CZECH UNIVERSITY OF LIFE SCIENCES PRAGUE

Faculty of Tropical AgriSciences

Modelling Mediterranean oak tree dieback using the Plant Phenology (PPI) vegetation index

MASTER'S THESIS

Prague 2024

Author: Gergely Söptei

Chief supervisor: doc. Ing. Radim Matula Ph.D.

CZECH UNIVERSITY OF LIFE SCIENCES PRAGUE

Faculty of Tropical AgriSciences

DIPLOMA THESIS ASSIGNMENT

Gergely Söptei

Tropical Forestry and Agroforestry

Thesis title

Modelling Mediterranean oak tree dieback using the Plant Phenology (PPI) vegetation index

---

Objectives of thesis

Hypothesis and aim:

The aim of the thesis is to map oak dieback using Plant Phenology Index (PPI), a specific vegetation index obtained from remote sensed satellite data. The Plant Phenology Index is derived from the radiative transfer equation, and it has a linear relationship with the Leaf Area Index (LAI). It was introduced in 2014 by Jin and Eklundh and it is available in the High-Resolution Vegetation Phenology and Productivity (HR-VPP) product suite as part of the European Space Agency's Copernicus Land Monitoring Service (CLMS) program, provided by the Sentinel 2 satellite constellation. Plant phenology is provided for 13 parameters, up to two growing seasons, with high spatial resolution. The hypothesis of this thesis is that PPI can be used as a predictor of present tree mortality and as an early warning indicator for the identification of areas where cork oak dieback might occur in the future. The practical application of the findings of this thesis could also be used as a pipeline for future predictions of tree decline processes in other areas, using the same remotely sensed data with similar field datasets.

Summary:

Vegetation indices have long been used in land use and land cover change identification. A commonly used vegetation index is the Normalized Vegetation Index (NDVI), but numerous others have been introduced and used in the past. The Plant Phenology Index (PPI), one of the many vegetation indices, was introduced in 2014 to overcome the limitations of other indices when used, for example, with evergreen vegetation or with vegetation affected by snow cover. It is a physically based vegetation index; it is derived from the radiative transfer equation and has a linear connection to the Leaf Area Index (LAI). Tree mortality changes land cover, and has been affecting the *montado*, a silvo-pastoral agroforestry system, of mainland Portugal for decades. Dead tree data was collected in Companhia das Lezírias, a state-owned property with high proportion of *montado* (dominated by *Quercus suber* – cork oak), from 2014 to 2019. This dataset, in conjunction with the High Resolution Vegetation Phenology and Productivity (HR-VPP) remote sensing product, was used in this thesis to determine a relationship between phenological patterns (PPI) and tree mortality. The modelling of the relationship between tree mortality and tree phenology will be done with one-class Support Vector Machine (SVM) models, because of the nature of the field dataset. SVMs are supervised learning models that are frequently used

for classification tasks. The choice for using a one-class classifier was made because the dataset contains only dead tree observations.

Rationale:

Cork oak (*Quercus suber*) is highly regarded in Portugal, both economically and culturally. It is a protected tree and it serves as the primary source for cork production on a global scale. Tree mortality has been affecting cork oak, not only in Portugal, but in the Iberian Peninsula for decades. Finding an early warning indicator would be highly useful in determining areas where oak decline might occur in the future. This could be applicable to other dryland agroforestry systems.

Justification:

Tree mortality has been a problem for decades in Portugal and the work done in this thesis tries to address this issue. Since the Plant Phenology Index is a rather novel vegetation index, its applications are yet to be explored more in detail. The aim of the work is to find an application of the Plant Phenology Index in areas affected by tree dieback, which later can be used to predict such events in other sites.

Methodology

Study area:

The study area is in Companhia das Lezírias, Central Portugal, in the NUTS3 region of the Alentejo region, Lezíria do Tejo.

Data collection:

The dead tree data was recorded by workers in the field, who selected the trees with complete loss of leaves. Each cork oak tree was identified, and its geographical position was recorded using a handheld GPS (Etrex Garmin). The datapoints were stored in a single shapefile, representing all the surveyed dead trees as point features. The original dataset contained more than 27000 datapoints after completion. Duplicate points were removed from the dataset as part of the data cleaning process.

The remotely sensed phenology dataset was part of a larger data acquisition process using the Copernicus WEkEO portal, acquiring data for the entire continental territory of Portugal. The country is covered by 17 mosaics, out of which only one – 29SND – was used for the writing of this thesis, with phenology data spanning over six years (from 2017 to 2022), with thirteen variables in high spatial resolution (10m x 10m). The initial data acquisition of the Sentinel 2 HR-VPP data, due to its large size, was done through the WEkEO portal's Harmonized Data Access REST-based API.

Data analysis:

The dead tree dataset was divided into individual years, based on the date of collection, using Quantum GIS (QGIS) in the form of point shape files. Raster data, for all thirteen phenological variables, was extracted for each of the point locations using the Rasterio library from Python and saved into yearly data frames using GeoPandas. The extracted data serves as the basis for creating machine learning models using a special, one-class classification Support Vector Machine algorithm (OCSVM) from the scikit-learn Python library.

Data analysis and cleaning was done using the Pandas and numpy libraries, while data visualization was done using the matplotlib library and QGIS software for mapping outputs.

For the validation of the machine learning models, points of live trees are to be used. These were acquired through visual identification based on images from Google Earth Pro, with the help of professionals familiar with the composition of the study area.

The proposed extent of the thesis

50-55 pages

Keywords

montado, *PPI*, *remote sensing*, *Sentinel 2*, *HR-VPP*, *oak decline*, *SVM*, *One-Class Classification*

Recommended information sources

Forrest J, Miller-Rushing AJ. Toward a synthetic understanding of the role of phenology in ecology and evolution. Philos Trans R Soc Lond B Biol Sci. 2010 Oct 12;365(1555):3101-12. doi: 10.1098/rstb.2010.0145. PMID: 20819806; PMCID: PMC2981948.

Hongxiao Jin, Lars Eklundh, A physically based vegetation index for improved monitoring of plant phenology, Remote Sensing of Environment, Volume 152, 2014, Pages 512-525, ISSN 0034-4257, https://doi.org/10.1016/j.rse.2014.07.010.

HR-VPP Product User Manual Seasonal Trajectories and VPP parameters. European Environment Agency

Huete, A.R. (2012), Vegetation Indices, Remote Sensing and Forest Monitoring. Geography Compass, 6: 513-532. https://doi.org/10.1111/j.1749-8198.2012.00507.x

Jinru Xue, Baofeng Su, "Significant Remote Sensing Vegetation Indices: A Review of Developments and Applications", Journal of Sensors, vol. 2017, Article ID 1353691, 17 pages, 2017. https://doi.org/10.1155/2017/1353691

Navarro, Ana, Joao Catalao, and Joao Calvao. 2019. "Assessing the Use of Sentinel-2 Time Series Data for Monitoring Cork Oak Decline in Portugal" Remote Sensing 11, no. 21: 2515. https://doi.org/10.3390/rs11212515

Pichler, Maximilian & Hartig, Florian. (2023). Machine learning and deep learning—A review for ecologists. Methods in Ecology and Evolution. 14. 10.1111/2041-210X.14061.

Pinto-Correia, T., Ribeiro, N. & Sá-Sousa, P. Introducing the montado, the cork and holm oak agroforestry system of Southern Portugal. Agroforest Syst 82, 99–104 (2011). https://doi.org/10.1007/s10457-011-9388-1

Qinghua Guo, Maggi Kelly, Catherine H. Graham, Support vector machines for predicting distribution of Sudden Oak Death in California, Ecological Modelling, Volume 182, Issue 1, 2005, Pages 75-90, ISSN 0304-3800, https://doi.org/10.1016/j.ecolmodel.2004.07.012.

Sebastian Goihl; Determining the usefulness of the Copernicus High-Resolution Vegetation Phenology and Productivity Product (HR-VPP) with official agricultural data on cropland in case of the 2018 drought in the Federal State of Saxony, Germany. Journal of Water and Climate Change 1 November 2023; 14 (11): 3931–3949. doi: https://doi.org/10.2166/wcc.2023.501

Expected date of thesis defence
Master's Thesis – SS 2022/23 – FTA

The Diploma Thesis Supervisor
doc. Ing. Radim Matula, Ph.D.

Supervising department
Department of Forest Ecology

Electronic approval: 31. 01. 2024

prof. Ing. Miroslav Svoboda, Ph.D.
Head of department

Electronic approval: 20. 02. 2024

prof. dr. ir. Patrick Van Damme
Dean

Prague on 25. 04. 2024

Declaration

I hereby declare that I have done this thesis entitled Modelling Mediterranean oak dieback using the Plant Phenology (PPI) index independently, all texts in this thesis are original, and all the sources have been quoted and acknowledged by means of complete references and according to Citation rules of the FTA.

In Prague, April 24th 2024

.....

Gergely Söptei

Acknowledgements

I would like to thank Radím Matula and Adriana Silva for their contributions to this thesis and their efforts to make it happen. I would like to thank Maria Alexandra Oliveira for her insights on machine learning and all other inputs along the way. I would like to thank my family and friends especially some, who are far away and yet they were able to keep my spirits up with their kind words of encouragement.

I would like to thank Companhia das Lezírias S.A. for the information on tree mortality and land management without which I could not have been able to work on such a challenging and interesting topic.

Abstract

Vegetation indices have long been used in land use and land cover change monitoring. The Plant Phenology Index (PPI), one of the many vegetation indices, was introduced in 2014 to overcome the limitations of other indices when used for phenology, for example, with evergreen vegetation or with vegetation affected by snow cover. It is a physically based vegetation index; it is derived from the radiative transfer equation and has a linear connection to the Leaf Area Index (LAI). Tree mortality changes land cover, and has been affecting the *montado*, a silvo-pastoral agroforestry system, of mainland Portugal for decades. Dead tree data was collected in Companhia das Lezírias, a state-owned property with high proportion of *montado* (dominated by *Quercus suber* - cork oak), from 2014 to 2019. This dataset, in conjunction with the High-Resolution Vegetation Phenology and Productivity (HR-VPP) remote sensing product with thirteen yearly phenology parameters, was used in this thesis to determine a relationship between phenological patterns (PPI) and tree mortality. The modelling of the relationship between tree mortality and tree phenology was done with one-class Support Vector Machine (SVM) models, because of the nature of the field dataset. SVMs are supervised learning models that are frequently used for classification tasks. However, the choice for using a one-class classifier was made, which is an unsupervised classification algorithm for outlier detection, because the dataset contained only dead tree observations. The initial dead tree dataset was divided on a yearly basis, including only the ones that overlapped with the HR-VPP product, 2017 and onwards. To validate our models, a dataset was created through visual observations using Google Earth Pro imagery in Quantum GIS. Before creating our final models, experimentation was done using different tools for hyperparameter search and for feature selection. After this experimentation process, to reduce complexity and to retrieve continuous values from the decision boundaries of the classifiers, pairs of phenology parameters were created, models were trained and tested with them using different sets of hyperparameters. We found that certain combination of hyperparameters and phenology parameters, especially ones that indicate the start and the end of the vegetation season, as features, were able to perform classification of dead trees and outliers with convincing accuracy for one-class classification.

Key words: One-Class Classification, Support Vector Machines, Sentinel 2 – HR-VPP, Oak Decline, Montado, Remote Sensing, Plant Phenology

Contents

1. Contents

List of tables

List of figures

List of the abbreviations

ANOVA	Analysis of Variance
CDL	Companhia das Lezírias
CLMS	Copernicus Land Monitoring Service
DVI	Difference Vegetation Index
EEA	European Economic Area
ESA	European Space Agency
EVI	Enhanced Vegetation Index
FAPAR	Fraction of Absorbed Photosynthetically Active Radiation
GEP	Google Earth Pro
GPP	Gross Primary Productivity
GPS	Global Positioning System
HR-VPP	High Resolution Vegetation Phenology and Productivity
LAI	Leaf Area Index
MI	Mutual Information
ML	Machine Learning
NDVI	Normalized Vegetation Index
NIR	Near-infrared
NPP	Net Primary Productivity
NUTS	Nomenclature of Territorial Units for Statistics
OCSVM	One-Class Support Vector Machine
PPI	Plan Phenology Index
QGIS	Quantum GIS
RBF	Radial Basis Function
ST	Seasonal Trajectories
SVM	Support Vector Machine

UAV	Unmanned Aerial Vehicle
VI	Vegetation Index
WMS	Web Map Service

2. Introduction and Literature Review

2.1. Introduction

Vegetation indices, such as the normalized vegetation index (NDVI), or the enhanced vegetation index (EVI) have long been in use in different fields, such as forestry and ecology, to monitor changes in the state of vegetation [1]. These indices are relatively easy to calculate based on remote sensed data, but have their limitations, when applied for certain landcover types with evergreen vegetation (such as boreal forests) or when the vegetation is affected, for example, by snow cover [2]. The use of remote sensing as a tool for vegetation monitoring (i.e., changes in the vegetation cover in a certain area) is a well-established practice not only among forestry professionals, but professionals from a broad spectrum of scientific fields such as geography, ecology, as well as other disciplines related to vegetation mapping. In ecology, remote sensing technologies have been implemented for decades now, with a wide variety of uses such as land use and land cover change, integrated ecosystem measurements and change detection (i.e., climate change, habitat loss etc.) [3]. The use of remote sensing can provide information and can be converted to estimates with ease across entire ecosystems. The plant phenology index (PPI) is a novel vegetation index aimed to overcome these limitations. It was proposed by Lin and Eklundh in 2014 2 to improve the effectiveness of monitoring plant phenology. It is a physically based index and has a linear relationship with the leaf area index (LAI). The European Space Agency's (ESA) Copernicus program offers the High-Resolution Vegetation Phenology and Productivity (HR-VPP) product suite, leveraging the technical abilities of the Sentinel 2 satellite constellation, with thirteen phenological parameters - derived from the PPI - in high spatial (10 m × 10 m) and temporal resolution (5 day return time) for the entire EEA39 (32 member states of the European Union plus cooperating countries). This data is openly available since 2017 [4]. The aim of this thesis was to determine the usefulness of the PPI in detecting cork oak decline, which has been affecting the *montado*, an agroforestry system part of the cultural landscape in continental Portugal, for decades, in the study area of the state-owned property of Companhia das Lezírias. We proposed that the PPI using the data from the HR-VPP product suite can be used as a predictor of present tree mortality. We obtained our results using a classification

algorithm and plotted the decision boundaries for the separate classes of mortality (alive/dead) with pairs of phenological parameters as feature sets. Our results indicate that the PPI can be used for such purposes and that the classification algorithm used in this thesis was able to differentiate between dead and living trees based on vegetation phenology parameters, most importantly the start and the end of the season date values. Based on this, we conclude that the PPI can be used for future monitoring and research of cork oak tree decline in agroforestry systems.

2.2. Literature review

2.2.1. Describing the *montado*

The oak dominated agroforestry system called *montado* has an elevated importance in mainland Portugal. It is protected by law [5] and it is part of the cultural landscape, predominantly in the southern part of the country. It has important economic and biological qualities, through production of goods and provision of ecosystem services, it contributes to the maintenance of ecosystem services such as a healthy soil ecosystem, which allows functional soil biome and prevents soil erosion, abundant pastures with biodiverse plant understory cover that provide habitat for wildlife (such as reptiles, birds, mammals, and pollinator insects), its diversified vertical structure of tree cover supports the regeneration of oak species, and the preservation and enhancement of habitats and landscape elements such as patches of shrubs, riparian galleries, ponds and other habitat types [6]. In recent years, the extent of the *montado* has reduced due to numerous factors such as increase of intensive farming, overgrazing and abandonment (a prevailing issue in the Portuguese countryside) and shrub encroachment [7]. The *montado* covers about 800,000 ha [8], predominantly in the Southern region of the country, Alentejo. It has its equivalent in Spain where it is called *dehesa*, and as such it is also acknowledged for its multifunctional properties, despite originally being valued for cork production and animal husbandry. Its savanna-like physiognomy is one of the main characteristics of the *montado*, where two species of oak, cork – *Quercus suber* and holm – *Quercus ilex rotundifolia* are prevalent – in varying densities – diffused in the landscape in a mosaic like fashion. Cork oak (*Quercus suber*) has been highly regarded for centuries in Portugal. According to the National Forest Inventory it is the second most abundant

species after eucalyptus covering about 23% of the forested areas of mainland Portugal [8]. Its importance has been recognized by law since the 13th century, and it was established as Portugal's national tree in 2011. This can be attributed to its relevance in multiple areas, such as social, economic, and environmental. Besides the centuries old trade of producing and harvesting cork, which accounts for 54 % of the global mean annual cork production, a wide range of other products - such as wood production and livestock raising - and ecosystem services are related to this agroforestry system [9]. The ecological functions of the *montado* include soil protection and water regulation mostly through vegetation cover (i.e., tree canopy cover), that has multiple roles in the protection of the soil and regulation the retention of water and rainfall interception [9]. Besides regulation, decomposing organic matter adds to the fertility of the soil. CO_2 retention is also an important feature of these agroforestry systems, Since the oak species occupying this type of system live long, they promote carbon storage over elongated periods of time [10]. Another important feat of the *montado* is the maintenance of biodiversity. According to Díaz-Villa et al. [11], about 135 vascular plants per 0.1 hectare can be found in an oak-savannah grassland like the *montado*, several of which are under protection. Besides plants, the *montado* system plays an important role in the life of animals, providing refuge and habitat, where 28 species have protection status [12].

2.2.2. Oak decline

Severe oak decline in the Mediterranean agro-silvopastoral systems have been observed since the second half of the 20th century and it has been reported from Portugal at least from the 19th century [13]. It has been observed in Spain and Portugal for decades, which raises concerns that these systems will not be able to maintain the balance between human land use and ecosystem protection in the face of a changing climate [14]. Often times this dieback is attributed to different types of pathogens like *Phytophtora* sp.. Although *Phytoptora* spp. are suspected to be the main culprits of oak decline in Portugal as well, there have been several other pathogens and pests associated. Pathogen species from different parts of the vegetation have been recovered such as *Brenneria quercina, Hypoxylon* sp. [13]. There are two main types of syndromes that have been associated with oak decline that have been observed through the years: (1) Characteristic fast dying of the tree crown followed by the sudden death of the tree, which could happen in one or two vegetation seasons and (2) a progressive decline, which is first characterized by the

dying of the top of the tree and more intense leaf drop. This could affect the whole crown or can be restricted to only some of the branches. These mortality events have been ascribed to complex events that involve abiotic stress factors related to the properties of the soil, but these stress factors might also include drought and inadequate silvicultural management practices [13].

Cork oak has been observed in recent decades to be significantly declining, due to numerous factors attributed to climate change and deficient management practices, such overgrazing [15].

2.2.3. Remote sensing, vegetation Indices, plant phenology and productivity

Vegetation indices are thoroughly used in vegetation monitoring due to their relative simplicity and good correlation in changes in the vegetation. Spectral vegetation indices (VIs) are established tools to monitor the states of forests and processes in the canopy [16]. Vegetation indices respond to upper sunlit leaves more than the lower ones, which results in a non-linear relationship with the leaf area index. Depending on the type of vegetation (broadleaf versus needleleaf canopy structures), this relationship can further vary [1]. These remote sensing derived indices can enhance the estimations of forest biophysical properties that are otherwise difficult to sample in situ. The most common way to perform vegetation monitoring is through the usage of vegetation indices, such as the Normalized Vegetation Index (NDVI) or the Enhanced Vegetation Index (EVI), which includes some corrections for atmospheric conditions, but there are many others [16]. Although NDVI is one of the most established vegetation indices in plant monitoring, it has its limitations. For example, NDVI (and EVI) are not optimal in areas where there is change in snow-cover (for example boreal areas) due to their sensitivity to such changes. NDVI is also sensitive to other atmospheric conditions such as clouds, haze, and aerosols, which might affect the reflectance of light in the visible and near-infrared (NIR) wavelengths [17]. As a practical application, high resolution remotely sensed imagery has been used in conjunction with different vegetation indices and machine learning algorithms to investigate the oak tree vitality in the Mediterranean [14].

In this thesis the Plant Phenology Index (PPI), a novel vegetation index was used to obtain relevant information on the change of tree canopies in agroforestry systems,

which itself is derived from another index, the leaf area index (LAI). Used in models and studies focused on forest processes, LAI is an important forest parameter. It is defined as the one-sided area of leaves in a canopy projected onto the ground (m^2/m^2) [1]. The PPI was developed rather recently [2], compared to those used in these previous studies and ecological studies in general. In this thesis, an effort was made to combine the novelty of the PPI and the HR-VPP Sentinel 2 product, both of which are discussed in depth in Chapter 3.

The definition of phenology, according to the Merriam-Webster dictionary, is "a branch of science dealing with relations between climate and periodic biological phenomena (such as bird migration or plant flowering)" or more succinctly, the "periodic biological phenomena that are correlated with climatic conditions" [18]. There is an elaborate interplay between an organism's genes and a variety of external environmental factors. This interplay influences phenological events such as onset of reproduction or entry or emergence from hibernation. Environmental factors like temperature and precipitation can directly affect the schedule of biological events. These factors can serve as cues for the organism's biological clock [19]. According to Abbe [20], phenology is the study of periodical phenomena of different organisms, which depend on the climate. Productivity of vegetation can be attributed to the leaf area, which regulates the development of plant biomass and the uptake of solar energy in a process of converting light into carbon through photosynthesis. Productivity means the growth of vegetation and it is often described as gross primary productivity (GPP), which is accumulation of biomass due to photosynthetic activity or net primary productivity (NPP), which can be formulated as GPP minus the respiration of vegetation, simply the net vegetation growth [4].

Jin and Eklundh introduced a physically based vegetation index [2] to improve the effectiveness of monitoring plant phenology, the PPI. It is approximately linear to the canopy green LAI. It is derived from the radiative transfer equation, and it is computed from red and Near-Infrared (NIR) reflectance. The following formula was conceived for the calculation of the Plant Phenology Index [2]:

$$PPI = -K * \ln\left(\frac{M - DVI}{M - DVI_S}\right)$$

Where the Difference Vegetation Index (*DVI*) is the difference between near-infrared (NIR) and red reflectances (sun-sensor geometry corrected), while DVI_S is the *DVI* of the soil. *M* is a site-specific canopy maximum *DVI*. This, in principle, could be estimated in several ways:

- Measuring non-sparse vegetation during a longer period (i.e., multiple years).
- Model simulations of canopy reflectance where the Leaf Area Index (LAI) is greater than 8 square meters.
- From measuring leaf single scattering albedo or absorbtance.
- From measured red and near infra-red reflectances for a site where the Leaf Area Index and the background reflectance is known.

K is a gain factor which is formulated in the following way:

$$K = \frac{0.25\ \cos(\theta)}{(1 - d_c)\ G + d_c\ \cos(\theta)} \frac{1 + M}{1 + M'}$$

Where θ is a sun zenith angle; d_c is an instantaneous diffuse fraction of solar radiation in case of clear sky and atmosphere (standard), when the sun has the zenith angle of θ; G is a geometric function of leaf angular distribution and M – as stated previously – a site specific maximum of DVI [17].

The reason to introduce the PPI was to overcome the many problems that beset traditional vegetation indices, that hamper the use of these indices, for example in higher latitudes and on evergreen vegetation, for example due to snow cover. The authors concluded that PPI has superiority to other popularly used vegetation indices such as Normalized Vegetation Index (NDVI) and EVI (Enhanced Vegetation Index) in the scenarios mentioned above [2]. The novelty of this vegetation index, combined with the high resolution of the Sentinel 2 HR-VPP product could be the source of important research in the future, although the application of this product is still being assessed. It is important to note here that there have already been studies that tried to make use of the PPI in situ. According to Goihl [21], vegetation phenology and productivity are not tangible values for stakeholders in agriculture such as practitioners and decision makers. In this study the author concludes that remote sensing data alone would not be able to provide valuable explanations (in this case on drought aid) on a farm level, and further data (i.e., ground truth data) is needed for a practical application. Others have found that when compared to flux tower observations (i.e. ground observations of plan phenology)

in boreal forests, PPI performs significantly better than other indices such as NDVI or EVI using remote sensing products with coarser spatial resolution [17]. Tian et. al. [22] conclude that the PPI has a higher potential, in conjunction with the capabilities of the Sentinel 2 satellite constellation, to monitor phenology on a continental scale at a 10 meter spatial resolution.

2.2.4. Application of machine learning

The application of machine learning – not just in ecology – came about with the growth of computing power, that allows usage of algorithms on machines available for larger audiences. During the ‘90s, the first wave of fundamental concepts and algorithms emerged (such as boosting, bagging, random forests and shrinkage estimation) that challenged for the first time, the supremacy of classical probability-based statistical models, used for data analysis and making predictions [23]. In a very simplistic way, we can define the objective of machine learning as building a model capable of good predictions [23]. This good predictive model means that it performs well on previously unseen data. Any algorithm capable of predicting certain tasks can be used for machine learning. Based on the nature of the task at hand, we can differentiate supervised and unsupervised machine learning tasks [23]. The former includes classification and regression tasks, while clustering, dimension reduction and anomaly detection constitute the latter. Practical use cases include species distribution modelling, identifying areas for conservation or restoration, forest protection, ecosystem service management, invasive species risk management, filling knowledge gaps in datasets, plant-pollinator networks and many more [23]. Machine learning, has been used to monitor and detect tree mortality in previous studies, using remote sensed data [14, 24] or environmental variables [25].

Support Vector Machines (SVMs) are supervised machine learning algorithms and are frequently used for classification problems such as binary or multi-class classification. SVMs look for the optimal hyperplane to separate the dataset into classes, maximizing the clearance distance between these classes [26]. Support Vector Machines use a set of mathematical functions that can be defined as the kernel [26,27]. Generally speaking the role of the kernel function is to map the data from a low-dimensional space to a space of higher dimension, to facilitate the classification task using linear decision surfaces (Figure 1.) [26]. A linear kernel can be used for linearly separable datasets, while

the polynomial kernel for data with nonlinear patterns. Figure 8. shows the hyperplane as decision boundary for linear SVM:

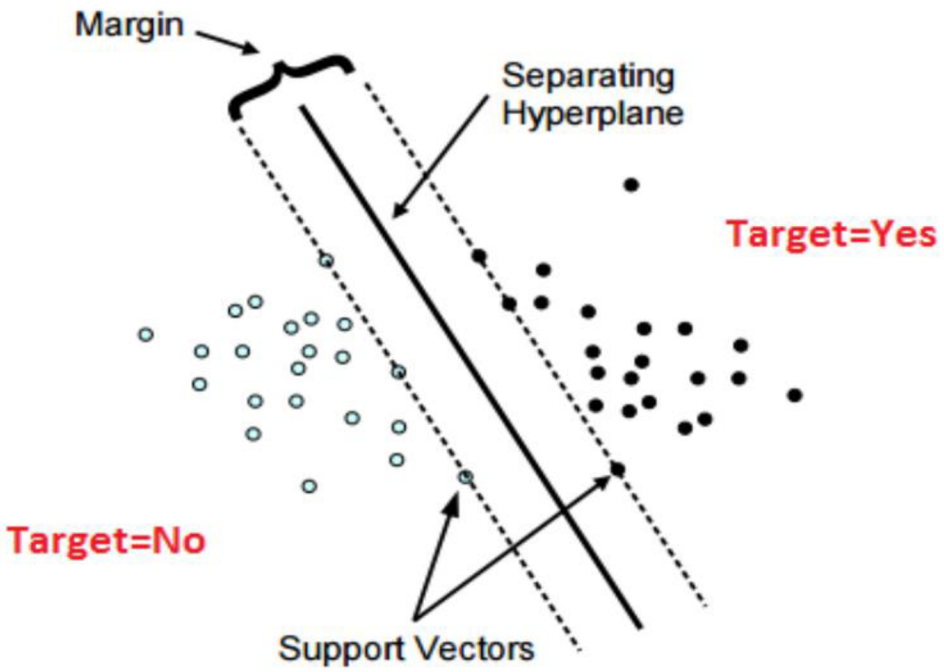

Figure 1. Data classification in SVMs (Source: Al Mejibli 2020)**.**

Figure 2. shows an example of how a non-linear kernel to transform the points of the dataset into higher dimensional feature space [26]:

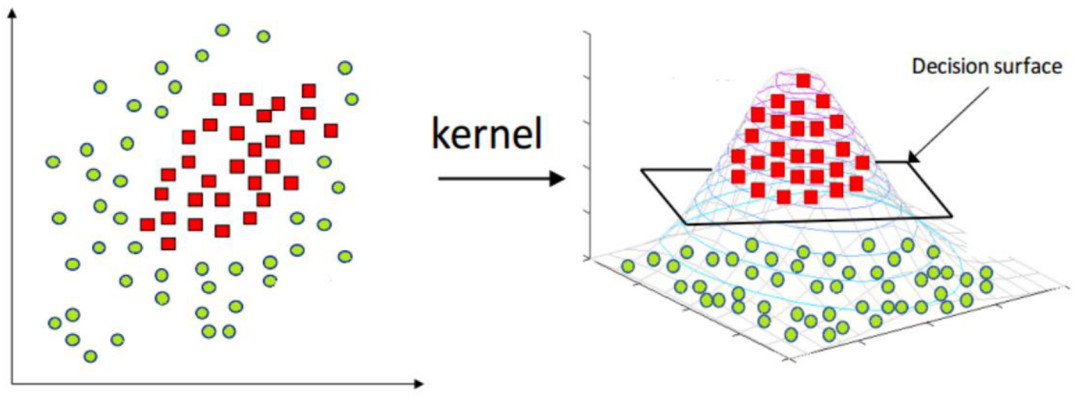

Figure 2. Data classification using kernel (Source: Al Mejibli 2020).

Table 1. shows the different kernel types used with SVMs [28,42]:

Table 1. Kernel types in SVMs

Kernel Type	Advantages	Disadvantages
Linear	Computationally effective - Works well for high-dimensional data	Limited to linearly separable data - May not capture complex relationships in nonlinear data
Radial Basis Function (RBF)	Good for complex nonlinear relationships	Sensitive to overfitting
Polynomial	Good for polynomial problems ,can also capture nonlinear relationships in the data	Prone to overfitting in high-degree polynomials, Sensitive to other parameters such as degree
Sigmoid	Can be effective in specific applications, such as neural networks	Application limited compared to other kernels
Custom Kernels	Tailored to specific domain or problem	Requires expertise and experimentation

When dealing with unbalanced data, or data that contains only one class, such as the data recorded in Companhia das Lezírias, a one-class classifier can be applied. These algorithms classify one class and look for outliers based on the training data. In this thesis a One-Class Support Vector Machine (OCSVM), which is an unsupervised algorithm, different from other SVMs, was used for the differentiation of dead trees from 'anything' else (i.e. possible living trees of the same species). One-Class classification has been used for example in species distribution mapping, where it does not make sense to record unsuccessful observations of a given species [28], in handwritten number recognition or in remote sensing [29]. Likewise, this lack of other labeled categories can be applied to other use cases, such as the distribution of sudden oak death [25]. When compared with methods that model the presence-only data directly, one-class SVMs have multiple advantages. They can represent different data distribution shapes in the feature space

using kernels (e.g., banana shapes, spheres, irregular shapes) [25]. Another advantage of one-class SVMs is that they aim to find boundaries of the hyperspace containing all or most of the training data and because of that, no assumptions on the probability density of the data are made [25].

Based on these use cases mentioned above, the nature of the dataset, and easy availability through the Python machine learning ecosystem, we decided to use the One-Class Support Vector Machine model from the scikit-learn library [28]. Python is the *lingua franca* of data science and machine learning and as such it makes working with algorithms such as the OCSVM classifier easier for people not having proper training or scientific background in the field of machine learning.

3. Aims of the Thesis

The aim of this thesis was to provide a link between the change in phenological characteristics of vegetation and the decline of cork oak in the study area, Companhia das Lezírias. For that we used PPI as an indicator of annual phenological changes and we hypothesized that the PPI could be potentially used as an indicator for evergreen oak dieback, due to its better performance on evergreen vegetation compared to more commonly used vegetation indices [2,4,22]. Moreover, we expect to understand if the PPI could be used as an early warning of oak dieback. To test our hypothesis on the applicability of the PPI in relation to oak dieback, a classification machine learning problem was solved, using a One-Class Support Vector Machine, to distinguish dead trees from living ones.

In many cases, it is hard to find proper environmental data to be used in such scenarios, such as high-resolution climatic data, even though many sources are available on the internet [30,31]. The work done in this thesis can be used to derive conclusions using only the openly available Sentinel 2 phenology data – the HR-VPP product suite - provided by the European Space Agency's (ESA) Copernicus program in high (10 m × 10 m) spatial resolution. The results of this thesis could be used in later research and monitoring efforts for the cork oak decline in the Iberian Peninsula or other areas where evergreen oak is affected by dieback.

4. Methods

4.1. Study site and data

The study area, Companhia das Lezírias is in Southwest Portugal, in the NUTS3 subregion of the Alentejo region, in Lezíra do Tejo. It is about 11,000 hectares of which 8,500 hectares are involved in forestry production. 77,2% (6570 hectares) of this area is covered by cork oak, the rest is covered by other species such as maritime pine, stone pine and eucalyptus [32,33]. Figure 3. shows the location of the study area in continental Portugal and in the Lezíria de Tejo region:

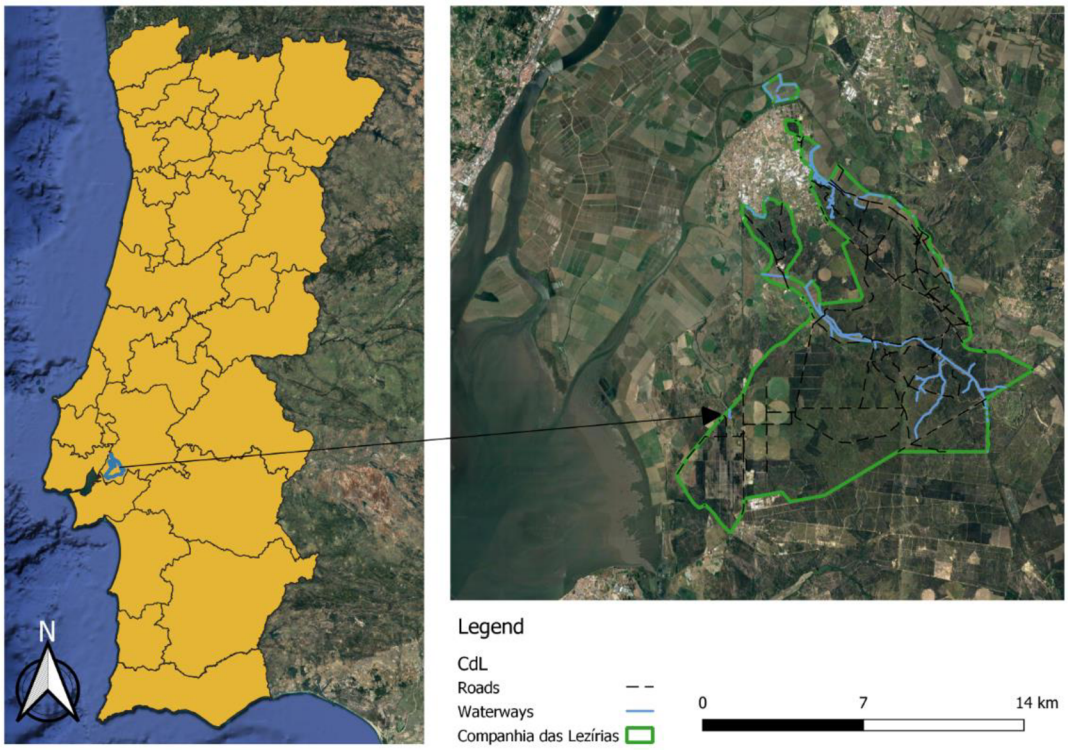

Figure 3. The location of the study area in continental Portugal.

Cork oak mortality in the study area was recorded from 2014 to 2019 after the dry season (i.e., from September to January) for the whole study area. The data were recorded by workers in the field, who selected the trees with complete loss of leaves even in the most remote places of the study area. Each cork oak tree was identified, and its geographical position was recorded using a handheld GPS (Etrex Garmin). The datapoints were stored in a single shapefile, representing all the surveyed dead trees as point features. The original database contained more than 27,000 observations. This number was later

reduced as the result of clipping the point dataset with the study area. This was done because the original dataset contained several points outside the study area boundaries. A yearly summary of dead trees recorded in Companhia das Lezírias is summarized in Table 2.:

Table 2. The summary of yearly recorded dead tree observations at Companhia das Lezírias

Year	No. Dead Trees Recorded
2014	1246
2015	8648
2016	2171
2017	6390
2018	3323
2019	4228
Total	26,006

To monitor the seasonal changes that take place in the study area, and to better understand the relation between tree mortality and plant phenology, the HR-VPP, a novel remote sensing product was used. The product used in this thesis is part of the European Union's Earth Observation program, also known as Copernicus Sentinel 2 [34]. The data is freely and openly available and accessible through the six different thematic Copernicus services. The High-Resolution Vegetation Phenology and Productivity product suite (HR-VPP) is provided by the Copernicus Land Monitoring Service (CLMS) as part of the Pan-European component at a high spatial (10 m × 10 m) and temporal (5-day revisit time) resolution. The HR-VPP product suite is derived from the Sentinel-2 satellite constellation (Sentinel-2A and Sentinel-2B). Products are generated for the entire EEA39, which includes the 32 member states of the European Union, the United Kingdom and 6 cooperating countries in the Western Balkans. The data is available from January 1, 2017, and onwards with different frequencies: daily, 10-daily and yearly [4].

The HR-VPP suite contains 3 product groups, 31 product types and 1522 files in more than 900,000 tiles per year. This amounts to more than 80 terra bytes of data each

year. These product groups include the raw Vegetation Indices (the VIs), which are generated near real-time, providing the status of the vegetation vigor for every pixel. The group includes three VIs: the Leaf Area Index (LAI), Fraction of Absorbed Photosynthetically Active Radiation (FAPAR), the Normalized Vegetation Index (NDVI) and the Plan Phenology Index (PPI). The second group contains the Seasonal Trajectories (STs). These products are provided yearly, after the end of each growing season, derived from the raw Plant Phenology Index by fitting a smoothing and gap filling function to it. Figure 4. shows the schematic representation of the HR-VPP product suite:

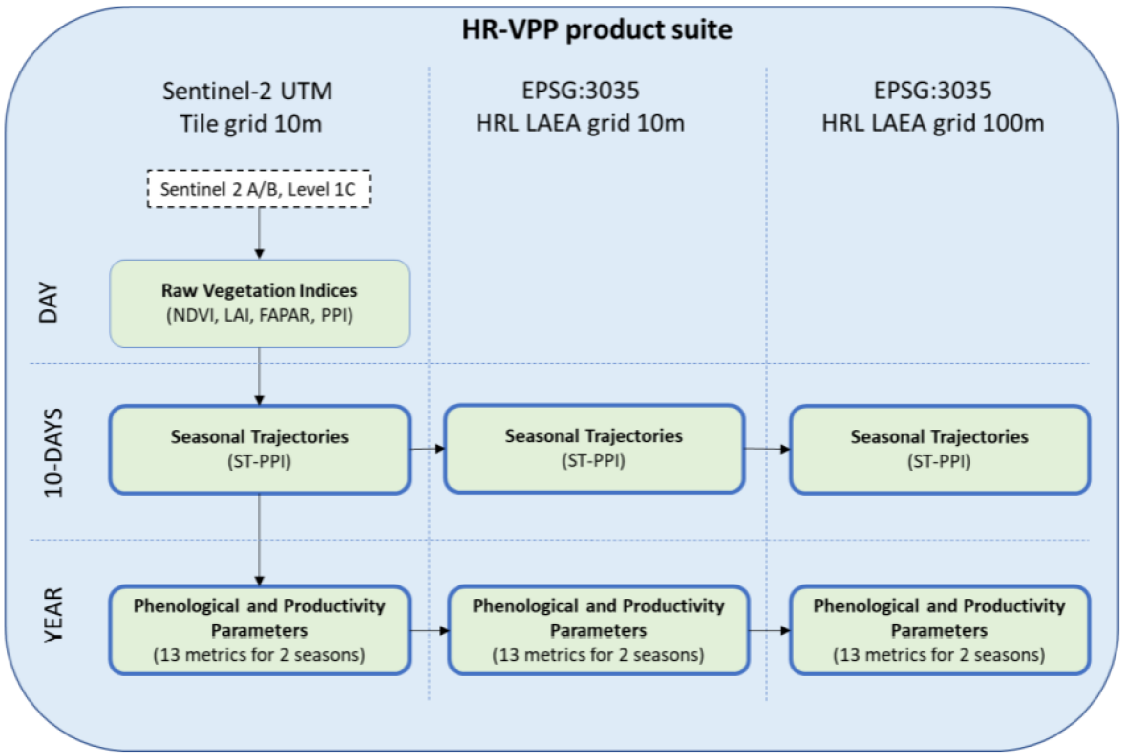

Figure 4. Overview of HR-VPP product collections (Source: HR-VPP User Manual)

The Vegetation Phenology Parameters are derived using the seasonal trajectories of the Plant Phenology Index. They are provided for up to two growing seasons, for such parameters as the start day and the end day of the vegetation season (with pertaining vegetation index values), length of the vegetation season, seasonal and total productivity, slope of the greenup and greendown periods, vegetation index maximum and minimum values with the day of the maximum and the minimum. For this thesis, data was extracted from the products available yearly, focusing on the first growing season, because of its length and significance. Table 3. lists the phenological parameters available in the phenology product group including some of their attributes [4]:

Table 3. HR-VPP phenology parameters

File_ID	File_ID description	Unit	Digital Range	No value
SOSD	Day of start-of-season	day-of-year	Format: YYDOY. E.g., 18030: DOY 30 in year 2018 16001 - 65365	0
EOSD	Day of end-of-season			
MAXD	Day of maximum-of-season			
SOSV	Vegetation index value at SOSD	PPI	see ST-PPI 0 to 3 physical range 0 to 30000 digital range	-32768
EOSV	Vegetation index value at EOSD			
MINV	Average vegetation index value of minima on left and right sides of each season			
MAXV	Vegetation index value at MAXD			
AMPL	Season amplitude (MAXV – MINV)			
LENGTH	Length of Season (number of days between start and end)	day	1 to 1096	0
LSLOPE	Slope of the greenup period	PPI × day-1	0.01 to 0.5 physical range 100 to 5000 digital range	-32768
RSLOPE	Slope of the greendown period			
SPROD	Seasonal productivity. The growing season integral computed as the sum of all daily values between SOSD and EOSD	PPI × day	0 to +1095 physical range 0 to 10950 digital range	65535
TPROD	Total productivity. The growing season integral computed as sum of all daily values minus their base level value.			

Figure 5. shows a schematic representation of the HR-VPP product bundle and aids the understanding of the phenology parameters and their values:

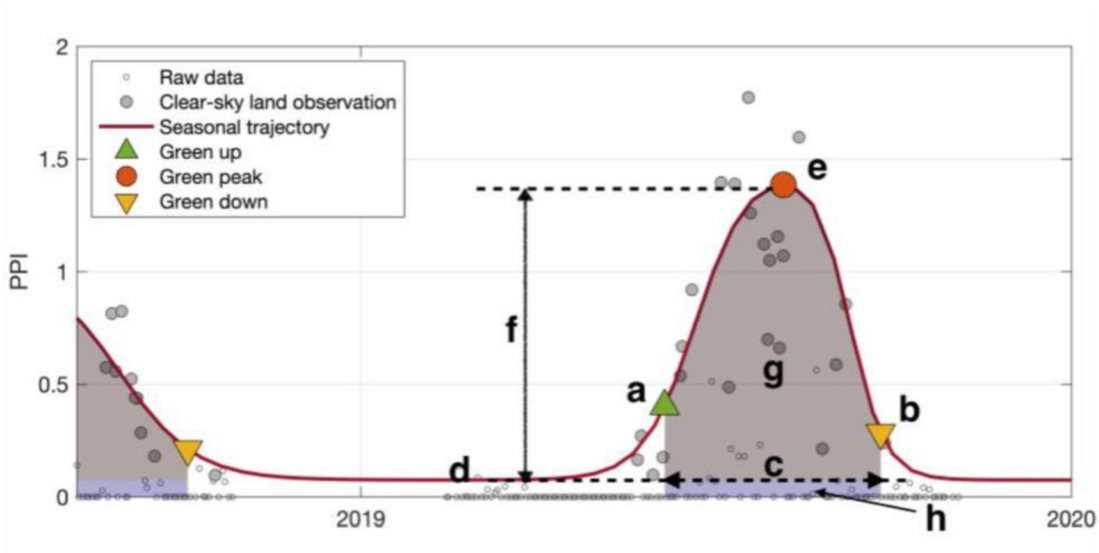

Figure 5. Schematic representation of the HR-VPP product bundle. Vegetation Phenology and Productivity parameters (VPPs) are: (a) start of season (date and PPI value), (f) amplitude, (g) small integrated value, (g+h) large integrated value. (Source: HR-VPP user manual)

The dataset was part of a larger data acquisition process using the Copernicus WEkEO portal [35], acquiring data for the entire continental territory of Portugal. The country is covered by 17 mosaics, out of which only one – 29SND – was used for the writing of this thesis, with phenology data spanning over six years (from 2017 to 2022).

Figures 6., 7., and 8. show the distribution of the dead tree datapoints for the years overlapping with the Sentinel 2 HR-VPP data:

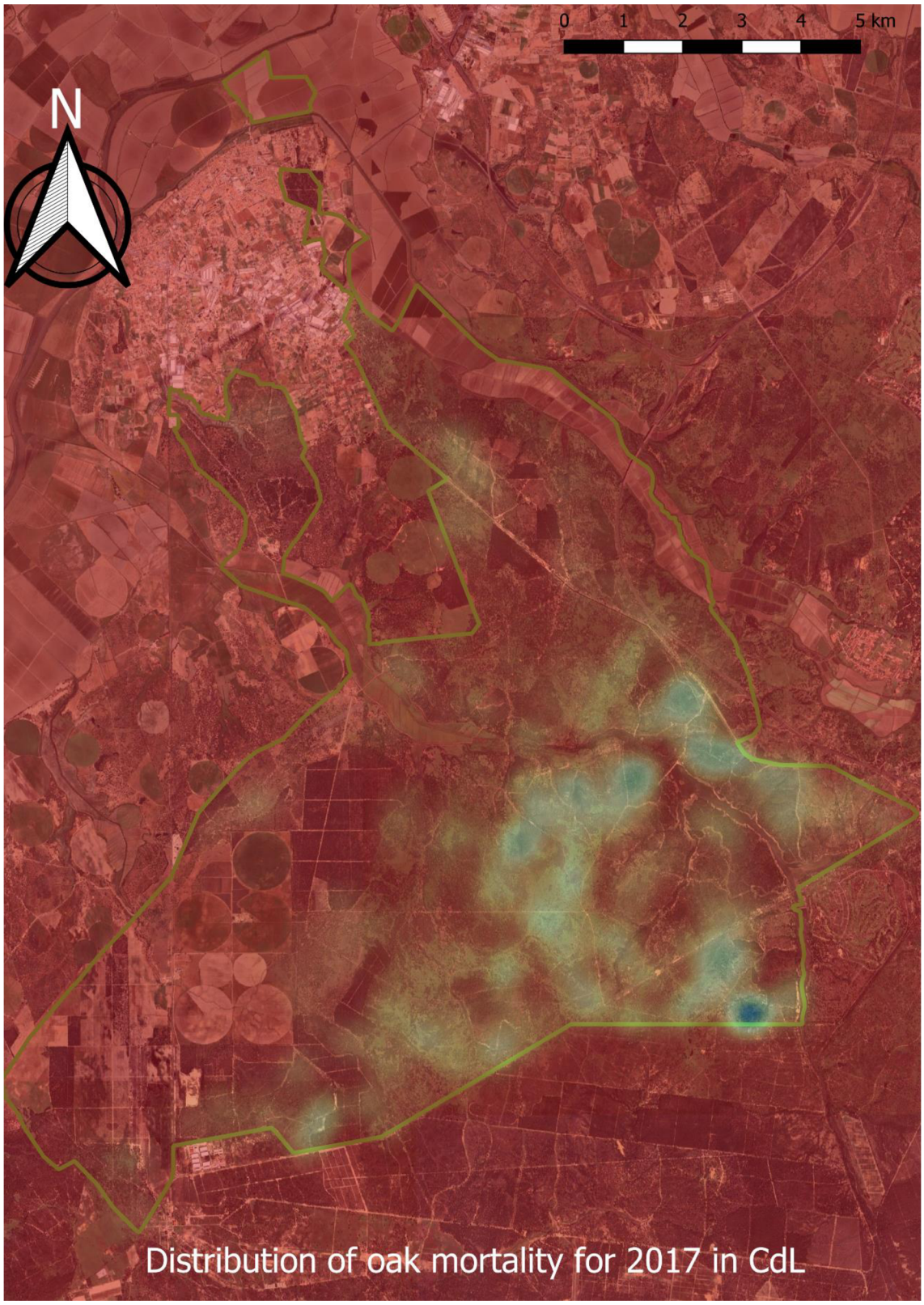

Figure 6. Distribution of dead tree datapoints for the base year - 2017

Figure 7. Distribution of the dead tree datapoints for validation dataset #1 - 2018

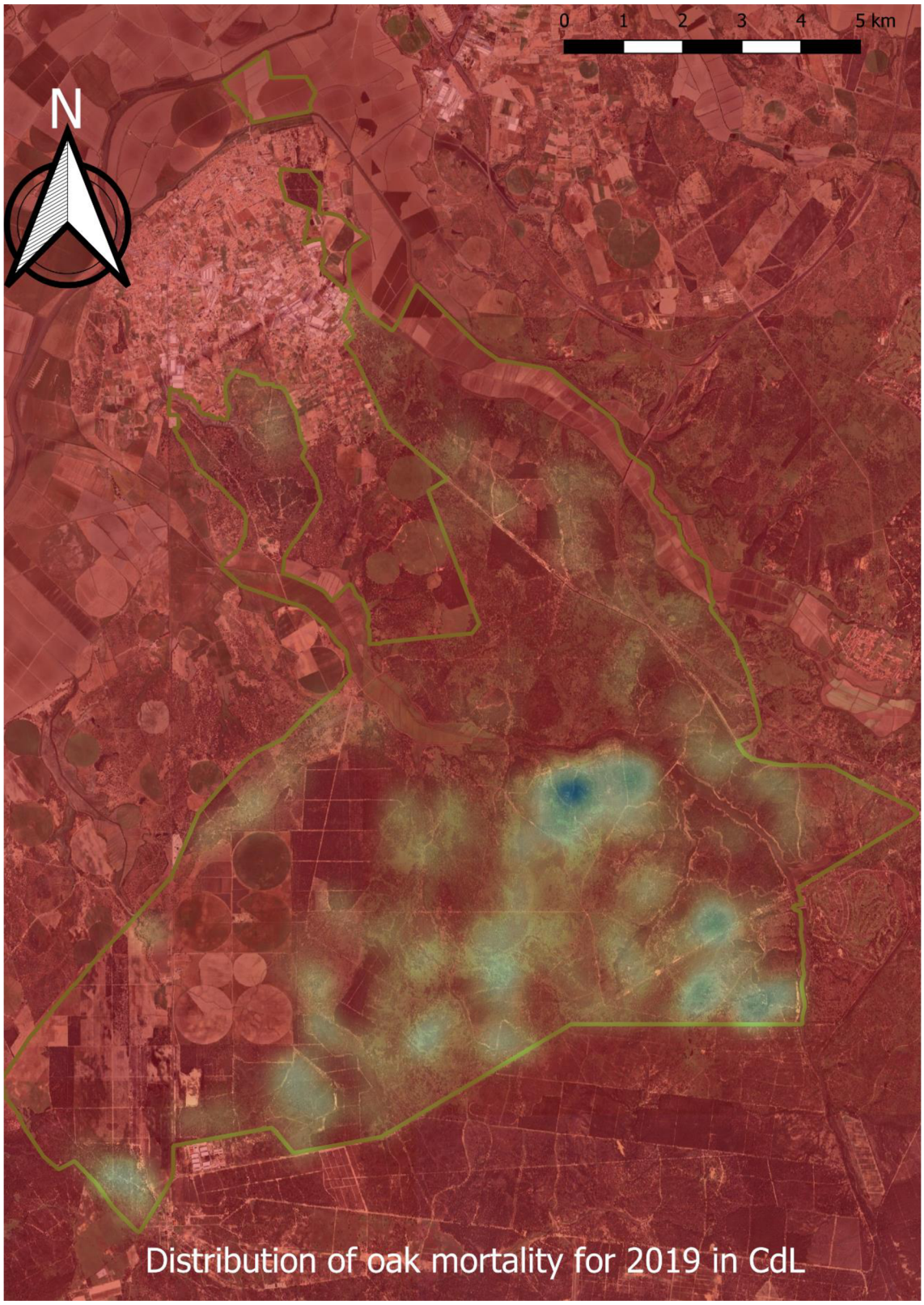

Figure 8. Distribution of the dead tree datapoints for validation dataset #2 - 2019

Living tree data was generated based on visual observations and manual digitalization of living trees using images from Google Earth Pro as a WMS layer in

Quantum GIS (QGIS). This dataset was used for testing our models trained and validated on the dead tree datasets obtained in situ. Other studies have used similar sources for model testing, such as Microsoft Bing satellite images, for visually identifying trees in similar agroforestry systems [14]. This was done based on a one tree per pixel basis, meaning that the point of reference for selecting the reference data was the pixel grid of the HR-VPP raster data. This grid was made in delineated areas, well known by people working on previous projects in Companhia das Lezírias, using the 'Create Grid' Vector Research tool in QGIS. The area was used based on previous knowledge of contributors and where it was possible to identify trees with a high accuracy.

Distribution of these points can be seen in Figure 9. below, with indication to the year of the phenology parameters used:

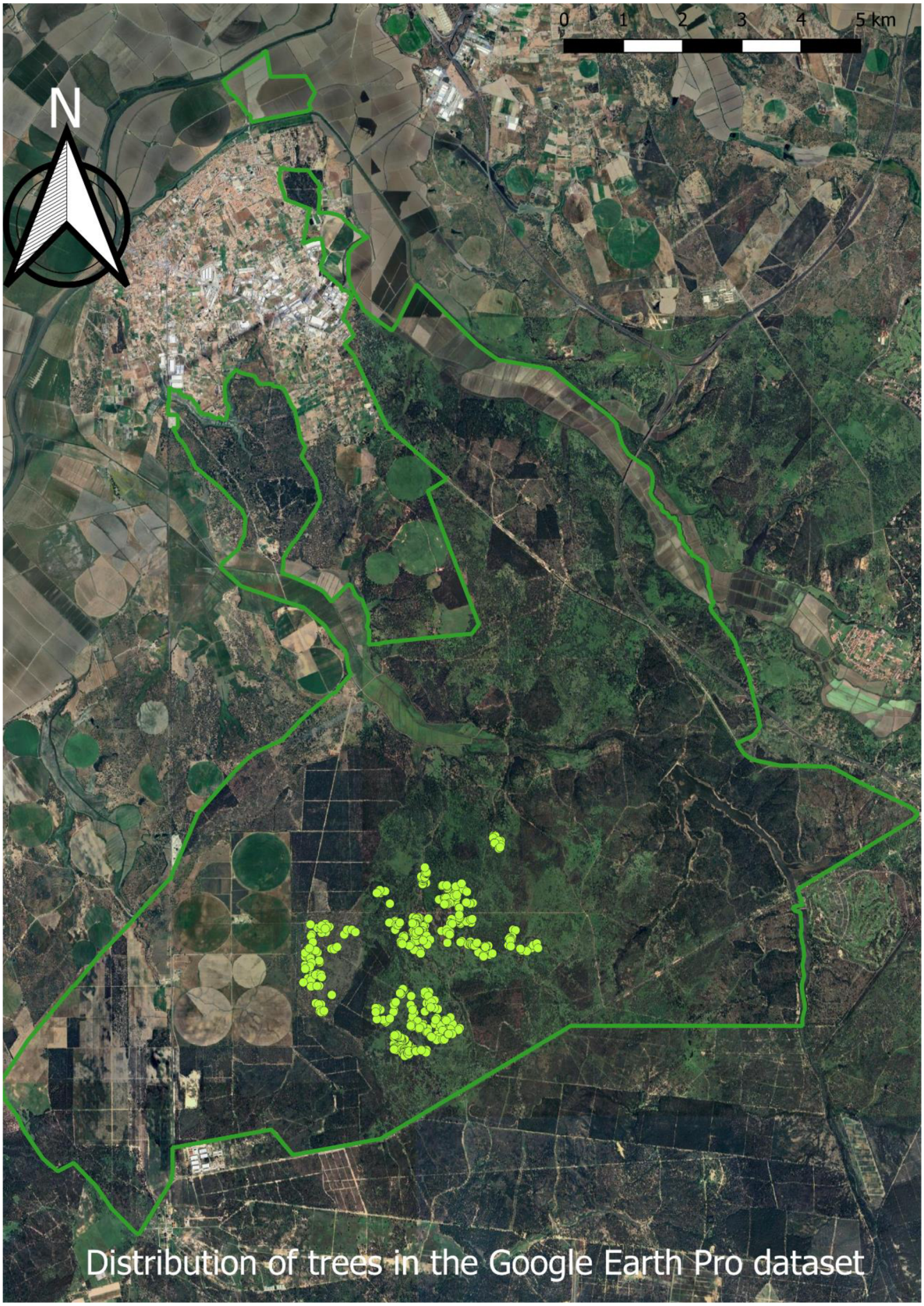

Figure 9. Distribution of datapoints for test dataset generated using Google Earth Pro - 2022

We chose 2022 as the year used for the Google Earth Pro observations, because it was the closest in time to the present-day observations in those images.

For data exploration, data processing, data analysis and visualization, various software products were used according to the task at hand. For partial data analysis and extraction, visualization and mapping, QGIS was used. For data extraction, data analysis, model training and predictions and data visualization the Python programming language was used with appropriate libraries such as GeoPandas, NumPy, Rasterio, seaborn, Matplotlib, scikit-learn among others. Table 4 shows the various software products and their applications:

Table 4. Software used in the making of this thesis**.**

Role	Software
Data cleaning	QGIS
	Jupyter Notebooks
Data analysis	Python
	Jupyter Notebooks
Visualization	QGIS (for mapping)
	Python (for plotting results)
Code development	Visual Studio Code
	Python

Jupyter Notebook was chosen as a tool for data exploration and data analysis because it offers a flexible and ideal way to manage the iterative nature of both processes. Later, the code was re-organized into multiple Python files in Visual Studio Code.

4.2. Workflow

4.2.1. Data analysis

To use the recorded tree mortality dataset in a meaningful way and to test the applicability of the PPI as we hypothesized, a classification problem [36] was solved, where dead trees were differentiated from everything else, based on a set of phenology variables as the feature set. The largest of the dead tree subset (year 2017) was used for training the classifier. Two other datasets (2018 and 2019) were used as validation datasets for the detection of dead trees based on the phenology traits. This allocation of the sub-datasets to different parts of the process was based on the number of records they contained. Since they only stored dead tree observations data and lacked any information on individual trees or environmental data, the logic was to use as many datapoints as possible for the model training. Due to the vast number of points – more than 13,000 points for the training and validation datasets – visual selection of training and validation data based on tree size was not considered as an option. The dead tree dataset was divided into individual years, based on the date of collection, in the form of point shape files. Raster data, for all thirteen phenological variables, was extracted for each of the point locations using the Rasterio Python library and saved into yearly data frames using GeoPandas, which is built upon the pandas and NumPy libraries. 'PROD_DIFF' was introduced as an extra parameter as the difference between total productivity ('TPROD') and seasonal productivity ('SPROD') and can be considered the baseline PPI value for seasonal productivity (i.e. values pertaining to the area designated as 'h' in Figure 3.). After data extraction, values were not normalized. We made this decision for a better practical application of the raster data and the interpretation of the results. All seasonal dates pertaining to certain phenological events, such as starting day of the phenological season, day of maximum value in the season and the day of the end of the phenological season were formatted for all datasets. 0 values indicate the beginning of the year in question (for which the phenology parameters were extracted). Minus values indicate that the phenological season started before the start of the year in question.

The number of records for each dataset after performing the data cleaning steps is summarized in Table 5.:

Table 5. Number of points after data cleaning

Dataset	No. Points
2017	6377
2018	3322
2019	4219
Google Earth Pro	498

4.2.2. Selection and tuning of hyperparameters

Before the application of a model on unseen data for any selected machine learning task, there is an important procedure where a special group of parameters, called hyperparameters are selected, tested, and tuned. Hyperparameters are not directly related to the data on which the model is trained, although they have a direct influence on accuracy of the chosen model and they specify the details of the learning process, such as the learning rate of the choice of the optimizer [37]. At first model training of One-Class SVMs might seem easy to perform and can be solved with numerous different approaches, but it is not self-explanatory even if there is previous knowledge regarding machine learning and the different steps that lead the eventual model training and later application (30). Because of this, hyperparameter selection and tuning is crucial since it affects the performance of the model. Another issue with hyperparameters is their co-dependency, meaning that some parameters will simply not work with other parameters. Due to the lack of a concrete methodology, consensus systematic framework and experience, this part of the machine learning task can be rather ad hoc [37].

Using 2017 as a base year (the training data for the classifier), the classification was performed using the other two years with overlapping data – 2018 and 2019 – as validation datasets. The data was first used to generate a set of models, with different hyperparameters, based on the predefined parameter matrix as part of the hyperparameter tuning. This parameter matrix was defined during the preliminary data analysis phase with the aid of the pertaining part of the scikit-learn documentation of the classifier [28]. A brief description of the hyperparameters used in this thesis can be found in Table 6.:

Table 6. Hyperparameters and their description (from the scikit-learn documentation)

Hyperparameters

Parameter	Description
kernel	{'linear', 'poly', 'rbf', 'sigmoid', 'precomputed'} or callable, default='rbf' Specifies the kernel type to be used in the algorithm. If none is given, 'rbf' will be used. If a callable is given it is used to precompute the kernel matrix.
gamma	{'scale', 'auto'} or float, default='scale' Kernel coefficient for 'rbf', 'poly' and 'sigmoid'. - if gamma='scale' (default) is passed then it uses 1 / (n_features * X.var()) as value of gamma, - if 'auto', uses 1 / n_features - if float, must be non-negative.
nu	float, default=0.5 An upper bound on the fraction of training errors and a lower bound of the fraction of support vectors. Should be in the interval (0, 1]. By default, 0.5 will be taken.

Table 7. shows the initial parameter matrix used for the experimentation phase of the thesis:

Table 7. Initial hyperparameters

gamma	kernel	nu
scale	linear	0.01
auto	poly	0.25
0.5	rbf	0.5
2.5	sigmoid	0.75

5.0	precomputed	0.99

Based on our experiences during manual experimentation and model testing we decided, for the sake of simplicity, to reduce the size of the original hyperparameter matrix. This reduced matrix is presented in the Chapter 4 as part of our results.

In order to obtain the best set of hyperparameters we used the GridSearchCV class from the scikit-learn library [38]. The GridSearchCV class performs an exhaustive search over specified parameter values for an estimator and returns the best possible combination of hyperparameters for the model. This process of cross-validation helps evaluate how well the model generalizes to unseen data and it reduces the risk of overfitting to a single train-test split. In the case of this thesis, the data was split manually (to individual years), while in machine learning practice it is common to randomly split your data into training and testing sets using a designated function [39]. Despite having separate datasets for training, validation and testing and not using a split function we still used this method to search for hyperparameters, because we wanted to see which hyperparameters affected the model predictions the most for the different datasets. The hyperparameter search was also done for comparison, on the Google Earth Pro observations. We used these results as indications on what parameters could work for classification of the different datasets and what could be a common ground in terms of classification accuracy for both living and dead trees. This part of the process also helped reduce the range of values used for different hyperparameters.

4.2.3. Feature selection

After we selected multiple sets of hyperparameters, models were then evaluated on the validation datasets in the next step, where feature importance scores were ranked and saved using the SelectKBest class from the same scikit-learn library to see which features are the most important in the classification of dead trees and to see which features made a difference when distinguishing dead trees from the live ones. For each model, feature importance was evaluated. Two functions from the feature_selection module of the scikit-learn library was used to report the feature importance scores: ‘f_classif’ and ‘mutual_info_classif’, both of which are used in classification tasks. The ‘f-classif’ function returns the ANOVA F-value while the ‘mutual_info_classif’ function returns the

mutual information (MI) between two random variables [40]. The F-value is the value used in the analysis of variance (ANOVA). It is calculated by dividing two mean squares. This calculation determines the ratio of explained variance to unexplained variance [41], which is a non-negative value, that measures the dependency between the variables. It is equal to zero if two random variables are independent, while higher values mean higher dependency. Because of the nature of the data analysis and to simplify our methodology, after thorough experimentation we decided to pair the phenology parameters in the form of sets (to avoid repetition). The pairs were made with the use of the 'itertools' python library using the 'combinations' class. These phenology parameter pairs were then used to subset the initial datasets and we used the 2017 data for these parameter pairs as training set, - in the fashion described earlier – 2018 and 2019 data for validation and on Google Earth Pro for testing. These pairs yielded the results that are presented in Chapter 4.

4.2.4. Decision boundaries

After conceiving the phenology parameter pairs in the manner described above, we used them with the selected hyperparameters to plot decision boundaries for each pair possible combination of phenology parameters and hyperparameters. These decision boundaries are technically the hyperplanes we described in the literature review section of the thesis. Classification algorithms separate the data based on these hyperplanes. We presented our findings in Chapter 4, where we further elaborated on them.

To automate workflow and the output generation, a python script was written. The script was divided into three files, containing the functions, the settings and a main function that runs the actual analysis respectively.

The visual summary of the workflow can be seen in Figure 10:

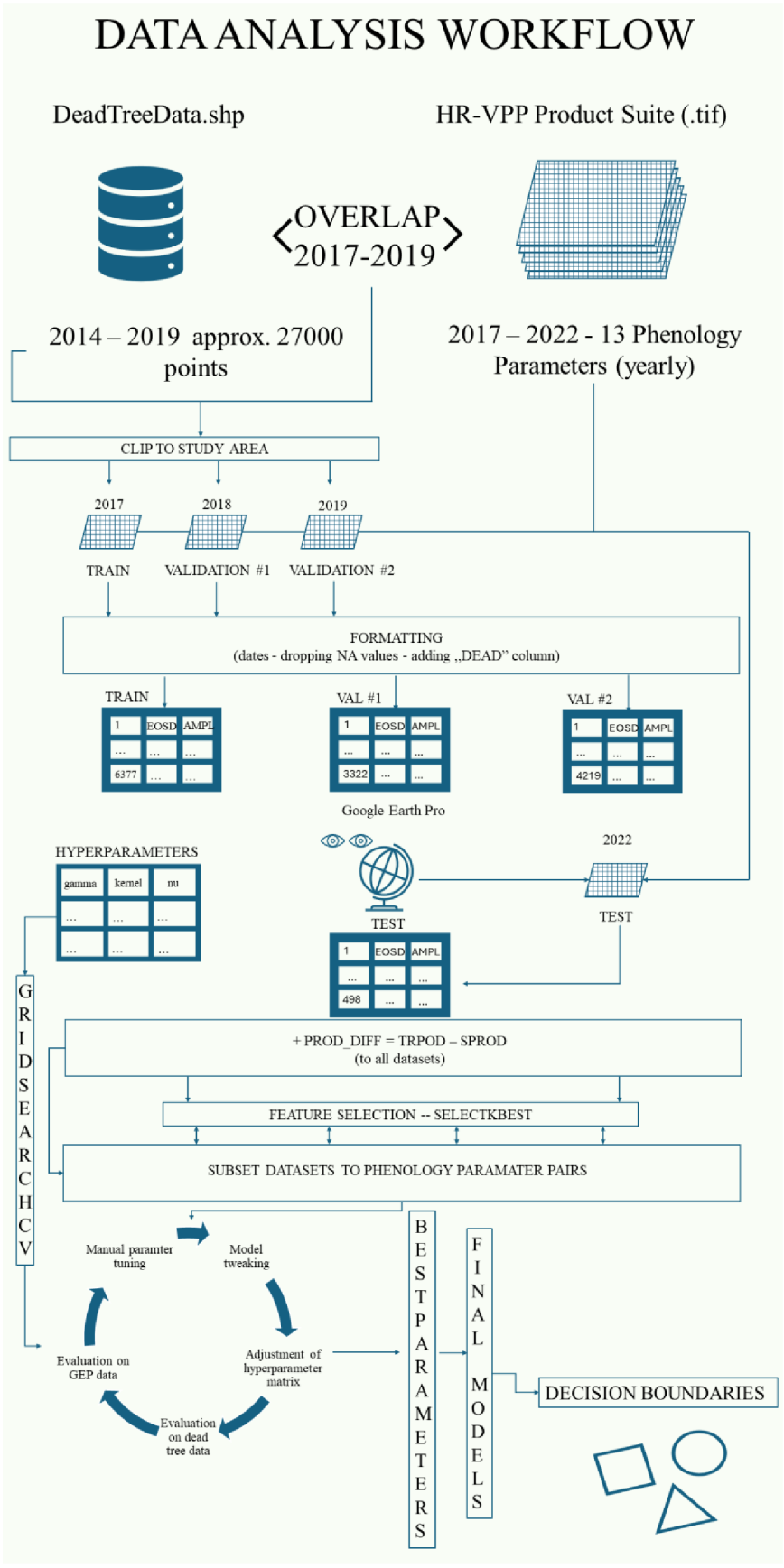

Figure 1. Summary of data analysis workflow

5. Results

5.1. The importance of hyperparameter tuning

The variety of hyperparameters to choose from, the range of their values, and the difficulty of selecting the appropriate phenology parameters made the process of hyperparameter tuning, feature selection and thus model selection and evaluation difficult and time consuming. The lack of exact, scientifically based method to tune hyperparameters also made this process rather experimental and heuristic [29]. Yet, it could not be avoided, since it affects model accuracy and applicability greatly. Hyperparameter selection affected the model predictions and the results regarding the decision boundaries (presented below) and classification accuracy and in the following section we elaborated on our findings when faced with this part of the machine learning task at hand, and we shared our experiences with different hyperparameters for our One-Class SVM.

The main difference between model accuracy between the two types of datasets was mainly caused by two hyperparameters: 'kernel' and 'nu'. These parameters define the decision function of the classifier and the boundaries set for the fraction of the support machines respectively. The 'nu' parameter sets the threshold for how many of the training samples can be classified erroneously and how many of the datapoints can be used as support vectors. Support vectors are the data points that are closer to the hyperplane (i.e. the decision boundary) and help to determine its position and orientation. A higher 'nu' value allowed more trees to be misclassified, which in turn reduced model accuracy for the validation datasets (for the 2018 dataset, especially for the models chosen as final ones) but allowed more 'freedom' for outlier/novelty detection (i.e. classification of living trees). It the training and validation datasets themselves (i.e., the phenological data pertaining to the year the dead tree data was recorded) affected the accuracy of the classifier, despite using the same hyperparameters. Although hyperparameters are technically independent from the data (they influence the performance of the model), they affect the overall accuracy of the model used for the specific task on specific data. As a result, we saw, that different sets of hyperparameters yielded higher classification accuracy for different datasets. When using the entire dataset of 13 + 1 variables, the

models with the highest accuracy, according to the results of the grid search, were the ones using a 'linear' kernel and a low value of the 'nu' parameter (0.01). These models were discarded as not appropriate for application on unseen data for two reasons: we did not assume that a linear delineation of the data (whole or subset) was possible and the fact that a very low 'nu' value does not allow a lot of room for the algorithm to properly differentiate dead trees from outliers. Using such a low value for 'nu' allowed no misclassification based on the training dataset and everything was classified as dead by the algorithm. For similar reasons, a very high 'nu' (0.5 and above), did not yield meaningful results on unseen data. Because of this, different kernels were used instead of 'linear', and more manual experimentation was done for a better approximation of the 'nu' value.

Model sensitivity to the kernel hyperparameter should not come as a surprise, since SVMs are very kernel dependent when making predictions [23]. When using all phenology parameters, the dead tree data was more responsive to 'linear' and 'sigmoid' or even 'poly' (polynomial) kernels and yielded better classification accuracy, while the living trees was more responsive to 'sigmoid' and 'rbf' (radial basis function) kernels after manual experimentation. This does not mean that the classifier did not perform well on the dead datasets with the latter two types of kernels, but that better results were obtained using the former two on the training and validation sets. The 'sigmoid' kernel, with the proper 'nu' values seemed to be a proper middle ground, but the nature of the decision boundary made it less applicable for result interpretation. In finding the imperfect classifier (i.e. one with lower accuracy for an undifferentiated training and validation datasets) for the dead trees, precision on living trees was considered crucial, as it determines future applicability of the developed methodology. We assumed that the relationship between these phenology parameters is non-linear (the decision plane cannot be separated by a linear line). Dropping the 'poly' kernel had two reasons: (1) other kernels had higher accuracy when tested on the Google Earth Pro data, (2) since we did not know the degree of the data, experimenting with it was time consuming and computationally heavy. We found that the Google Earth Pro data responded better to kernels 'rbf' and 'sigmoid' – which also worked well for our training data according to GridSearchCV – and a higher value of the 'nu' parameter (<0.20). After removing 'linear' and 'poly' kernels, 'rbf' seemed to be a good candidate to produce meaningful results from the data, because of the nature of our datasets and the good preliminary results

obtained from manual experimentation. The ‘rbf’ kernel is used for data where prior knowledge on the data is not available and it can be used to capture complex relationships between the data [42] and it is the default parameter for the OCSVM classifier.

Hyperparameter ‘gamma’ determines the influence the individual datapoints have on the decision boundary, defining the width or slope of the kernel function. When the value for gamma is low, the decision boundary's curve becomes very low, making the decision region broad. Conversely, when set for higher values, the curve of the decision boundary becomes high, which creates island of decision boundaries around the datapoints [26]. For the sake of simplicity, only two values were used ‘auto’ and ‘scale’, and although numerical values were used in the initial phase of the development of the code but were later removed. We found that in contrast to Probst et. al [37], who – along with the ‘kernel’ – found this parameter the most tuneable, the value of ‘gamma’ did not influence our results significantly, but a more rigorous approach should be devised to ascertain this.

After manually experimenting with different parameter ranges, a subset of the original hyperparameters was created. This simplified matrix, seen in Table 8, only contained parameters that provided meaningful results for both types of datasets, using different pairs of phenology parameters:

Table 8. Simplified hyperparameter matrix

gamma	kernel	nu
scale	rbf	0.15
scale	rbf	0.16
auto	sigmoid	0.19

5.2. Notes on feature selection

For the ease of interpretation and to map decision boundaries for presenting results and to retrieve continuous values from the classifiers, we settled for using every

phenological parameter in combinations of two and train new classifiers using only these combinations instead of the whole set of 13 + 1 variables. Experimenting with feature selection and reduction, hyperparameter search and tuning, our results gravitated towards a set of phenology parameter pairs with a set of model hyperparameters with acceptable prediction accuracy for all the datasets. It is important to note that for solely predicting dead trees, productivity parameters were considered more important by the functions used for feature importance scoring, than for example the onset of seasonal phenological events, although these parameters are inevitably interrelated with one another. After trying multiple approaches to subset features based on their importance, we decided to use all phenology parameters in the manner described in the workflow. The main reason behind this was that feature importance scores depended on the hyperparameters and dataset, so in the end we decided to use all features in the manner described in the workflow part of this thesis.

5.3. Final models

During the evaluation of the different models produced by the workflow presented in Chapter 2 and the final python script, only accuracy was used as a metric of model accuracy. This ratio was the number of correct classifications divided by the total number of datapoints. In our case this could be either 1, meaning dead tree or –1, meaning everything else. In our case –1 was the desired output for the presumed live point dataset described later in this section. The choice was made for this evaluation because the number of correct predictions (for the ‘dead’ and ‘other’ classes) was already given, so we found that there was no need for further elaboration. Accuracy measures the frequency of how often the model correctly predicts the outcome. It is calculated by dividing the number of correct predictions and the total number of predictions [43]:

$$Accuracy = \frac{Correct\ predictions}{All\ predictions}$$

I the end we found that the classifier was able to make classifications with decent accuracies despite the undifferentiated nature of the dead tree datasets.

Table 9. shows our finding that the phenology parameters 'SOSD' – start of season day and 'MAXD' – day of maximum for the PPI - were the best for the model to determine the difference between dead trees and outliers with the highest accuracy:

Table 9. Results - SOSD - MAXD

SOSD - MAXD			ACCURACY (%)		
gamma	kernel	nu	2018	2019	Google Earth Pro
scale	rbf	0.15	77.00	91.32	90.16
scale	rbf	0.16	76.22	90,66	91.97

Another pair of phenology parameters with the same hyperparameter set was the 'EOSD' – end of season day and 'SOSD', producing slightly lower accuracies for all datasets, summarized in Table 12.:

Table 10. Results - SOSD - EOSD

SOSD - EOSD			ACCURACY (%)		
gamma	kernel	nu	2018	2019	Google Earth Pro
scale	rbf	0.15	74.95	88.01	86.14
scale	rbf	0.16	73.90	87.39	87.15

A third combination of phenology parameters with the same hyperparameter set was found with 'PROD_DIFF' as the second phenology parameter. As stated previously, 'PROD_DIFF' is the difference of total productivity ('TPROD') and seasonal productivity ('SPROD'). The accuracies for these models can be found in Table 13.:

Table 1. Results - SOSD -PROD_DIFF

SOSD – PROD_DIFF			ACCURACY (%)		
gamma	kernel	nu	2018	2019	Google Earth Pro
scale	rbf	0.15	82.30	91.30	81.93
scale	rbf	0.16	81.22	90.02	82.93

The accuracy in the case of 2018 and 2019 signified how many of the trees were considered dead by the pertaining model and for the Google Earth Pro data, it showed the percentage of the trees classified as outliers (i.e. not dead). As seen in the tables above, a slight change in the 'nu' parameter affected accuracy of the model predictions, and a slightly higher value worked better for the Google Earth Pro dataset in terms of prediction accuracy. For the size of each dataset please see Table 4.

In the following section, the results of the model validation (for 2018 and 2019) and testing (GEP) are presented in the form of decision boundaries. The yellow areas defined the decision boundaries for the two-parameter subset of the original dataset. Yellow circles represented the datapoints that were classified as dead trees, while blue ones are the outliers/novelties. The representation of the decision boundary in such manner provided a better interpretation and further application of the results. As stated previously, the 'sigmoid' kernel provided decent accuracies for some of the datasets, but we excluded them from the results of this thesis due to inconsistency (i.e. not showing up as best results for all datasets) and the difficulty in interpreting the decision boundaries.

In Figures 13-18 the decision boundaries between the phenology parameters 'MAXD' and 'SOSD' when used as feature set for the models above are presented. See Appendix 1. for the rest of the decision boundaries for the other phenology parameter combinations. As stated in the methodology, date values have been reformatted from the original raster values and '0' represents the start of the year in question:

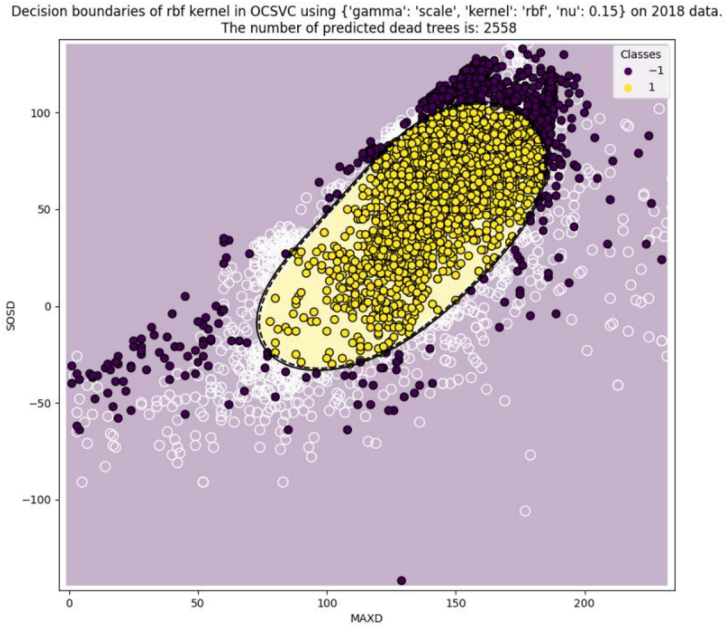

Figure 2. Decision boundary for One-Class SVM for feature set of ['SOSD', 'MAXD'], using hyperparameters: {'gamma':'scale', 'kernel':'rbf', 'nu':0.15} on validation dataset #1 (2018). '1': dead tree prediction, '-1': outlier. Accuracy: 77.00 %

Figures 11. and 12. show the decision boundaries between dead trees and outliers for the phenology parameters start of season day and the maximum day of the PPI value for 2018 for 3322 datapoints. The classifier had the following hyperparameters: {'gamma':'scale', 'kernel':'rbf', 'nu':0.15} and {'gamma':'scale', 'kernel':'rbf', 'nu':0.16} respectively. Out of 3322 datapoints, 2558 and 2532 were classified correctly, which gave the models 77.00 % and 76.21 % accuracies. The 2018 dataset lagged behind in accuracies when compared to the rest of the data used for validation or testing. This could be due to various reasons, on which we elaborated on in the Discussion.

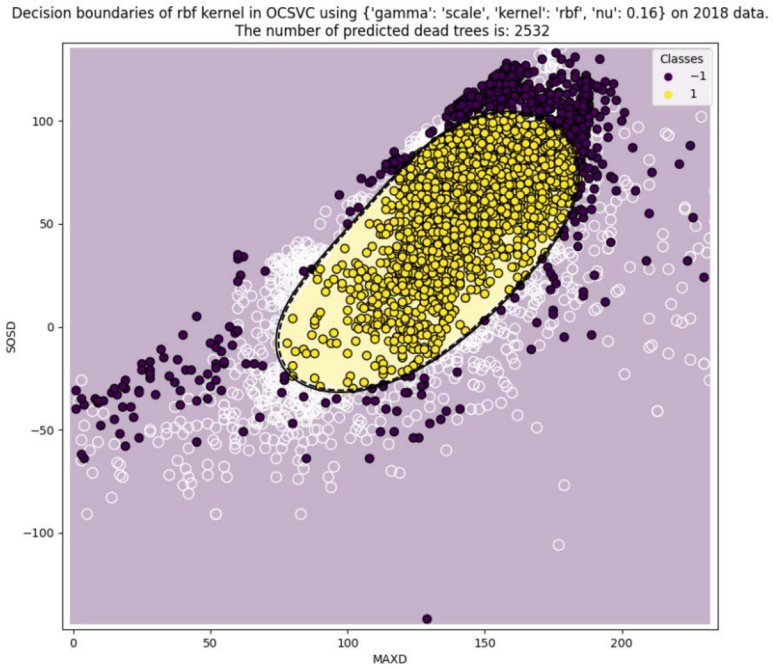

Figure 3. Decision boundary for One-Class SVM for feature set of ['SOSD', 'MAXD'], using hyperparameters: {'gamma':'scale', 'kernel':'rbf', 'nu':0.16} on validation dataset #1 (2018). '1': dead tree prediction, '-1': outlier. Accuracy: 76.21 %

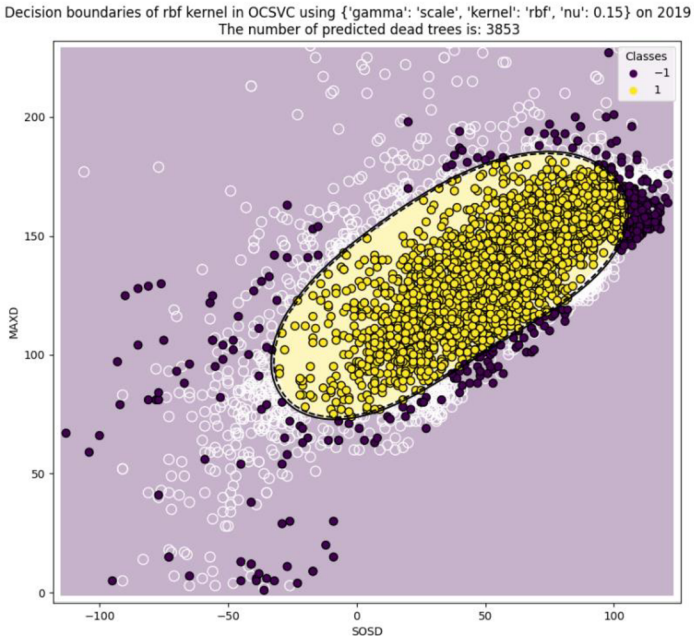

Figure 4. Decision boundary for One-Class SVM for feature set of ['SOSD', 'MAXD'], using hyperparameters: {'gamma':'scale', 'kernel':'rbf', 'nu':0.15} on validation dataset #2 (2019). '1': dead tree prediction, '-1': outlier. Accuracy: 91.32 %

Figures 13. and 14. show the decision boundaries between dead trees and outliers for the phenology parameters start of season day and the maximum day of the PPI value for 2019 for 4219 dead tree observations. The classifier had the following hyperparameters: {'gamma':'scale', 'kernel':'rbf', 'nu':0.15} and {'gamma':'scale', 'kernel':'rbf', 'nu':0.16} respectively. From the 4219 datapoints, 3853 and 3825 were classified correctly, which gave the models 91.32 % and 90.66 % accuracies respectively.

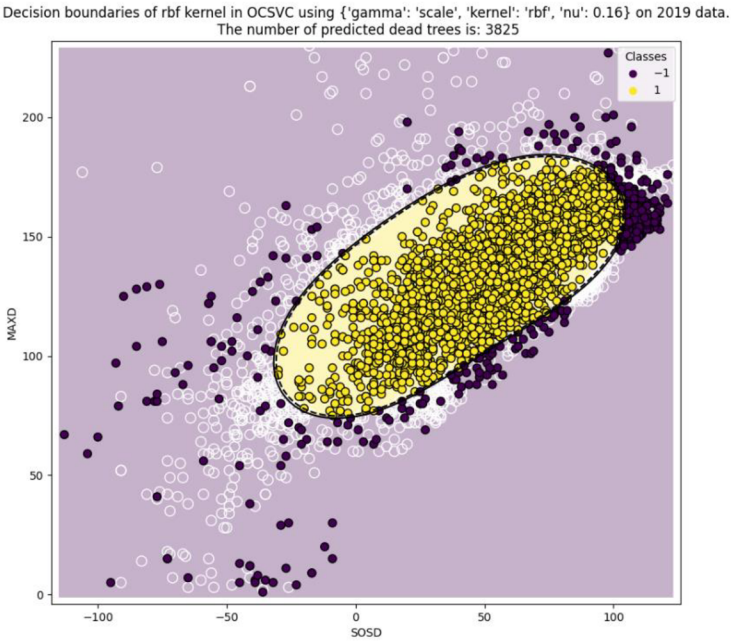

Figure 5. Decision boundary for One-Class SVM for feature set of ['SOSD', 'MAXD'], using hyperparameters: {'gamma':'scale', 'kernel':'rbf', 'nu':0.16} on validation dataset #2 (2019). '1': dead tree prediction, '-1': outlier. Accuracy: 90.66 %

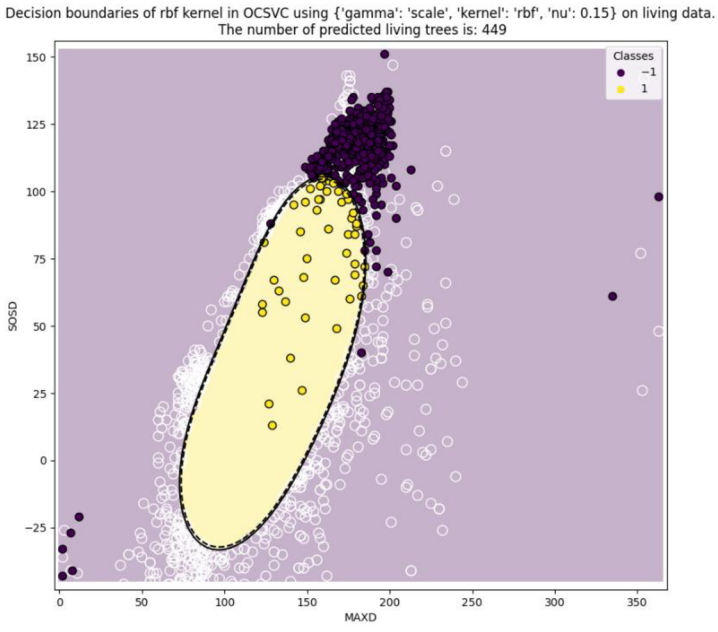

Figure 6. Decision boundary for One-Class SVM for feature set of ['SOSD', 'MAXD'], using hyperparameters: {'gamma':'scale', 'kernel':'rbf', 'nu':0.15} on test dataset (GEP - 2022). '1': dead tree prediction, '-1': outlier. Accuracy: 90.16 %

Figures 15. and 16. show the decision boundaries between dead trees and outliers for the phenology parameters start of season day and the maximum day of the PPI value for the GEP data for 2022 for 498 observations. The classifier had the following hyperparameters: {‘gamma’:’scale’, ‘kernel’:’rbf’, ‘nu’:0.15} and {‘gamma’:’scale’, ‘kernel’:’rbf’, ‘nu’:0.16} respectively. Out of 498 observations, 449 and 458 were classified correctly (as outliers or not dead), which gave the models 90.16 % and 91.97 % classification accuracies respectively.

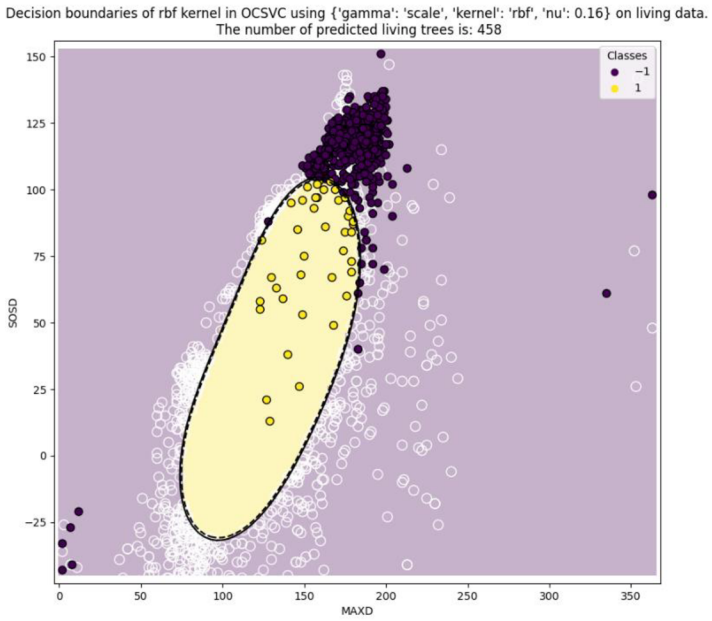

Figure 7. Decision boundary for One-Class SVM for feature set of ['SOSD', 'MAXD'], using hyperparameters: {'gamma':'scale', 'kernel':'rbf', 'nu':0.16} on test dataset (GEP - 2022). ‘1’: dead tree prediction, ‘-1’: outlier. Accuracy: 91.97 %

We found it interesting how the observations accumulated in particular areas of the decision boundary for the unseen data (GEP) compared to the datasets used for validation of the models. In our opinion, these decision boundaries could be further used for example in raster reclassification by using range values obtained from these decision boundary plots pertaining to these areas. After plotting the decision boundaries their values could be saved into variables which we found could be useful in output in potential future application of our approach. This could mean using these boundaries in mapping

potential areas that are affected by oak decline, making the PPI useful in mapping and monitoring oak dieback in the future for agroforestry system such as the *montado*.

6. Discussion

In this thesis we hypothesized that the PPI can be used as an indicator to map cork oak decline using data from the HR-VPP remote sensing product suite and a OCSVM classifier. We found that the OCSVM classifier of the scikit-learn library was able to make classifications with significant predictions when used with different 'kernel' and 'nu' hyperparameter values. SVMs are innately kernel dependent, but not all the available ones worked well when we tried to differentiate dead trees from living ones with a significant accuracy. The results we presented were for the default kernel of the classifier ('rbf'). Other kernels such as 'sigmoid' showed promising results, but we excluded it from the results, because the results with these models were not as consistent as the ones presented for 'rbf'. Nonetheless, we suggest further experimentation using the 'sigmoid' kernel as we found it to have good potential when using it for differentiating living evergreen oaks from dead ones.

Our results showed that the most important phenology features to determine the difference between dead and living cork oak trees, when using the dead tree observations data are related to the start and the end of season dates. Besides these parameters, the date of the PPI maximum day showed useful results (in conjunction with the start of the season date). A third pair of parameters included the difference between total and seasonal productivity which we named 'PROD_DIFF' and represents the baseline signal for the PPI with the starting day of the phenology season. We presented two sets of hyperparameters that worked well on unseen data. Using these models, we plotted the decision boundaries of these phenology parameter pairs. We suggested that using these decision boundaries, areas that are potentially affected by oak decline could me mapped using range values based on the decision boundaries. The decision boundaries, combined with other environmental variables, such as climatic data, could be further improved and used for creating maps of areas that are potentially affected by oak dieback. Based on these findings we suggested that the PPI, a novel vegetation index could be used to map tree mortality in our study area and potentially can be used in similar areas where cork oak is also the dominant tree species.

Although, the OCSVM classifier was able to make accurate predictions on tree mortality, but the value of the classification accuracy had slight changes between years. Further studies should include tree mortality data and its relationship with tree size and PPI pixel size to understand if there is a relationship between them and how that can affect the yearly results of productivity. The makeup of each yearly dataset could vary based on the size of the trees, which might cause a difference in classification accuracy, as smaller trees cannot be accounted for in the spatial resolution of the HR-VPP product suite. We suggest that besides tree size, climatic differences between individual years might have also affected the results. The following figure shows the years (datasets) in a coordinate system of deviation from average annual temperature and precipitation [44]:

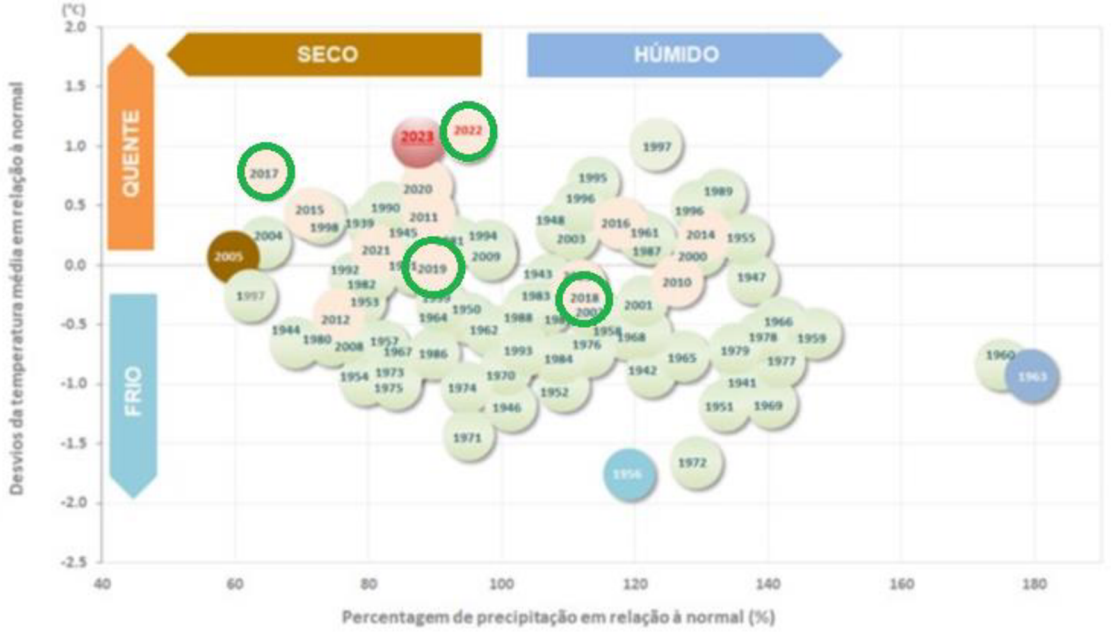

Figure 8. Climatic classification of years in Portugal based on the deviation from annual average precipitation and temperature (Source: IPMA [44])

Figure 17. suggests that three out of four of our datasets can be considered drier and warmer than the average (middle of the figure), while the 2018 dataset is closer to the average. These differences found in accuracy related to the 2018 dataset could be justified by these climatic differences.

Defining the starting and ending points of a growing season for the HR-VPP product suite was based on the moment in time when the seasonal amplitude exceeds a user-defined value. These threshold values, according to Tian et al. [23], are defined as 25 % and 15 % for start of season and for the end of the vegetation season respectively.

We found that the 'SOSD' value is one of the best phenology parameters in the differentiation of dead trees from living ones. In our view, if the 'AMPL' value of a pixel exceeds the thresholds discussed above, and the pixel contains a tree with a right size, we can use this information to speculate, that the tree is alive, and it is producing biomass. Of course, this works only for a certain tree size with a certain size of canopy. Further work would need to be done to have a better conclusion, but our findings enforce the importance of the choosing dates (datasets) with the same climatic conditions to have comparable results.

The lack of living tree observations or any other type of data further limited the usability and application of the recorded data, but our models yielded useful results despite this limitation. Because the points for the Google Earth Pro data were identified using a grid based on the resolution of the rasters, each datapoints represents a tree that fits a 10-meter by 10-meter pixel. On the other hand, the training and validation sets, had a more random spatial distribution. Another limitation that can be attributed to the spatial resolution of the HR-VPP product is the model's inability to work with smaller trees and woodlands/forests with lower canopy closure.

As previously discussed, SVMs depend highly on the set of hyperparameters used to train the model, especially on the 'kernel' and 'nu' parameters. This was apparent during the parameter tuning process, and the difficulty faced when trying to obtain meaningful results, applicable for all types of data used in the thesis.

7. Conclusions

In our work we hypothesized that the Plant Phenology Index (PPI) can be used as an indicator to map cork oak decline using data from the Copernicus Sentinel 2 High Resolution Vegetation Phenology and Productivity remote sensing product suite and a One-Class Support Vector Machine classifier. We concluded that the classifier handled the classification task of dead and living trees with applicable accuracy, and we presented different hyperparameter and phenology parameter pair combinations as our results. We concluded that based on these results, the PPI can be used as an indicator of annual phenological changes and has a potential as an indicator to cork oak dieback.

In our view, the results presented in this thesis could be used in future research and monitoring efforts. The models with the given phenology feature combinations could yield reasonable classification results when using the HR-VPP product suite for classification of evergreen oak dieback based on their remote sensed phenology parameters. Using these results and data, the script developed can be a useful tool in decision boundary delimitation for data for both living and dead trees. The script, with further improvements, can be used to automate the whole process from data acquisition – through the WEkEO portal's REST API – to hyperparameter search, feature importance evaluation, decision boundary delineation and producing potential mapping outputs in the form of shapefiles or raster reclassification based on decision boundary values at the very end of the pipeline. The decision boundary coordinates could be saved into variables and returned for further calculations, if necessary, for example in the raster classification mentioned above. Such practical application would be useful, especially in the Iberian Peninsula, where evergreen oak has been affected by dieback for decades and it is expected to worsen due to climate change.

Besides our results, we also presented the limitations of our work and included our comments and suggestions on how to improve and tackle such limitations in the future when using these results. Further research is suggested on the usage of different kernel types for classification tasks using the PPI. We also suggested taking climatic variables into consideration when using the PPI in such manner as in this thesis.

8. References

1. Huete AR (2012) Vegetation Indices, Remote Sensing and Forest Monitoring. Geography Compass 6(9): 513–32.

2. Jin H, Eklundh L (2014) A physically based vegetation index for improved monitoring of plant phenology. Remote Sensing of Environment 152:512.

3. Kerr JT, Ostrovsky M (2003) From space to species: ecological applications for remote sensing. Trends in Ecology and Evolution 6: 299–305.

4. HR-VPP Product User Manual Seasonal Trajectories and VPP parameters, issue 2.3 - Copernicus Land Monitoring Service

5. Ministério Da Agricultura Do Desenvolvimento Rural E Das Pescas. (2001) Diário da República – I Série-A N.° 121 – 15 de Maio de 2001 Artigo 13.°. Lisboa.

6. Helena Guimarães M, Pinto-Correia T, De Belém Costa Freitas M, Ferraz-de-Oliveira I, Sales-Baptista E, Da Veiga JFF, et al (2023) Farming for nature in the Montado: the application of ecosystem services in a results-based model. Ecosystem Services 61:101524.

7. Cerqueira Y (2014) Social-ecology of rural abandonment: farmers´ perceptions to ecosystem services [PhD.]. Porto: Faculdade de Ciências da Universidade do Porto. 220 p.

8. ICNF (2013) IFN6 – Áreas dos usos do solo e das espécies florestais de Portugal continental. Resultados preliminares. , 34 pp, Instituto da Conservação da Natureza e das Florestas. Lisboa.

9. Pinto-Correia T, Ribeiro N, Sa-Sousa P (2011) Introducing the montado, the cork and holm oak agroforestry system of Southern Portugal. Agroforestry Systems 82: 99-104.

10. Laporta L. It's a keeper (2021) Valuing the carbon storage service of Agroforestry ecosystems in the context of CAP Eco-Schemes. Land Use Policy 109 (C).

11. Díaz-Villa MD, Marañón T, Arroyo J, Garrido B (2003) Soil seed bank and floristic diversity in a forest-grassland mosaic in southern Spain. Journal of Vegetation Science 14(5): 701–709.

12. Batista T, de Mascarenhas JM, Mendes P (2017) Montado's ecosystem

functions and services: the case study of Alentejo Central – Portugal. The Problems of Landscape Ecology 44: 15-27.

13. De Sampaio E Paiva Camilo-Alves C, Da Clara MIE, De Almeida Ribeiro NMC (2013) Decline of Mediterranean oak trees and its association with Phytophthora cinnamomi: a review. European Journal of Forest Research. 132(3): 411–32.

14. Tilly N, Reddig F, Lussem U, Bareth G (2020) First investigation of mediterranean oak tree vitality with high-resolution WorldView-3 satellite data: comparing ten vegetation indices and three machine learning classifiers. The International Archive of the Photogrammetry, Remote Sensing and Spatial Information Sciences. XLIII-B3-2020:1069–76.

15. Navarro A, Catalao J, Calvao J (2019) Assessing the Use of Sentinel-2 Time Series Data for Monitoring Cork Oak Decline in Portugal. Remote Sensing. 11(21): 2515.

16. Xue J, Su B (2017) Significant Remote Sensing Vegetation Indices: A Review of Developments and Applications. Journal of Sensors 2017: 1-17.

17. Karkauskaite P, Tagesson T, Fensholt R (2017) Evaluation of the Plant Phenology Index (PPI), NDVI and EVI for Start-of-Season Trend Analysis of the Northern Hemisphere Boreal Zone. Remote Sensing 9(5): 485.

18. Phenology - Merriam-Webster dictionary. In. Available from: https://www.merriam-webster.com/dictionary/phenology: Accessed 2024-04-25.

19. Forrest J, Miller-Rushing AJ (2010) Toward a synthetic understanding of the role of phenology in ecology and evolution. Philosophical Transactions of the Royal Society B 365: 3101–3112.

20. Abbe C (1905), United States. Department of Agriculture., United States. Weather Bureau. A first report on the relations between climates and crops. Vol. no.36 (1905). Washington: Govt. Print. Off; 1905. Available at: https://www.biodiversitylibrary.org/item/51388: Accessed 2024-04-25.

21. Determining the usefulness of the Copernicus High-Resolution Vegetation Phenology and Productivity Product (HR-VPP) with official agricultural data on cropland in case of the 2018 drought in the Federal State of Saxony (2023) Germany. Journal of Water and Climate Change 14(11): 3931–49.

22. Tian F, Cai Z, Jin H, Hufkens K, Scheifinger H, Tagesson T, et al (2021) Calibrating vegetation phenology from Sentinel-2 using eddy covariance, PhenoCam, and

PEP725 networks across Europe. Remote Sensing of Environment 260:112456.

23. Pichler M, Hartig F (2023) Machine learning and deep learning—A review for ecologists. Methods in Ecology and Evolution 14(4): 994–1016.

24. Byer S, Jin Y (2017) Detecting Drought-Induced Tree Mortality in Sierra Nevada Forests with Time Series of Satellite Data. Remote Sensing 9(9): 929.

25. Guo Q, Kelly M, Graham CH (2005) Support vector machines for predicting distribution of Sudden Oak Death in California. Ecological Modelling 182(1): 75–90.

26. Al-Mejibli IS, Alwan JK, Abd DH (2020) The effect of gamma value on support vector machine performance with different kernels. International Journal of Electrical and Computer Engineering (IJECE) 10(5): 5497.

27. Patle A, Chouhan DS (2013) SVM kernel functions for classification. International Conference on Advances in Technology and Engineering (ICATE). Mumbai, India: IEEE p. 1–9. Available from: http://ieeexplore.ieee.org/document/6524743/

28. OneClassSVM - Unsupervised Outlier Detection. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.svm.OneClassSVM.html: Accessed 2024-04-25.

29. Mack B (2017) Applied One-Class Classification of Remote Sensing Data [PhD.]. Berlin: Institut für Geographische Wissenschaften der Freien Universität Berlin. 80 p.

30. easyclimate: Easy Access to High-Resolution Daily Climate Data for Europe - an R package. Available at: https://cran.r-project.org/web/packages/easyclimate/index.html: Accessed 2024-04-25.

31. Meteostat Python Package. Available at: https://github.com/meteostat/meteostat-python/blob/master/README.md: Accessed 2024-04-25.

32. Carta Administrativa Oficial de Portugal - CAOP2023 (Continente) . Available at: https://dados.gov.pt/pt/datasets/carta-administrativa-oficial-de-portugal-caop2023-continente/: Accessed 2024-04-25.

33. Compania das Lezírias - Valores Naturais. Available at: https://www.cl.pt/storage/pdfs/companhialezirias-apresentacao-compressed.pdf/: Accessed 2024-04-25.

34. Sentinel-2 - In succession to SPOT and Landsat. Available at: https://dataspace.copernicus.eu/explore-data/data-collections/sentinel-data/sentinel-2/: Accessed 2024-04-25.

35. New HR Vegetation Phenology and Productivity service. Available at: https://www.wekeo.eu/use-cases/hr-vegetation-phenology-and-productivity-service/: Accessed 2024-04-25.

36. Overview of Supervised Learning model SVM (support vector machines). Available at: https://medium.com/@hakobavjyan/overview-of-supervised-learning-model-svm-support-vector-machines-20b683a4eaf/: Accessed 2024-04-25.

37. Probst P, Boulesteix AL, Bischl B. Tunability: Importance of Hyperparameters of Machine Learning Algorithms.

38. GridSearchCV - Exhaustive search over specified parameter values for an estimator. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html/: Accessed 2024-04-25.

39. Split arrays or matrices into random train and test subsets. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html/: Accessed 2024-04-25.

40. Mutal Information. In. Available at: https://en.wikipedia.org/wiki/Mutual_information/: Accessed 2024-04-25.

41. ANOVA - F-value. Available at: https://www.ibm.com/docs/en/cognos-analytics/11.1.0?topic=terms-f-value/: Accessed 2024-04-25.

42. How to Select the Type of Kernel for a SVM?. Available at: https://www.baeldung.com/cs/svm-choose-kernel/: Accessed 2024-04-25.

43. Lorena LHN (2014). Seleção De Atributos Em Problemas De Classificação Unária [PhD.]. Universidade Federal De São Paulo Instituto de Ciência e Tecnologia 64 p.

44. IPMA. Available at: https://www.ipma.pt/pt/publicacoes/boletins.jsp?cmbDep=cli&cmbTema=pcl&idDep=cli&idTema=pcl&curAno=-1/: Accessed 2024-04-25.

9. Appendices

9.1. Appendix 1. - Decision Boundaries

The following plots show the decision boundaries between the phenology parameters ‘SOSD’ and ‘EOSD’ when used as feature set for the models deemed the best:

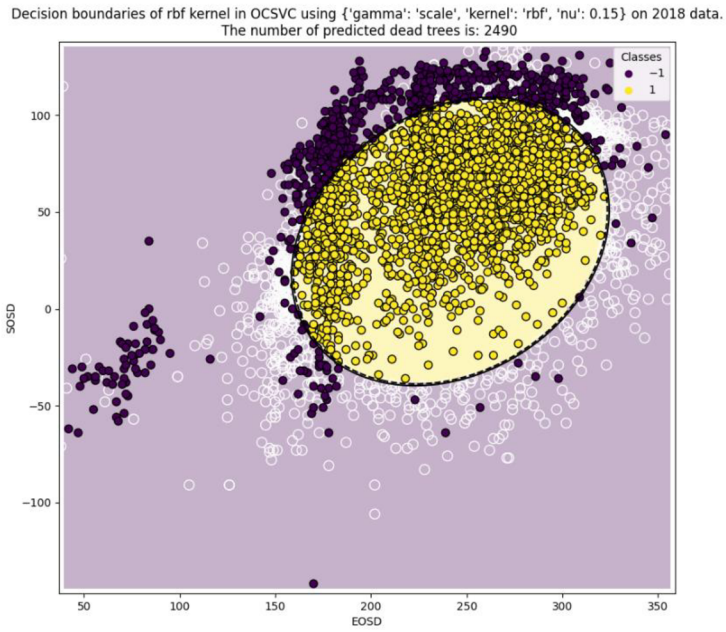

Figure A1. 1. Decision boundary for One-Class SVM for feature set of ['SOSD', 'EOSD'], using hyperparameters: {'gamma':'scale', 'kernel':'rbf', 'nu':0.15} on validation dataset #1 (2018). ‘1’: dead tree prediction, ‘-1’: outlier. Accuracy: 74.95 %

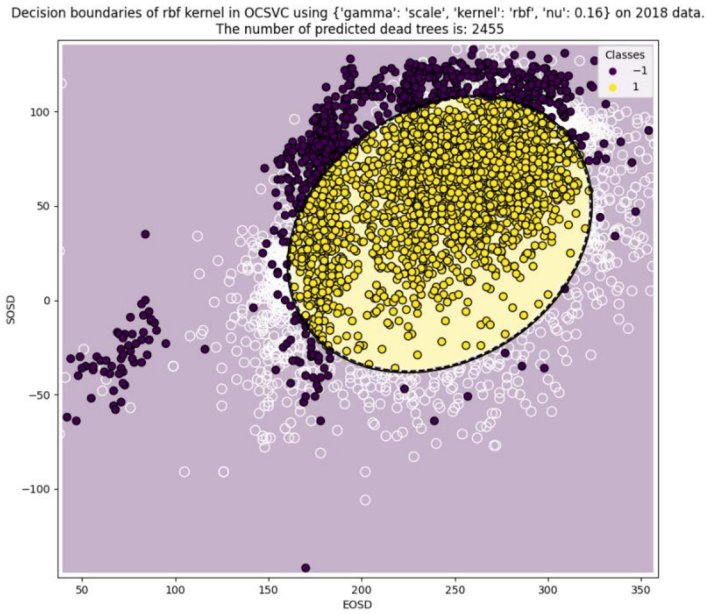

Figure A1. 2. Decision boundary for One-Class SVM for feature set of ['SOSD', 'EOSD'], using hyperparameters: {'gamma':'scale', 'kernel':'rbf', 'nu':0.16} on validation dataset #1 (2018). '1': dead tree prediction, '-1': outlier. Accuracy: 73.90 %

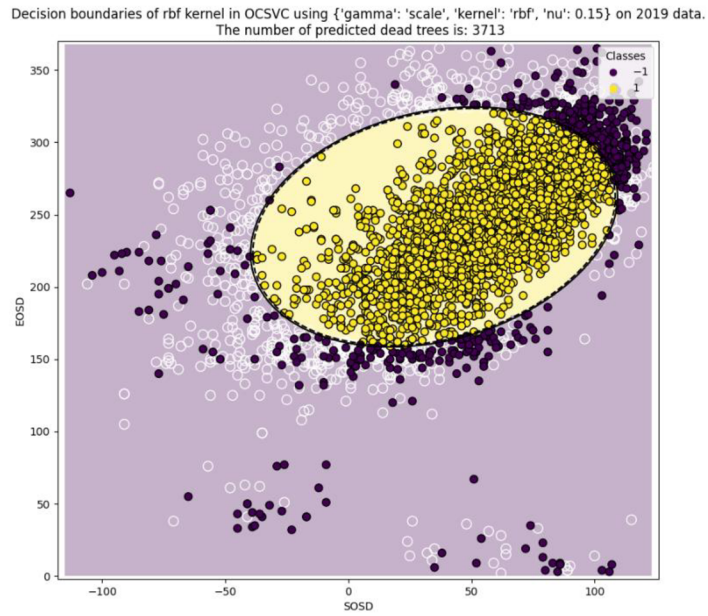

Figure A1. 3. Decision boundary for One-Class SVM for feature set of ['SOSD', 'EOSD'], using hyperparameters: {'gamma':'scale', 'kernel':'rbf', 'nu':0.15} on validation dataset #2 (2019). '1': dead tree prediction, '-1': outlier. Accuracy: 87.82 %

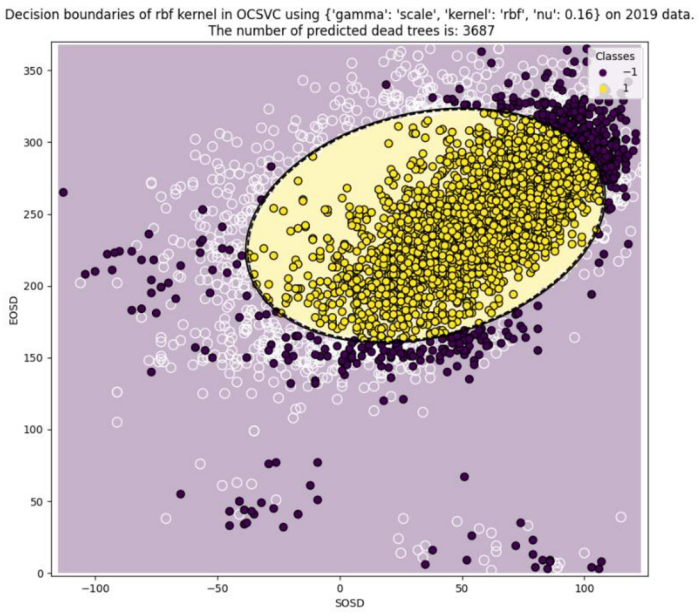

Figure A1. 4. Decision boundary for One-Class SVM for feature set of ['SOSD', 'EOSD'], using hyperparameters: {'gamma':'scale', 'kernel':'rbf', 'nu':0.16} on validation dataset #2 (2019). '1': dead tree prediction, '-1': outlier. Accuracy: 87.20 %

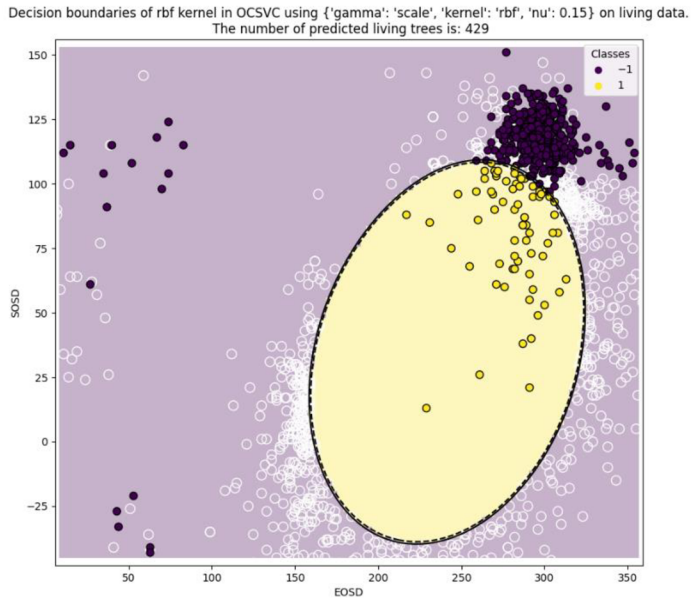

Figure A1. 5. Decision boundary for One-Class SVM for feature set of ['SOSD', 'EOSD'], using hyperparameters: {'gamma':'scale', 'kernel':'rbf', 'nu':0.15} on test dataset (GEP - 2022). '1': dead tree prediction, '-1': outlier. Accuracy: 86.14 %

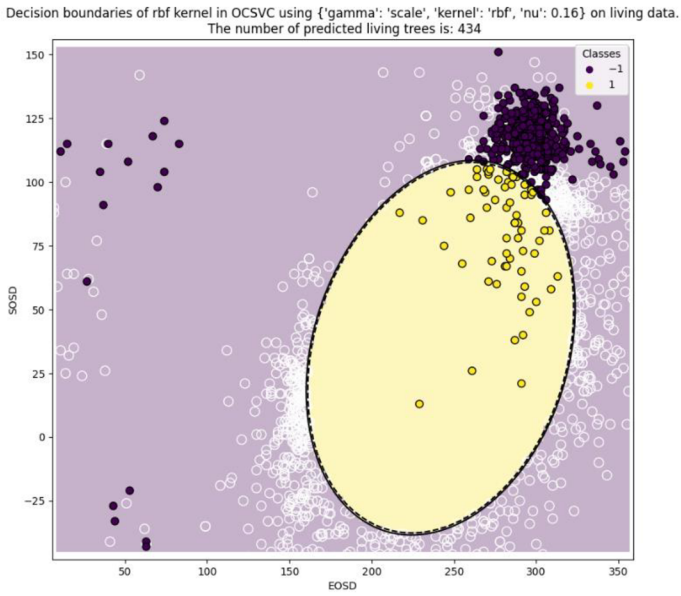

Figure A1. 6. Decision boundary for One-Class SVM for feature set of ['SOSD', 'EOSD'], using hyperparameters: {'gamma':'scale', 'kernel':'rbf', 'nu':0.16} on test dataset (GEP - 2022). '1': dead tree prediction, '-1': outlier. Accuracy: 87.14 %

The following plots show the decision boundaries between the phenology parameters 'SOSD' and 'PROD_DIFF' when used as feature set for the models deemed the best:

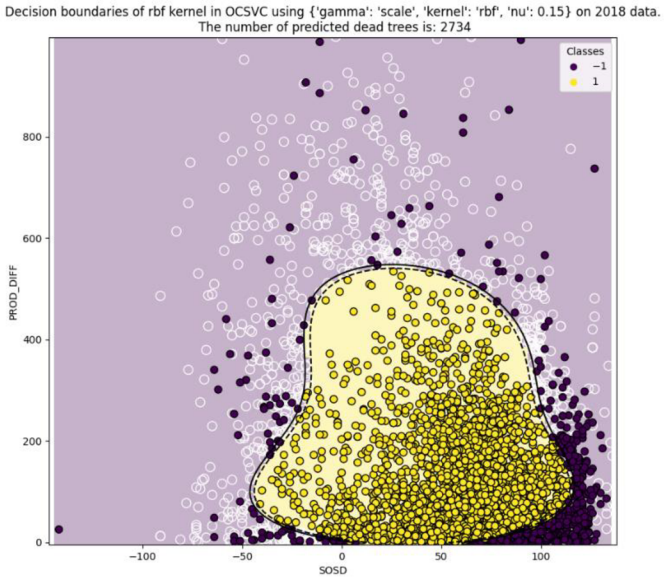

Figure A1. 7. Decision boundary for One-Class SVM for feature set of ['SOSD', 'PROD_DIFF'], using hyperparameters: {'gamma':'scale', 'kernel':'rbf', 'nu':0.15} on validation dataset #1 (2018). Accuracy: '1': dead tree prediction, '-1': outlier. 82.30 %

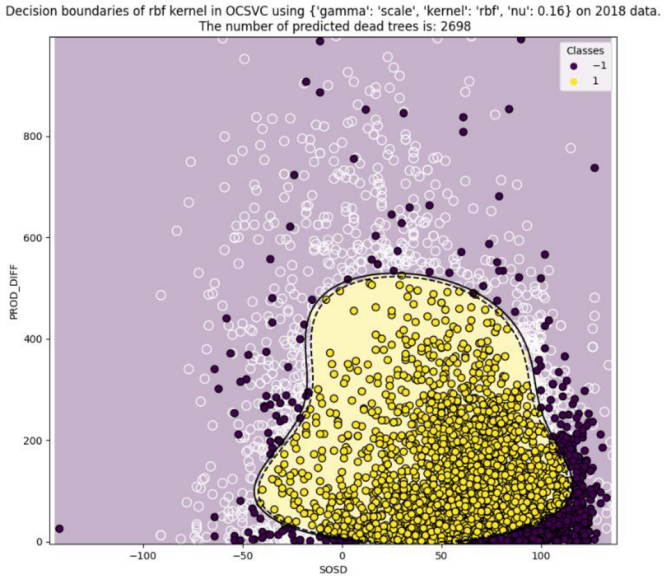

Figure A1. 8. Decision boundary for One-Class SVM for feature set of ['SOSD', 'PROD_DIFF'], using hyperparameters: {'gamma':'scale', 'kernel':'rbf', 'nu':0.16} on validation dataset #1 (2018). '1': dead tree prediction, '-1': outlier. Accuracy: 81.22 %

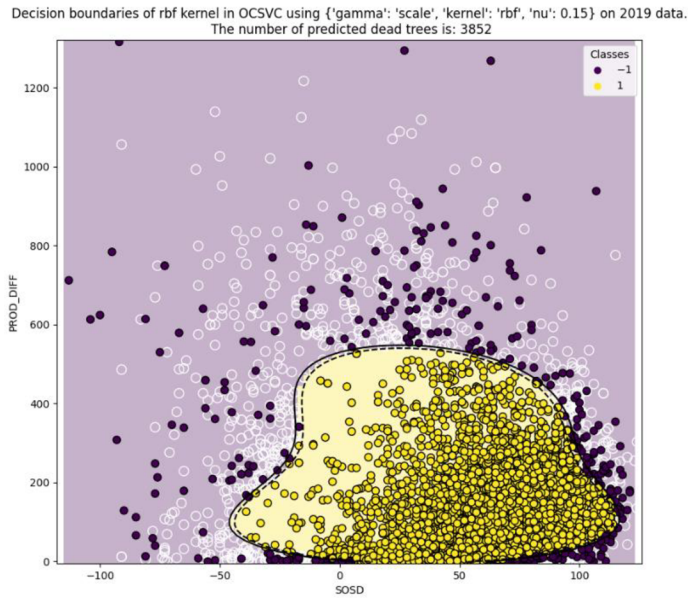

Figure A1. 9. Decision boundary for One-Class SVM for feature set of ['SOSD', 'PROD_DIFF'], using hyperparameters: {'gamma':'scale', 'kernel':'rbf', 'nu':0.15} on validation dataset #2 (2019). '1': dead tree prediction, '-1': outlier. Accuracy: 91.30 %

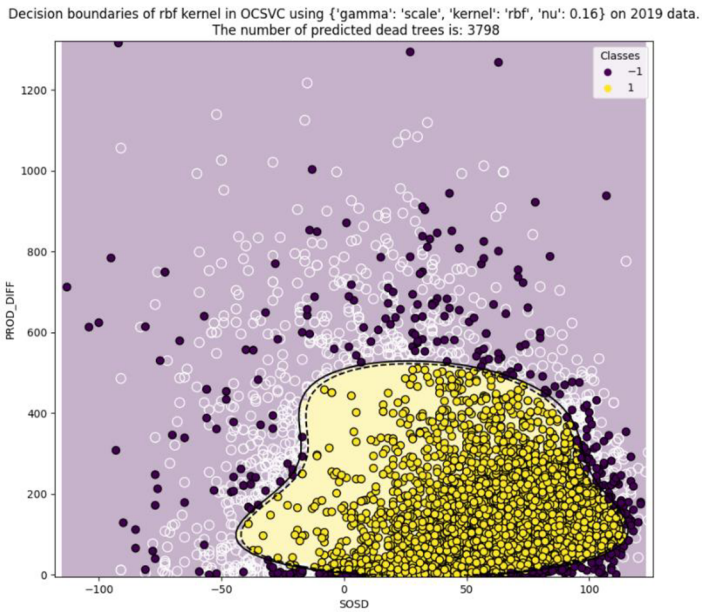

Figure A1. 10. Decision boundary for One-Class SVM for feature set of ['SOSD', 'PROD_DIFF'], using hyperparameters: {'gamma':'scale', 'kernel':'rbf', 'nu':0.16} on validation dataset #2 (2019). '1': dead tree prediction, '-1': outlier. Accuracy: 90.02 %

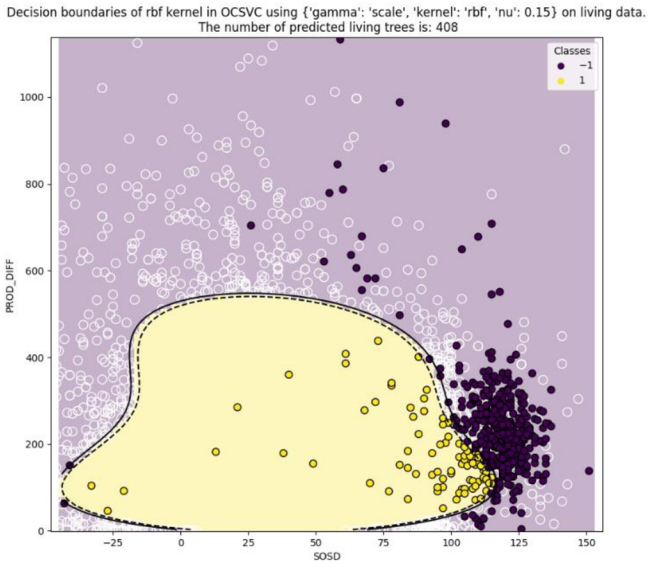

Figure A1. 11. Decision boundary for One-Class SVM for feature set of ['SOSD', 'PROD_DIFF'], using hyperparameters: {'gamma':'scale', 'kernel':'rbf', 'nu':0.15} on test dataset (GEP - 2022). '1': dead tree prediction, '-1': outlier. Accuracy: 81.93 %

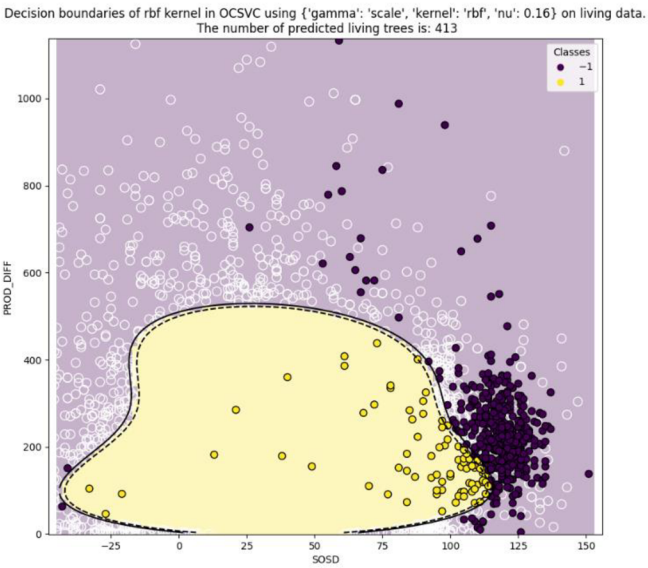

Figure A1. 12. Decision boundary for One-Class SVM for feature set of ['SOSD', 'PROD_DIFF'], using hyperparameters: {'gamma':'scale', 'kernel':'rbf', 'nu':0.16} on test dataset (GEP - 2022). '1': dead tree prediction, '-1': outlier. Accuracy: 82.93 %