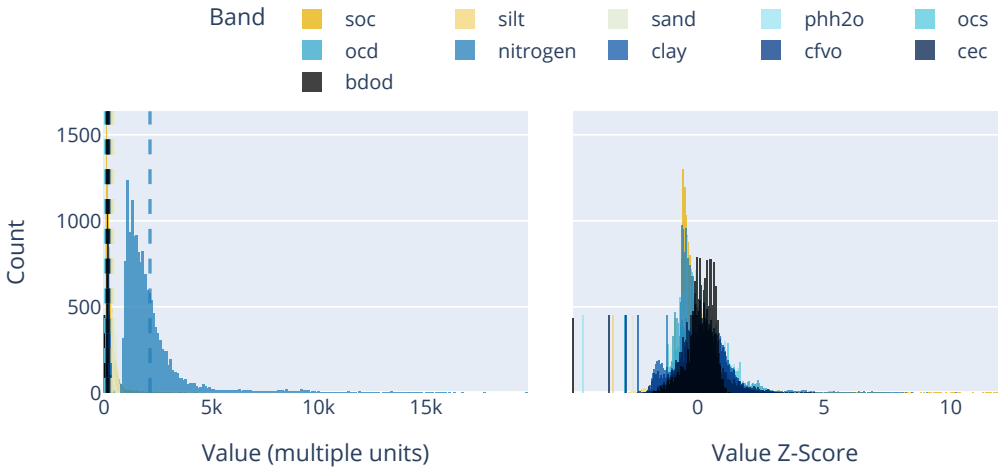

Figure 1.12: Distributions for elevation (left), including a sample means as a dotted line, and z-score normalisation (right) for the Copernicus DEM GLO-30 dataset.

Fig. 1.12 displays the distribution of various soil properties present in the SoilGrids dataset.

The x-axis represents the value of these properties in multiple units, while the y-axis represents the count of observations. Additionally, the z-score distribution is shown to standardise these values.

The distribution of soil properties varies, reflecting the diversity of soil types across the study area. The z-score distribution standardises these values, bringing them onto a common scale, which is essential for integrating soil data with spectral and elevation data in the classification model.

Chapter 2

Feature Selection

2.1 Sentinel-2 Seasons

Fig. 2.1 shows the performance of the CNN model across different seasons, with metrics including recall, precision, weighted f1-score, precision-recall curve area under the curve (PRC), and receiver operating characteristic area under the curve (AUC). Filled boxes indicate the combination of seasons used to train and validate the model. Blue boxes indicate a combination of 3D CNN and fully-connected layers and the orange box indicates the use of a similar model but with the introduction Long Short-Term Memory (LSTM) alongside 3D convolutions and fully-connected layers.

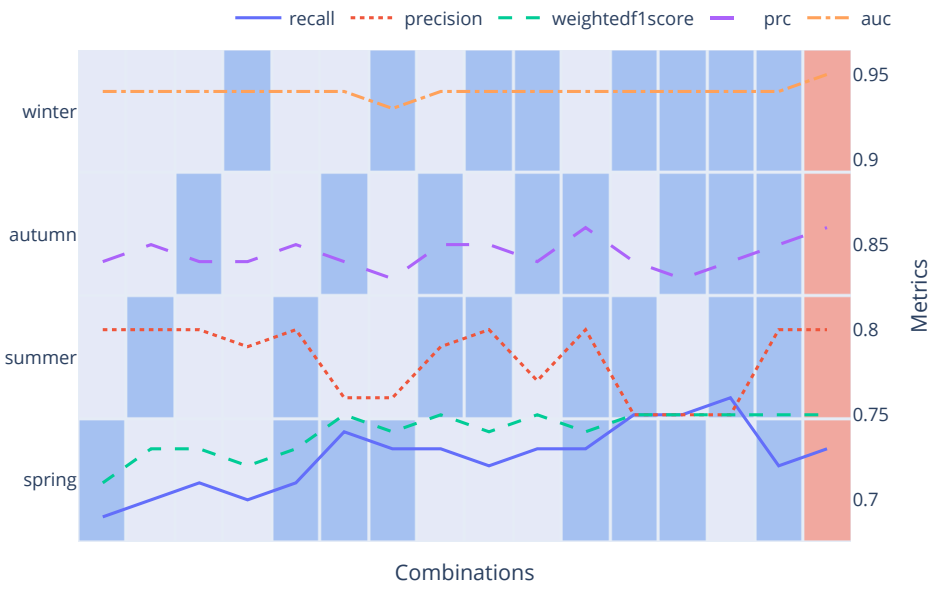

Figure 2.1: Seasonal Sentinel-2 analysis for all season combinations using a 3D CNN.

3D convolutions were selected for this model due to their suitability for this problem as they can effectively capture the spatial and spectral dependencies in the multi-temporal Sentinel-2 data. By considering the additional temporal dimension, 3D convolutions can leverage seasonal variations and changes in vegetation phenology, which are crucial for accurate tree genus classification.

LSTM was introduced to the model alongside 2D convolutions because they process the spatial structure of the Sentinel-2 images through convolution operations while simultaneously capturing temporal sequences. This dual capability allows the model to learn intricate spatial patterns within each image and understand how these patterns evolve over time.

The selected metrics used to create Fig. 2.1 are well-suited for handling class imbalance in the seasonal analysis of the CNN model. Recall ensures the model captures as many instances of the minority classes as possible, which is critical when dealing with imbalanced datasets. Precision assesses the accuracy of the model's positive predictions, reducing the impact of false positives. The weighted f1-score balances precision and recall, accounting for class imbalance by considering the support of each class. PRC focuses on the trade-off between precision and recall, highlighting the model's performance on minority classes. Lastly, AUC measures overall model performance across all thresholds, providing a comprehensive view of its ability to distinguish between classes.

For individual seasons, the model performs best in summer and autumn overall across the metrics. This suggests that the CNN model is more effective at classifying tree genera during these times, likely due to clearer and more distinct spectral signatures in the data collected during summer and autumn. The lower performance in spring and winter might be attributed to less distinct spectral signatures or more challenging environmental conditions, such as cloud cover and snow, which can affect data quality.

Based on the weighted f1-scores shown in Fig. 2.1, adding more seasons does not seem to offer significant benefits. For instance, some two-season combinations, such as summer and autumn, performed on par with the more complex four-season models. Additionally, single-season models were only a few percentage points below the top-performing models.

Based on these results, further analysis focused solely on summer seasons. This approach benefits from faster model training and reduced storage requirements, as adding an extra season nearly doubles the storage needs, a challenge that intensifies with the extension of the analysis over additional years. Despite these adjustments, a complete Sentinel-2 dataset for a single season still requires nearly 200GB, or approximately 1MB per location.

2.2 Sentinel-2 Bands

In addition to the seasonal analysis in Section 2.1, another analysis was conducted to identify the most effective band combinations. Due to the large number of possible combinations, a direct analysis was impractical. Instead, Sentinel-2 bands were divided into three groups based on summer correlation groupings shown in Fig. 1.10: B2, B3, B4, and B5; B6, B7, B8, and B8A; and B11 and B12.

The resulting analysis, shown in Fig. 2.2, indicates that NIR and SWIR bands perform slightly better overall. Based on these results, another group was selected: B2, B3, B6, B8, and B11. These bands represented the best combinations that resulted in a practical analysis within the available timeframe.

Fig. 2.3 shows that most selected combinations performed relatively well, as indicated by their proximity to the weighted f1-score for all bands. Based on these results, the most reasonable choices appear to be B3, B8, and B11, or the same but with B6 in addition. As the B6 combination displays better recall, precision, and weighted f1-score, as well as taking a fraction of storage compared to 10m bands due to being 20m. For the 250,000 samples, this combination shoulld take roughly 50 GB of storage. As such, the combination B3, B6, B8, and B11, was used for the remainder of this study.

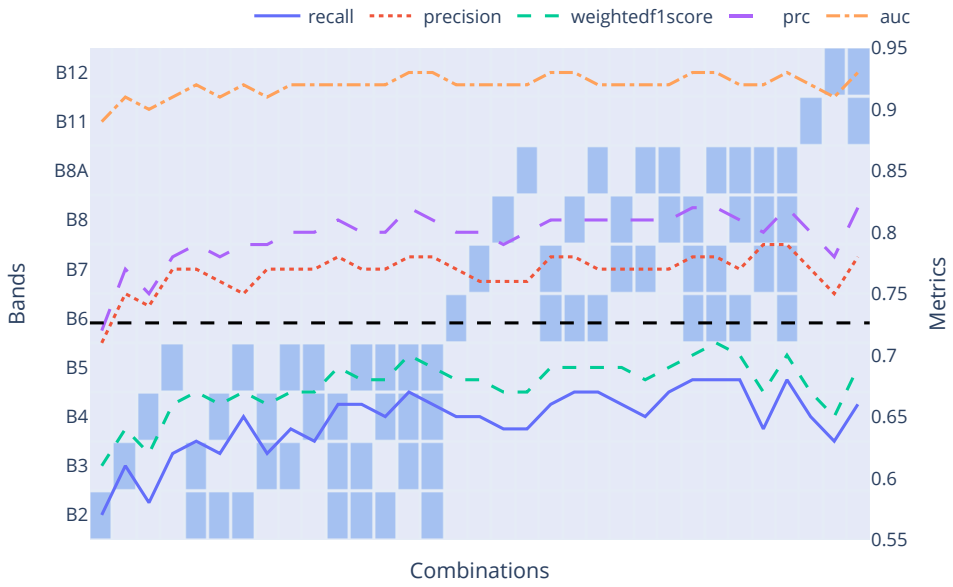

Figure 2.2: Sentinel-2 analysis for all combinations within three band groups using a CNN. The horizontal black line represents the weighted f1-score for all bands.

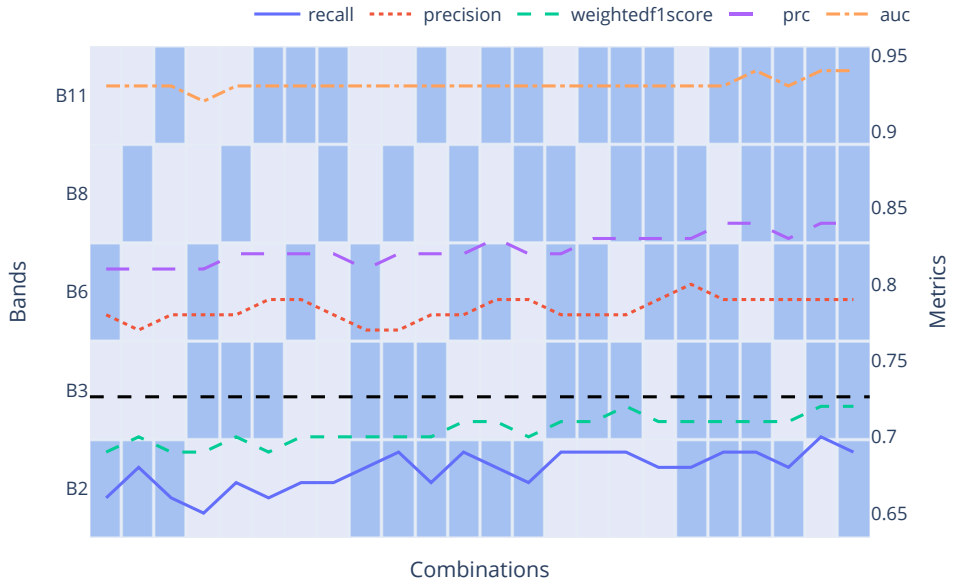

Figure 2.3: Sentinel-2 analysis for selected combinations between each of the three groups using a CNN. The horizontal black line represents the weighted f1-score for all bands.

2.3 Soil and Elevation

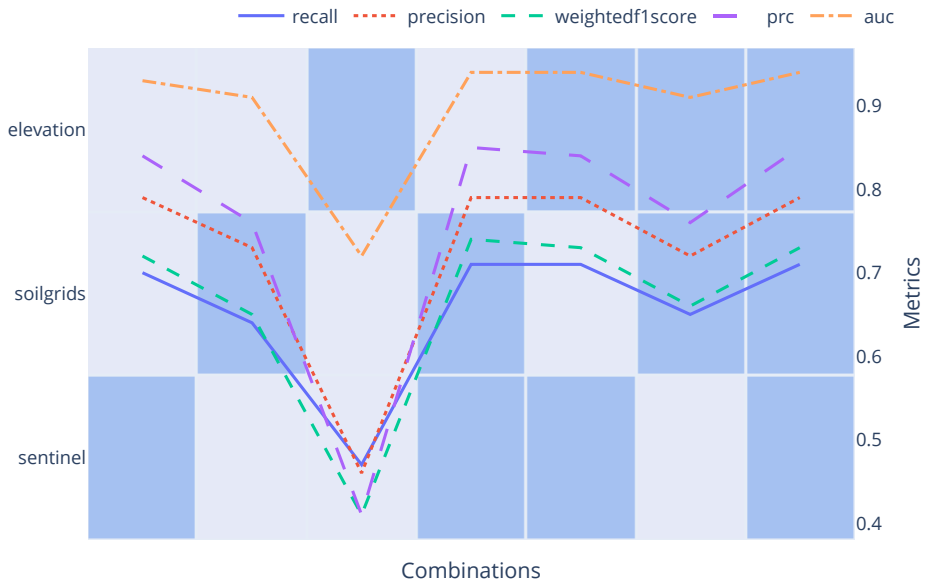

Figure 2.4: Analysis of SoilGrids and elevation data integration.

Fig. 2.4 demonstrates the performance metrics for different combinations of Sentinel-2, SoilGrids, and elevation data in tree genus classification. Sentinel-2 data alone provides a solid baseline, attributed to its high resolution and rich spectral information. When considering SoilGrids data alone, the metrics are lower than those for Sentinel-2, reflecting the coarse resolution and potentially limited predictive power for this specific task. Similarly, elevation data alone shows lower performance metrics, indicating that while elevation is a relevant feature, it does not provide sufficient information by itself for accurate classification of tree genera. The combined use of these datasets shows marginal improvements, suggesting that while additional data may contribute some value, Sentinel-2's high-resolution spectral data is the most significant factor in the model's performance.

Given that Sentinel-2 data alone provides strong results, further enhancements and optimisations were solely focused on this dataset.

Chapter 3

Neural Network Configuration

3.1 Hyperparameter Optimisation

Table 3.1 provides a summary of the search space for the initial hyperparameter tuning process. These parameters are critical as they influence the model's learning process, generalisation ability, and susceptibility to issues such as overfitting or bias. The chosen values in the search space represent a balance between exploring a wide range of options and focusing on promising areas based on prior knowledge or domain-specific considerations.
In the most successful trial, the model utilised training data from the medians of 2017, 2018, and 2019, applied L1L2 kernel regularization, used a spatial dropout rate of 0.5, and employed Leaky ReLU activation. The pool size was set to 4, dropout to 0.1, and bias initialization was disabled. The learning rate was 0.01, and the loss function was binary crossentropy, resulting in a score of 0.69.
While some of the initial trial's hyperparameters do not show significant overall improvement, some parameters do stand out. This is illustrated in Fig 3.1, particularly regarding training years, kernel regularization, activation function, and learning rate.

Table 3.1: Summary of initial hyperparameter search space.

name	values
class_weight	[true, false]
training_years	['2017_2018_2019', '2017']
kernel_regularizer	['l1l2', 'l1', 'l2']
spatial_dropout	[0.3, 0.1, 0.5]
activation	['leaky_relu', 'relu']
pool_size	[4, 2]
dropout	[0.3, 0.1, 0.5]
bias_initializer	[true, false]
learning_rate	[0.001, 0.0001, 0.01]
loss_function	['binary_focal_crossentropy', 'binary_crossentropy']

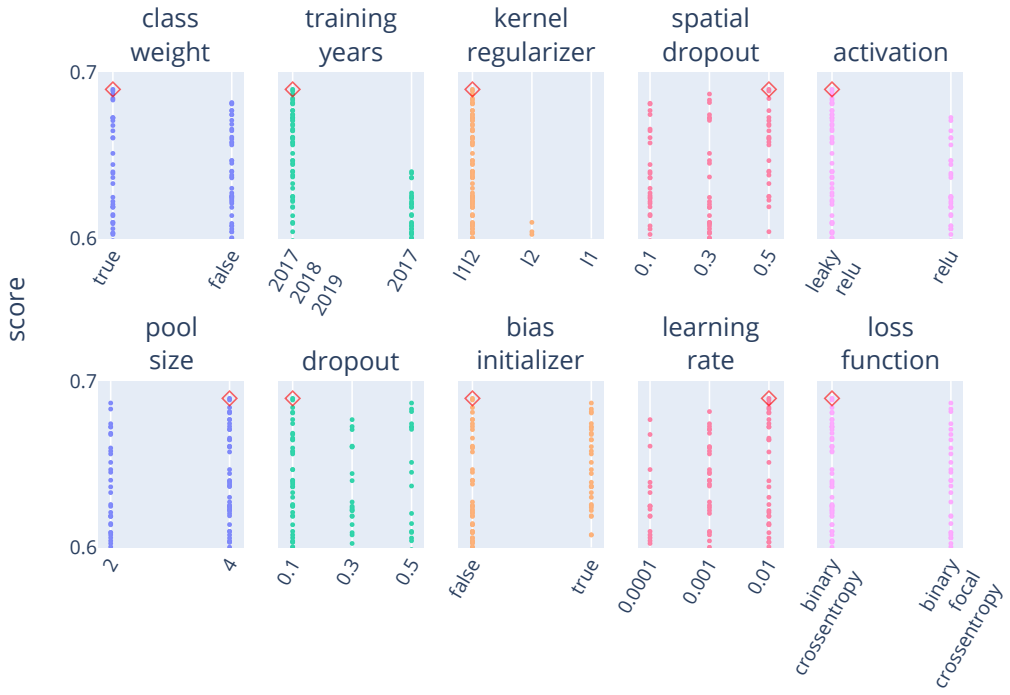

Figure 3.1: The top f2-scores for the initial hyperparameter tuning process are highlighted, with the best score marked by a diamond shape.

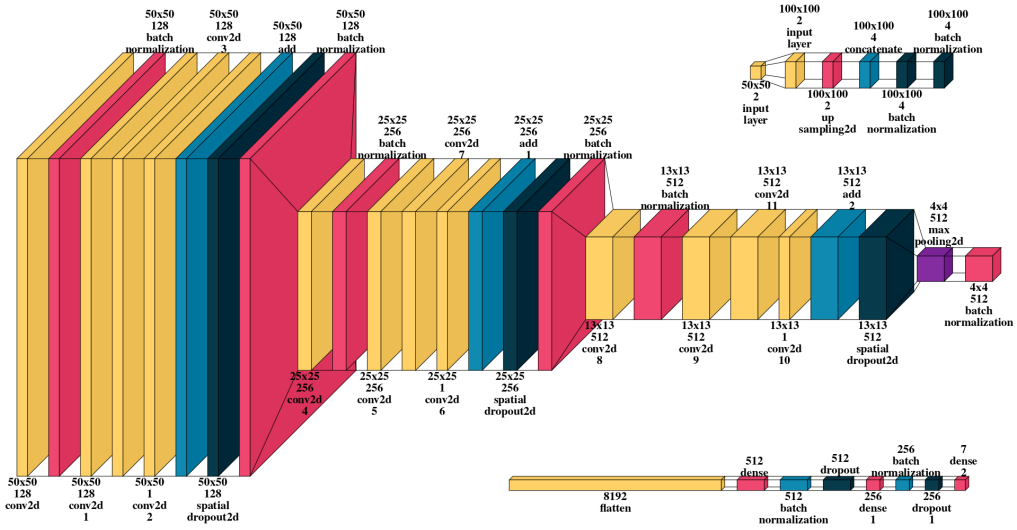

Figure 3.2: The model's layered architecture is depicted. The top-right shows the input layers, the middle section displays the convolutional layers, and the bottom-right illustrates the fully-connected layers. The layered views were generated using Gavrikov (2020).

Chapter 4

Climate Analysis

4.1 ERA5 Data

References

Gavrikov, P. (2020), 'visualkeras', https://github.com/paulgavrikov/visualkeras.

Mauri, A., Strona, G. and San-Miguel-Ayanz, J. (2017), 'EU-Forest, a high-resolution tree occurrence dataset for Europe', *Scientific Data* **4**(1), 160123.
URL: *https://doi.org/10.1038/sdata.2016.123*

Poggio, L., de Sousa, L. M., Batjes, N. H., Heuvelink, G. B. M., Kempen, B., Ribeiro, E. and Rossiter, D. (2021), 'Soilgrids 2.0: producing soil information for the globe with quantified spatial uncertainty', *SOIL* **7**(1), 217–240.
URL: *https://soil.copernicus.org/articles/7/217/2021/*