Predicting biodiverse semi-natural grasslands through satellite imagery and machine learning

Adrian Baggström

Department of Physical Geography
Degree Project in Physical Geography and Quaternary Geology, 60 HE credits
NKA296
Master's Programme in Geomatics with Remote Sensing and GIS (120 credits)
Spring term 2021
Supervisor: Ian Brown; Co-supervisor: Jessica Lindgren

Stockholms universitet

Department of Physical Geography

Stockholm University

Preface

This Master's thesis is Adrian Baggström's degree project in Physical Geography and Quaternary Geology at the Department of Physical Geography, Stockholm University. The Master's thesis comprises 60 credits (two terms of full-time studies).

Supervisors have been Ian Brown and Jessica Lindgren at the Department of Physical Geography, Stockholm University. Examiner has been Gustaf Hugelius at the Department of Physical Geography, Stockholm University.

The author is responsible for the contents of this thesis.

Stockholm, 21 June 2021

Björn Gunnarson
Vice Director of studies

Predicting biodiverse semi-natural grasslands through satellite imagery and machine learning

Adrian Baggström

Abstract

Semi-natural grasslands are amongst the most biodiverse ecosystems in Europe, though their importance they are experiencing a declining trend. To monitor and assess the health of these ecosystems is generally costly, personnel demanding and time-consuming. With satellite imagery and machine learning becoming more accessible, this can offer a cheap and effective way to gain ecological information about semi-natural grasslands.

This thesis explores the possibilities to predict plant species richness in semi-natural grasslands with high resolution satellite imagery through machine learning. Five different machine learning models were employed with various subsets of spectral- and geographical features to see how they performed and why. The study area was in southern Sweden with satellite and survey data from the summer of 2019.

Geographical features were the features that influenced the machine learning models most. This can be explained by the geographical spread of the semi-natural grasslands, as well as difficulties in finding correlations in the relatively noisy satellite data. The most important spectral features were found in the red edge- and the short-wave infrared spectrums. These spectrums represent leaf chlorophyll content and water content in vegetation, respectively. The most accurate machine learning model was Random Forest when it was trained using with all the spectral- and geographical features. The other models; Logistic Regression, Support Vector Machine, Voting Classifier and Neural Network, showed general inabilities to interpret feature subsets containing the spectral data.

This thesis shows that with deeper knowledge about the satellite-biodiversity relationship and how to apply it with machine learning have the possibilities of cheaper, more efficient and standardized monitoring of ecologically valuable areas such as semi-natural grasslands.

Keywords

Machine learning, remote sensing, biodiversity, plant species richness, semi-natural grasslands, Sentinel-2.

Contents

Introduction

Semi-natural grasslands, i.e. pastures or meadows that has a tradition of cattle grazing, mowing or hay-cutting, are areas of particularly high plant species richness (Wilson et al., 2012) and provides important ecosystem services such as pollination, erosion control and carbon storage (Bengtsson et al., 2019). Even though their ecological importance, these ecosystems are currently on decline. Most semi-natural grasslands in parts of Europe have disappeared during the last century, mainly replaced by modern croplands and forest (Biró et al., 2018; Cousins et al., 2015). Authorities like the EU recognize these threats and states in "EU Biodiversity Strategy for 2030" that low-intensity grasslands should be long-term sustainable (European Commission, 2020). The Swedish Environmental Protection Agency have one of 16 environmental objectives as "A varied agricultural landscape" and states the importance that *"biological diversity and cultural heritage assets are preserved and strengthened."* (Naturvårdsverket, 2018). A recent follow up of the environmental objectives shows that both "A varied agricultural landscape" and the related one "A rich diversity of plant and animal life" have declining trends (Naturvårdsverket, 2019).

Monitoring and assessing the status of biodiversity is vital to be able to manage ecosystems well. Long-term data of ecosystems health can be used by policymakers and scientists to make sound conclusions regarding ecologically important questions. Data regarding ecosystems health is traditionally gathered *in situ.* Whilst you get precise data about the ecosystem, this method does not upscale well. It is costly, time-consuming, demands educated personnel and though through sampling strategies. With satellite imagery becoming more accessible, and with higher temporal- and spatial resolution, it has the potential for cheap, efficient and unbiased information about ecosystems on a larger scale.

A suggested technique to monitor biodiversity through satellite imagery is the spectral variation hypothesis (SVH). SVH assumes that if there is a spatial variation in the reflection, there is spatial variation in the environment (Palmer et al., 2002). This is based on the ecological concept that there is a relationship between environmental heterogeneity and biodiversity (Tamme et al., 2010). Spatial variation in spectral reflection, hereafter spectral variation, can be measured in different way, e.g. variation in specific satellite bands, vegetation indices like normalized difference vegetation index (NDVI) or principal components (Rocchini et al., 2010). SVH have been applied to various measures of biodiversity, including species diversity (Duro et al., 2014), Shannon- and Simpson Index (Fauvel et al., 2020; Oldeland et al., 2010) and plant species richness (Hall et al., 2010; Warren et al., 2014). Choosing the spatial resolution in which to measure the spectral variation can be complicated. Pixels from satellite imagery are mixed, it contains spectral information about all the things contained within that pixel (Fisher, 1997). It is rarely pure information in a pixel about a specific class, e.g., pure "forest" pixels. To capture the spectral variation in a good way, the spatial resolution should be as fine as the physical elements contained within the studied area. Using a very high-resolution image could result in that pixels are biased by shadows or partial elements such as half a tree crown, so the pixels does not generalize the studied habitat sufficiently (Rocchini et al., 2010). On the other hand, using a too coarse resolution will smooth much of the variation. Schmidtlein & Fassnacht (2017) found that the SVH does not comply on a landscape scale using the MODIS sensor with 500 meters spatial resolution, though noted that SVH might have applicability on more local scales. The Sentinel-

2 satellites, managed by the European Space Agency (European Space Agency, 2015), have spatial resolutions down to 10 meters and shows potentials of being able to capture the spectral variation of small scale parcels like semi-natural grasslands (Fauvel et al., 2020).

Semi-natural grasslands are areas represented mainly by herbaceous plants, grasses, sparsely growing woody plants and small groves or scattered trees. The soils are usually nutrient poor and rocky since other land that was richer in nutrients and more easily managed, were traditionally used for crops and housing. Grazing, mowing and hay-cutting transfer of nutrients away from the land which additionally contributes to poorer soils. Due to the continuous removal of plants, the competition over space is largely eliminated and suppresses opportunistic plant species to overtake the habitat. This allows a larger variety of plant species a chance to occupy the grassland. Should these kinds of disturbances stop, the grasslands would most likely get overgrown by shrubs and trees.

Managing grasslands in this fashion has probably been done in Europe since the Bronze Age (Feurdean et al., 2018), and became widespread during the Iron Age due to the introduction of iron tools (Eriksson, 2020). Though semi-natural grasslands are bound to human activity, there are natural grasslands in Europe that are being kept from overgrowth by poor and rocky soils, wildfires or grazing herbivores (Feurdean et al., 2018). It is through historical natural grasslands the associated biota has adopted to though time (Pärtel et al., 2005; Retallack, 2001). With the current human expansion and shift in land use for the last centuries, the semi-natural grasslands provides an important habitat to the species that once evolved in the natural grasslands (Pärtel et al., 2005). Semi-natural grasslands are intertwined with the local cultural history (Cousins et al., 2009; Eriksson, 2020; Eriksson & Cousins, 2014). Grasslands that have a long tradition of cattle grazing or hay-cutting has recreational values, not only for its beauty but they also reflect the cultural heritage and identity of the area (Eriksson & Cousins, 2014). It is also seen that semi-natural grasslands with long-term traditional management shows higher plant species richness than those that has a more recent history (Cousins et al., 2009; Cousins & Eriksson, 2002).

Machine learning is becoming an integrated part in land classification from remotely sensed data, largely due to the ability to accurately handle large amounts of complex data (Abdi, 2020; Maxwell et al., 2018). Common machine learning algorithms used for different remote sensing tasks are for example Random Forest (Belgiu & Drăgu, 2016) and Support Vector Machines (Mountrakis et al., 2011). Trying to find correlations between biodiversity and remote sensing data, with methodologies like SVH, has been an area of research for some while (Rocchini et al., 2010). To this date, few studies have implemented machine learning to investigate these relationships. Fauvel et al. (2020) uses machine learning, Sentinel data and species richness indices to predict biodiversity in grasslands. However, their focus is the prominent use of the Sentinel-1 and -2 satellites and does not investigate the performance of the specific machine learning models.

In this thesis I investigate the possibilities to predict plant species richness in semi-natural grasslands by using Sentinel-2 imagery and machine learning. Five machine learning models was employed to see if they could find correlations between plant species richness proxies and various satellite- and geographical features in southern Sweden. The questions I will try to answer in this thesis are following:

1. Which machine learning model(s) is suited to interpret the correlations between positive indicator species and the various features from semi-natural grasslands.
2. What model performance measures are viable to get the most honest results.
3. What spectral- and geographical features are important for the machine learning models and why.
4. How applicable are the spectral, geographical and ecological features chosen for this task.

I will also present the code for a program that works as a basic workflow for processing, calculating and predicting biodiverse semi-natural grasslands.

Theoretical background of machine learning

Machine learning is in essence the ability for a computer to improve on a specific task through experience. A classic explanatory example of machine learning is that of creating a spam filter for e-mail services. E-mails labelled either spam or non-spam are given to the machine learning model, from which it learns to recognize patterns of the mail content. That could be specific length of the mail or words frequently uses, maybe such as “LOTTERY” or “PHARMACY”. This could of course be detected and implemented by a human, but since the spammers will try find ways around the spam filter it will be a continuous game of cat and mouse. If instead we have a machine that learns itself and updates the filter, the e-mail service could redirect its human resources on other tasks. To evaluate how good a machine learning model is, the data is split up into a training set and a test set. The training set is given to the model for it to learn from and the test set is to test how well the model preforms on data it has not seen before.

Nowadays when an abundance of various data is available and powerful computers are relatively cheap, machine learning is an effective and cost-efficient tool to solve suitable problems. The use of machine learning algorithms in remote sensing tasks is widely adopted (Reichstein et al., 2019), that so commonly used GIS applications such as ArcGIS and QGIS have machine learning techniques such as Random Forest and Support Vector Machine integrated in them. The most common use of machine learning in remote sensing is for land use classification (Maxwell et al., 2018), but has many other uses , e.g. tree detection (Li et al., 2016) or predicting biophysical parameters of vegetation (Verrelst et al., 2012).

Five commonly used machine learning algorithms (Géron, 2019; Lawrence & Moran, 2015; Maxwell et al., 2018; Yang et al., 2019) were used in this study to research the capability of predicting indicator species in semi-natural grasslands from satellite driven data. Random Forest, Support Vector Machine, Logistic Regression and Voting Classifier are integrated in the Python library scikit-learn (Pedregosa et al., 2011) that has excellent documentation (scikit-learn, 2021). Neural networks is made comprehensible using TensorFlow (Abadi et al., 2016; TensorFlow, 2021).

Random Forest

First introduced by Breiman (2001), Random Forest is an ensemble method using Decision Trees. A Decision Tree is a series of true/false decisions that is the foundation of the tree and the leaves are the predicted classes (fig. 1). Random Forest uses randomized subsets of the original data to create several Decision Trees, and outputs either the majority vote (classification tasks) or mean value (regression tasks) of all Decision Tree predictions. The number of trees used in Random Forests is set

by the user and range from a few up to several thousand. There is usually a moment of trial and error to get right number of trees for the task. The strengths of Random Forest in remote sensing are that it has relatively low computational costs, handles high dimensional data well and frequently delivers higher accuracy than most other models (Belgiu & Drăgu, 2016; Maxwell et al., 2018).

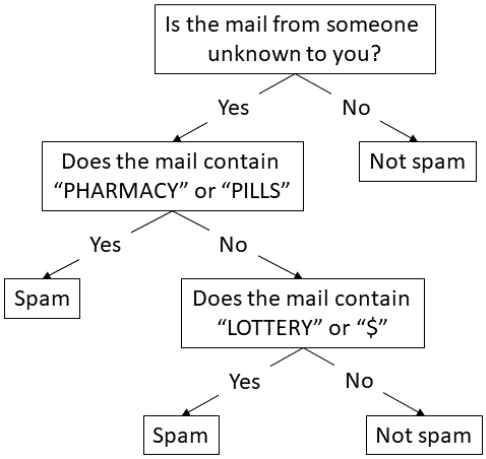

Figure 1. A simplified Decision Tree model of a not-so-good spam filter. The statements build the tree, and the prediction makes the leaves. Random Forest uses the results from several different Decision Trees to draw conclusions about the task on hand. Each Decision Tree is given a subset of the data to build it statements upon.

Support Vector Machine

Support vector machines (SVM) models are binary linear classifiers meaning that given two classes, SVMs aims to design a border (hyperplane) that maximizes the distances between the instances in both classes (Boser et al., 1992). This means that the margin to the nearest instances in both classes (the support vectors) to the hyperplane is equally large. An unclassified new instance is going to be classified depending on which side of the hyperplane it appears. To be able to handle outliers and overlaps a soft margin method can be implemented (Cortes & Vapnik, 1995). A method to overcome some of the problems with linear classification in SVM is called a kernel trick. The kernel trick transform the original data in to a higher dimensional feature space to optimize the fit of the hyperplane (Kavzoglu & Colkesen, 2009). SVM are commonly used models for classifications tasks in remote sensing (Mountrakis et al., 2011), and can provide high accuracies similar to Random Forest models (Maxwell et al., 2018).

Logistic Regression

Logistic regression is a statistical method similar to linear regression, but by adding a logistic function it can handle binary classification better. The limit range is infinite using linear regression which is problematic working with binary data since it does not scale well with extreme values (fig. 2). By adding a logistic function, it limits the value range to between 0 and 1 (fig. 2). The default threshold is at 0.5 for separating binary classes, e.g., a mail that returns with a probability of 0.6 will be classified as spam.

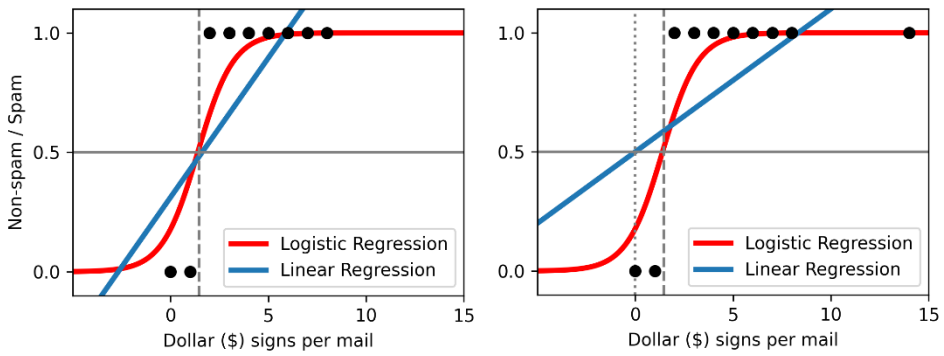

Figure 2. A comparison of linear- and logistic regression. The graphs show the number of dollar signs in mails. All mails that have two or more dollar signs are spam. In the left graph both models find a threshold (dashed line) that separates both classes successfully. In the right graph we see that if a spam mail with 14 dollar signs is included, the linear model's threshold (dotted line) shifts to 0 and will classify all mails as spam.

Voting Classifier

The Voting classifier is not a model per se, but it takes other models as inputs and evaluates the results from these and outputs a unified result. If we have three classifiers, Random Forest, SVM and Logistic Regression, and two of these classify an instance as spam and one non-spam, the Voting classifier will output it as spam. This is an example of a *hard* voting, the output is the majority vote (table 1). As of *soft* voting, the Voting classifiers takes regard to the probabilities output by each model and average the result. The output will be the class which gets the highest score (table 1).

Table 1. Probabilities for model to predict a certain class, resulting in the average score. A soft Voting classifier would classify this instance as spam, though a hard Voting classifier would have classified it as non-spam.

	Spam	**Non-spam**
Random Forest	0.4	0.6
SVM	0.8	0.2
Logistic Regression	0.4	0.6
Average score	0.53	0.47

Neural Network

Neural networks, or *artificial* neural networks, are inspired by the biological neurons in that sense that if one neuron is activated by an impulse, it fires a response to other neurons connected to it (Géron, 2019). Artificial neural networks are built up by layers, which is itself built up by neurons. The neurons in each layer has an activation function connected to it, so if it is activated, the output response is not binary but on a scale. Depending on how powerful the output response is will influence the neurons in the next layer. The response will go through the network and end with a layer consisting of a single neuron, and the response from that single neuron will result in a classification. To be able to learn and improve, the artificial neural networks iterates the responses back and forth (Rumelhart et al., 1986) and tune the weights (that adjusts the power of the response) between neurons to get as high performance as possible.

Method

Study area

The study area is in southern Sweden and covers the Stockholm region in the north-east to the region of Skåne in the south, the island of Öland included (fig. 3A). It is around 180,000 km^2 and the centre is located at 15°2’7’’ E and 57°42’57’’ N. The mean temperature is 15°C in July and -3°C in January and the annual precipitation at around 500-600 mm. Most of Sweden’s agricultural lands lies within the study area, even though the predominant land use is forestry. Here we find 80% of the semi-natural grasslands inventoried in 2019 for the national survey and the TUVA database (fig. 3B). A total of 2293 semi-natural grasslands were included in this study. The mean parcel size for each site is 26,600 m^2 and median size is 12,600 m^2.

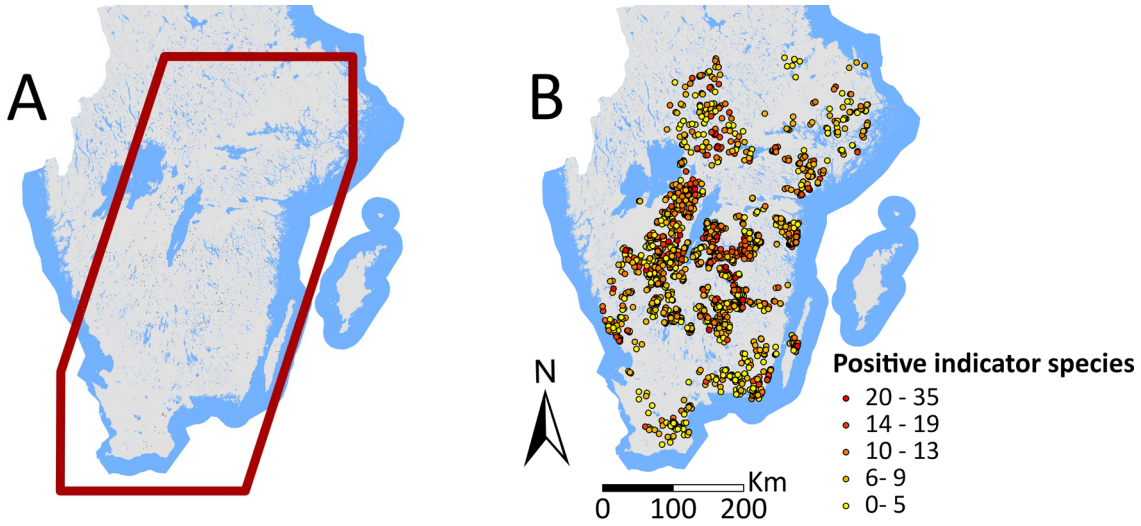

Figure 3. A) The study area in southern Sweden. The diagonal shape is due to the Sentinel-2 satellite’s path. B) The locations and number of positive indicator species for all semi-natural grasslands included in this study.

The TUVA database and positive indicator species

Between 2002-2004 a national survey of semi-natural grasslands was made by the Swedish Board of Agriculture and the County Administrative Boards on assignment by the Swedish government (Jordbruksverket, 2005). The objective was to map semi-natural grasslands with natural- and cultural values across Sweden and to present the information in an accessible database to act as a basis for various evaluations and follow-ups (Jordbruksverket, 2005). For a semi-natural grassland to get inventoried for the survey, the area should indicate qualities of long-time traditional management for pastures or meadows. Such qualities could be a representative flora, landscape historical elements, particularly old trees or characteristically pruned trees used for winter fodder (Jordbruksverket, 2017). Each inventoried instance in the database has 1866 columns containing information about location, inventory dates, management status, type of nature, cultural elements, flora, fauna and more. The survey has been updated two additional times, between 2007-2013 and 2017-2020 (Jordbruksverket, 2017), and the detailed information is accessible at the open database TUVA (Jordbruksverket, 2019).

The TUVA database data used for this thesis is the indicator species among vascular plants. Vascular plants that are listed as indicator species (Jordbruksverket, 2017: appendix 2 p. 39) are noted for

each registered semi-natural grassland (fig. 4). The list of vascular plants contains 80 species that are split up into positive- (70) and negative (10) indicator species. Positive indicator species for semi-natural grasslands are assumed to indicate that there is or has recently been long-term traditional management of the pasture or meadow (Jordbruksverket, 2005). Negative species on the list indicates a disruption of some kind in the traditional management (Jordbruksverket, 2005). Ekstam & Forshed (1997) goes into detail in how specific plant species indicates land use history in Swedish semi-natural grasslands.

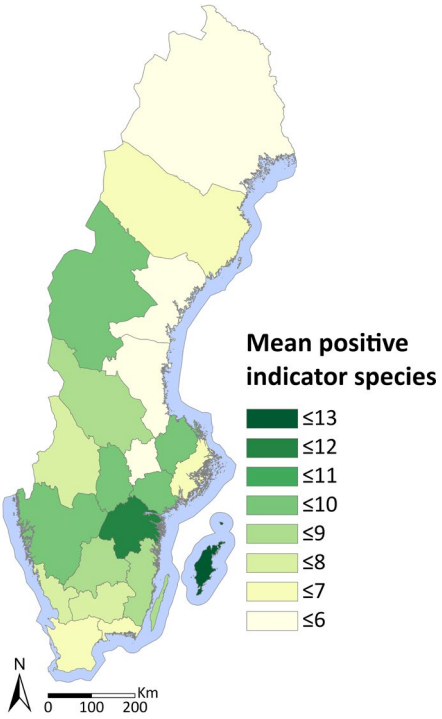

Figure 4. County map over Sweden showing the average amount of positive indicator species per semi-natural grassland. The figure is based on the inventories done in 2019 for the TUVA database, and might not be representative for every county. For example, the mean calculated for Norrbottens län is based on only two semi-natural grasslands.

Brunbjerg et al. (2018) shows that vascular plant species richness can be used as a predictor for species richness in other taxonomic groups, and thus useful for biodiversity monitoring. Since positive indicator species are selected to indicate a history of traditional management, it can be suitable to be used as a proxy for plant species richness in semi-natural grassland. Not only because of the indication of traditional management, but also that the amount of indicator species partially reflects plant species richness.

Sentinel-2 imagery and cloud mask

The Sentinel-2 satellite mission, managed by the European Space Agency (ESA), has opened the possibilities to monitor and assess detailed information about ecosystems. The mission is focused on observing change in land and costal conditions. It contains of two satellites, Sentinel-2A and 2B, launched in June 2015 and March 2017 respectively (European Space Agency, 2015). The satellites have a synchronized orbit around the earth and are phased 180° to each other. They have a revisit frequency of 10 days per satellite and combined they capture most of the land and coastal areas every 5 days. Each satellite is equipped with a multi-spectral instrument (MSI) that covers 13 bands in

the visible, near infrared (NIR) and shortwave infrared (SWIR) spectrums in 10m, 20m and 60m resolution (table 2). The swath width is 290 km. With its high temporal and spatial resolution, good access and emphasis on vegetation monitoring, the Sentinel-2 satellites are well suited for task of gathering information about small scale ecosystems such as semi-natural grasslands.

Table 2. Information about the Sentinel-2A bands. NIR = Near infrared, SWIR = Shortwave infrared.

	Targeted spectrum	Spatial resolution (m)	Central wavelength (nm)	Bandwidth (nm)
Band 1	Costal aerosol	60	442.7	21
Band 2	Blue	10	492.4	66
Band 3	Green	10	559.8	36
Band 4	Red	10	664.6	31
Band 5	Vegetation red edge	20	704.1	15
Band 6	Vegetation red edge	20	740.5	15
Band 7	Vegetation red edge	20	782.8	20
Band 8	NIR	10	832.8	106
Band 8A	Narrow NIR	20	864.7	21
Band 9	Water vapor	60	945.1	20
Band 10	SWIR Cirrus	60	1373.5	31
Band 11	SWIR	20	1613.7	91
Band 12	SWIR	20	2202.4	175

ESA delivers the Sentinel-2 images in 100 km^2 tiles at two different processing levels, 1C and 2A. Both levels are orthorectified, radiometrically and geometrically corrected, as well as spatial registered on the UTM/WGS84 reference system (European Space Agency, 2015). The level 2A product is derived from 1C, but in addition it is delivered with bottom-of-atmosphere (BOA) reflectance and a scene classification map (Main-Knorn et al., 2017). To have BOA reflectance instead of top-of-atmosphere, as in 1C, is a necessity for comparing images at ground level as they normalize atmospheric effects on surface reflectance. The scene classification map is a 20m pixel-based map with classes such as vegetation, water, snow, shadows, different cloud probabilities etc. The classes are derived from various thresholds based on reflection, band ratios and indices (European Space Agency, 2020b). All Sentinel data is open to the public and free to download from The Copernicus Open Access Hub (European Space Agency, 2020a)

Three dates in 2019 (6 June, 26 July, 25 August) and corresponding Sentinel-2A level 2A images was chosen with season and cloud coverages in regard. 2019 was a reasonable year to study since the new TUVA inventories were on-going, both Sentinel-2 satellites was operating, and the summer temperatures were not as abnormal as in 2018. The studied period is during summer so the

vegetation in Sweden is relatively vigorous and reflects representative spectral signals. No date captured during 2019 was cloud free over the study area. A selection algorithm was made to avoid faulty values from semi-natural grasslands affected by clouds, cloud shadows or other non-vegetated areas. The algorithm uses the scene classification map included in the level 2A products and removes semi-natural grasslands that contains less than 90% vegetated pixels (See appendix B of the code for details).

Band statistics and feature selection

The bands investigated in this study are the bands 2-4 and 8 in the 10 meters resolution and bands 5-7, 8A, 11 and 12 in the 20 meters resolution (table 2). For every semi-natural grassland, the parcel median reflectance and standard deviation of reflectance were calculated for all bands as well as the NDVI mean and standard deviation. The band median value was chosen over the mean to minimize the effect of outliers. There are 25 features in total, 20 for each band and statistic, 2 for both NDVI values, longitude, latitude and area. Each feature was calculated for all three dates and then averaged to get a somewhat representative value for each semi-natural grassland during summer.

Six subsets of the features were chosen to be inputs to the machine learning models (appendix A). One subset with only geographical data (3 features), and one that NDVI std was added (4 features). The third subset used Pearson's correlation coefficient (also Pearson's r), where all features chosen had a higher linear correlation with positive indicator species than a feature with only random numbers (PCC, 10 features). Similarly, there was a subset based on all features that had a higher mutual information (Ross, 2014) than a random number feature (MI, 12 features). The fifth subset was with the geographical data and standard deviations (std, 14 features) and the sixth with all 25 features.

Machine learning and the process to the results

The machine learning models used in this study are Random Forest, Support Vector Machine, Logistic Regression, Voting Classifier and Neural Network. All feature subsets were used as to train each model to find correlations between the features and the amount of positive indicator species. 20% of the data was held out from model training to test how accurate the machine learning models were. It is on this data the models tried to predict the amount of positive indicator species and evaluated based on the accuracy of these predictions. To make the results more comprehensible for both man and machine, a binary classification approach was applied. So, the target for each model was to try and learn to predict if a semi-natural grassland had 8 or more positive indicator species or not. The threshold was set at 8 or more since it was the closes amount to have both classes of equal size. This results in a divide of 1278 (55.7%) of the semi-natural grasslands that has 8 or more positive indicator species and 1015 (44.3%) that has less. All data is managed, calculated and processed in Python 3.7 (Python Software Foundation, 2021), and to see the code and the workflow for this thesis please see appendix B. ArcGIS Pro 2.5.2 (ESRI, 2021) was used to visualize the geographical differences in the results.

Model performance measures

All classifiers mentioned do binary classifications, and a powerful tool to evaluate how well a binary classifier works is the confusion matrix. A confusion matrix is created based on how a model classifies the unseen data from the test set. Since it is a binary task, the model predicts the data from the test

set as either True or False. A confusion matrix is a table with the classes of the predicted and the actual data on the axes (fig. 5). This results in four possible outcomes for each classified instance; true positive, false positive, true negative and false negative. If the instance is true positive, it is True as well as predicted as True. A false positive is predicted as True, though acually False. The same goes for the negatives. These four parameters are the foundation of several equations that can evaluate binary classifiers preformance (equations 1.1-1.5).

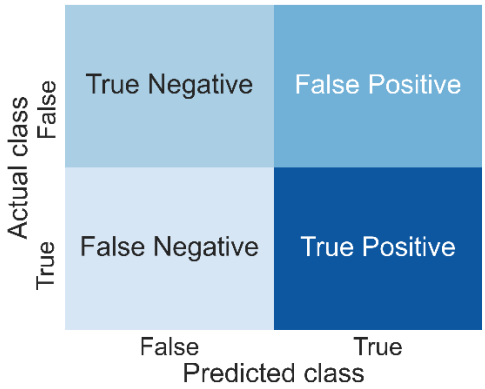

Figure 5. The design and labels of a binary classification confusion matrix.

One of the most used preformance measure is accuracy (eq. 1.1), which tells us how large portion of the data that has been rightfully classified. Two measures that shows how the model classifies the data is precision (eq. 1.2) and recall (eq. 1.3). Precision, also called positive predictive value, shows the fraction of how many of the instances that the model predicted positive acually was postitive. Recall, or sensitivity, shows the fraction of how many of the instances that acually are positive was classified as positive. F1 score (eq. 1.4) is the harmonic mean of precision and recall. F1 scores ranges between 0 and 1, with 1 if both precision and recall is perfect. Chicco & Jurman (2020) critizes both accuracy and F1 for giving misleading scores in data sets that are not balanced and suggests the use of Matthews correlation coefficient (MCC; eq. 1.5) instead. MCC give scores in between 1 and -1, where 1 is perfect predicitons, 0 totally random and -1 when not a prediction is right, i.e., perfectly wrong. Lastly there is Receiver operating characteristic curves (ROC) that are graphs with recall (eq. 1.3) plotted against the false positive rate (FPR; eq 1.6) for various thresholds. The plot shows at what rate at specific thresholds the classifiers get a true positive over a false positive. By calulating the area that is under the curve you get the Area under curve (AUC), and this value is used as a model performance measure.

Equations 1.1-1.5. Measures derived from a confusion matrix. TP = True positive, TN = True negative, FP = False positive, FN = False negative, MCC = Matthews correlation coefficient, FPR = False positive rate.

$$1.1.\ Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$1.2.\ Precision = \frac{TP}{TP + FP}$$

$$1.3.\ Recall = \frac{TP}{TP + FN}$$

$$1.4.\ F1 = 2 * \frac{Precision * Recall}{Precision + Recall}$$

$$1.5.\ MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

$$1.6\ FPR = \frac{FP}{FP + TN}$$

Permutation feature importance is a measure which randomly shuffles the values of a feature, and calculates how the performance of a model decreases. It shows how dependent a model is on particular features. The results are given with R^2 scores for each feature.

Results

Model, feature subset and performance measure comparisons

Random Forest scored the highest of all machine learning models; 0.645 in accuracy, 0.722 in F1 score, 0.680 in ROC AUC and 0.275 in MCC. These results indicate that they are better than random predictions. Random predictions would have given accuracy scores at around 0.557, ROC AUC scores at 0.500 and MCC scores at 0.000. All the metrics score ranges between 1 as best and 0 as worst, apart from MCC that has between 1 (best) and -1 (worst). MCC is the preferred accuracy performance score in this thesis. Though, other scores will not be ignored since they contribute with additional information.

The Random Forest high scores were made with the 25 features as input, but for the ROC AUC scores where the 3-feature subset worked best (table 3). Voting classifier with the 3-feature subset predicted second best, with MCC scores at 0.255. With the other feature subsets, it scored close or better than the models that was included in the voting, i.e., Logistic Regression, SVM and Random Forest, but only if those had similar accuracy performance scores. But if Random Forest had much better scores than Logistic Regression and SVM, the Voting Classifier could not perform as well. Logistic Regression steadily decreased in MCC with increasing features. SVM varied in performance scores and the only model that performed better with the PCC features than the 4-feature subset. Neural Network had all subsets performing better than the one with 25 features, which had the lowest accuracy and MCC score of all models. The recall scores were higher than the precision scores for all models and feature subsets.

For all models but Random Forest, the 3 features, 4 features and PCC feature subsets all performed better than the one with all 25 features (table 3). The MI-feature and std-feature subsets had similar performances for all models but Neural Network and to some extent Random Forest.

The most feature-sensitive model was the Neural Network models, that had the highest standard deviation across the feature subsets and performance measures. SVM followed with the second highest standard deviation. Logistic regression had the lowest standard deviation, followed by Voting Classifier.

Table 3. Six performance measure scores for each machine learning model and feature subset. The first four statistics measures the model classification performances. Precision and recall indicate how each model classifies. The different feature subsets are selected based on common statistical and geographical traits. To see what features each subset contains, see Appendix A. PCC = Pearson's correlation coefficient, MI = Mutual

information, std = Standard deviation, ROC AUC = Receiver Operating Characteristic, Area Under Curve, MCC = Matthew's correlation coefficient.

	3 features	4 features	PCC (10) features	MI (12) features	std (14) features	All (25) features	
Accuracy							**std**
Logistic Regression	0.623	0.608	0.601	0.601	0.601	0.597	0.009
Random Forest	0.625	0.632	0.614	0.636	0.632	0.645	0.009
SVM	0.619	0.606	0.614	0.584	0.584	0.586	0.015
Voting Classifier	0.636	0.630	0.627	0.612	0.612	0.617	0.009
Neural Network	0.630	0.617	0.582	0.586	0.593	0.564	0.022
F1							
Logistic Regression	0.717	0.707	0.694	0.700	0.696	0.687	0.010
Random Forest	0.692	0.697	0.694	0.712	0.711	0.722	0.011
SVM	0.701	0.690	0.693	0.661	0.668	0.656	0.017
Voting Classifier	0.716	0.708	0.716	0.701	0.703	0.713	0.006
Neural Network	0.701	0.705	0.699	0.694	0.721	0.670	0.015
ROC AUC							
Logistic Regression	0.616	0.605	0.609	0.596	0.608	0.602	0.006
Random Forest	0.680	0.672	0.657	0.659	0.641	0.641	0.014
SVM	0.652	0.641	0.631	0.620	0.607	0.595	0.019
Voting Classifier	0.676	0.666	0.652	0.645	0.636	0.636	0.015
Neural Network	0.656	0.642	0.604	0.594	0.596	0.562	0.032
MCC							
Logistic Regression	0.229	0.193	0.176	0.176	0.176	0.166	0.021
Random Forest	0.231	0.245	0.206	0.254	0.245	0.275	0.021
SVM	0.215	0.186	0.206	0.142	0.139	0.149	0.031
Voting Classifier	0.255	0.240	0.238	0.201	0.201	0.214	0.021
Neural Network	0.252	0.239	0.129	0.139	0.178	0.088	0.059
Precision							
Logistic Regression	0.612	0.601	0.601	0.598	0.601	0.601	0.004
Random Forest	0.633	0.638	0.617	0.632	0.627	0.635	0.007
SVM	0.617	0.609	0.617	0.600	0.596	0.605	0.008
Voting Classifier	0.629	0.626	0.617	0.609	0.608	0.607	0.009
Neural Network	0.632	0.612	0.579	0.586	0.579	0.575	0.021
Recall							
Logistic Regression	0.866	0.858	0.822	0.842	0.826	0.802	0.022
Random Forest	0.763	0.767	0.794	0.814	0.822	0.838	0.028
SVM	0.810	0.794	0.791	0.735	0.759	0.715	0.034
Voting Classifier	0.830	0.814	0.854	0.826	0.834	0.866	0.017
Neural Network	0.787	0.830	0.881	0.850	0.957	0.802	0.056

Regarding how fast each model works, Logistic Regression both fits and predicts the data the fastest (table 4) and Voting Classifier the slowest. Voting Classifiers times in both fitting and predicting are close to Logistic Regression, Random Forest and SVM when added together. Random Forest and Neural Network has similar prediction times, with both faster than SVM and Voting Classifier.

Table 4. Time to fit and predict each model with all 25 features. The time is measured in seconds. When a model is fitted it processes the training data and adjusts the model parameters to get as accurate as possible. After fitted, a model can predict how the unseen instances of the test data should be classified.

	Time to fit the model	Time to predict
Logistic Regression	0.0597	0.0003
Random Forest	2.1532	0.0621
SVM	3.6820	0.0918
Voting Classifier	5.8184	0.1534
Neural Network	0.1735	0.0576

Feature comparison

As seen in figure 6, the geographical features are the most important features for the machine learning models. Latitude was the feature that influences the models most, followed by longitude and area. For the spectral features, the red edge and SWIR bands scores high in importance. Both the standard deviation and median values for the bands in the red edge spectrum (5, 6, 7) scored among the highest across all band values, except the median value for band 5. The median values for the short-wave infrared, i.e., band 11 and 12, also scores relatively high. Bands in the visible spectrum and near infrared bands were least the least important features, except for median values for band 8 and standard deviation of band 4. The mean NDVI is of higher importance than the standard deviation. If the two most influential features latitude and longitude (fig. 6) were removed from the feature subsets, most models scored worse (appendix D).

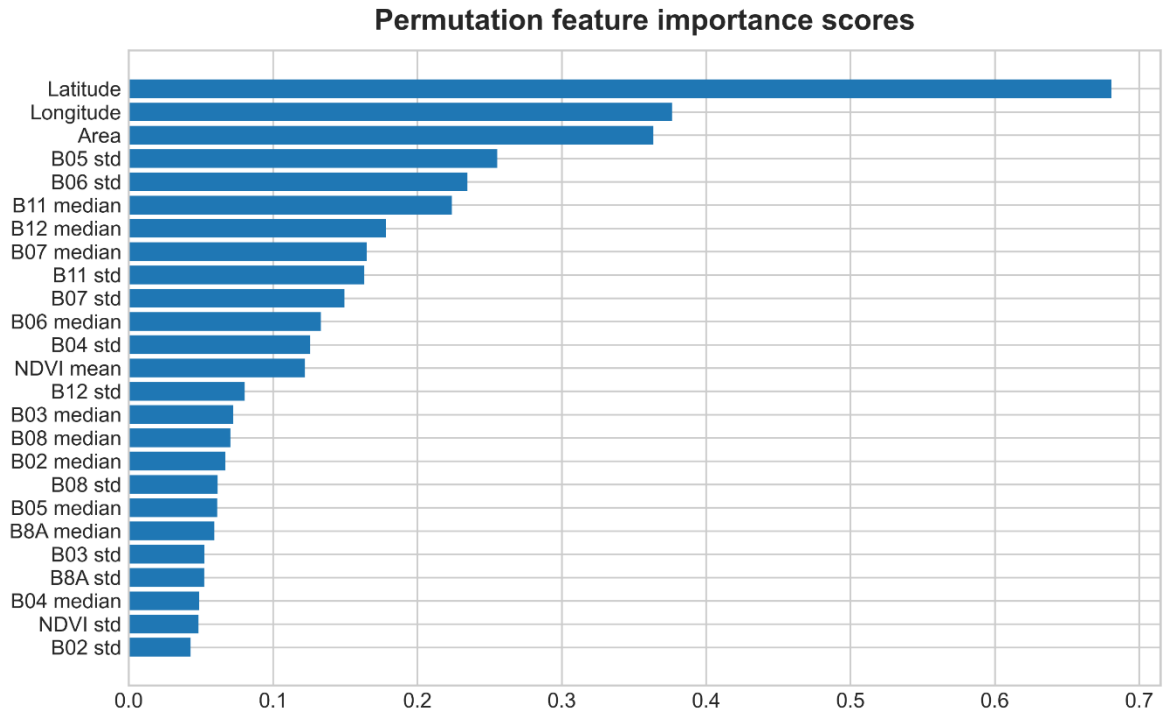

Figure 6. Permutation feature importance for all features included in this study. It was measured with each model fitted with all 25 features. The scores are the cumulative sum for the models normalized permutation feature importance. For each model's individual score, see appendix C. Voting classifier was not included to avoid overrepresentation, since it is exclusively based on other models. B = Band, std = Standard deviation, NDVI = Normalized difference vegetation index.

Geographical differences in classification

Regarding the models' error classifications, there is a general geographical pattern that all different models have most of their False negatives in the southern parts of the study area (fig. 7-8). SVM and Random Forest are the only models with False negatives in the most northern parts (fig. 7). For the models' False positives, they are more frequently located to the north (fig. 7-8). Note that this north-south error gradient is pronounced for Voting Classifier (fig. 7), which is based on Random Forest, SVM and Logistic Regression.

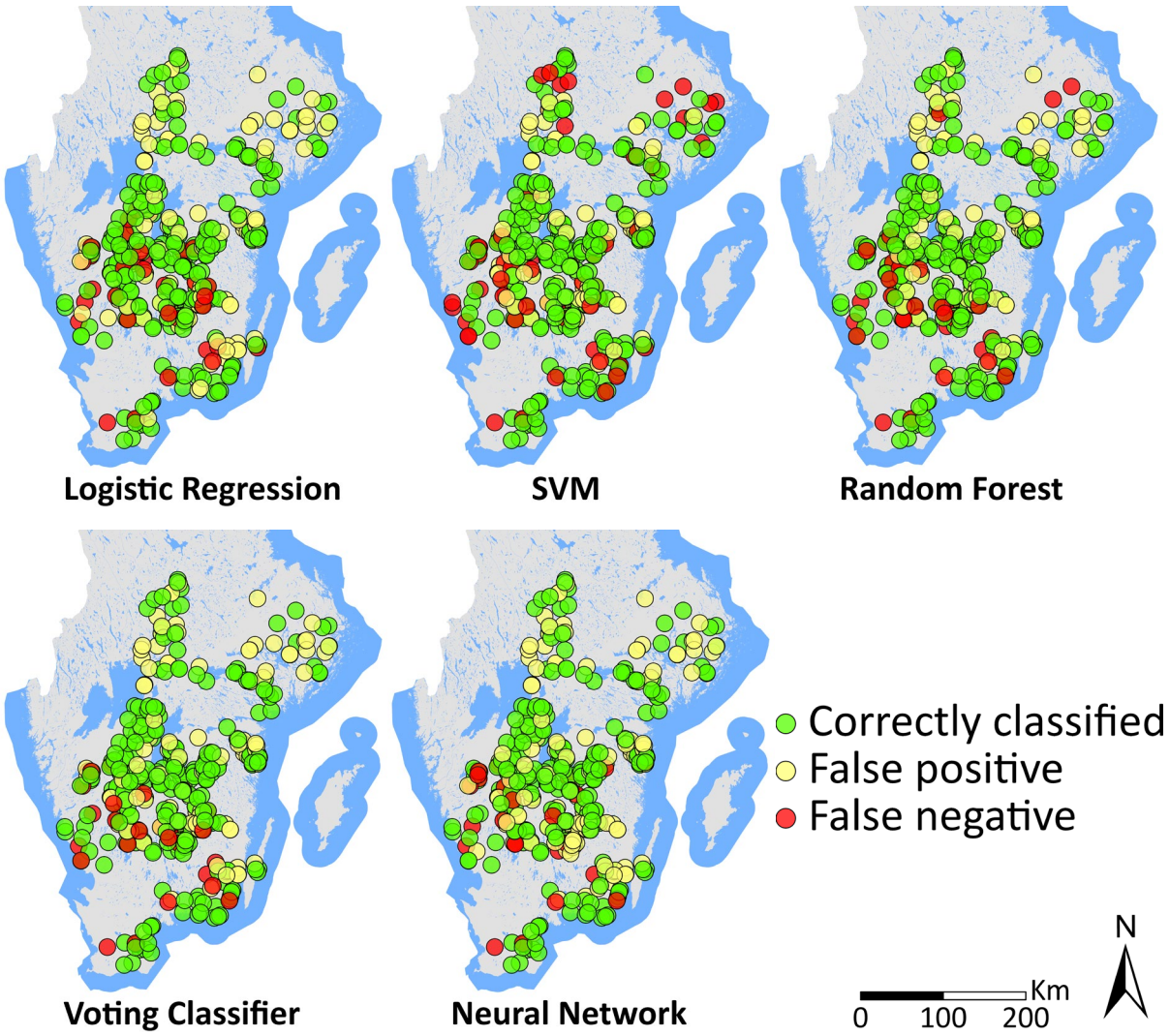

Figure 7. Predictions on the test data for all five models with the 25 features subset. A false positive (yellow dot) is when the model predicts that a semi-natural grassland has 8 or more positive indicator species, but it actually has less than 8. A false negative (red dot) is the opposite, classified as less than 8 but actually is 8 or more

.

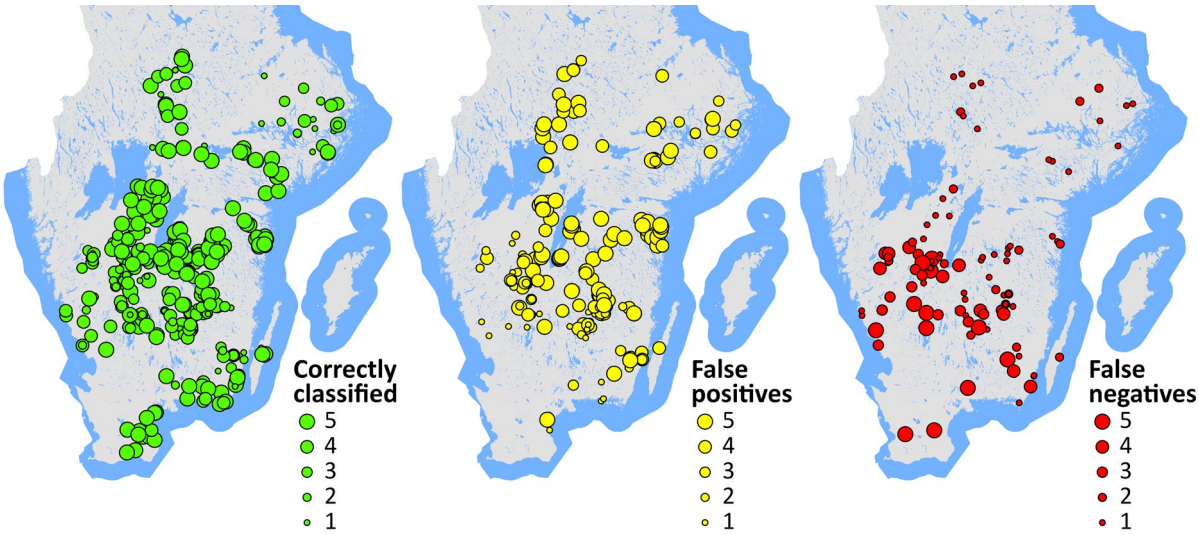

Figure 8. The cumulative classification results for all five machine learning models, fitted with the 25 features subset. A false positive is when the model predicts that a semi-natural grassland has 8 or more positive indicator species, but it actually has less than 8. A false negative is the opposite, classified as less than 8 but actually is 8 or more.

Discussion

Features importance for the machine learning models

That the coordinate data are the top important features might be contributed to that these features are noise-free. Each feature numbers exclusively represent that feature. The satellites band data does not aim to represent a defined measure, only a spectrum, and therefore each band contains multitudes of levels of spectral information about the targeted area. Therefore, the band data will be inherently noisy, regardless of one's objective. Regarding this, the machine learning models will more successfully find patterns in latitude, longitude and area related to positive indicator species than in the satellite data. That latitude is in this case more important than longitude can be because of the range of values is larger and more varied, i.e., more data to find correlations in.

As seen in figure 5B, most of the semi-natural grasslands are clustered in the middle and west parts of the study area, whilst getting sparser north, south and on the east coast. Four of five machine learning models showed tendencies to classify false positives to the north and false negatives to south (fig. 7-8). This indicate that the models are performing well on the instances where it most of the data is spatially located, though does not generalize as well on semi-natural grasslands further away. The gradient pattern, where the models predict that there is more positive indicator species to the north of the study area and less in south, can be seen in the actual data as well (fig. 3B, 4). It could be that the models overinterpret this data. Since the models are more likely to find patterns in the non-noisy coordinate data than in the satellite data, the importance is put heavily on these features to predict positive indicator species, and this generalizing spatial pattern seen in figures 7 and 8 emerges.

In a sense, area is a noise-free feature, similar to the coordinate data, when used as a parameter for the machine learning models. The correlations to area that the models finds could be related to the species-area relationship (Scheiner, 2003), that with increasing area sample size, more species will be

found. These factors contribute to the high importance of area as a feature. But area also serves as a function as for which pixels from the satellite data are to be calculated on for each semi-natural grassland. Depending on the shape and location of the semi-natural grasslands, noise can be induced to the satellite data, e.g., bordering objects such as trees, water bodies and buildings can affect the reflection registered by the satellite (Richter et al., 2006). To lower the risk of adjacency effects, only pixels that was completely contained within the borders of the semi-natural grassland was included in the calculations.

That both the median and standard deviation of the three red edge bands are important features (fig. 6) could be connected to that in this spectrum, 680nm-750nm, vegetation has a sharp increase in reflection (Horler et al., 1983). As Horler et al. (1983) states, the change in reflection from vegetation in the "red edge" spectrum is related mainly to leaf chlorophyll content. The standard deviation in these bands shows how steep the change in reflection for each semi-natural grassland. It might be that since the median values of band 5 are relatively unimportant and the standard deviation of the band the most important feature, is that in this spectrum where we see the initial increase for the red edge change. So, the median values for band 5 might be quite similar semi-natural grasslands alike, but a high value of standard deviation could indicate that a grassland has this property of the initial increase. Bands 6 and 7 spectrums should capture the continuous increase of reflectance, so high values in both median and standard deviation indicate high chlorophyll content, i.e., much and/or healthy vegetation. Band 7 might also capture where the increase in reflectance levels off, so low scores in standard deviation indicate that this have happened.

The shortwave infrared (SWIR) region between 1400nm-3000nm can be used to assess information about the water content in vegetation (Ceccato et al., 2001; Gao, 1996). High reflections in parts of the SWIR spectrum indicates low water contents in vegetation (Ceccato et al., 2001; Cheng et al., 2011). That band 11 and 12 shows importance can be a function of the vegetation water content in the semi-natural grasslands. The median values of the SWIR bands could indicate the general amount of water in the vegetation of each semi-natural grasslands, and standard deviation the variation of vegetation water content over the area.

The low importance of the NDVI standard deviation can be that the semi-natural grasslands in the study have similar vegetation, so the spectral variation of NDVI in the grasslands is not especially comparable. That the mean NDVI values are of some importance seems more intuitive, that the vigour and density of the vegetation between semi-natural grasslands have correlations to positive indicator species. NDVI mean indicate, to some extent, the same information as the red edge bands. The red edge bands spectra are located between band 4 and band 8, which NDVI is based upon. The mean value of NDVI scores higher than both median values for band 4 and band 8 (fig. 6). This could help in minimizing the number of features without losing too much information in future machine learning tasks. Since semi-natural grasslands are relatively nutrient-poor ecosystems, it is not necessary that the correlations between positive indicator species and NDVI is positive. To investigate the relationships of correlations between features and biodiversity proxies such as positive indicator species has not been examined in this thesis.

The machine learning model performance

That the Random Forest model is the only model that generally increases in accuracy and MCC score with more features, whilst the other models decrease (table 3; appendix D). This shows the

robustness of Random Forest, that even if the satellite data is noisy, it has the capability to find correlations that increases the model's performance. As Maxwell et al. (2018) concluded, Random Forest, as well as SVM, are in general robust and accurate machine learning models for remote sensing tasks. Though in this thesis SVM is one of the worst performing models. Foody et al. (2016) shows that SVM can be sensitive to noisy and mislabelled data, which is what could be the case here. Both SVM and Logistic Regression are algorithms that tries to fit borders and thresholds to be able to classify the data. Noisy data, such as the satellites band data, will be hard to decipher the levels of spectral data that is correlated to the targeted feature. These models get confused when trying to find patterns in noisy satellite features and therefore will decrease in performance. Differences between Logistic Regression and SVM is that SVM varied more in performance between feature subsets. This could be that Logistic Regression is not as flexible as SVM as to find complex relationships. That we see a slight increase in accuracy score with SVM for PCC features over the 4-feature subset could because PCC is based on linear correlations, and the flexibility allows SVM to find complex correlations between these features. As for voting classifier performance, if the included models have similar performance scores, the voting classifier seems to outperform each of them (table 3, see 3 features and PCC (10) features). Otherwise, it generally is worse than the best performing model, but better than the worst.

Similar to Logistic Regression and SVM, the Neural Network models' performance scores decreased in most cases when features increased. These models are more volatile and varies the most between feature subsets, which indicate sensitivity to input features. The F1 scores for the Neural Network models with more features are relatively high. As F1 is the harmonic mean of recall and precision (eq. 1.4), the high recall score for the Neural Network models raises the F1 score. The precision scores are close to 0.557, the percentage of actual semi-natural grasslands that have 8 or more positive indicator species. This could be because it captures almost all true positives, though the rest is false positives. The recall is generally high, e.g., 0.957 with the 14-feature subset, i.e., many true positives and very few false negatives. High recall and low precision will tell us that these models classify close to all the semi-natural grasslands as to have 8 or more positive indicator species. These recall and precision score patterns are similar for most other models (table 3), and in fig. 7 and 8 we see that the models generally have more false positives than false negatives.

Both accuracy and MCC have similar pattern between model scores (table 3). Since the data is slightly imbalanced with 55.7% and 44.3% split, accuracy may give misleading results. For example, the model with the worst accuracy has 0.564 (Neural Network with 25 features; table 3) and could give a false sense that it performs with 6.4 percentage points better than random. In reality it only performs 0.7 percentage points better than random. This is in line with what Chicco & Jurman (2020) points out and they argue that MCC should be used since it takes both true positives and true negatives, as well as works on imbalanced datasets. The MCC score of the same worst performing Neural Network model is 0.088, which tells us both intuitively and explicitly that the performance is only just better than random (score of 0). ROC AUC is critiqued as a model performance measure, e.g. is produces inconsistent values between classifiers (Hand, 2009) and includes values that are unlikely used in the calculations (Lobo et al., 2008), and thus unsuitable for this task. To measure model performance in tasks similar to this, MCC is recommended as an accuracy metric. Precision and recall are also important scores to evaluate to get information how the models classify.

The processing times for the different models with data on this scale is almost neglectable. Though, using larger dataset, computational cost could be a factor. Since voting classifier processing times are the sum of the included models, this should be considered if the gain in performance is worth the added processing time. Logistic Regression process the data fast, but this algorithm fails to capture the complexity of the satellite data and thus is not viable for the task. SVM is neither viable since the data deems to noisy for this algorithm as well as it is computationally heavy. Neural Network process the data faster than SVM, though have similarly low performance scores. Random Forest is the machine learning model that shows most promises to perform well at a moderate computational cost, as well as it is accessible and relatively easy to use.

Biogeographical issues

The positive indicator species used in this study as proxies for plant species richness in semi-natural grassland has been treated as equally important. There are two intertwined issues in this. The first issue is that indicator species has different preferences and sensitivities towards the environmental factors within the semi-natural grassland ecosystem (Ekstam & Forshed, 1997). For example, the positive indicator species of *Leucanthemum vulgare*, a flower that is commonly seen in semi-natural grasslands in most parts of Sweden, is treated with equal weight as *Primula farinosa*, a red-listed flower that grows in wet, calcareous semi-natural grasslands (Ekstam & Forshed, 1997; SLU Artdatabanken, 2020). Though, as this study aims to generalize plant species richness it is not a problem per se, but it translates into a biogeographical issue when applied on a larger spatial scale. As the first law of geography states, everything is connected, but near things have more in common than distant things (Tobler, 1970); semi-natural grasslands within a region will have different biophysical parameters to more distant semi-natural grasslands. A regions habitat can be more or less suited to hold a larger pool of positive indicator species than another region, depending on factors such as management history (Cousins et al., 2009; Cousins & Eriksson, 2002), surrounding landscape (Schmucki et al., 2012) and environmental factors (Klimek et al., 2007; Wellstein et al., 2007). This issue can induce bias when trying to compare plant species richness in semi-natural grasslands located far from each other.

Another potential bias that influences the amount of positive indicator species in the semi-natural grasslands geographically is local differences in inventories. Different personnel in regions can give varied results in the inventory depending on things like expertise, experience and daily form. Since the TUVA database is not solely concentrated on vascular plants as positive indicator species, but also registers cultural elements, birds, fungi, trees, etc. (Jordbruksverket, 2017), there is a demand on a broad knowledgebase on the personnel. It cannot be expected that the people employed for the survey have equal expertise in all different parameters that TUVA registers. It can be that the same person or group of people have done the inventories in local regions, and thus within these regions the inventories are fairly standardised but will differ from other regions.

It may be that these different issues are underlying factors that influences the geographical patterns in fig. 7 and 8 as discussed previously. That semi-natural grasslands within regions have similar prerequisites that differ from regions farther away. Many of the semi-natural grasslands included in this study are spatially concentrated (fig. 3B) and this could result in that the machine learning models are more accurate in these areas (fig. 7-8).

Practical use and potential future studies

The use of positive indicator species as a proxy for plant species richness is not entirely viable since they are in fact two different metrics that share similarities. Both metrics could be used for similar

tasks like the one done in this thesis, depending on what the desired outcome is. For example, plant species richness for getting a sense for the overall biodiversity in an area. But using positive indicator species on more specific tasks like as a basis for funding semi-natural grasslands. That positive indicator species is used in this thesis is because of the extensive data collected in the TUVA database, and machine learning models require large amounts of data to be able to perform well (Halevy et al., 2009). Up to this date, there is to my knowledge no dataset over Sweden as extensive for ecologically valuable areas as the TUVA database. One way to improve the quality of the dataset is to add more dates to get even more representative spectral reflection values. To add other years can improve the predictive accuracy of the models, though it is restricted to the years when surveys are done. It could be that a large and noisy dataset can be replaced by a smaller but more detailed dataset with concise measures of plant species richness and give better results for feature-sensitive models like Neural Networks or SVM.

The machine learning models used in this thesis are not well suited for the practical use of prediction plant species richness in semi-natural grasslands. There are too many question marks regarding the potential biases in the TUVA data and noise of the satellite data that could be investigated and optimized. Though, this study shows that there are underlying potentials to develop machine learning models that could predict ecologically valuable semi-natural grasslands. To further investigate the capabilities of remotely sensed biodiversity proxies in semi-natural grasslands, a suggested area of study is to investigate the relationships between the different Sentinel-2 bands and plant species richness. Especially those bands with spectrums in the red edge and short-wave infrared regions. This could increase the quality of the data, thus make it easier for the machine learning models to interpret it. Another area of interest would be to run this methodology on more local study areas based on different biophysical properties. This could potentially minimize the biogeographical issues mentioned, as well as see how different machine learning models behaves with less amount of training data.

A code for a program was produced that employs this methodology and process, calculates and make predictions based on the input data (appendix B). To make a program like the one presented here work better, the most important issue is the quality of the data. With better quality data, the machine learning models will have it easier to find correlations. The data could be improved by investigating the satellite-biodiversity relationship more, taking local biophysical parameters into account or having more suitable proxies for biodiversity.

To have a deeper knowledge about the satellite-biodiversity relationship would now only be beneficial to gain assess the status of semi-natural grasslands, but might be applicable to other ecosystems as well. That combined with a good strategy to sample biophysically similar sites within an ecosystem could mean cheaper, more efficient and standardized monitoring of ecologically valuable areas around the globe.

Conclusions

The possibilities of using Sentinel-2 data, geographical features and machine learning to predict biodiversity in semi-natural grasslands have been investigated in this thesis. Random Forest is the machine learning model that shows the best results for predicting if a semi-natural grassland has high plant species richness or not. This could be due to Random Forests ability to find relationships in the relatively noisy spectral data. Whilst other models decreased in accuracy with increasing number of spectral features, they might perform better in similar tasks that uses less noisy or smaller

datasets. No model worked well enough for practical use. Based on binary classification confusion matrix scores, Matthew's correlations coefficient was the most prominent accuracy measure to see the predictive performances for machine learning models with regards to find biodiverse ecosystems. This because MCC takes all metrics produced by the confusion matrices into account as well as works well on imbalanced datasets.

Latitude, longitude and area are most influential features for the machine learning models, which can be rooted in the noise-free nature of these features. There were geographical patterns of predictive errors made by the models, with a tendency to overestimate plant species richness in the study areas northern parts and underestimate in the southern parts. Regarding the spectral features from the Sentinel-2 satellite, the red edge and short-wave infrared bands scored high in feature importance. The three red edge bands are intended to capture the sharp increase in spectral reflection seen from vegetation, and these values from semi-natural grasslands can indicate how much and/or healthy the vegetation is. For the two short-wave infrared bands, they can indicate information about the water content in the vegetation of semi-natural grasslands.

Acknowledgments

I want to thank Ian Brown for excellent supervising and guidance through this thesis. Your knowledge and advices always seemed to come in the right time. Thank you, Jessica Lindgren, my co-supervisor, for the ecological expertise as well as insightful and encouraging discussions. Thanks to Karl Samuelsson for the structural and communicative advices and Lukas Rimondini for the constant support and enlightening debates.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., … Zheng, X. (2016). *TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems*. www.tensorflow.org.

Abdi, A. M. (2020). Land cover and land use classification performance of machine learning algorithms in a boreal landscape using Sentinel-2 data. *GIScience and Remote Sensing*, *57*(1), 1–20. https://doi.org/10.1080/15481603.2019.1650447

Belgiu, M., & Drăgu, L. (2016). Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing*, *114*, 24–31. https://doi.org/10.1016/j.isprsjprs.2016.01.011

Bengtsson, J., Bullock, J. M., Egoh, B., Everson, C., Everson, T., O'Connor, T., O'Farrell, P. J., Smith, H. G., & Lindborg, R. (2019). Grasslands-more important for ecosystem services than you might think. *Ecosphere*, *10*(2), e02582. https://doi.org/10.1002/ecs2.2582

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. *Proceedings of the Fifth Annual Workshop on Computational Learning Theory - COLT '92*, 144–152. https://doi.org/10.1145/130385.130401

Brunbjerg, A. K., Bruun, H. H., Dalby, L., Fløjgaard, C., Frøslev, T. G., Høye, T. T., Goldberg, I., Læssøe, T., Hansen, M. D. D., Brøndum, L., Skipper, L., Fog, K., & Ejrnæs, R. (2018). Vascular plant species richness and bioindication predict multi-taxon species richness. *Methods in Ecology and Evolution*, *9*(12), 2372–2382. https://doi.org/10.1111/2041-210X.13087

Ceccato, P., Flasse, S., Tarantola, S., Jacquemoud, S., & Grégoire, J.-M. (2001). Detecting vegetation leaf water content using reflectance in the optical domain. *Remote Sensing of Environment*, *77*(1), 22–33. https://doi.org/10.1016/S0034-4257(01)00191-2

Cheng, T., Rivard, B., & Sánchez-Azofeifa, A. (2011). Spectroscopic determination of leaf water content using continuous wavelet analysis. *Remote Sensing of Environment*, *115*(2), 659–670. https://doi.org/10.1016/j.rse.2010.11.001

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, *21*(1), 6. https://doi.org/10.1186/s12864-019-6413-7

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/BF00994018

Cousins, S. A. O., & Eriksson, O. (2002). The influence of management history and habitat on plant species richness in a rural hemiboreal landscape, Sweden. *Landscape Ecology*, *17*(6), 517–529. https://doi.org/https://doi.org/10.1023/A:1021400513256

Cousins, S. A. O., Lindborg, R., & Mattsson, S. (2009). Land use history and site location are more important for grassland species richness than local soil properties. *Nordic Journal of Botany*, *27*(6), 483–489. https://doi.org/10.1111/j.1756-1051.2009.00472.x

Duro, D. C., Girard, J., King, D. J., Fahrig, L., Mitchell, S., Lindsay, K., & Tischendorf, L. (2014). Predicting species diversity in agricultural environments using Landsat TM imagery. *Remote Sensing of Environment*, *144*, 214–225. https://doi.org/10.1016/j.rse.2014.01.001

Ekstam, U., & Forshed, N. (1997). *Om hävden upphör: kärlväxter som indikatorarter i ängs- och*

hagmarker. Naturvårdsverket.

Eriksson, O. (2020). Origin and Development of Managed Meadows in Sweden: A Review. *Rural Landscapes: Society, Environment, History*, *7*(1), 1–23. https://doi.org/10.16993/rl.51

Eriksson, O., & Cousins, S. (2014). Historical Landscape Perspectives on Grasslands in Sweden and the Baltic Region. *Land*, *3*(1), 300–321. https://doi.org/10.3390/land3010300

ESRI. (2021, May 31). *ArcGIS Pro*. https://www.esri.com/en-us/arcgis/products/arcgis-pro/overview

European Commission. (2020). *Biodiversity Strategy for 2030, Bringing nature back into our lives* (Issue May). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52020DC0380

European Space Agency. (2015). *Sentinel-2 User Handbook*. https://sentinel.esa.int/documents/247904/685211/Sentinel-2_User_Handbook

European Space Agency. (2020a). *Copernicus Open Access Hub*. Copernicus Open Access Hub (Previously Known as Sentinels Scientific Data Hub). https://scihub.copernicus.eu/

European Space Agency. (2020b). *Sen2Cor Configuration and User Manual*. *1*, 53. https://step.esa.int/thirdparties/sen2cor/2.4.0/Sen2Cor_240_Documenation_PDF/S2-PDGS-MPC-L2A-SUM-V2.4.0.pdf

Fauvel, M., Lopes, M., Dubo, T., Rivers-Moore, J., Frison, P.-L., Gross, N., & Ouin, A. (2020). Prediction of plant diversity in grasslands using Sentinel-1 and -2 satellite image time series. *Remote Sensing of Environment*, *237*(July 2019), 111536. https://doi.org/10.1016/j.rse.2019.111536

Fisher, P. (1997). The pixel: A snare and a delusion. *International Journal of Remote Sensing*, *18*(3), 679–685. https://doi.org/10.1080/014311697219015

Foody, G., Pal, M., Rocchini, D., Garzon-Lopez, C., & Bastin, L. (2016). The Sensitivity of Mapping Methods to Reference Data Quality: Training Supervised Image Classifications with Imperfect Reference Data. *ISPRS International Journal of Geo-Information*, *5*(11), 199. https://doi.org/10.3390/ijgi5110199

Gao, B. (1996). NDWI—A normalized difference water index for remote sensing of vegetation liquid water from space. *Remote Sensing of Environment*, *58*(3), 257–266. https://doi.org/10.1016/S0034-4257(96)00067-3

Géron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. O'Reilly Media.

Halevy, A., Norvig, P., & Pereira, F. (2009). The Unreasonable Effectiveness of Data. *IEEE Intelligent Systems*, *24*(2), 8–12. https://doi.org/10.1109/MIS.2009.36

Hall, K., Johansson, L. J., Sykes, M. T., Reitalu, T., Larsson, K., & Prentice, H. C. (2010). Inventorying management status and plant species richness in semi-natural grasslands using high spatial resolution imagery. *Applied Vegetation Science*, *13*(2), 221–233. https://doi.org/10.1111/j.1654-109X.2009.01063.x

Hand, D. J. (2009). Measuring classifier performance: a coherent alternative to the area under the ROC curve. *Mach Learn*, *77*, 103–123. https://doi.org/10.1007/s10994-009-5119-5

Horler, D. N. H., Dockray, M., & Barber, J. (1983). The red edge of plant leaf reflectance. *International Journal of Remote Sensing*, *4*(2), 273–288. https://doi.org/10.1080/01431168308948546

Jordbruksverket. (2005). *Ängs- Och Betesmarks- inventeringen 2002–2004*. https://www2.jordbruksverket.se/webdav/files/SJV/trycksaker/Pdf_rapporter/ra05_1.pdf

Jordbruksverket. (2017). *Ängs- och betesmarksinventeringen, Metodik för inventering från och med 2016*. https://www2.jordbruksverket.se/download/18.48a7452e15c7b4a5a65a3a6b/1496908244029/ra17_9.pdf

Jordbruksverket. (2019). *Databasen TUVA*. https://etjanst.sjv.se/tuvaut/site/webapp/tuvaut.html

Kavzoglu, T., & Colkesen, I. (2009). A kernel functions analysis for support vector machines for land cover classification. *International Journal of Applied Earth Observation and Geoinformation*, *11*(5), 352–359. https://doi.org/10.1016/j.jag.2009.06.002

Klimek, S., Richter gen. Kemmermann, A., Hofmann, M., & Isselstein, J. (2007). Plant species richness and composition in managed grasslands: The relative importance of field management and environmental factors. *Biological Conservation*, *134*(4), 559–570. https://doi.org/10.1016/j.biocon.2006.09.007

Lawrence, R. L., & Moran, C. J. (2015). The AmericaView classification methods accuracy comparison project: A rigorous approach for model selection. *Remote Sensing of Environment*, *170*, 115–120. https://doi.org/10.1016/j.rse.2015.09.008

Li, W., Fu, H., Yu, L., & Cracknell, A. (2016). Deep Learning Based Oil Palm Tree Detection and Counting for High-Resolution Remote Sensing Images. *Remote Sensing*, *9*(1), 22. https://doi.org/10.3390/rs9010022

Lobo, J. M., Jiménez-Valverde, A., & Real, R. (2008). AUC: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography*, *17*(2), 145–151. https://doi.org/10.1111/j.1466-8238.2007.00358.x

Main-Knorn, M., Pflug, B., Louis, J., Debaecker, V., Müller-Wilm, U., & Gascon, F. (2017). Sen2Cor for Sentinel-2. In L. Bruzzone, F. Bovolo, & J. A. Benediktsson (Eds.), *Image and Signal Processing for Remote Sensing XXIII* (Vol. 1042704, Issue October 2017, p. 3). SPIE. https://doi.org/10.1117/12.2278218

Maxwell, A. E., Warner, T. A., & Fang, F. (2018). Implementation of machine-learning classification in remote sensing: an applied review. *International Journal of Remote Sensing*, *39*(9), 2784–2817. https://doi.org/10.1080/01431161.2018.1433343

Mountrakis, G., Im, J., & Ogole, C. (2011). Support vector machines in remote sensing: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, *66*(3), 247–259. https://doi.org/10.1016/j.isprsjprs.2010.11.001

Oldeland, J., Wesuls, D., Rocchini, D., Schmidt, M., & Jürgens, N. (2010). Does using species abundance data improve estimates of species diversity from remotely sensed spectral heterogeneity? *Ecological Indicators*, *10*(2), 390–396. https://doi.org/10.1016/j.ecolind.2009.07.012

Palmer, M. W., Earls, P. G., Hoagland, B. W., White, P. S., & Wohlgemuth, T. (2002). Quantitative tools for perfecting species lists. *Environmetrics*, *13*(2), 121–137. https://doi.org/10.1002/env.516

Pärtel, M., Bruun, H. H., & Sammul, M. (2005). Biodiversity in temperate European grasslands: origin and conservation. *Grassland Science in Europe*, *10*, 1–14. http://lup.lub.lu.se/record/532202/file/625284.pdf

Pedregosa, F., Michel, V., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Vanderplas, J., Cournapeau, D., Pedregosa, F., Varoquaux, G., Gramfort, A., Thirion, B., Grisel, O., Dubourg, V.,

Passos, A., Brucher, M., Perrot andÉdouardand, M., Duchesnay, A., & Duchesnay EDOUARDDUCHESNAY, Fré. (2011). Scikit-learn: Machine Learning in Python. In *Journal of Machine Learning Research* (Vol. 12). http://scikit-learn.sourceforge.net.

Python Software Foundation. (2021, May 19). *Python*. https://www.python.org/

Retallack, G. J. (2001). Cenozoic Expansion of Grasslands and Climatic Cooling. *The Journal of Geology*, *109*(4), 407–426. https://doi.org/10.1086/320791

Richter, R., Bachmann, M., Dorigo, W., & Muller, A. (2006). Influence of the Adjacency Effect on Ground Reflectance Measurements. *IEEE Geoscience and Remote Sensing Letters*, *3*(4), 565–569. https://doi.org/10.1109/LGRS.2006.882146

Rocchini, D., Balkenhol, N., Carter, G. A., Foody, G. M., Gillespie, T. W., He, K. S., Kark, S., Levin, N., Lucas, K., Luoto, M., Nagendra, H., Oldeland, J., Ricotta, C., Southworth, J., & Neteler, M. (2010). Remotely sensed spectral heterogeneity as a proxy of species diversity: Recent advances and open challenges. *Ecological Informatics*, *5*(5), 318–329. https://doi.org/10.1016/j.ecoinf.2010.06.001

Ross, B. C. (2014). Mutual Information between Discrete and Continuous Data Sets. *PLoS ONE*, *9*(2), e87357. https://doi.org/10.1371/journal.pone.0087357

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533–536. https://doi.org/10.1038/323533a0

Scheiner, S. M. (2003). Six types of species-area curves. *Global Ecology and Biogeography*, *12*(6), 441–447. https://doi.org/10.1046/j.1466-822X.2003.00061.x

Schmidtlein, S., & Fassnacht, F. E. (2017). The spectral variability hypothesis does not hold across landscapes. *Remote Sensing of Environment*, *192*, 114–125. https://doi.org/10.1016/j.rse.2017.01.036

Schmucki, R., Reimark, J., Lindborg, R., & Cousins, S. A. O. (2012). Landscape context and management regime structure plant diversity in grassland communities. *Journal of Ecology*, *100*(5), 1164–1173. https://doi.org/10.1111/j.1365-2745.2012.01988.x

scikit-learn. (2021). *scikit-learn: machine learning in Python — scikit-learn 0.24.2 documentation*. https://scikit-learn.org/stable/

SLU Artdatabanken. (2020). *Rödlistade arter i Sverige 2020*. SLU, Uppsala. https://www.artdatabanken.se/publikationer/bestall-publikationer/bestall-rodlista-2020/

Tamme, R., Hiiesalu, I., Laanisto, L., Szava-Kovats, R., & Pärtel, M. (2010). Environmental heterogeneity, species diversity and co-existence at different spatial scales. *Journal of Vegetation Science*, *21*(4), 796–801. https://doi.org/10.1111/j.1654-1103.2010.01185.x

TensorFlow. (2021). *TensorFlow*. https://www.tensorflow.org/

Tobler, W. R. (1970). A Computer Movie Simulating Urban Growth in the Detroit Region. *Economic Geography*, *46*, 234. https://doi.org/10.2307/143141

Verrelst, J., Muñoz, J., Alonso, L., Delegido, J., Rivera, J. P., Camps-Valls, G., & Moreno, J. (2012). Machine learning regression algorithms for biophysical parameter retrieval: Opportunities for Sentinel-2 and -3. *Remote Sensing of Environment*, *118*, 127–139. https://doi.org/10.1016/j.rse.2011.11.002

Warren, S. D., Alt, M., Olson, K. D., Irl, S. D. H., Steinbauer, M. J., & Jentsch, A. (2014). The

relationship between the spectral diversity of satellite imagery, habitat heterogeneity, and plant species richness. *Ecological Informatics*, *24*, 160–168. https://doi.org/10.1016/j.ecoinf.2014.08.006

Wellstein, C., Otte, A., & Waldhardt, R. (2007). Impact of site and management on the diversity of central European mesic grassland. *Agriculture, Ecosystems & Environment*, *122*(2), 203–210. https://doi.org/10.1016/j.agee.2006.12.033

Wilson, J. B., Peet, R. K., Dengler, J., & Pärtel, M. (2012). Plant species richness: the world records. *Journal of Vegetation Science*, *23*(4), 796–802. https://doi.org/10.1111/j.1654-1103.2012.01400.x

Yang, L., Cervone, G., Loia, V. B., & Liping, Y. (2019). Analysis of remote sensing imagery for disaster assessment using deep learning: a case study of flooding event. *Soft Computing*, *23*, 13393–13408. https://doi.org/10.1007/s00500-019-03878-8

Appendix A

The table shows the six feature subsets and what features that are included. PCC = Pearson's correlation coefficient, MI = Mutual information, B = band, std = Standard deviation, NDVI = Normalized difference vegetation index.

	3 features	4 features	PCC (10) features	MI (12) features	std (14) features	All (25) features
B2 median				X		X
B3 median				X		X
B4 median				X		X
B5 median				X		X
B6 median						X
B7 median						X
B8A median						X
B8 median				X		X
B11 median			X			X
B12 median			X	X		X
B2 std			X	X	X	X
B3 std					X	X
B4 std					X	X
B5 std			X	X	X	X
B6 std					X	X
B7 std					X	X
B8A std					X	X
B8 std				X	X	X
B11 std			X		X	X
B12 std			X		X	X
NDVI mean						X
NDVI std		X	X		X	X
Longitude	X	X	X	X	X	X
Latitude	X	X	X	X	X	X
Area	X	X	X	X	X	X

Appendix B

The code presented here is a 4000+ line program that process, calculates and gives predictions about biodiversity in targeted areas. Inputs are in the example code from Sentinel-2 optical images and data about semi-natural grasslands inventoried in 2019 for the TUVA database. The code is written in Python 3.7 and uses the libraries numpy, pandas, geopandas, rasterio and rasterstats to manage and process the data. For the machine learning the libraries of scikit-learn and Tensorflow are used.

The code is divided into sections: Import data, Calculations, Managing and Exploring the data and finally Machine Learning. The Calculation sections include the cloud mask, band median and standard deviation calculations for the semi-natural grasslands, NDVI calculations and obtaining geographical and biological information. The Manging and Exploring section process and visualizes the information gained from the calculations. It also sets the feature subsets. The final sections of Machine Learning sets the data up for the machine learning models, does hyperparameter tuning, trains, predicts and produces interpretable results based on confusion matrices.

The entire code is uploaded at GitHub and can be found at:

https://github.com/AdrianBaggstrom/master_thesis/blob/main/python_code.ipynb

Appendix C

Tables that shows the feature permutation importance for all machine learning models fitted with all 25 features. The mean importance is given with R^2 scores. B = band, std = Standard deviation, NDVI = Normalized difference vegetation index.

Logistic Regression

Feature	*Importance*	
	mean	*std*
B11 median	0.074	0.008
B12 median	0.056	0.009
B11 std	0.053	0.009
B05 std	0.048	0.008
B06 std	0.044	0.007
Latitude	0.043	0.008
B07 std	0.036	0.006
B04 std	0.034	0.007
B07 median	0.024	0.007
B06 median	0.021	0.006
B12 std	0.015	0.006
Area	0.014	0.004
Longitude	0.011	0.006
B05 median	0.007	0.006
B08 std	0.005	0.005

Random Forest

Feature	*Importance*	
	mean	*std*
Latitude	0.139	0.007
Area	0.084	0.006
Longitude	0.083	0.005
NDVI mean	0.019	0.003
NDVI std	0.016	0.002
B02 median	0.013	0.002
B03 std	0.012	0.002
B08 std	0.011	0.002
B11 std	0.011	0.002
B06 std	0.011	0.002
B08 median	0.011	0.002
B02 std	0.01	0.002
B07 std	0.01	0.002
B04 median	0.01	0.002
B12 median	0.01	0.002
B12 std	0.01	0.002
B05 median	0.009	0.002
B05 std	0.009	0.002
B07 median	0.009	0.002
B8A median	0.009	0.002
B8A std	0.008	0.002
B06 median	0.008	0.001
B04 std	0.007	0.002
B03 median	0.007	0.002
B11 median	0.007	0.002

SVM

Feature	*Importance*	
	mean	*std*
Latitude	0.104	0.008
B06 median	0.09	0.008
B03 median	0.071	0.008
B11 median	0.07	0.006
B07 std	0.068	0.005
B08 median	0.06	0.006
Longitude	0.054	0.006
B12 median	0.053	0.006
B02 median	0.051	0.007

Voting Classifier

Feature	*Importance*	
	mean	*std*
Latitude	0.126	0.008
B11 median	0.077	0.007
Area	0.067	0.004
Longitude	0.065	0.005
B12 median	0.057	0.006
B11 std	0.05	0.006
B06 std	0.046	0.005
B05 std	0.039	0.006
B04 std	0.032	0.004

B8A median	0.051	0.006	B06 median	0.031	0.005
B04 std	0.051	0.007	B07 std	0.03	0.004
B06 std	0.05	0.004	B02 median	0.026	0.004
B05 std	0.045	0.006	B8A std	0.024	0.004
B8A std	0.045	0.005	B12 std	0.023	0.003
Area	0.041	0.005	NDVI mean	0.021	0.003
B11 std	0.04	0.006	B08 median	0.02	0.003
B12 std	0.037	0.005	B05 median	0.02	0.004
B08 std	0.037	0.006	B04 median	0.02	0.004
B05 median	0.036	0.005	B08 std	0.018	0.004
B04 median	0.036	0.004	NDVI std	0.017	0.002
B03 std	0.036	0.004	B03 std	0.017	0.003
B02 std	0.029	0.005	B07 median	0.016	0.003
NDVI std	0.022	0.005	B03 median	0.016	0.004
NDVI mean	0.018	0.004	B02 std	0.013	0.003
B07 median	0.008	0.003	B8A median	0.01	0.002

Neural Network

Feature	*Importance*	
	mean	*std*
Latitude	0.024	0.006
Longitude	0.015	0.006
Area	0.014	0.006
B05 std	0.010	0.005
B07 median	0.009	0.004
B06 std	0.008	0.004
NDVI mean	0.007	0.004
B08 median	0.007	0.003
B05 median	0.004	0.004

Appendix D

The results produced without the feature of longitude and latitude included. PCC = Pearson's correlation coefficient, MI = Mutual information, std = Standard deviation, ROC AUC = Receiver Operating Characteristic, Area Under Curve, MCC = Matthew's correlation coefficient.

	area	NDVI std + area	PCC (8) features	MI (10) features	std (12) features	All (23) features	
Accuracy							**std**
Logistic Regression	0.551	0.551	0.547	0.553	0.556	0.545	0.004
Random Forest	0.551	0.529	0.529	0.532	0.523	0.521	0.010
SVM	0.551	0.551	0.547	0.536	0.542	0.514	0.013
Voting Classifier	0.549	0.538	0.545	0.532	0.545	0.542	0.006
Neural Network	0.547	0.575	0.547	0.534	0.527	0.501	0.022
F1							
Logistic Regression	0.711	0.711	0.699	0.708	0.698	0.671	0.014
Random Forest	0.645	0.620	0.648	0.642	0.640	0.643	0.009
SVM	0.710	0.710	0.704	0.694	0.691	0.635	0.026
Voting Classifier	0.700	0.692	0.698	0.692	0.697	0.692	0.004
Neural Network	0.644	0.690	0.700	0.691	0.675	0.516	0.064
ROC AUC							
Logistic Regression	0.575	0.549	0.572	0.545	0.560	0.554	0.011
Random Forest	0.562	0.544	0.528	0.528	0.533	0.543	0.012
SVM	0.449	0.529	0.490	0.507	0.496	0.494	0.024
Voting Classifier	0.575	0.548	0.549	0.544	0.549	0.565	0.011
Neural Network	0.575	0.563	0.514	0.532	0.503	0.491	0.031
MCC							
Logistic Regression	0.000	0.000	0.003	0.030	0.046	0.029	0.018
Random Forest	0.065	0.022	0.000	0.012	-0.013	-0.022	0.029
SVM	0.007	0.007	-0.015	-0.067	-0.006	-0.035	0.026
Voting Classifier	0.012	-0.035	-0.012	-0.092	-0.006	-0.008	0.034
Neural Network	0.054	0.111	-0.002	-0.064	-0.047	0.006	0.059
Precision							
Logistic Regression	0.551	0.551	0.551	0.554	0.558	0.558	0.003
Random Forest	0.572	0.559	0.551	0.555	0.548	0.545	0.009
SVM	0.551	0.551	0.550	0.545	0.550	0.542	0.004
Voting Classifier	0.553	0.547	0.550	0.543	0.550	0.550	0.003
Neural Network	0.568	0.577	0.551	0.544	0.543	0.555	0.012
Recall							
Logistic Regression	1.000	1.000	0.953	0.980	0.933	0.842	0.055
Random Forest	0.739	0.696	0.787	0.763	0.771	0.783	0.031
SVM	0.996	0.996	0.976	0.957	0.929	0.767	0.080
Voting Classifier	0.957	0.941	0.957	0.953	0.949	0.933	0.009
Neural Network	0.743	0.858	0.960	0.945	0.889	0.482	0.164

Stockholm University

SE-106 91 Stockholm

Phone: 08 – 16 20 00

www.su.se

Stockholms universitet