Crop identification with Sentinel-2 satellite imagery in Finland

Joona Laine

School of Engineering

Thesis submitted for examination for the degree of Master of Science in Technology.
Espoo 28.7.2018

Supervisor

Prof. Miina Rautiainen

Advisor

D.Sc. Jaana Mäkelä

Aalto University
School of Engineering

Aalto University
School of Engineering

Aalto University, P.O. BOX 11000, 00076 AALTO
www.aalto.fi
Abstract of the master's thesis

Author Joona Laine

Title Crop identification with Sentinel-2 satellite imagery in Finland

Degree programme Geoinformatics

Major Geoinformatics **Code of major** ENG22

Supervisor Prof. Miina Rautiainen

Advisor D.Sc. Jaana Mäkelä

Date 28.7.2018 **Number of pages** 82+2 **Language** English

Abstract

European Union member countries are obligated to control the validity of Common Agricultural Policy subsidy applications. Each member country performs manual inspection for at least 5% of these subsidy applications. This is both expensive and a considerable administrative burden. According to European Union, the crop type identification process in Common Agricultural Policy could be carried out using remote sensing or orthophoto imagery for an alternative to physical inspections by competent authorities. Automated crop type identification would reduce the costs significantly.

This master's thesis addressed the crop identification with optical Sentinel-2 satellite imagery in Finland. The aim was to investigate whether it was possible to reliably identify the crop growing in land parcels by using machine learning classification methods. This thesis presented an automated approach of identifying crops. Multiple different machine learning classification algorithms were trained and tested to find out the most suitable processing method, time period and classification algorithm by utilizing the land parcels obtained from the Finnish Agency for Rural Affairs.

The developed processing method and most of the tested classification algorithms were able to perform relatively well in crop identification in cloudy growth period 2017 of Finland. Therefore, the developed method could be applied to different use cases and cloudy weather conditions. The further development and training of the classification algorithms could make it possible to utilize this approach in Finland as well as in other EU countries for the Common Agricultural Policy control and possibly in numerous other tasks.

Keywords crop identification, remote sensing, Sentinel-2, machine learning

Aalto University
School of Engineering

Aalto-yliopisto, PL 11000, 00076 AALTO
www.aalto.fi
Diplomityön tiivistelmä

Tekijä Joona Laine

Työn nimi Viljelykasvien tunnistaminen Sentinel-2 -satelliittikuvien avulla Suomessa

Koulutusohjelma Geoinformatics

Pääaine Geoinformatics **Pääaineen koodi** ENG22

Työn valvoja Professori Miina Rautiainen

Työn ohjaaja Tekniikan tohtori Jaana Mäkelä

Päivämäärä 28.7.2018 **Sivumäärä** 82+2 **Kieli** Englanti

Tiivistelmä

Euroopan Unionin jäsenmaiden on noudatettava yhteisen maatalouspolitiikan tarjoamien maataloustukihakemusten valvomista. Jokainen jäsenmaa suorittaa manuaalisen valvonnan vähintään 5% tukihakemuksista. Tämä on sekä kallista, että huomattava hallinnollinen taakka. Euroopan Unionin mukaan viljelykasvin tunnistamisprosessin voisi suorittaa kaukokartoitus- tai ortokuvien avulla paikan päällä tehtävien tarkastuksien sijaan. Automaattinen viljelykasvin tunnistaminen vähentäisi valvonnan kustannuksia huomattavasti.

Tämä diplomityö käsitteli viljelykasvien tunnistamista optisten Sentinel-2 satelliittikuvien avulla Suomessa. Tarkoitus oli tutkia, pystyttäisiinkö koneoppimista hyödyntävien luokittelualgortimien avulla tunnistamaan pelloilla kasvavia maatalouskasveja. Tämä diplomityö esitteli automaattisen lähestymistavan viljelykasvin tunnistamiselle. Useaa erilaisia luokittelualgrotimia opetettiin ja testattiin kaikkein sopivimman prosessointimenetelmän, ajankohdan ja luokittelualgoritmin löytämiseksi Suomen oloihin Maaseutuviraston tarjoamien peltolohkojen avulla.

Kehitetty prosessointimenetelmä ja suurin osa testatuista luokittelualgoritmeistä suoriutuivat suhteellisen hyvin viljelykasvin tunnistamisesta Suomen vuoden 2017 pilvisellä kasvukaudella. Tämän vuoksi on mahdollista, että kehitetty prosessointimenetelmää voisi hyödyntää myös erilaisissa ilmastoissa ja eri käyttötapauksissa. Jatkokehityksen ja lisäopetuksen avulla luokittelumenetelmät voisivat mahdollistaa tämän lähestymistavan hyödyntämistä yleisen maatalouspolitiikan maataloustukihakemusten valvontaan Suomessa ja myös muissa EU-maissa muiden käyttötapausten lisäksi.

Avainsanat Viljelykasvin tunnistaminen, kaukokartoitus, Sentinel-2, koneoppiminen

Foreword

This master's thesis is written under Spatineo Inc, where I have had the opportunity to get familiar with machine learning and deepen my understanding in remote sensing and spatial data processing.

First I would like to thank the best possible pair of advisor and supervisor - Jaana Mäkelä and Miina Rautiainen - for all the invaluable help and encouragement during the research and writing process. I would also like to thank my manager Sampo Savolainen for all the help, fresh ideas, motivation and IT consulting during the project. I would like to thank Finnish Agency for Rural Affairs (MAVI) and Ministry of Agriculture and Forestry for the project funding. MAVI has also been an active partner in the project and has provided all land parcels used in the study.

Furthermore, I would like to thank Aalto University for the unique opportunity to study fascinating field of Geoinformatics in an inspirational environment with good friends. Finally, I would like to thank my whole family and especially my wife Outi for her everlasting support and faith that she has had in me and my dog Hilkka who was there to warm my feet even in the darkest of hours during the long nights of writing.

Helsinki, 28.7.2018

Joona Laine

Contents

Nomenclature and abbreviations

Symbols

γ	Kernel bandwidth of Support Vector Machine classifier
λ	Reliability level
ν	Clear pixel count
ρ	Cloud pixel percentage
ϱ	Shadow pixel percentage
ϕ	Days between valid images
φ	Total number of valid images
ω	True class of the classified object
C	Regularization parameter of Support Vector Machine classifier
$f1$	f1 Score
fp	Amount of false positives
fn	Amount of false negatives
$\mathbf{h}$	Vectors of neurons of a hidden layer
K	Cohen's Kappa coefficient
P	Posterior probability
p_e	Expected agreement ratio
tp	Amount of true positives
tn	Amount of true negatives
$\mathbf{W}$	Weights of the neural network
x	Object from input data data for the classifiers
$\mathbf{x}$	Input data for the classifiers
X_1	Supervised parcels that contain one of the 10 crop classes used in the study
X_2	Parcels digitized by farmers that contain one of the 10 crop classes used in the study
$\hat{y}$	Class labels, i.e. output of the classifiers

Abbreviations

ACP	Automatic Classification Proportion
AOI	Area Of Interest
AWS	Amazon Web Services
BOA	Bottom-Of-Atmosphere
CAP	Common Agricultural Policy
CNN	Convolutional Neural Network
CPP	Cloud Pixel Percentage
ConvRNN	Convolutional Recurrent Neural Network
CV	Cross-validation
DN	Digital Number
DOY	Day Of the Year
EO	Earth Observation
EU	European Union
FNN	Multilayer Feedforward Neural Networks
GNDVI	Green Normalized Vegetation Index
IACS	Integrated Administration and Control System
KELM	Kernel-based Extreme Learning Machine
L1C	Sentinel-2 Level 1C
L2A	Sentinel-2 Level 2A
LCC	Land Cover Classification
LPIS	Land Parcel Identification System
MAVI	Finnish Agency for Rural Affairs / Maaseutuvirasto
MCARI	Modified Chlorophyll Absorption in Reflectance Index
MGRS	Military Grid Reference System
MLP	Multilayer Perceptron
MSI	Multi-spectral Instrument
NDVI	Normalized Difference Vegetation Index
OA	Overall Accuracy
OAA	One-against-all
OBIA	Object-based Image Analysis
PA	Producer's Accuracy
PBIA	Pixel-based Image Analysis
PBC	Parcel-based Classification
PCA	Principal Component Analysis
RBF	Radial Basis Function
RMSE	Root Mean Square Error
RF	Random Forest
RNN	Recurrent Neural Network
RS	Remote Sensing
SAR	Synthetic-aperture Radar
SVM	Support Vector Machine
TA1	The first test area covering the whole Finland
TA2	The second test are covering one tile in Finland
TOA	Top-Of-Atmosphere
TWDTW	Time-Weighted Dynamic Time Warping
UA	User's Accuracy
VI	Vegetation Index

1 Introduction

1.1 Background and motivation

There are 22 million farmers and agricultural workers in the European Union (EU) agri-food sector. To ensure a decent standard of living to the farmers, EU is supporting them with the common agricultural policy (CAP). CAP aids farmers with direct income support and market measures and it also ensures sustainable rural development individually for each EU country according to its needs. (European Comission 2017b.) The budget of CAP is significant part of the EU yearly budget. In 2016, the amount of CAP direct subsidy payments was over 40 billion euros (European Comission 2017a).

The system for the management and control of payments to the farmers is called the Integrated Administration and Control System (IACS) (European Comission 2017c). IACS includes a computer database for storing all agricultural areas eligible for a direct payment in a Member State. This database is called Land Parcel Identification System (LPIS). LPIS ensures that the payments are distributed according to the correct area of the parcels. (European Comission 2018.) LPIS contains geo-referenced polygons of land parcels (later parcels), land cover information as well as the identifier of the crowing crop in each parcel among with other information. The parcel geometries are digitized, and the parcel information is announced by the farmers when applying for the subsidies. (Hart et al. 2017.)

In order to ensure that the subsidies are divided equally and without misuse, the CAP subsidies must be controlled by the local authorities within IACS. At least 5% of the agricultural parcels, that the subsidies are applied for, have to be monitored for each year. In Finland the Finnish Agency for Rural Affairs (MAVI) is responsible for organizing the monitoring. (Maaseutuvirasto 2018b.) The controlling of agricultural parcels includes identifying of the crop, measuring of the parcel area and checking that the farm has followed the other relevant subsidy conditions (Maaseutuvirasto 2018a). The supervised parcels are then digitized again to LPIS with the checked boundaries. Until now the monitoring has been done by visiting the farms or manually studying ortophotos or using other similar methods.

In 2014 European Comission regulation No. 807/2014 the new technologies are suggested to be used for the control. One of the suggested methods is the monitoring using Remote Sensing (RS) imagery (European Comission 2014). The Finnish Parliament statement about the regulation, suggests multiple alternatives for the in-situ monitoring. One of the alternatives is using the Sentinel-1 and Sentinel-2 satellite imagery. Sentinel-1 synthetic-aperture radar (SAR) sensors are able to produce imagery despite the weather conditions and would therefore produce certain time series of data. However, the optical sensors of Sentinel-2 produce images with higher spatial resolution allow more accurate interpretation. Therefore, the use of combination of Sentinel-1 and Sentinel-2 imagery is recommended by the Finnish Parliament. (Finnish Parliament 2017.)

With the help of classification algorithms, it is possible to identify crops automatically. There are multiple modern studies covering the crop identification based on RS data (Peña et al. 2014, Lussem et al. 2016) and even in the context of CAP (Schmedtmann and Campagnolo 2015, GISAT s.r.o 2017). Most of these studies have used other data sources than Sentinel-2, but there are also several that have utilized Sentinel-2 as well (Belgiu and Csillik 2018, GISAT s.r.o 2017), Rußwurm and Korner 2018). There are however, no such classification crop identification study published in Finland where the cloudy years cause real challenges for the monitoring during the growth period.

1.2 Aim of the study

The aim of the study was to investigate the reliability of identifying parcel crops based on Sentinel-2 imagery in Finland using different machine learning classification methods. The aim was to find the most suitable classification method and the most suitable time range inside the crop growth period for the task. The study also investigated which crop classes could be best identified among the chosen crop classes.

The study investigated what methods have been used in RS crop identification in previous studies to gain a wide perspective on the issue. A case study was performed for the year 2017 for identifying some of the most important crops growing in Finland. The thesis focused on the use of Sentinel-2 imagery, since there are 2 available satellites and good temporal resolution of the images available free of charge. The other IACS tasks, such as measuring the parcel area, were outside the scope of this thesis.

The study was performed using Python programming language and Amazon Web Service (AWS) cloud computing resources (Amazon Web Services 2018). The partial spatial LPIS database for the year 2017 provided by MAVI including the agricultural parcels, was used for training and testing of the classification methods. The results of the case study were compared to the results of the previous studies aiming to find the most reliable method for the weather conditions of Finland.

1.3 Structure of the thesis

This thesis is divided in to seven chapters. Chapter 2 presents the study methods and main outcomes of existing crop identification studies using satellite data and machine learning methods. Chapter 2 discusses also the key findings and problems found in the literature.

The study area and all the data sets used in the study are described in detail in Chapter 3. The methods used in the study are presented in Chapter 4. The Chapter covers gathering and preprocessing of the data and calibration of the classification

models as well as the brief theory of the used several processing and classification methods.

Chapter 5 presents the collected results and key findings for all the methods of the study regarding crop identification. The further discussion of the results and findings are discussed in Chapter 6. Finally, the conclusions of the study are wrapped in Chapter 7.

2 Literature review

In RS approach of agricultural crop identification, the aim is to recognize and correctly classify the crop growing in the parcel using the RS imagery. This is not a new area of research, since in 1970 Haralick et al. (1970) utilized three different statistical methods in order to discriminate crops from the radar RS imagery. In this study however, to aim was to use optical RS imagery for the task. Every material has a unique spectral signature, which is the reflectance as a function of wavelength, and it is therefore possible to recognize the different materials, or crops, from each other using optical RS imagery (NASA 2013). The crop identification, or classification, is a subset of the problem domain of land cover classification (LCC), which aims to classify every pixel in RS images to a land cover class. In LCC there have been studies focusing in only mono-temporal imagery, meaning that the classification is performed for a single image acquisition date. (Rußwurm and Korner 2017.)

Crops, like all periodically growing vegetation, have species specific changes of the spectral and textural appearance depending on the growth cycle. This make crops difficult to be classified with mono-temporal approach. Also, the weather can make the crop development start earlier or later, so the amount and correct timing of the RS images are important factors for the crop identification. (Foerster et al. 2012.) The study of the vegetation specific growth cycles caused by the specie's biology is called phenology (Rußwurm and Korner 2017).

In crop identification the aim is to generate a classification model into which the satellite imagery and parcel geometries can be fed as an input and a crop class will be received as an output for each parcel. The aim is to maximize the ratio of correctly classified parcels in the classification process. There are lot of aspects affecting the performance of the classification: chosen classification method, spectral characteristic similarities of the crops (Schmedtmann and Campagnolo 2015) and weather conditions, such as cloudiness (Belgiu and Csillik 2018), temperature and precipitation of the study area (Foerster et al. 2012). Therefore, an identification method working well in one agro-region, i.e. area with similar climate, soil and agro-technical conditions, cannot be assumed to work as well in another, different environment (Foerster et al. 2012).

In this study, the term "classification" means supervised classification where the model learns to classify the unseen samples into classes based on the trained samples containing all the classes. The other classification approach is unsupervised classification, or clustering, where the model finds patterns, or clusters, from the data. The classification is a form of supervised machine learning. (Hutson 2017.) In crop identification the supervised classification seems to be preferred, since the unsupervised classification classes could give multiple classes inside one crop and leave some crops out completely.

In recent studies of crop identification some of the most popular and most promising machine learning classification methods have been Support Vector Machine (SVM)

(Schmedtmann and Campagnolo 2015, Peña et al. 2014, GISAT s.r.o 2017), Random Forest (RF) (Duro et al. 2012, Niculescu et al. 2018) and Time-Weighted Dynamic Time Warping (TWDTW) (Belgiu and Csillik 2018). In addition to these traditional classification systems (Rußwurm and Korner 2017), there are also studies done with deep learning classification algorithms, such as Multilayer Perceptron (MPL) (Peña et al. 2014), Convolutional Neural Network (CNN) (Rußwurm and Korner 2017), Recurrent Neural Network (RNN) (Ienco et al. 2017, Niculescu et al. 2018, Rußwurm and Korner 2017) and Convolutional Recurrent Neural Network (ConvRNN) (Rußwurm and Korner 2018). The selected classification methods are described in more detail in Chapter 4.4. Some of the studies that were closest to the aims of this study are listed in Table 2.

2.1 Study areas and data sets in previous studies

As can be seen from Table 2, there were multiple different sized test areas in the studies. Usually the study area was covered by at most few satellite images, other had multiple Area of Interests (AOI) to generalize the results (Ienco et al. 2017). Since climate affects crop production greatly (Iizumi and Ramankutty 2015), every model should be trained with the similar climate. The climate and annual temperatures and precipitation varied greatly among the studies. Time period of the studies also varied. In Czech Republic and Germany, where the climate was more similar to the climate of Finland, the study period for optical images were between January and September (GISAT s.r.o 2017) , and December and September (Rußwurm and Korner 2017).

In order to do the crop classification using satellite images, one needs the satellite data and information about the crops crowing in the parcels of the study area. The RS data could be gathered from multiple different optical sources, like RapidEye (Lussem et al. 2016), Landsat 7 or 8 (Schmedtmann and Campagnolo 2015), (GISAT s.r.o 2017), Sentinel-2 (Belgiu and Csillik 2018, GISAT s.r.o 2017, Rußwurm and Korner 2017, Rußwurm and Korner 2018) or ASTER (Peña et al. 2014). Different satellite sensors have different characteristics, such as spatial and spectral resolution, which make the choice of used sensor matter. Apart from optical satellite imagery, it is possible to gather images from other sources. Honkavaara et al. (2017) even suggested the use of unmanned aerial vehicles for precision tasks in addition to Sentinel-2. Some of the studies have proved that SAR-sensors, like Sentinel-1, are also strong addition to the optical ones (Lussem et al. 2016, GISAT s.r.o 2017, Niculescu et al. 2018). Other optical and SAR satellite sensors than Sentinel-2 were however outside the scope of this study.

The number of different image acquisition dates were between 1 and 247 (Rußwurm and Korner 2018). Usually the images chosen to cover the study area were free or relatively free of clouds. Belgiu and Csillik (2018) set the cloud coverage limit for the images used to the maximum of 9%. GISAT s.r.o (2017) used cloud-free image mosaics by applying the multi-temporal image compositing. They also generated

cloud and snow masks in order to identify clouds and snow to be replaced with clear pixels from the images of the nearest dates. Rußwurm and Korner (2018) used every image available for the study area during 2016 and 2017 that had the cloud coverage less than 80%. In Finland, the majority of the cloud free images of the year have typically been acquired between May and November (Saarinen et al. 2018). However, in 2016, the monthly average of cloud cover was 30% on May and October and 50-60% on April, June, July, August and September in one study area in Finland (Honkavaara et al. 2017).

The classification of the crops requires known information about the geometries of the parcels and about the crop growing in the parcels to get the ground truth data. To get the geometries of the parcels by classifying the land use from the RS imagery, two general image analysis approaches can be used: pixel-based image analysis (PBIA) and object-based image analysis (OBIA) (Duro et al. 2012). For PBIA, every pixel is used individually, whereas in OBIA the objects can be gathered from the pixels automatically or semi-automatically using methods like segmentation (Duro et al. 2012). However, segmentation has found to be challenging task as it requires defining the optimal parameters (Belgiu and Csillik 2018).

In addition to these two, Schmedtmann and Campagnolo 2015 used term "parcel-based classification" (PBC) to address the situation where the parcel geometries are gathered from other sources without need of segmentation. The crop crowing in the parcels could be identified using existing information from spatial databases, such as LPIS (Lussem et al. 2016, Schmedtmann and Campagnolo 2015) or by visiting the site and identifying the crop manually (Peña et al. 2014). Peña et al. (2014) drew borders for selected 350 parcels manually for the ground truth reference.

The spatial resolution of the satellite sensors set the restriction for the parcels used in the classification. Some studies set the minimum area threshold for the parcel areas, GISAT s.r.o (2017) used the threshold of 1ha that cut off 0.9% of the total arable land from the LPIS used. Crop groups could be used instead of crops as a ground truth classes by aggrecating some of the similar crops together. The study areas might contain some other crops, but the parcels would then be dropped for having the crop outside the study. (Rußwurm and Korner 2018.) The number of classes, or labels, i.e. different crops or crop groups, used in the studies typically varies between 5 (Peña et al. 2014) to 18 (Rußwurm and Korner 2017).

2.2 Processing of the satellite data in previous studies

The pixel values for each sensor bands are called Digital Numbers (DN) and they do not represent any physical value. Every satellite product has a method of converting DN into more descriptive value such as radiance or reflectance. However, processing and multiple corrections, such as radiometric corrections, and ortirectification have to be made to the raw RS imagery in order to extract the most meaningful quantity, reflectance out of it (SUHET 2015). The processing of the RS data is defined as

several processing levels. For different satellite sensors, the processing differs, but the first general levels are described in Table 1. Sentinel-2 processing levels were described in more detail in Chapter 3.2.

Table 1: The processing levels of earth observation data (NASA 1986)

Level Name	**Description**
Level-0	Full resolution reconstructed data
Level-1A	Time referenced, full resolution and annotated data including additional information. Radiometric and geometric coefficients and georeferencing parameters computed but not applied
Level-1B	A data, such as radiance, that has been processed to sensor units. This level is not necessarily included with all sensors
Level-2	Derived environmental variables (i.e. reflectance) from the Level 1 data

Top-of-Atmosphere (TOA) reflectance requires Level-1 processing and it can be calculated using TOA radiance. TOA reflcetance has the effects of the atmosphere included, but after Level-2 atmospheric correction, which removes the effects of the atmoshpere, it can be turned into the Bottom-of-Atmosphere (BOA) reflectance. (NASA 2013). Images of some of the satellite sensors, like Landsat 7 and Sentinel-2, are served to include DN that can be converted to TOA reflectance values. The atmospheric correction has to be done by the user (NASA 2013, SUHET 2015). Sentinel-2 images that have DN corresponding to the BOA reflecance as Level-2 processed product, are also partially available to the users (SUHET 2015). As can be seen in Table 2, most of the studes used BOA reflectance values, while few got along with TOA reflectance.

The RS images are georeferenced to some coordinate systems. Sometimes it is essential to project the images to other coordinate systems so that they can be used with other spatial data, like parcels. However, reprojection of the RS images lead to degradation of the fidelity of the data (Roy et al. 2016). Other alternative is to project the other spatial data to the coordinate systems of the images.

Since the satellite sensor bands have different spatial resolution (pixel sizes), most of the studies suggest resampling all the used RS data and bands to a specific spatial resolution. This means applying some resampling method, such as nearest-neighbor or area- or distance-weighted resampling, in order to achieve the smaller or larger spatial resolution for the images (Galbraith et al. 2003). Lussem et al. (2016) and Peña et al. (2014) resampled the 30m bands to 15m with nearest-neighbor method while GISAT s.r.o (2017) resampled 10m and 20m bands to 20m. (Rußwurm and Korner 2018) resampled 10m, 20m and 60m bands to 10m using bilinear interpolation. Resampling causes distortions to the data (Roy et al. 2016) and also makes the physical size of the images bigger. However, the studies have not reported resampling affecting the results negatively under our knowledge.

2.3 Creating the data for the classification

The usual aim is to get the RS data as a singular stacked multitemporal multiband image (Lussem et al. 2016, Schmedtmann and Campagnolo 2015). To extract just the parcel values, the multitermporal multiband image could be masked with parcel boundaries (Lussem et al. 2016). The data would then consist multiple pixel values for each parcel for each bands in the multitemporal image. Each pixel value could be used on their own resulting to pixel-based classification, or some bandwise statistical, or textural features (Peña et al. 2014), could be calculated and used to represent the parcel resulting object-based classification, or PBC. If multiple crops grow on a single parcel, it will cause problems when using object-based classification (GISAT s.r.o 2017). Pixel-based classification does not suffer from this problem.

The number of optical bands utilized from the sensors varied between studies as shown in Table 2. For Sentinel-2 data, Sonobe et al. (2017) recommended using just 10m and 20m bands stating that the 60m bands were left out due to their mainly dedication to cloud detection and atmospheric correction. Rußwurm and Korner (2018) on the other hand used all the 13 bands, while Belgiu and Csillik (2018) used only four bands, three visible bands (B2, B3 and B4) and near-infrared band (B8). The Sentinel-2 bands are described in more detail in Table 4.

Although the band reflectance values could be used as is as input data for the classifier, many of the studies suggested using some vegetation index (VI), like Normalized Difference Vegetation Index (NDVI) (Duro et al. 2012, Ienco et al. 2017, Belgiu and Csillik 2018). Peña et al. (2014) even suggests using of ten different VIs to get the maximum classification results. VIs are spectral transformations that consists two or more bands. The purpose of VIs is to emphasize the input of vegetation properties. With VIs it is possible to monitor and compare temporal changes of vegetation structural, phenological and biophysical parameters. (Huete et al. 2002.)

If cloudy images were to be used with masks excluding some of the pixels and images for some of the parcels, the amount of data for parcels might vary. The classifiers usually expect the length of the input data vectors to be equal and thus some processing is required. Multitemporal linear interpolation for each parcel and each band is one way of getting same sized data vectors (Ienco et al. 2017). Other way is to use day of the year (DOY) as one of the variable and filling the rest of the variables with values that the models understand to leave unnoticed (Rußwurm and Korner 2017, Rußwurm and Korner 2018).

If the model performs well whit the data it has been tested on, but fails to generalize to the unseen data, the model might have overfitted to the data (Buitinck et al. 2013). Reduction of the dimensionality in the data is a suggested processing step in supervised classification because it reduces the risk of overfitting and possible redundancy. The use of textural features in representing the whole parcel is one way of reducing the dimensionality. Schmedtmann and Campagnolo (2015) investigated whether the set of all pixel reflectances could be replaced by the average of the set by investigating whether the variability between parcels was higher than within parcels.

They calculated the F-value as shown in Equation 1 to find out the variability. To find out whether the simple average would be a suitable statistic, they calculated Root Mean Square Error (RMSE) as shown in Equation 2.

$$F_{\text{parcels}} = \frac{\text{between-parcel mean square}}{\text{within-parcel mean square}}. \tag{1}$$

$$\text{RMSE} = \sqrt{\text{within-parcel mean square}}. \tag{2}$$

The high F-value means that the single parcels are homogeneous and low $RMSE$ suggests that average could indeed replace the set of pixel reflectances. Their investigation revealed that average could be used without losing much of the relevant information in their study area in Portugal. (Schmedtmann and Campagnolo 2015.)

Schmedtmann and Campagnolo (2015) also tried to find the most significant sensor bands for reducing the variables. They used Principal Component Analysis (PCA) to detect linear relations and utilized the method of Jolliffe (2002) to trace back the original, most significant, variables. The classification was computed with both all sensor bands and with two most significant sensor bands for each image date. The difference in classification accuracy was only 0.29 percentage points with all sensor bands giving slightly better results. The results suggest that the original variable set could be replaced with the set with reduced dimensionality. (Schmedtmann and Campagnolo 2015.)

2.4 Classification in crop identification

Before the classification the processed data has to be divided to the training and testing sets. If there is lot of tuning in the model, the data is often divided into three sets instead: training, testing and validation (Rußwurm and Korner 2017, Rußwurm and Korner 2018). The training set is used to train the classification model, while testing and validation sets are used to validate the model with unseen data. However, the testing set is used during the model tuning and the validation dataset is used only after the model is in its final phase to ensure that the information of the test set does not leak into the model and the model would not be able to perform well with unseen data. One way to avoid splitting the data into three sets, and thus reducing samples in the training set, but still prevent overfitting during optimization, is to use k-folded Cross-validation (CV). In CV the model tuning and optimization is done only with the training set by folding it k times. Each fold of data acts as a training data, while some other fold acts as an evaluation data. Each training sample then acts both as part of training and testing sample. The testing set is then used as a final model evaluation. (Buitinck et al. 2013.) The amount of the training samples for each crop or crop group class should also be sufficient in order to algorithm to learn enouch (Lussem et al. 2016, Peña et al. 2014).

The most of the supervised classification methods used in the studies are parametric and thus require setting of one or more parameters. These parameters are called hyperparameters and setting those is called hyperparameter tuning or optimization. In order to ensure receiving the best possible classification results for the data with given method, the studies suggest tuning the hyperparameters with different parameter sets and validating the results with k-folded CV. Peña et al. 2014 suggested 5-folded and Schmedtmann and Campagnolo 2015 suggested 10-fold CV for hyperparameter tuning. Some of the parameters were found to be more effective than others for crop identification, e.g. kernel function Radial Basis Function (RBF) is often preferred over linear kernel function in SVM even without CV (Schmedtmann and Campagnolo 2015).

There were two popular hyperparameter optimization methods among that are used in the studies: grid search optimization (Bergstra and Bengio 2012), used by (Rußwurm and Korner 2017, Peña et al. 2014, Schmedtmann and Campagnolo 2015) and Bayesian optimization (Sollich 2002), used by (Sonobe et al. 2017). In both methods all testable values for all tunable parameters are given. For Bayesian optimization, the values could be continuous distributions as well. Optimizers then combine systematically the parameters and train the model using CV. Grid search method enumerates trough all possible parameter combinations when trying to find the best parameter combination. Bayesian optimization, however, does not try all the possible parameters, but it utilizes a predictive model to model the whole hyperparameter combination search space and tries to find good values as quickly as possible. (Head et al. 2018.) Third alternative optimization method is random search, which selects the subset of parameter combinations in a random way. Random search is preferred over grid search when the search space has large number of different parameter values. (Bergstra and Bengio 2012.)

The classification can consist of one or multiple classifiers. Peña et al. (2014) used hierarchical classification in order to differentiate woody and herbaceous crops from one another and the second classification to actually identify the crop. One approach of hierarchical classification is to train individual classifier for each class in one-against-all (OAA) manner. In OAA each of the K classifiers tries to classify the class K against other (K-1) classes. The results can be then joined together. (Marir et al. 2018.) It is also possible to chain multiple classification algorithms together, i.e. Ienco et al. (2017) used the data from RNN second final layer as an input to SVM classifier.

The classification decision means that a classifier decides that the object x of input data $\mathbf{x}$ belongs to a class based on the variables of the input vector in the classification process. Many classification algorithms, such as RF, MLP and SVM, output an individual estimated posterior probability of the object x belonging to each class ω. Those posterior probabilities are denoted as $P(\omega_j|x)$, e.g. $P(\text{grass}|x)$ is the posterior probability of crop being grass with object signature x. The highest probability estimate for the i-th object is denoted as p_i and decides the final class to that the classification decision assigns the object. (Schmedtmann and Campagnolo 2015.)

GISAT s.r.o (2017) used independent classification for optical and SAR time series and combined the results afterwards into a single product using maximum posterior probability values p_i of each classifier.

2.4.1 Quality assurance

The common ways to assess the quality and performance of the crop identification model are confusion matrix (also called error matrix), overall accuracy (OA), user's accuracy (UA), producer's accuracy (PA), f1 score (f1) and Kappa coefficient (K) (Lussem et al. 2016). In literature, UA is also often called recall and PA is called precision. Confusion matrix compares for each class the relationship between the classification results and the known ground truth data. Metrics OA, UA, PA and f1 can be calculated using it. As can be seen from Figure 1, the columns represent the ground truth classes and the rows represent predicted classes. The cell value can be absolute value of how many objects that have ground truth of the column are predicted to belong to the class of the row. Often the cell value is however UA, and the the values of the row would then sum up into 1.0 and thus giving. Cell values could then be interpreted for each ground truth class as percentage of objects that were classified into class of the column. In Figure 1 cell values represent UA.

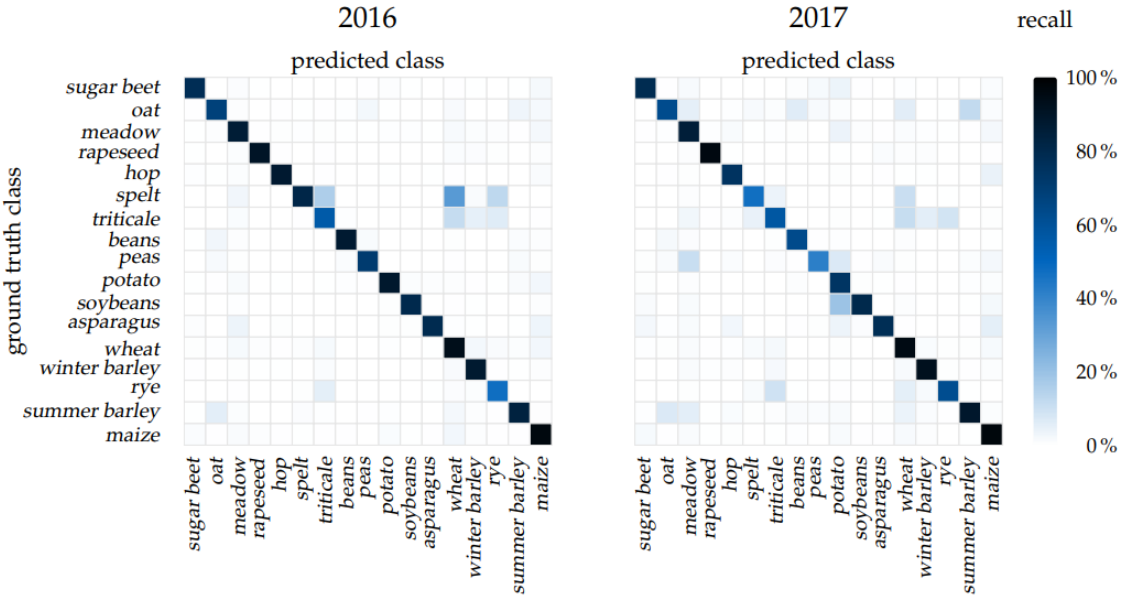


Figure 1: Example of a confusion matrix. Cell color corresponds to the user's accuracy of the class. (Rußwurm and Korner 2018.)

For each class, every predicted object falls into one of the four categories: true positives *tp*, true negatives *tn*, false positives *fp* and false negatives *fn*. *tp* includes the parcels that were rightfully predicted to belong to the class. *tn* includes parcels that were rightfully predicted to belong to some other class. *fp* and *fn* include the

parcels that were wrongly predicted to belong to the class and not to belong to the class respectively. (Buitinck et al. 2013.)

OA represents the total accuracy of the model in one value and is calculated simply by using the Equation 3. However, inside each class more meaningful values are given by PA and UA. PA can be calculated using Equation 4 and it tells how well validation objects of the class are classified, i.e. how many percentage of objects that were predicted to belong to the class actually belong to the class. UA, that can be calculated with Equation 5, on the other hand tells the percentage of how many samples that belong to the class were predicted correctly. f1 tries to join the information value of UA and PA by being the harmonic mean of the both as can be seen from the Equation 6. (GISAT s.r.o 2017, Buitinck et al. 2013.)

Many of the studies suffered from imbalance, Rußwurm and Korner (2017) had largest class with 28000 parcels and smallest with less than 600 parcels, for Ienco et al. (2017) the respective numbers were 3882 and 81. Non-uniform or imbalanced crop class distribution is not considered optimal, since it skews the OA towards the high frequency classes. Cohen's Kappa coefficient K can compensate the imbalanced distributions and is therefore good to include if there is imbalance between classes. (Rußwurm and Korner 2018). K is defined using Equation 7, where p_e is the expected agreement ratio (Jacob 1960).

$$\mathrm{OA} = \frac{N_{\mathrm{correct}}}{N_{\mathrm{total}}}. \tag{3}$$

$$PA = \frac{tp}{tp + fp}. \tag{4}$$

$$UA = \frac{tp}{tp + fn}. \tag{5}$$

$$f1 = 2\frac{PA * UA}{PA + UA}. \tag{6}$$

$$K = \frac{OA - p_e}{1 - p_e} \tag{7}$$

OA, UA and f1 are class specific, so in order to show just one value of those metrics, some aggregating is needed. Two of the most common methods are micro and macro averaging. Micro averaging counts sums of tp, fn and fp from all the classes for calculating the metrics, while macro averaging calculates the unweighted average of the metric of each class. (Buitinck et al. 2013.) For imbalanced datasets, it is common to use micro averaging to include the weights (Rußwurm and Korner 2018). If the aim is to predict each class equally well, the macro metric could be more suitable.

2.4.2 Classification calibration

It is possible to calibrate the results of the classification with the help of the posterior probabilities $P(\omega_j|x)$. Schmedtmann and Campagnolo (2015) wanted to have high classification accuracy for each class in exchange of lowering the proportion of the accepted classified parcels for each class. This proportion is called Automatic Classification Proportion (ACP). Schmedtmann and Campagnolo (2015) utilized reliability level λ to set a limit of $UA >= \lambda$ for each class. This could be achieved by excluding parcels that had low posterior probabilities until the limit was reached. The limit for the posterior probabilities could be different for each class. The limits are denoted as $\{q_1, ..., q_c\}$, where c is a number of classes. Schmedtmann and Campagnolo (2015) managed to rise OA from 68.1% to 84.1% with overall ACP of 55.4%, i.e. they excluded 44.6% of their validation data to get significantly better classification accuracy.

2.4.3 Comparison of classification methods in previous crop identification studies

Most of the studies were satisfied with the identification results that they achieved. However, there were differences between the classification performances. The best preformed classification methods in the order of best OA to worst based on the Table 2 were KELM, SVM, multilayer feedforward neural networks (FNN), RF, TWDTW (OBIA version of the study Belgiu and Csillik 2018), ConvRNN and MLP. Since number of images, bands, crop classes, bands and parcels varied between studies and different sensors were used, comparing just OA scores of the classification does not tell the whole truth. Also most of the studies had imbalanced datasets, which makes comparing just OA scores short-sighted. Since the studies were performed with varying agro-regions and different sized study areas, one algorithm might as well work better in one agro-region than in another. It was therefore really common to compare different classification algorithms to find out which algorithm would best in the AOI at hand.

Many studies compared both traditional machine learning and deep learning classification approaches. Often the deep learning outperformed the traditional methods (Rußwurm and Korner 2017, Ienco et al. 2017), but they were also sometimes equally good (Sonobe et al. 2017), or even worse than traditional methods (Ienco et al. 2017). It would therefore be beneficial to use both approaches in comparison.

In studies, both pixel-based and object-based classification approaches were used. Duro et al. (2012) and Belgiu and Csillik (2018) compared both approaches with same classification algorithms. Duro et al. (2012) did not find any statistical difference between approaches, but they stated that object-based approaches were significantly more time consuming to produce, but based on OA, the choice of particular approach would make no difference. However, Belgiu and Csillik (2018) concluded that object-based classification outperformed the pixel-based counterparts

in every AOI.

2.5 The approach in this study

In this study, the aim was to compare different machine learning classification methods. Out of the best performing classification methods, four methods were chosen for the case study of this thesis: SVM, RF, MLP and ConvRNN. SVM and RF represent the traditional or standard machine learning classification strategies while MLP and ConvRNN represent neural network and deep learning strategies.

SVM was chosen because of it is the most popular algorithm in the studies and it has performed relatively well in many cases. It is also relatively easy to use and optimize using Scikit-learn implementation (Pedregosa et al. 2011). SVM has also only two parameters that are tuned based on the studies, so hyperparameter tuning for crop identification is relatively straight-forward (Schmedtmann and Campagnolo 2015). SVM has been described to be a base classifier, which has been used to compare other methods with (Rußwurm and Korner 2017, Peña et al. 2014). RF was chosen also because of its popularity and relatively good performance in the studies. Also because of its structure, RF can handle thousands of dependent input variables and should be robust against overfitting (Sonobe et al. 2017). This makes it suitable method for large datasets.

For neural networks, MLP was chosen to represent a simple, deep learning classification method. MLP can approximate any smooth, measurable function between input and output (Gardner and Dorling 1998). It would be thus possible, that MLP could find spectral similarities that would be impossible to find using traditional machine learning method. MLP has also been widely used in various RS problems (Peña et al. 2014). ConvRNN was chosen, since it is one of the newest, most promising methods in the field of crop identification. Rußwurm and Korner (2018) developed ConvRNN with Sentinel-2 images in mind to the use of LCC. ConvRNN was able to detect and ignore clouds from the images, so no cloud masking was required. One downside of the ConvRNN is that it takes long time to train. Rußwurm and Korner (2018) trained two variations for 58h and 51h for an AOI of size $420km^2$.

Since the growing seasons of Finland tend to be cloudy (Finnish Meteorological Institute 2018a), as many Sentinel-2 images as possible were tried to use with every method to ensure that every parcel would have enough data. As there would be images with partial cloud coverage, the ConvRNN naturally learns to ignore the clouds but for other algorithm, masking of clouds was adopted as a processing step.

Since there were ready geometries of the parcels available, as explained in Chapter 3.3, the PBC was selected as the classification approach for methods SVM, RF and MLP. ConvRNN is a pixel-based classification method by nature, so both approaches could be tested.

Table 2: Overview about the recent studies. The accuracy means Overall Accuracy (OA) for the crop or land cover classification. Explanations: RE=RapidEye, Sx=Sentinel-x. Lx=Landsat-x, R=Radiometric, O=Orthorectification, A=Atmospheric, R(x)= resampled to x m, P=reprojection. ? means that that information was not available from the study

Study	Study area	Sensor	Preprocessing	Features	Classifier and accuracy (%)	# of images/ bands/classes/ objects or pixels
Lussem et al. 2016	Rur-Watershed, Germany	RE S1	R O A R(15) P	BOA and C-band	SVM: 91, MLC: 83	1+3/?/5/?
Peña et al. 2014	Yolo County, USA	ASTER	R A R(15)	10 VIs and Textural features	SVM+SVM: 89, SVM: 88, MLP: 88, LR: 86	2/8/9/1007
GISAT s.r.o 2017	Whole Czech Republic	S1 S2 LS7 LS8	R O A R(20)	C-band and BOA	SVM: 87	240/?/7/4491
Schmedtmann and Campagnolo 2015	6390km^2 in Portugal	LS7	R A	BOA mean	SVM:68	6/6/12/11852
Rußwurm and Korner 2017	4284km^2 in Germany	S2	R A	BOA	LSTM:76, RNN:75, CNN:60, SVM:32	26/6/18/137000
Rußwurm and Korner 2018	4284km^2 in Germany	S2	R R(10)	TOA	ConvRNN:90	247/13/17/137000
Ienco et al. 2017	420km^2 in France	Pléiades	R O	NDVI and BOA textural features	RF:84, SVM:72, LSTM:76, LSTM+RF:76, LSTM+SVM:74	122/4/11/15196
Ienco et al. 2017	226km^2 in Reunion Island (France)	L8	R A	3 Vis and BOA	RF:81, SVM:82, LSTM:86, LSTM+RF:86, LSTM+SVM:86	23/7/9/37900
Belgiu and Csillik 2018	640km^2 in Romania	S2	A P	BOA NDVI	OBIA TWDTW:95	13/4/6/601
Sonobe et al. 2017	Hokkaido, Japan	S1 S2	R(10)	C-band and TOA average	SVM:96, RF:96, FNN:96, KELM:97	5+1/1+10/6/4719

3 Materials

3.1 Study Area Description

The aim in this study was to find the suitable method in whole Finland and therefore the goal was to utilize every available parcel. As can be seen from Table 2, GISAT s.r.o (2017) used the whole Czech Republic as an AOI. Inspired by that, the study area of whole Finland that has any parcels, was selected as test area 1 (TA1). Since Sentinel-2 images follow the Military Grid Reference System (MGRS) tiles (NGA 2006), the tiles where there were no parcels are outside of the study. The total amount of tiles with parcels was 57. The training of the ConvRNN was extremely time consuming, as mentioned in Section 2.5. Therefore, only one tile (34VFN) was selected as a test area 2 (TA2) for evaluating the performance of ConvRNN method as can be seen from Figure 2. There were thus actually two AOIs, one for whole Finland and one for one tile. Three classification methods SVM, RF and MLP were compared in TA1 and all four methods were compare in TA2.

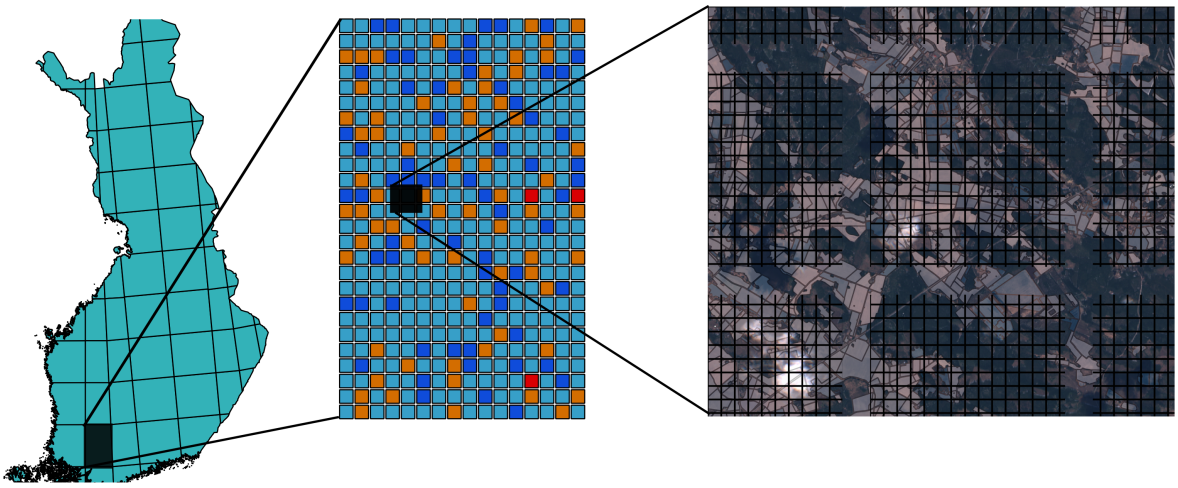


Figure 2: MGRS tiles (blue) that intersect with area of Finland covering test area 1 and test area 2. For test area 2, light blue blocks are for the training, blue are for testing and orange are for validation. Three red blocks are blocks that do not contain any parcels, so they are not assigned to any of the sets. The area of Finland was provided by Statistics Finland (2017) and MGRS tiles by National Geospatial-Intelligence Agency (2012)

Area for TA1 is 3769963km^2 and is bounded by 19°19’37" to 31°24’4"E and 59°48’36" to 70°5’20.400"N. Area for TA2 is 6057km^2 and is bounded by 22°49’15" to 24°0’36"E and 60°24’36" to 61°19’19"N.

Since climate is a major factor in all components of crop production (Iizumi and

Ramankutty 2015), the annual average and temperature and precipitation of Finland are presented. The annual average temperature range from 5.5°C to under -2°C decreasing towards northeast. Similarly, annual precipitation varies depending on the location. In southern and central Finland, the annual average precipitation varies between 600 and 700mm, whereas in northern Finland it is only about 600mm. The annual cloud coverage increases from the northwest towards southeast from about 65% to 85%. Most of the days have cloud cover over 80% and most of the clear days (cloud cover under 20%) occur most frequently in May and June. (Finnish Meteorological Institute 2018d, Finnish Meteorological Institute 2018a.)

3.2 Satellite Data

In this study, Sentinel-2 images were used as RS data because of its suitable spatial and temporal resolutions. The first atmospherically corrected Sentinel-2 images were not distributed until March 2017 (ESA 2017), so this study got an opportunity to use some of the first corrected Sentinel-2 images. Sentinel-2 products follow the similar processing levels as described in Table 1. The Sentinel-2 end user products were described in Table 3. To get the TOA or BOA reflectances from DN, so called quantification value is used to divide DN values. In practice quantification value is usually 10000. (ESA 2018b.) The Level 1C (L1C) and Level2-A (L2A) images are distributed with the naming following Military Grid Reference System (MGRS) tiles (NGA 2006) so that the left upper corner of the images are the same as the left upper corner of the tile. The images are 100x100 km^2 in width and length (SUHET 2015), but the sizes of MGRS grid vary. Sentinel-2 images often span to neighboring tiles as well so there is often some redundancy with the neighboring tiles. The MGRS tiles were acquired from National Geospatial-Intelligence Agency (National Geospatial-Intelligence Agency 2012).

Table 3: The Sentinel-2 processing and product types (ESA 2018b), (SUHET 2015)

Name	**Description**	**Available for end users**
Level-0	Compressed raw image data	No
Level 1A	Uncompressed raw data with coarse coregistration and appended ancillary data	No
Level-1B	Provides TOA radiances in sensor geometry	No
Level-1C	Provides TOA reflectances in cartographic geometry and also cloud and land/water masks	Yes
Level-2A	Provides BOA reflectances in cartographic geometry. The product also includes other spatial data such as a Scene Classification map and Snow masks	Partially

Both Sentinel-2 satellites, Sentinel-2A and Sentinel-2B carry a single payload, Multi-spectral Instrument (MSI). Sentinel-2's MSI have 13 spectral sensor bands, 290 km swath width and high revisit frequency, i.e. temporal resolution. (SUHET 2015.) The bands, their most common usage, wavelengths and spatial resolutions are described in Table 4. 60m bands play important role in generating the cloud masks and cloud shadow masks as discussed in Chapter 4. In addition to the satellite imagery, the L1C cloud masks and L2A snow masks are used in the study as well.

Table 4: Information about the Sentinel-2 spectral bands (Gatti and Naud 2017)

Band	**Use**	**Central wavelength S2A nm**	**Bandwidth S2A nm**	**Central wavelength S2B nm**	**Bandwidth S2B nm**	**Spatial resolution m**
B1	Aerosols	443.9	27	442.3	45	60
B2	Blue	496.6	98	492.1	98	10
B3	Green	560.0	45	559	46	10
B4	Red	664.5	38	665	39	10
B5	Red Edge 1	703.9	19	703.8	20	20
B6	Red Edge 2	740.2	18	739.1	18	20
B7	Red Edge 3	782.5	28	779.7	28	20
B8	NIR	835.1	145	833	133	10
B8a	Narrow NIR	864.8	33	864	32	20
B9	Water vapor	945.0	26	943.2	27	60
B10	Cirrus	1373.5	75	1376.9	76	60
B11	SWIR 1	1613.7	143	1610.4	141	20
B12	SWIR 2	2202.4	242	2185.7	238	20

For SVM, RF and MLP, the Sentinel-2 L2A bands utilized are four 10m bands (B2, B3, B4 and B8) and six 20m bands (B5, B6, B7, B8a, B11 and B12) according to Sonobe et al. (2017). Three bandwise textural features were computed for each parcel for each image: statistical mean (Ienco et al. 2017, Duro et al. 2012), standard deviation (Ienco et al. 2017, Duro et al. 2012, Peña et al. 2014) and median. In addition to these values, also pixel counts for the cloud, cloud shadow, snow and no-data were calculated. ConvRNN method utilizes only the raw pixels of Sentinel-2 L1C bands including three 60m bands (B1, B9 and B10).

The Sentinel Hub AWS S3 storage (Sinergise 2018) was chosen as the download portal for Sentinel-2 images because it includes both L1C and L2A products, which leaves the atmospheric corrections out of the workflow. S3 storage also enabled fast download of images to AWS instances. The atmospherically corrected L2A products of Sentinel-2B satellite were not available for the growing season of 2017 because they were not distributed before 17 December 2017 (ESA 2018c). That is why only Sentinel-2A L2A images were used for methods SVM, RF and MLP. Since ConvRNN used Sentinel-2 L1C TOA reflectance, also Sentinel-2B L1C images were used with training and evaluating of the method.

In this study, the temporal range was limited by the thermal growing season. Thermal

growing season in Finland starts when the snow has melted from expanse areas and the average temperature of the day stays above +5 °C for 10 days. Thermal growing season ends on autumn as the average temperature of the day stays under +5 °C or when the permanent snow or night frosts appear. (Finnish Meteorological Institute 2018c.) As can be seen in Figure 3, the thermal growing season 2017 started at the earliest on first of May and at the latest on 13 June and it ended at the earliest on 27 September and at the latest on 25 October (Finnish Meteorological Institute 2017).

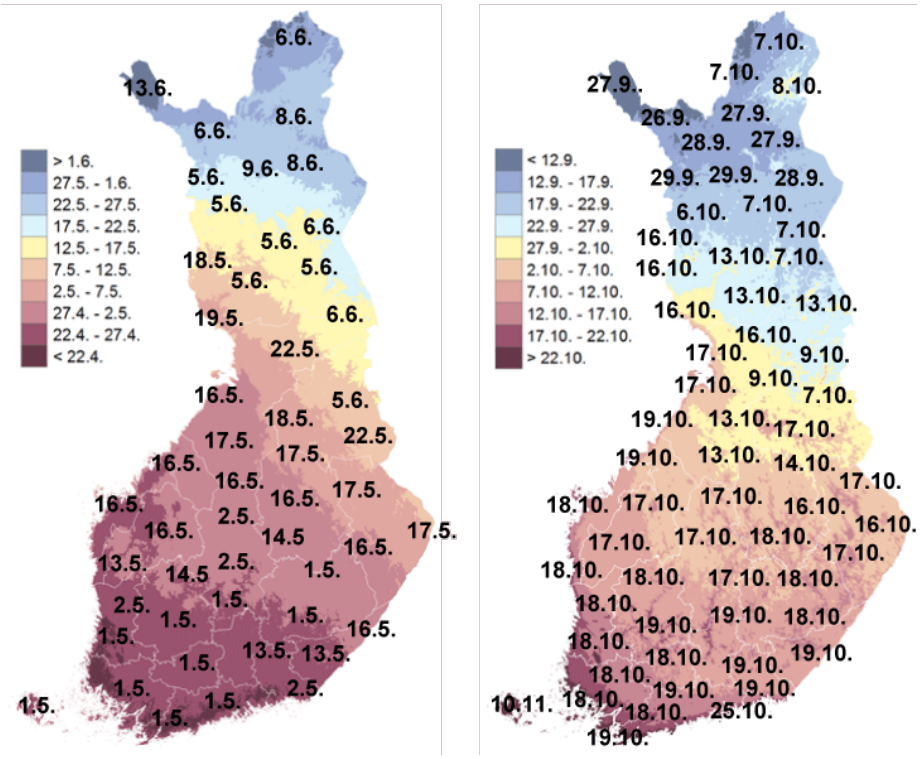


Figure 3: Starting dates (left) and ending dates of the thermal growing season 2017. The colors represent the average thermal growing season start and end ranges based on seasons 1981-2010. The dates in the maps are dates for the season 2017. Modified from Finnish Meteorological Institute 2017.

Based on the thermal growing season 2017, the study focused on the Sentinel-2 images between May and October of 2017. The number of available Sentinel-2A images between April and October 2017 is 3919 but only 3712 images are available for download in the Sentinel Hub S3. For Sentinel-2B the number of available images is 2299. To further describe the availability of the usable Sentinel-2A images, Figure 4 presents the tile availability for each MGRS tile. The minimum monthly image amount was 1, the maximum amount was 13 and the per tile monthly average was 6.6.

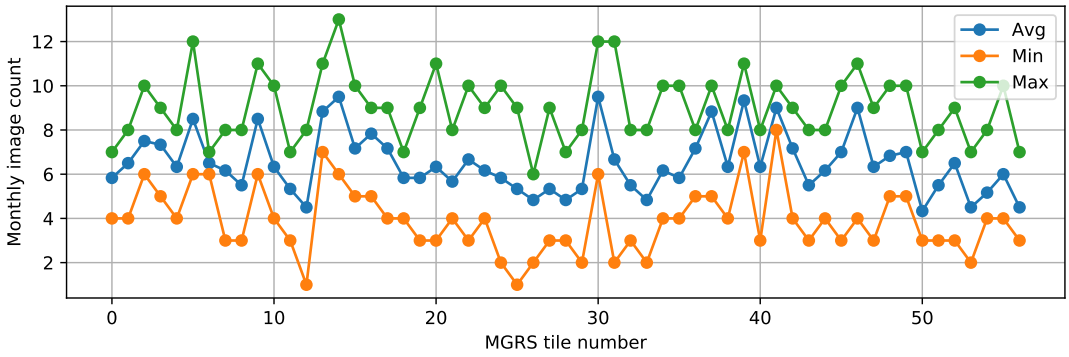


Figure 4: Average, minimum and maximum monthly counts of the Sentinel-2A images for each 57 MGRS tile between May and October 2017 with maximum cloud coverage of 90%

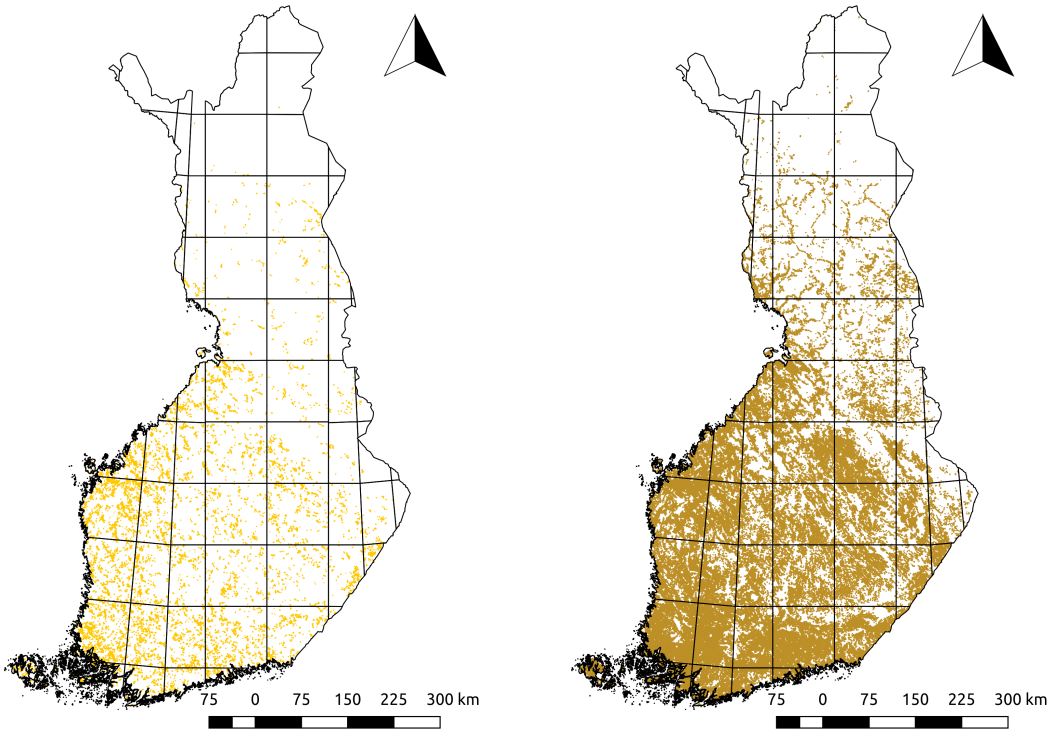
3.3 Land parcels

The parcels were received from Finnish LPIS for the year 2017. The data consists of all the parcels for which CAP subsidies were applied for and the supervised parcels that were digitized and checked by authorities and contain the correct crop with 100% certainty. The supervised parcels and the parcels digitized by farmers that contain the crop classes of the study are denoted as X_1 and X_2 respectively. Parcels in X_2 are not 100% guaranteed to have the crop that they are labeled to have but one can still assume so. The parcels for X_2 and X_1 are shown in Figure 5, where it can be seen that X_1 covers approximately 5% of X_2. The classification methods could use X_1 as training parcels and X_2 as validation parcels or vice versa.

According to MAVI, the 10 most important crop classes in context of CAP were broad bean, pea, beet, fallow, spring rapeseed, spring cereal, grass, potato, turnip rape and winter cereal. The classes are described in more detail in Table 5 and the crops included in each class are described are found in Appendix A in Table A1 in Finnish. Because of the cloudy conditions of Finland and the processing filters in described in Chapter 4.2, the total area and count of parcels used in classification was smaller than described in Table 5.

3.4 Training, testing and evaluation data

In this study, the parcels from X_1 were used for training and testing for the optimization of the processing steps of SVM, RF and MLP methods for TA1. The training set was 80% and the test set was 20% of the X_1, as suggested by Camps-Valls and Bruzzone (2009). After the preprocessing, the final evaluations were done by using the whole X_2 as the validation data. To compensate the small size of TA2, parcels from X_2 were used as training and evaluation data. As instructed by Rußwurm and Korner (2018), TA2 was divided into 3.84km x 3.84km squared blocks with 480m margins to ensure that the same parcel would not be in both training and testing



(a) Supervised parcels X_1 (b) Parcels digitized by farmers X_2

Figure 5: Parcels containing one of the 10 crop classes in Finland with MGRS grid on top

sets. Each block was randomly assigned to be either training, testing or validation block in a ratio of 4:1:1, as can be seen in Figure 2. For ConvRNN, each block was further divided to 16 240m x 240m squares and each square acted as a input vector for the network.

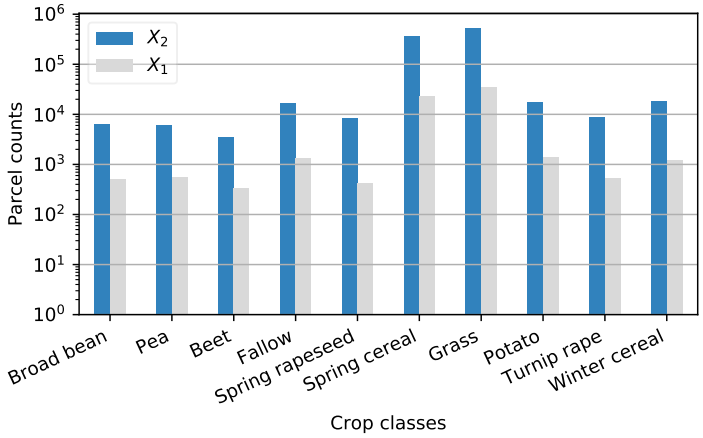


Figure 6: The distribution of the parcel classes on a logarithmic scale in Finland for the year 2017

Table 5: Information about the crop classes used in the study. The average area includes the standard deviation after ±.

Crop Class	Class Label	# of Parcels in X_1	Total Area (ha)	Average Area (ha)
Broad bean	BEA	503	1316	2.6 ± 2.9
Pea	PEA	565	1130	2.0 ± 2.7
Beet	BEE	333	830	2.5 ± 2.4
Fallow	FAL	1326	1987	1.5 ± 2.3
Spring rapeseed	SRA	428	1624	3.8 ± 4.0
Spring cereal	SCE	22969	54803	2.4 ± 2.6
Grass	GRA	35380	54760	1.5 ± 2.0
Potato	POT	1418	1474	1.0 ± 1.5
Turnip rape	TRA	522	1512	2.9 ± 3.3
Winter cereal	WCE	1216	3540	2.9 ± 3.0
Crop Class	Class Label	# of Parcels in X_2	Total Area (ha)	Average Area (ha)
Broad bean	BEA	6522	22228	3.4 ± 3.4
Pea	PEA	6226	15444	2.5 ± 3.2
Beet	BEE	3433	12223	3.6 ± 3.4
Fallow	FAL	17012	27642	1.6 ± 2.4
Spring rapeseed	SRA	8544	35340	4.1 ± 4.0
Spring cereal	SCE	361517	1005394	2.8 ± 3.0
Grass	GRA	538784	948748	1.8 ± 2.2
Potato	POT	17516	21702	1.2 ± 1.8
Turnip rape	TRA	8712	28012	3.2 ± 3.3
Winter cereal	WCE	18779	67284	3.6 ± 3.6

4 Methods

4.1 Classification methods

4.1.1 Support vector machine

SVM tries to find the optimum hyperplane in n-dimensional space, to separate the classes. The n corresponds to the dimension of variables, or number of bands in the case of this study. The two dimensional example of this is presented in Figure 7. The side of the plane determines the class which the test sample is getting. (Ray 2017.) The dashed lines in Figure 7 are maximum separation margins and SVM tries to fit the hyperplane in a way that maximizes the size of the margins (Sonobe et al. 2017).

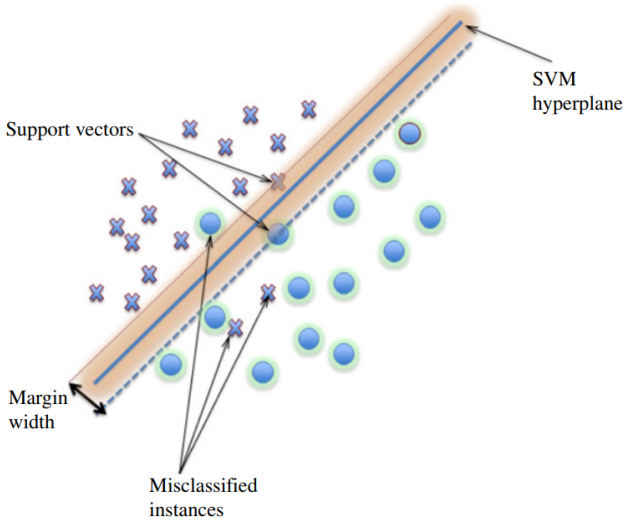


Figure 7: SVM hyperplane and maximum separation margins (dashed) separating two two dimensional classes (Mountrakis et al. 2011)

Since it would be impossible to fit linear hyperplane to separate nonlinear data, the "kernel trick" was applied instead of attempting to fit a nonlinear model (Sonobe et al. 2017). "kernel trick" are functions, called kernels, that map low dimensional input space into higher dimensional space in order to make the problem separable. The kernels could be linear, polynomic sigmoidic or RBF (Ray 2017.)

In this study, the following tunable hyperparameters of SVM are tuned: the regularization parameter, or penalty parameter (Buitinck et al. 2013) C, and the kernel bandwidth γ. High C values lead to high penalties, which may cause overfitting. On the other hand, low C values can lead to underfitting, which means that the algorithm does not learn as well as it could (Sonobe et al. 2017). High γ values try to fit to the training data as exactly as possible while low values try to generalize to the data. High γ values may cause overfitting as well. (Ray 2017.) Scikit-learn implementation default parameters are kernel=RBF, $C = 1.0$ and $\gamma = 1/n_{\text{features}}$ (Buitinck et al. 2013).

4.1.2 Random forest

RF is an ensemble learning technique and it works by building several trees by taking random samples of the training data. The nodes of the trees are then divided into two or more sub-nodes in a process called splitting. The best split is determined from a cluster of variables selected randomly. The best split is then used to split the nodes of the trees. The majority vote of the classification trees determines the final output of the RF and thus gives a probability estimate. (Sonobe et al. 2017.) The classification process and the majority vote of RF is shown in Figure 8.

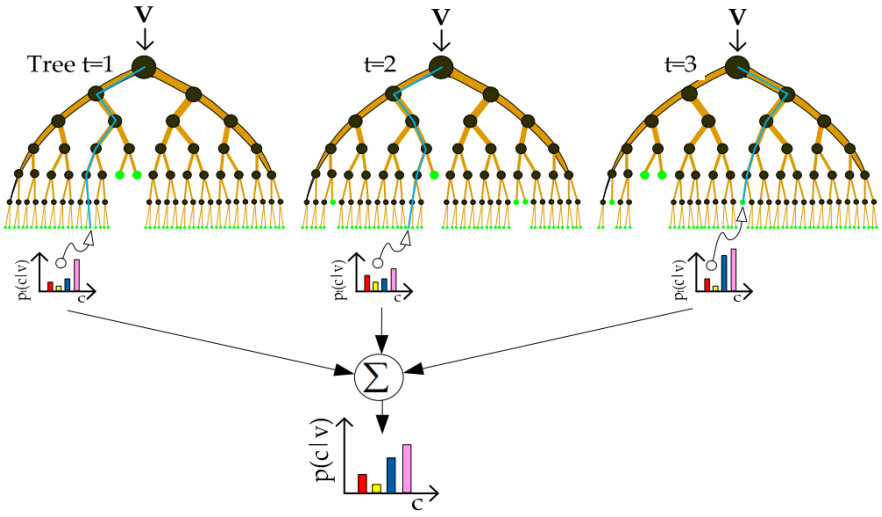


Figure 8: RF classification process (Sun et al. 2017)

According to Sonobe et al. (2017), RF has following important tunable hyperparameters for crop identification: the number of trees in the forest (n_tree), the number of features to consider with the best split (max_features), the maximum depth of the tree (max_depth), the minimum number samples in splits (min_samples_split), number of nodes in a leaf (max_leaf_nodes).

4.1.3 Multilayer perceptron

MLP forms one type of artificial neural networks, which belong to the research of artificial intelligence. MLP avoid making prior assumptions of data distribution and it can be used to model non-linear functions. MLP has the ability to accurately generalize to the unseen, new data. As can be seen in Figure 9, MLP consists of structure of interconnected neurons that form a nonlinear paths between the neurons of the input vector and output vector. Each node is connected to other nodes by weights **W** and output signals. The output signal is a function of summed input weights modified by nonlinear activation function. (Gardner and Dorling 1998.) In this study, rectified linear unit (Nair and Hinton 2010) is used as an activation function.

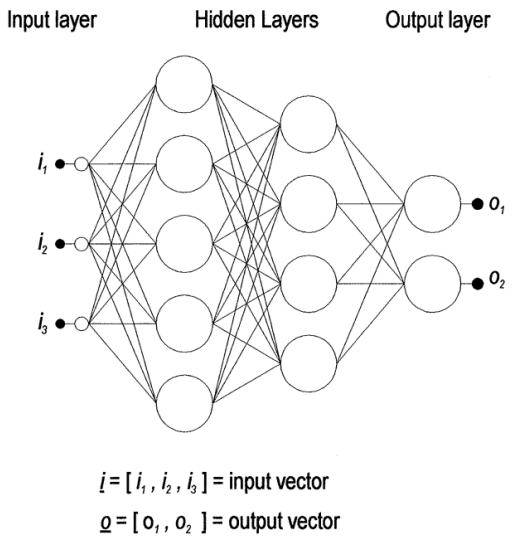


Figure 9: Multilayer perceptron using two hidden layers (Gardner and Dorling 1998)

The node output is fed to the node of the next layer after being scaled by the weight of the connection between them. Since the direction of the information is from input to output, MLP is a type of feed-forward neural network. The input layer nodes correspond to the vector of input data $\mathbf{x}$ and thus, dimensions of the input data vector determine the amount of input layer neurons. The output layer on the other hand, has as many neurons as there are different classes in the data, for $\hat{y}$ represent class label. Between input and output layer there can be one or more hidden layers with arbitrary amount of neurons. Hidden layer vector including its neurons is marked as $\mathbf{h}$. (Gardner and Dorling 1998.) Since there are 10 crop classes in this study, there would be 10 neurons in output layer. Artificial neural networks can be seen to approximate a function from Equation 8 (Rußwurm and Korner 2018).

$$\hat{y} = f(\mathbf{x}; \mathbf{W}) \tag{8}$$

In training of MLP networks, the training data is presented to the network in batches and weights of neuron connections are adjusted until the optimal input-output mapping is found. There are many algorithms that can be used as optimizers for finding the correct weights. (Gardner and Dorling 1998.) The scikit-learn implementation of Adam optimizer (Kingma and Ba 2014) is used in this study. The number of hidden layers and the number of neurons in each layer are used as tunable parameters in hyperparameter tuning of MLP.

4.1.4 Convolutional recurrent neural network

Out of four classification methods used in this study, ConvRNN is the only one that was actually designed to identify vegetation from RS imagery. ConvRNN was designed to be able to approximate phenological model for vegetation classes from

series of Sentinel-2 images. ConvRNN, like MLP, is a neural network. Unlike MLP, the network is recurrent, so the information does not only go to one way from input to output. ConvRNN uses convolutional recurrent layers of cells for inspecting the sequential, or temporal, aspects of the input data. The temporal samples of input data are fed to the network both in sequential order and in reversed order for ensuring that the final stage of the samples does not affect the end result as much. The type of convolutional recurrent cells are either LSTM or gated recurrent unit (GRU) cells. (Rußwurm and Korner 2018.) The schematic illustration of the ConvRNN is presented in Figure 10.

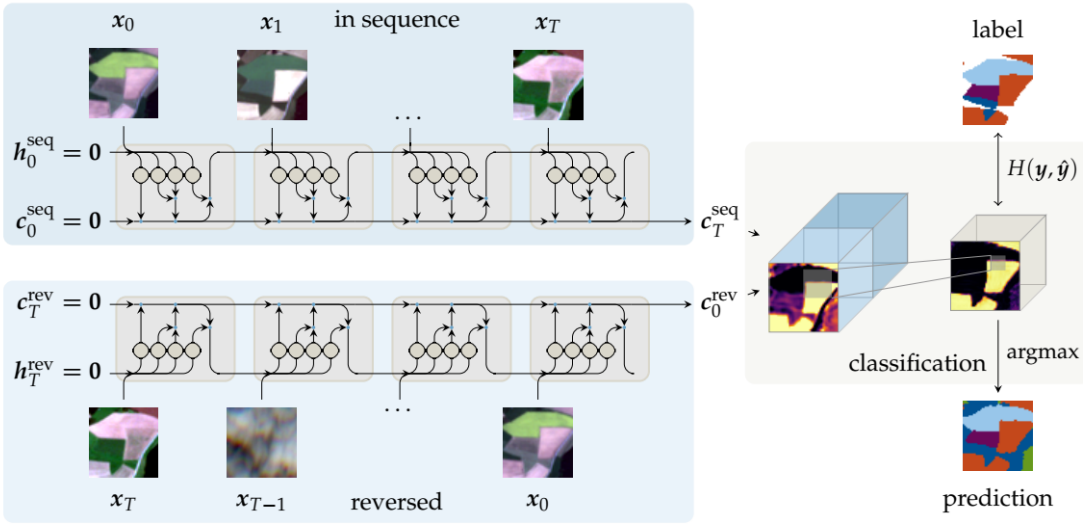


Figure 10: Convolutional recurrent neural network architecture and classification process of parcels (Rußwurm and Korner 2018)

ConvRNN does not require much preprocessing of the images and since it is designed to analyze time series of images, it is able to detect anomalities. Rußwurm and Korner (2018) observed that some of the cells were able to detect clouds and filtering them off from the weights. For this reason, the cloudy images could be used as well, and no external cloud masking is needed (Rußwurm and Korner 2018).

The tunable parameters of ConvRNN are the type of convolutional recurrent cells, number of convolutional recurrent cells, the convolutional recurrent cell kernel sizes, the final convolutional cell kernel size and the number of temporal observations to take the samples on (Rußwurm and Korner 2018). In addition, the batch size of the input data affects the speed of the training and the length of the training can be determined by number of epochs, which means how many times the network has seen all the training data.

SVM, RF and MLP classification algorithms were used and modified as implemented in Python 3.5 Scikit-learn version 0.19.1 (Pedregosa et al. 2011). ConvRNN model compiling and training were done using Python 2.7 and Tensorflow 1.4 (Abadi et al. 2015) using the code provided by Rußwurm and Korner (2018).

4.2 Processing of the Sentinel-2 and parcel data

In this study, multiple processing steps were adopted for transforming the Sentinel-2 images and parcels to the multitemporal input vectors for the classifiers. Some steps were included also to ensure that the quality and performance of the crop identification would be good enough. The preprocessing data flow is visible in Figure 11 and the related RS quantities, such as reflectance and radiance, were used according to definitions from Schaepman-Strub et al. (2006). It consisted of generating the masks, resampling the Sentinel-2 imagery, rasterizing the parcels, extracting the reflectance values from the images and producing the data for the classification.

The processing steps for SVM, RF and MLP are similar with each other, but ConvRNN has a complete different processing chain. This was due to the fact that ConvRNN uses pixel-based classification, while other three methods used parcel-based classification. The processing steps for ConvRNN are mentioned separately in the suitable sections.

Some steps required comparing of the classification accuracies to optimize the process. Since the training of every of the four selected classification methods for every comparison would have been computationally expensive and time consuming, the SVM method was chosen as the base classifier which was used in optimizing of the processing steps. Since it would have been also computationally expensive to tune the hyperparameters of SVM for every step, the default Scikit-learn SVM implementation parameters ($C = 1.0$ and $\gamma = 1/n_{\text{features}}$) were used. The optimal results are then used with other classifiers as well.

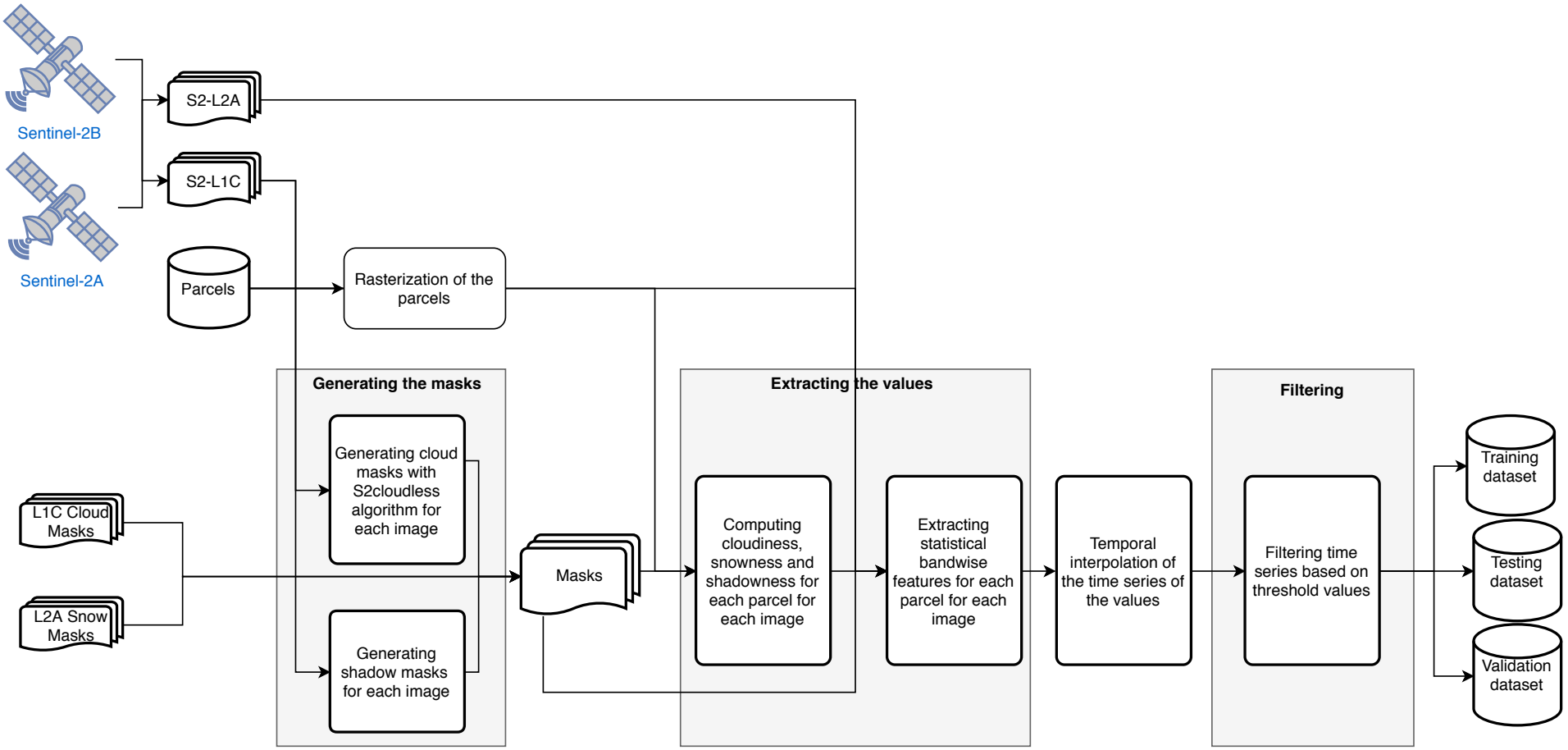


Figure 11: Data flow and processing steps

4.2.1 Generating the masks

In addition of the Sentinel-2 L1C cloud masks and L2A snow masks, another cloud mask and cloud shadow mask were generated for each image. This was done to ensure that the parcel reflectance values would not contain clouds or cloud shadows, since after visual inspection it was decided that the cloud mask provided with L1C product was not accurate enough.

The cloud masks were generated using Sentinel Hub Cloud Detector S2cloudless (Zupanc 2017) algorithm, which uses machine learning approach to detect the clouds from Sentinel-2 images. The algorithm was chosen because it seemed to outperform other cloud masking algorithms Fmask (Zhu et al. 2015), MAJA (CNES 2018) and Sen2Cor, which is used as a processing tool for creating official Sentinel-2A L2A products (ESA 2018a). (Zupanc 2017.) The algorithm used Sentinel-2 image bands 1, 2, 4, 5, 8, 8A, 9, 10, 11 and 12 as input bands resampled into the same spatial resolution beforehand. In this study, L1C TOA reflectance bands were resampled to 80m using suggested nearest-neighbor resampling Roy et al. (2016). S2cloudless could produce several products, like cloud probability maps, but for the purposes of this study, default parameters were used to generate binary cloud masks. The no-data value was added to the cloud mask as well to mask out the areas of the image that did not have reflectance values.

For the cloud shadow mask, the Fmask (Zhu et al. 2015) algorithm and L2A scene classification images were evaluated, but the results were not satisfactory enough after visual interpretation because both algorithms showed cloud shadows on fields among other places and failed to mask actual shadows. The more suitable algorithm after visual interpretation seemed to be the algorithm presented at Murphy et al. (2017). The algorithm takes the list of approximate cloud heights in meters, cloud mask, azimuth, zenith and L1C bands 3 and 12 as an input data and produces the cloud shadow binary mask as an output. In this study, the cloud height list of 500, 1000, 1500, 2000 and 2500 were used by adopting cloud heights used by Murphy et al. (2017) and presented in Finnish Meteorological Institute (2018b).

4.2.2 Resampling

To harmonize the raster data dimensions, all Sentinel-2 bands with spatial resolution of 20m (B5, B6, B7, B8a, B11 and B12) were resampled to 10m using nearest-neighbor interpolation method to be used with 10m bands (B2, B3, B4 and B8) following the example of Rußwurm and Korner (2018) and Sonobe et al. (2017), who resampled Sentinel-2 bands to 10m. The nearest-neighbor method was adopted, since it is used successfully by Peña et al. (2014) and it is recommended for Sentinel-2 imagery as best practices by Roy et al. (2016). For the ConvRNN method, all 20m and 60m (B1, B9 and B10) bands were resampled to 10m resolution using bilinear interpolation, since it was the default interpolation method for image analysis in Tensorflow (Abadi et al. 2015).

4.2.3 Rasterization

The parcels polygons were rasterized to raster images in order to perform masking with image bands. The pixel size of rasterized parcels was set to 10m, to match the spatial resolution of resampled Sentinel-2 images. The rasterization was made once for parcels of every tile so that the raster size of the rasterized parcels of each tile matched the Sentinel-2 tile size on top left corner. This ensured that the corresponding pixels matched with certainty. During rasterization, each parcel was reprojected to the projection of the tile and the resulting tile parcel raster would have the tile projection as well. Each rasterized parcel contained the id of the parcel as band value, so it was possible to track down the parcel information after masking. The parcel value was bit shifted before and after the masking so no information was lost during the process.

Since the bandwise statistics were calculated, it is important, that the parcel pixels do contain only the relevant information. Since there often are irrelevant pixels, like tree shadows, subsurface drains or trenches, at the borders of the parcel, it was decided to reduce the areas of the parcels in exchange of losing small parcels. To reduce the parcel pixels from the outer borders, the negative 5m buffering is applied during the rasterization process. Boudewyn et al. (2000) stated that buffering would not affect greatly to OA of classification, but other reason for buffering was to cut off the smallest parcels. The Figure 12 swows the original polygon parcel and the buffered raster parcel. Only the parcels that have an area large enough, are taken into account. For ConvRNN processing chain, the parcels were also rasterized but without any buffer to keep the processing as close to the original processing done by Rußwurm and Korner (2018).

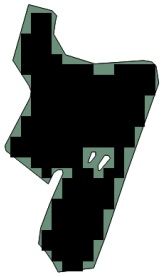


Figure 12: Parcel polygon and rasterized parcel with -5m buffer

4.2.4 Extracting the values

In the reflectance extraction process, the mask pixel values were calculated for parcels. The masks were also used to mask out the pixels containing cloud, cloud shadow, snow or no-data. In calculation of the pixelwise reflectance statistics, only the pixels that did not contain any of the masked values were taken into account. When processing the time series, every parcel went through stages of filters to ensure that

the classification data was clean. The filters are shown in Figure 13. For ConvRNN image extraction, each block was used as is, without any filtering or aggregating, but DOY and year of the images were used in addition to the pixel reflectance values.

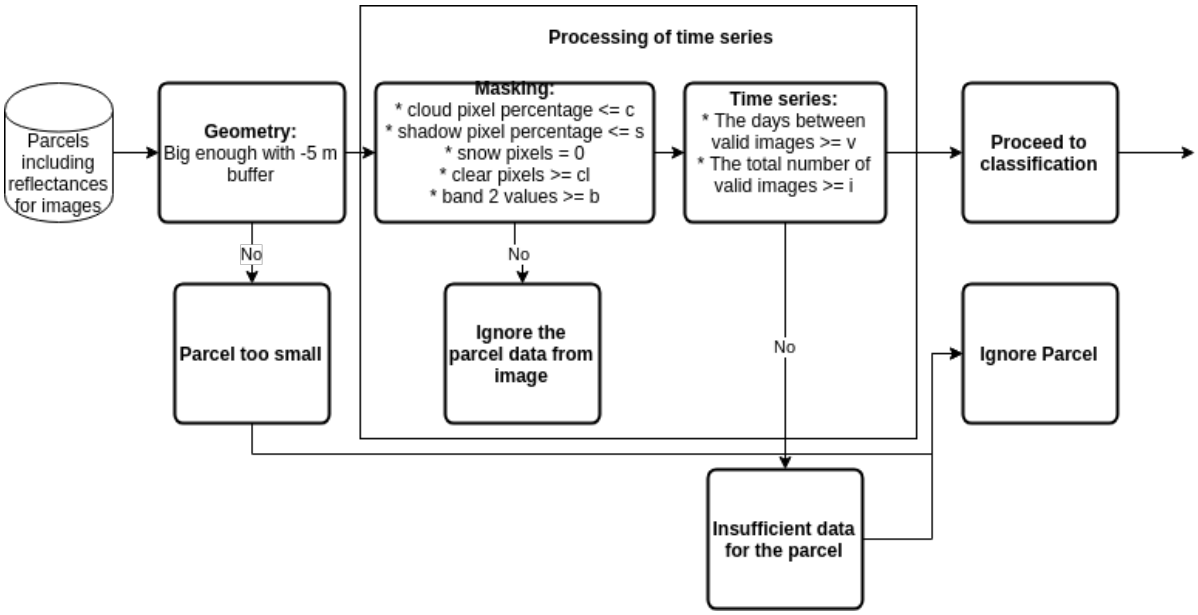


Figure 13: The filtering flow of the parcel data

The Geometry filter means just that some of the parcels are left out during the rasterization process. Since it was not certain that the masks are perfect and catch every value, the Masking filter was applied. In Masking filter, there were multiple criteria: cloud pixel percentage ρ and shadow pixel percentage ϱ had to be smaller than a threshold, the absolute snow pixel count had to be 0, the absolute clear pixel count ν (no cloud, shadow, snow or no-data) had to be greater than a threshold. Also, the value of B2 had to be above a 0.001. This was to leave out all extremely small values that would have been probably caused by a shadow or water. The Time series filter checked that the days between valid images ϕ was smaller than a threshold and also that a total number of valid images φ was above a threshold. Every parcel that did not pass these filters were left out from the training data in order to keep the data clean. The calibration of these threshold values is explained in Chapter 4.2.8.

4.2.5 Decomposing and scaling of the classification data

With 3 textural features for 10 bands and n time steps, the dimension of the data vector grew to be $30*n$. For classifying this would have meant that the training would have taken a long time for each model. In this study, there were lot of comparisons between different methods and preprocessing steps and there was only limited amount of time and resources to use. Therefore, PCA decomposition method was adopted for reducing the dimensionality. PCA decomposition projects the data to lower

dimensions using Singular Value Decomposition (SVD). It is possible to configure the PCA decomposition with exact number of wanted components or to give the wanted ratio of the variance of the original data that can be explained with these new components. (Halko et al. 2009, Buitinck et al. 2013.) In the following comparisons in Chapter 4.2.6, PCA decomposition was used to allow faster training times. The comparison between multiple PCA decomposition ratios is done in Chapter 4.2.7.

Many classification algorithms, like SVM, expects the data vectors to be standardized to look like standard normally distributed data. In standard scaling, the mean of the features is scaled to 0 and the features are scaled to unit variance. (Buitinck et al. 2013.) In this study, SVM, RF and MLP classification methods had the standard scaling as a preprocessing step.

PCA decomposition was not tested with ConvRNN method, but DOY was divided with 365 and year was subtracted with 2017 in order to have the similar scale with reflectance values (Rußwurm and Korner 2018).

4.2.6 Temporal interpolation

To address the problem of imbalanced distribution of temporal cloud free parcel values, the interpolation of values was adopted for each multispectral band independently. Since the statistical features were extracted from the pixel values, instead of pixelwise interpolation like used by Ienco et al. (2017) the bandwise interpolation was utilized for each textural value independently. The linear interpolation method was adopted to the task according to Ienco et al. (2017).

The interpolation brings a problem with some of the parcels having several cloud free images and some having only few. Therefore, the concept of resampling of the interpolated values was adopted. In this study, the sample size of the resampling of the interpolated values is called interpolation step. The classification results with different interpolation steps were compared in order to find out the optimal value for the study.

One of the objectives of the study was to find the optimal temporal range for the identification task. Therefore, different classification results were compared with different starting and ending dates. The number of interpolation steps had an effect in the temporal range so rather than comparing the results of different absolute interpolation steps, the comparison was made with interpolation steps per month. The starting day of the case study was 1st May 2017 and the ending date was 31st October 2017. The comparisons were made with 18 date ranges shown in Table 6. Some of the starting times were after the start of the thermal growing season because the classification might have found the most significant variances between classes in temporal features that are middle of the growing season. Based on the observations of the Chapter 3.2 and especially on Figure 4, the monthly interpolation steps that were compared for each date range were 2, 3, 4, 5, 6 and 7 in Chapter 5.2.2. The number of interpolation comparisons was thus 109. From those models,

the temporal range and interpolation steps of the one with best accuracy was selected and further analysis was done with those values.

The number of parcels might differ between time ranges, but it would be constant within the same temporal range between interpolation steps. To be able to compare the classification results reliably, the train and test splits were chosen among the parcels of X_1 that were available with every time range. For consistency, the train set was 80% and the test set was 20% of the data, as suggested by Camps-Valls and Bruzzone (2009). The same split of train and test data was used with every model. Only the parcel bandwise average was used from textural features, according to Schmedtmann and Campagnolo (2015) and Duro et al. (2012) and no temporal restriction was set during the data extraction for the models. Also, the mask pixel percentages was set to 100% to get as much data as possible for the comparisons.

To be able to run over 100 models in relatively fast time, the input classification data was decomposed with PCA decomposition first. For each model the PCA decomposition ratio was set to 99%, i.e. the amount of PCA components are automatically chosen so that the explained variance of the new components would be over 99%.

Table 6: The dates of the interpolation

Name	Start	End	Months
t-int1	1st May 2017	1st June 2017	1
t-int2	1st May 2017	1st July 2017	2
t-int3	1st May 2017	1st August 2017	3
t-int4	1st May 2017	1st September 2017	4
t-int5	1st May 2017	1st October 2017	5
t-int6	1st May 2017	31st October 2017	6
t-int7	1st June 2017	1st July 2017	1
t-int8	1st June 2017	1st August 2017	2
t-int9	1st June 2017	1st September 2017	3
t-int10	1st June 2017	1st October 2017	4
t-int11	1st June 2017	31st October 2017	5
t-int12	1st July 2017	1st August 2017	1
t-int13	1st July 2017	1st September 2017	2
t-int14	1st July 2017	1st October 2017	3
t-int15	1st July 2017	31st October 2017	4
t-int16	1st August 2017	1st September 2017	1
t-int17	1st August 2017	1st October 2017	2
t-int18	1st August 2017	31st October 2017	3

4.2.7 Decomposition comparison

To have an idea of how much effect the PCA decomposition had to the results, the different decomposition ratios of PCA decomposition were compared in Chapter 5.2.3. The best model configuration and common parcels for all models was used from the Chapter 4.2.6. However, to reduce the effect of overfitting, the 5-fold CV for the whole data was used in training of the models. In the comparisons, following ratios were used: ratios $\in \{0.00, 0.25, 0.50, 0.75, 0.99\}$. The ratio 0.00 means here that the PCA decomposition is not used and the original feature vectors were used instead. The best ratio in the comparisons would be the one with relatively high accuracies and high training speed. The PCA decomposition ratio used in the further experiments was adapted based on these results.

4.2.8 Processing threshold calibration

The following threshold ranges were selected for the threshold calibration: $\rho \in \{0, 25, 50, 75\}$, $\nu \in \{1, 10, 20\}$, $\phi \in \{10, 20, 30, 365\}$ and $\varphi \in \{1, months, 3{*}months, 5{*}months, 7 * months\}$. The ϱ values were linked with ρ to save computation time. ρ values were percentage values and φ values were number of images per month but there was also an absolute value of 1. The final φ range was determined according to the number of months of the results from Chapter 4.2.6.

The calibration processing was done by changing one variable at the time while keeping the other thresholds at the most liberal value, i.e. at the value that would let most of the parcels through. The liberal values for the thresholds were: ρ at 75%, ν at 3, ϕ at 365 and φ at 1. Since the calibration of the thresholds affected the number of passed parcels, another thing to observe in addition of the model accuracies was the count of the passed parcels. The goal was to use many parcels but at the same time getting better accuracies. If accuracies would not be negatively affected from keeping the thresholds liberal, the models with the most parcels with a good accuracy would be chosen for further analysis. For this comparison, total of 13 models were made, one with all thresholds liberal, and each of different threshold changed. Based on the results, some further comparisons with combined parameter threshold values were executed. Based on the results of these comparisons in Chapter 5.2.4, the filtering thresholds were adopted for further experiments.

For the calibration process, the data consisted of bandwise average of the X_1. Since the sizes of the parcels would vary drastically between runs, there was no sense in using same train and test data for all the classification models, as in Chapter 4.2.6. Instead the 5-folded CV for the whole data was used in training of the models. This ensured that the results would be reliable for comparison.

4.3 Variable selection

4.3.1 Statistical features

As 10 bands with three statistical features make up 30 values for each interpolation step. This combination would not necessarily be the optimal one and to determine which textural features would give the best results, six classification comparisons were made as presented in Table 7. The training set was 80% and the test set was 20% of the parcels in X_1 and if not mentioned, the same split of train and test data was used for the classifications in following Chapters. The results of the comparisons are presented in 5.3.1.

Table 7: The comparison runs of the textural feature selection

Name	Description
AVG	Average of the parcel pixels for each band
MEDIAN	Median of the parcel pixels for each band
STD	Standard deviation of the parcel pixels for each band
AVG+STD	Average and standard deviation values of the parcel pixels for each band
AVG+MEDIAN	Average and median values of the parcel pixels for each band
MEDIAN+STD	Median and standard deviation values of the parcel pixels for each band
AVG+MEDIAN+STD	Average, median and standard deviation values of the parcel pixels for each band

4.3.2 Vegetation indices

Since many of the studies calculated VIs, the comparisons between VIs was made. Three VIs were selected among the indices used in previous crop identification studies: NDVI (Peña et al. 2014, Ienco et al. 2017, Belgiu and Csillik 2018), Green Normalized Vegetation Index (GNDVI) (Peña et al. 2014) and Modified Chlorophyll Absorption in Reflectance Index (MCARI) (Peña et al. 2014). The equations of calculating the VIs from Sentinel-2 bands are shown in Equations 9-11 (Frampton et al. 2013).

$$\text{NDVI} = \frac{B7 - B4}{B7 + B4} \tag{9}$$

$$\text{GNDVI} = \frac{B7 - B3}{B7 + B3} \tag{10}$$

$$\text{MCARI} = [(B5 - B4) - 0.2(B5 - B3)] * (B5 - B4) \tag{11}$$

NDVI was selected because of it is one of the most-used VI to study vegetation phenology. It has ability to reduce the spectral noise caused cloud shadows, topographic differences and certain illumination conditions. (Belgiu and Csillik 2018.) GNDVI was selected since chlorophyll content acts as one of the essential indicators for accessing crop growth (Liang et al. 2016) and GNDVI has been reported to be much more sensitive to chlorophyll-a concentration than NDVI (Frampton et al. 2013). MCARI was also selected due to its sensitivity to chlorophyll and its use in agricultural applications (Haboudane et al. 2002).

In the previous studies pixelwise VI values were used. However, here the calculations were performed for the statistical features for computational reasons. The classification runs were made with each VI combination in Chapter 5.3.2. Each of the VI was calculated using median and average. Based on the results, the additional runs were made with statistical features combined with VIs and the very best configuration was used in further experiments. For the visual inspection, the NDVI time series of each crop class was also plotted.

4.3.3 Most meaningful components

With optimal combination of statistical features and VIs, a data vector would comprise of all the components for each interpolated image date. It was interesting to know, if some of the components would be more important and meaningful for one date and some other components for some other date and whether the data vectors comprised of set of the most meaningful components overall would perform better in classification than a data vector with all the components. Inspired by Schmedtmann and Campagnolo (2015) and Jolliffe (2002), PCA was used to find out the most meaningful components.

Based on the method presented by Otterbach (2016), the most important components were identified by applying PCA decomposition inversion to the identity matrix and calculating the mean of each row. The variables with the largest absolute mean would be the most meaningful original variables. The comparisons were made with different amounts of original meaningful variables and the results were compared to the PCA decomposed variables in Chapter 5.3.3.

4.4 Classification step

4.4.1 Imbalance of the crop classes

The distribution of crop classes was highly imbalanced, as can be seen from Figure 6 and Table 5. One reason for this was that some of the crops were farmed more often than others. One of the ways to cope with imbalanced distribution is to over-sample or under-sample the classification data. The purpose of over-sampling is to generate new samples to the training data for minor classes while under-sampling does the

opposite by reducing the samples of the major classes in training data. The two popular over-sampling algorithms were chosen for the comparison: Random Over Sampler (ROS) and Synthetic Minority Oversampling Technique (SMOTE). Similarly, the following under-sampling algorithms were chosen: Clustering Centroids (CC), Random Under Sampler (RUS), Instance Hardness Threshold (IHT), Tomek Links (TL), AllKn (AKNN), Condensed Nearest Neighbour (CNN), One Sided Selection (OSS) and Neighbourhood Cleaning Rule (NCL). (Lemaître et al. 2017.)

ROS works by randomly duplicating some existing samples back to the training data, while SMOTE uses nearest neighbor interpolation of existing samples to generate new ones. In under-sampling, CC generates new samples by utilizing K-means clustering centroids coordinates of the majority classes instead of the original samples. Other methods select the samples from the original distribution. RUS selects samples randomly, TL detects if two samples from different classes are each other's nearest neighbours and removes the one from majority class. AKNN uses nearest neighbor clustering to find iteratively increasing number of neighbors and removes the ones, that are too different from their neighbours. CNN uses one nearest neigbour rule iteratively to decide if the sample should be removed or not. CNN may include noisy samples, so OSS uses TM to remove the noise. NCL cleans the data using the union of nearest neighbour removal and 3 nearest neighbour rule. IHT is a different from the rest, since it trains the classifier and and removes the samples with lowest probabilities of the major classes of the training data. (Lemaître et al. 2017.) For IHT, SVM algorithm with default parameters was used as a classifier.

The comparison with different over- and under-sampling methods was made in Chapter 5.4.1. The over-sampling ratios were set so that the number corresponds to the aimed minimum amount of features in a class. For example, if ROS ratio is 600 and class BEA has 500 samples and PEA has 650, after resampling BEA would have 600 and PEA 650 samples. The under-sampling ratio works in the same manner but using minimum amount of features instead of maximum. If the ratio is set to "auto", the algorithm does not accept set class ratios and the resampling is done using all but maximum class in over-sampling and all but minimum class in under-sampling. (Lemaître et al. 2017.)

The over-sampling ratios based on the approximate training data class sizes were tried so that the over-sampling ratio was set to second smallest class, third smallest class etc. until the largest class was achieved. In under-sampling, the similar approach was used for comparisons. The over-sampling and under-sampling can be used at the same time as well and comparison with best over-sampler and under-sampler were performed based on the best resampling methods.

Other way to address the problem of imbalanced distribution was to use class weight parameter of SVM, which can be used to assign weights to the penalty parameter C for each class. There is a possibility to assign weights in balanced manner, inversely proportional to class frequencies as shown in Equation 12. (Buitinck et al. 2013.)

The comparison was made with balanced weighting without additional resampling.

$$W_{\text{class}} = \frac{n_{\text{samples}}}{n_{\text{classes}} n_{\text{class samples}}} \tag{12}$$

4.4.2 Hyperparameter tuning

The final optimization step was to tune the hyperparameters of the classifiers. The previous sections have used only SVM classifier for optimizing all processing steps, but now the classifiers RF and MLP were tuned as well. The results of these three classifiers are tested with the 20% test split of X_1 sampled in Chapter 4.3.1.

With the right hyperparameters that are suitable for the data, it would be possible to improve accuracies. It is possible to optimize the hyperparameters based on any accuracy metric. Improving parameters based on OA would probably lead to a situation where two of the most dominating classes SCE and GRA would be well predicted while the other 8 would be predicted poorly. Since the aim was to be able to identify all classes as well as possible, f1 value with macro averaging was selected to be the optimized metric. The results of the hyperparameter tuning are presented in Chapter 5.4.2.

For SVM, the kernel function was set to RBF based on the literature (Schmedtmann and Campagnolo 2015, Sonobe et al. 2017, Peña et al. 2014) and Baysian optimization, as implemented in skopt by (Head et al. 2018), was used to optimize hyperparameters C and γ. The search spaces were: $C \in \{log_10(e^{-6}), ..., log_10(e^{6})\}$ and $\gamma \in \{log_10(e^{-6}), ..., log_10(e^{3})\}$ according to Head et al. (2018). For RF, the following continuous parameters were tested: n_estimators $\in \{(10, ..., 1000\}$, max_depth $\in \{2, ..., 1000\}$, max_features $\in \{2, ..., 10\}$ and min_samples_split $\in \{2, ..., 20\}$ using Bayesian optimization.

For MLP the random search was used to optimize the hidden layers because there were lot of alternatives. According to Peña et al. (2014), the hidden layers should be tested with values $h = (i + c)/2$ with c being the number of classes and $i \in \{50, 100, 150, 200, 250, 300\}$. In addition there are possibility to add multiple layers after the first one. Every possible hidden layer combination is tested with up to 5 number of layers. This made total of 9330 combinations for the parameters.

After noticing that training of ConvRNN was extremely time consuming and computationally expensive, the hyperparameter tuning was not done for ConvRNN. ConvRNN was just trained with 64 convolutional recurrent LSTM cells with kernel sizes of $k_{\text{rnn}} = k_{\text{final convolutional cell}} = 3$. The subset of 20 observations was chosen out of maximum 31 images that were available for the single block. The batch size of 16 was used and the model was trained for 60 epochs. All the parameters except LSTM cell, were the default parameters set by Rußwurm and Korner (2018). LSTM cell was chosen because of its popularity in other studies, like Ienco et al. (2017).

4.5 Comparisons method comparison

In previous Chapters, the processing and models have been optimized and trained with X_1 data. For comparing the methods with large dataset, the evaluation was done with X_2 in Chapter 5.5. Confusion matrices and accuracy metrics were compared for the models. For TA2 comparison, ConvRNN method was trained and optimized with training and testing blocks and evaluated with validation blocks. All the other methods were trained with the parcels from X_2 that intersected with the training blocks and evaluated with parcels that intersected with the validation blocks. SVM, MLP and RF were also hyperparameter optimized before the evaluation for TA2.

4.6 Calibration of the classification results

The purpose of the crop identification is to be a reliable method. Therefore, the classification calibration method from Schmedtmann and Campagnolo 2015, described in Chapter 2.4.2 was adopted to the work flow. Similarly to Schmedtmann and Campagnolo 2015, the different values of λ were tested, ranging from 50% to 100% with a 5% step. For each λ, ACP for each class and overall ACP are estimated. The best method of TA1 comparison was selected for the calibration step.

5 Results

In the Tables and Figures of this Chapter UA, PA and f1 values are macro averaged between classes, since the aim was to be able to have a good classification performance for every class. There is one exception of this in Table 23, where UA, PA and f1 values were calculated for each class individually and the macro averaged values are presented below them. In the Tables, the bold values mean the best value of the metric.

5.1 Mask generation and satellite image assesment

Using different masks helped to automate the process, since there was no need to identify clouds, snow or cloud shadow manually. S2cloudless cloud mask, L1C cloud mask, L2A snow mask and cloud shadow mask used in the processing were calculated as described in Chapter 4.2.1. There was a notable difference between the S2cloudless mask and L1C cloud mask and thus it seemed like a good decision to use both cloud masks to cover all clouds (Figure 14). With the help of generated masks, it was possible to evaluate the cloudiness of the Sentinel-2A images for TA1. The cloud coverage assesment, i.e. the ratio of cloudy pixels, of the all Sentinel-2A images used in the study was calculated with the average of every image in three days period and with monthly average. For three days average, there were no date where the cloud coverage was less than 25% and in many cases it was over 80% (Figure 15). For monthly average, the cloud coverage was over 50% for every month and on august, it was over 70% (Figure 16).

5.2 Preprocessing

5.2.1 Rasterization

As expected, the rasterization process left some of the parcels out because the 10m pixel size might have been too large for the parcel with -5m buffering and the geometries became invalid. Total of 2645 parcels were left out from parcels in X_1 and 64660 parcels were left out from X_2 for TA1 (Table 8). The further experiments and model evaluations were performed with the rest of the parcels.

5.2.2 Temporal interpolation

Temporal interpolation found the optimal time period to be between 1st May 2017 and 1st September 2017 with 4 monthly interpolation steps. The temporal interpolation was performed with 109 models and temporal range of t-int4 from Table 6 ("05-09" in Figure 17), gives best OA and K scores with 90.9% and 84.6% with four interpolation steps and the second best values were for t-int4 with 3 interpolation steps (Figures 6

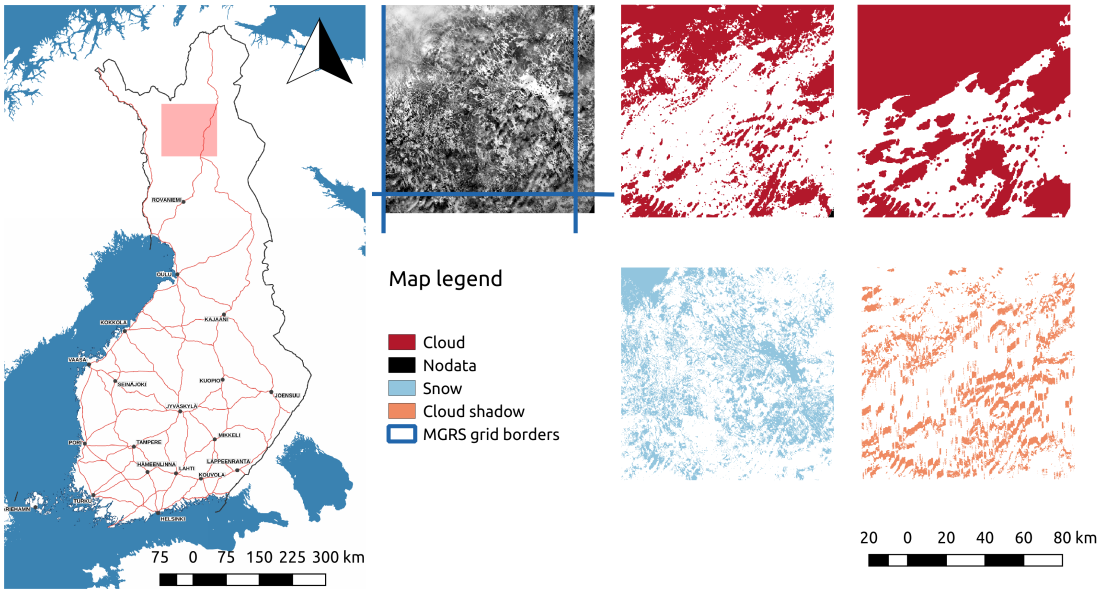


Figure 14: Sentinel-2 band B2 image for 27th April 2017 of the tile 35WMR and different masks. The masks from left upper corner to right lower corner: S2cloudless cloud mask, L1C product cloud mask, L2A snow mask and generated cloud shadow mask. The basemap is provided by National Land Survey of Finland (2018).

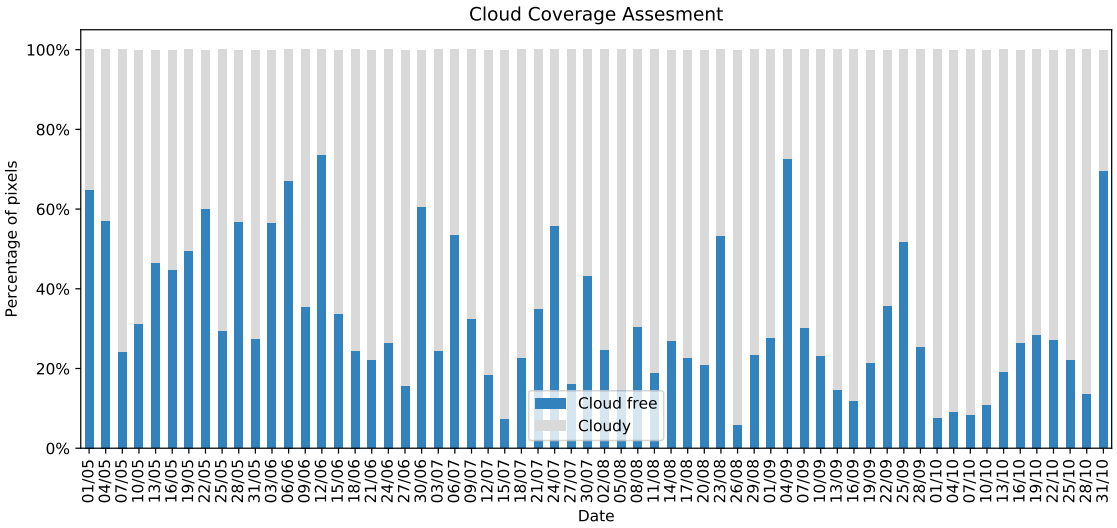


Figure 15: Three days average of cloud cover for Sentinel-2 images between May and October 2017 in Finland.

and 18). t-int4 was chosen for the further analysis, which means using the temporal period between 1st May 2017 and 1st September 2017 and 4 monthly interpolation steps, which corresponds to total of 16 interpolated dates. 16 dates in that range meant approximately one date in every 7 days. Since 16 does not divide evenly between 1st May 2017 and 1st September 2017, the actual ending of the range was 29th August 2017.

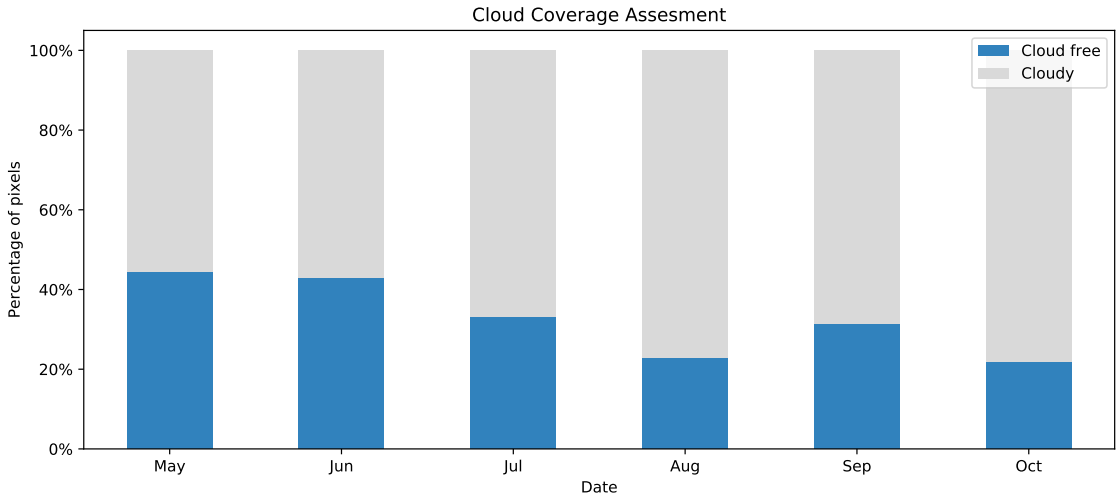


Figure 16: Monthly average of cloud cover for Sentinel-2 images between May and October 2017 in Finland.

5.2.3 Decomposition comparison

PCA expected variance ratio of 0.99 was found to be the most balanced ratio between training time and accuracy. The results of comparison between 5 PCA decomposition ratios show that the best OA and K was achieved with PCA ratio of 0.00, i.e. by using the original components of the data vector (Table 9). However, with ratio of 0.99, the execution time, f1, PA and UA were notably better with only slight difference in OA and K. Therefore the PCA ratio of 0.99 was selected for the further experiments. The time shown in Table 9 corresponds to the total cross-validated training time. With PCA ratio over 0.0, the PCA fitting and transforming was also included in the time so with larger dataset the time differences would be more significant.

5.2.4 Threshold calibration

With liberal filtering thresholds, it was possible to asses how many images there are to analyze for each parcel. The total amount of images and the monthly amount of images between 1st May 2017 and 1st September 2017 was calculated for parcels in X_1 and in X_2. Some of the parcels had only one or just couple of images available during the whole growth period (Figures 19 and). This would make it impossible to identify the crops properly. Based on the monthly image availabilities, the choise of 4 monthly interpolation steps seems well justified (Figures 21 and 22).

The threshold combination $\rho = 75, \nu = 10, \phi = 365, \varphi = 4$ were found to be a good balance between good prediction accuracy and amount of kept parcels. Using the findings from Chapter 5.2.2, 13 classification models were trained for threshold calibration. Two of the 13 models cut the parcel count so low, that the models could

Table 8: Parcels that were left out during rasterization process with -5m buffer. The Average area includes the standard deviation after ± symbol.

Class	# of Parcels in X_1	Total Area (ha)	Average Area (ha)
BEA	8	1	0.1 ± 0.1
PEA	21	2	0.1 ± 0.1
BEE	17	1	0.1 ± 0.1
FAL	125	102	0.8 ± 1.9
SRA	5	0	0.1 ± 0.1
SCE	643	916	1.4 ± 2.1
GRA	1522	1619	1.1 ± 1.7
POT	298	27	0.1 ± 0.3
TRA	3	0	0.0 ± 0.0
WCE	3	0	0.1 ± 0.0
Class	# of Parcels in X_2	Total Area (ha)	Average Area (ha)
BEA	503	1316	2.6 ± 2.9
PEA	565	1130	2.0 ± 2.7
BEE	333	830	2.5 ± 2.4
FAL	1326	1987	1.5 ± 2.3
SRA	428	1624	3.8 ± 4.0
SCE	22969	54803	2.4 ± 2.6
GRA	35380	54760	1.5 ± 2.0
POT	1418	1474	1.0 ± 1.5
TRA	522	1512	2.9 ± 3.3
WCE	1216	3540	2.9 ± 3.0

Table 9: The results of PCA decomposition ratio comparison

PCA ratio	Time (s)	OA	K	f1	PA	UA
0.00	3.6	**85.7**	**75.7**	51.7	58.9	48.8
0.25	1.6	56.9	18.6	11.4	10.7	12.4
0.50	1.7	60.5	26.0	12.6	21.6	13.5
0.75	**1.5**	71.0	47.3	21.1	35.4	20.7
0.99	2.1	85.4	75.2	**55.4**	**62.0**	**51.9**

not be run. They were φ with values 20 and 28. Changing two thresholds seemed to improve the accuracies without losing much of the parcels: $\nu = 10$ or $\nu = 20$ and $\varphi = 4$ (Table 10). Based on these observations, two other comparison runs were made: one with $\nu = 10, \varphi = 4$ and one with $\nu = 20, \varphi = 4$ keeping the liberal values with other thresholds. The $\nu = 10$ lost less parcels and gave better f1, PA and UA values than $\nu = 20$, so therefore the following threshold values were chosen: $\rho = 75, \nu = 10, \phi = 365, \varphi = 4$ (Table 11). To verify that the distribution of the classes was still similar to the liberal one, the class distributions were calculated.

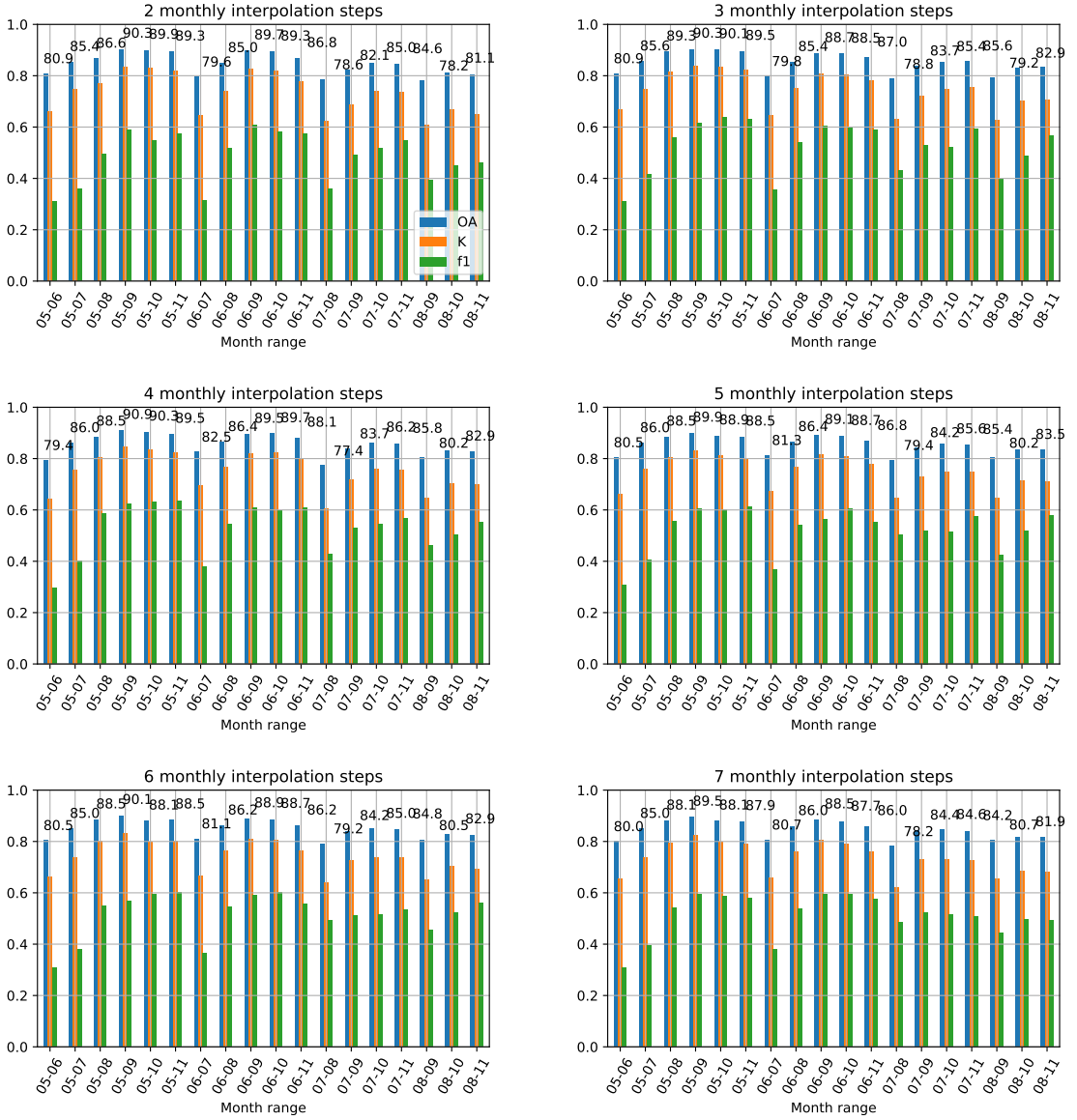


Figure 17: Accuracies of different temporal ranges and interpolation steps. The x-axis represents the months of the starting date and ending date

The distribution ended up being similar enough, so those parameters were used in the further analysis (Table 12).

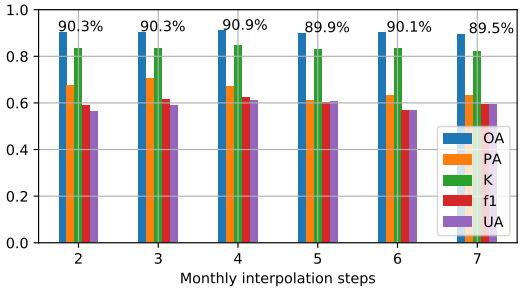


Figure 18: Accuracies of different interpolation steps between 1st May 2017 and 1st September 2017

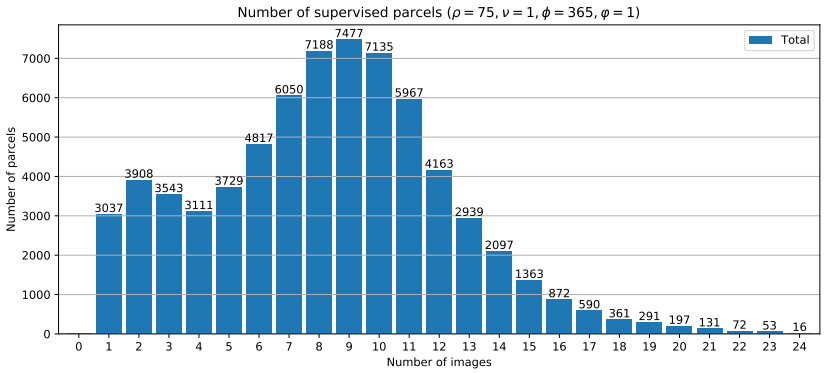


Figure 19: Parcels histogram showing the amount of Sentinel-2 images for X_1 between 1st May 2017 and 1st September 2017. The preprocessing filtering thresholds are with liberal values.

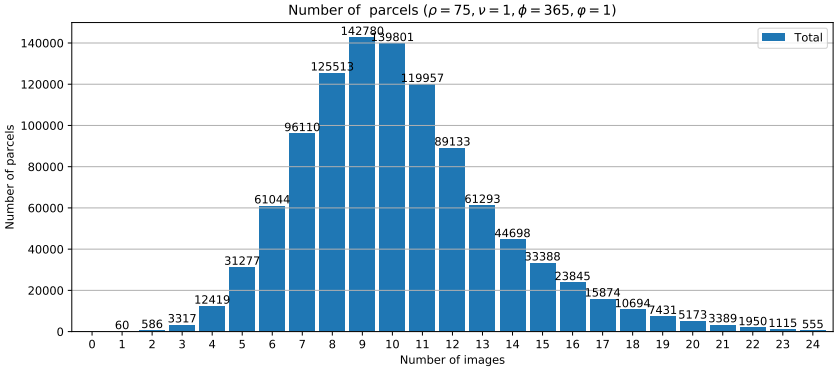


Figure 20: Parcels histogram showing the amount of Sentinel-2 images for X_2 parcels between 1st May 2017 and 1st September 2017. The preprocessing filtering thresholds are with liberal values.

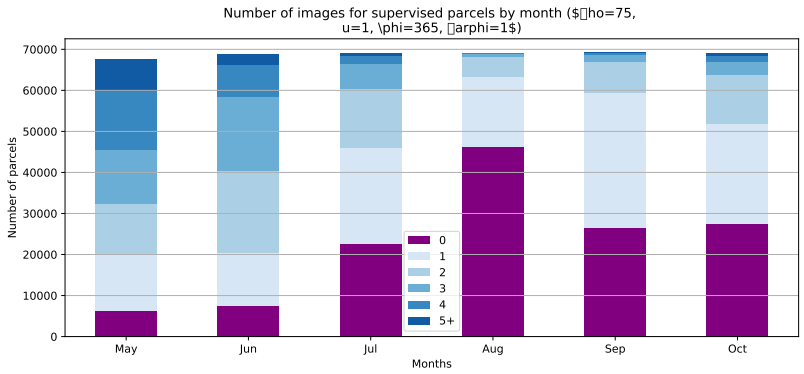


Figure 21: Parcels histogram showing the amount of Sentinel-2 images for X_1 parcels by month.

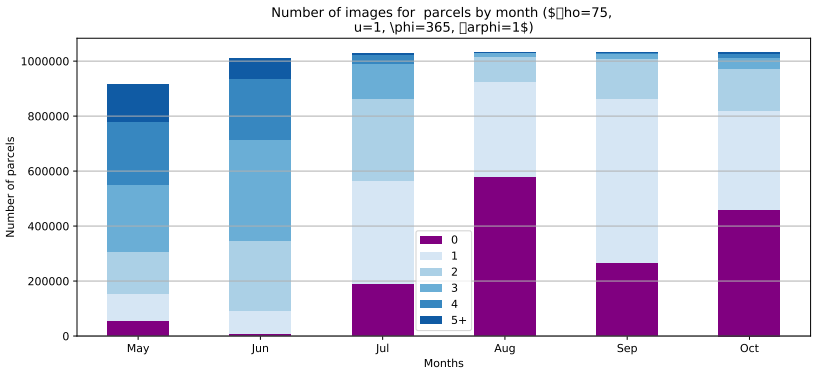


Figure 22: Parcels histogram showing the amount of Sentinel-2 images for X_2 parcels by month.

Table 10: The results of a threshold calibration. Explanations: ν = Clear pixel count, ρ = Cloud pixel percentage, ϱ = Shadow pixel percentage, ϕ = Days between valid images and φ =Total number of valid images

Threshold	Value	OA	K	f1	PA	UA	Parcels
Liberal	0	83.8	70.5	42.0	52.6	37.2	**53852**
ρ	0	83.4	69.7	39.9	52.0	35.2	53188
ρ	25	83.8	70.5	42.0	52.6	37.2	**53852**
ρ	50	83.8	70.5	42.0	52.6	37.2	**53852**
ϕ	10	80.2	65.7	**53.6**	**70.4**	47.9	800
ϕ	20	84.6	71.7	44.1	52.3	40.3	7354
ϕ	30	**85.7**	**73.9**	46.3	55.1	41.7	18581
φ	4	85.1	73.0	43.6	53.6	39.0	48150
φ	12	81.7	69.8	51.5	56.0	**49.0**	3199
ν	10	84.2	71.4	42.1	52.6	37.4	49868
ν	20	84.5	72.1	42.6	53.0	38.1	45908

Table 11: The results of a second threshold calibration. Explanations: ν = Clear pixel count, ρ = Cloud pixel percentage, ϱ = Shadow pixel percentage, ϕ = Days between valid images and φ =Total number of valid images

Threshold	OA	K	f1	PA	UA	Parcels
$\rho = 75\nu = 10\phi = 365\varphi = 4$	85.5	73.9	**43.8**	**53.3**	**39.5**	**44634**
$\rho = 75\nu = 20\phi = 365\varphi = 4$	**85.8**	**74.6**	43.5	52.7	39.3	41180

Table 12: The class distributions of liberal threshold values and potentially better threshold values.

	BEA	PEA	BEE	FAL	SRA	SCE	GRA	POT	TRA	WCE
Liberal	412	468	311	1036	381	19623	29166	995	420	1040
$\rho = 75\nu = 10\phi = 365\varphi = 4$	373	360	301	810	340	17010	23442	722	358	918

5.3 Variable Selection

5.3.1 Statistical features

Using just bandwise median values proved to provide best balance betnween accuracy and efficiency. The comparison of 7 combinations of statistical features presented in Chapter 4.3.1 found out that the combinations MEDIAN and AVG+MEDIAN seemed to perform the best but there was no large difference between them (Table 13). However, combining the bandwise average pixels with median pixels, the dimensionality would double and that would mean more complex models and longer training times. Since the difference was not that significant, the median statistical feature was selected as the feature for the rest of the models.

Table 13: The results of a statistical feature selection.

Name	OA	K	f1	PA	UA
AVG	88.8	80.0	58.6	76.4	51.3
MEDIAN	**89.0**	**80.3**	58.4	76.5	51.4
STD	78.3	60.1	30.6	46.0	26.8
AVG+STD	87.5	77.5	53.7	70.2	47.1
AVG+MEDIAN	88.5	79.3	**59.1**	**77.8**	**51.8**
MEDIAN+STD	87.5	77.6	54.3	71.1	47.5
AVG+MEDIAN+STD	87.2	77.0	54.4	71.7	47.7

5.3.2 Vegetation indices

The best classification accuracy using just VIs was found to be with NDVI calculated from bandwise median (later "median <VI>"), average GNDVI calculated from bandwise average, and MCARI, but when combined VIs and bandwise median, the best accuracy was received using the combination of bandwise median, median NDVI and median GNDVI (Tables 14 and 15). The 26 VI combination were calculated as described in Chapter 4.3.2. Some of the best classification accuracies were received by combining VIs calculated with both bandwise average and bandwise median reflectance values and by combining multiple VIs (Table 14). For another iteration, some of the best working VIs and bandwise median were combined. The combination of bandwise median, median NDVI and median GNDVI was selected, because it had the best OA, K and f1 values (Table 15). Calculation of the VIs allowed the opportunity to visually inspect the similarity of the crop classes. Median NDVI time series of each crop class was calculated for studying the similarities in spectral signature. Some of the classes, like WCE and POT have unique looking NDVI signature based on the mean of all X_1 parcels, while other classes like PEA, SRA, SCE and TRA have similar looking NDVI signature (Figure 23).

Table 14: The results of a Vegetation Index comparison. Indices without _AVG suffix were calculated using median features.

Name	OA	K	f1	PA	UA
GNDVI	81.3	65.8	27.9	52.3	25.6
GNDVI+MCARI	82.7	68.3	32.6	52.0	28.6
GNDVI+MCARI_AVG	84.9	72.6	40.9	74.1	34.6
GNDVI_AVG	81.4	65.9	28.7	79.1	25.2
GNDVI_AVG+MCARI	84.9	72.5	40.3	73.7	34.1
GNDVI_AVG+MCARI_AVG	84.8	72.5	40.3	73.8	34.0
MCARI	79.1	61.6	19.9	26.0	19.5
MCARI_AVG	80.0	63.3	19.6	72.0	19.3
NDVI	83.5	69.5	30.7	70.6	27.5
NDVI+GNDVI	85.3	73.6	42.8	64.9	37.0
NDVI+GNDVI+MCARI	85.4	73.7	42.5	63.6	36.8
NDVI+MCARI	83.3	69.6	34.0	59.6	30.7
NDVI+GNDVI+MCARI_AVG	86.9	76.3	49.0	79.3	41.3
NDVI+GNDVI_AVG	85.4	73.6	41.6	79.4	35.2
NDVI+GNDVI_AVG+MCARI	**86.9**	**76.4**	**49.3**	78.9	**41.6**
NDVI+GNDVI_AVG+MCARI_AVG	86.9	76.3	49.2	79.3	41.5
NDVI+MCARI_AVG	85.0	72.7	38.8	75.8	32.7
NDVI_AVG	82.8	68.6	29.7	**82.0**	25.9
NDVI_AVG+GNDVI	85.4	73.6	41.5	79.8	35.1
NDVI_AVG+GNDVI+MCARI	86.9	76.3	48.8	79.2	41.2
NDVI_AVG+GNDVI+MCARI_AVG	86.9	76.3	49.1	79.0	41.6
NDVI_AVG+GNDVI_AVG	85.4	73.6	41.5	80.2	35.1
NDVI_AVG+GNDVI_AVG+MCARI	86.9	76.3	49.1	79.1	41.4
NDVI_AVG+GNDVI_AVG+MCARI_AVG	86.9	76.3	48.7	79.3	41.1
NDVI_AVG+MCARI	85.0	72.7	38.6	75.9	32.5
NDVI_AVG+MCARI_AVG	85.0	72.7	38.7	75.9	32.6

Table 15: The results of a Vegetation Index joined with median comparison. Indices without _AVG suffix are calculated using median features

Name	OA	K	f1	PA	UA
MEDIAN+NDVI	90.0	82.2	61.3	73.5	56.6
MEDIAN+GNDVI	89.3	80.9	58.9	76.5	52.3
MEDIAN+GNDVI_AVG	89.3	80.9	61.5	74.5	56.2
MEDIAN+NDVI+GNDVI	**90.1**	**82.3**	**62.9**	75.0	**58.4**
MEDIAN+NDVI+GNDVI_AVG	89.5	81.2	59.0	74.0	52.5
MEDIAN+NDVI+GNDVI+MCARI	89.8	81.8	61.9	**77.2**	56.2
MEDIAN+NDVI+GNDVI_AVG+MCARI	89.0	80.2	58.4	71.8	53.2

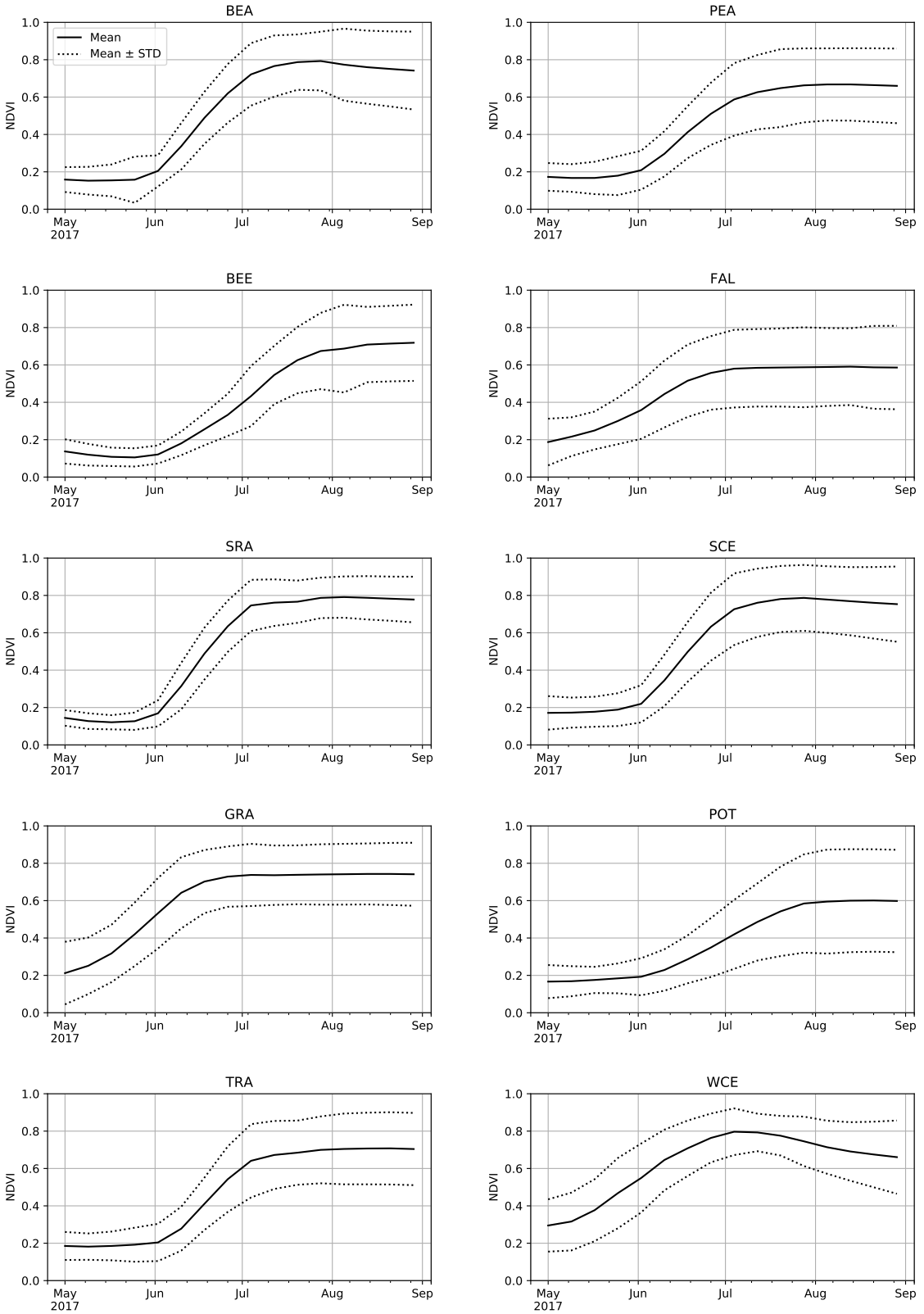


Figure 23: Time series of NDVI mean and standard deviation values for each crop class. The data is based on the supervised parcels extracted with parameters $\rho = 75, \nu = 10, \phi = 365, \varphi = 4$

5.3.3 Most meaningful original components

After identifying most meaningful original components that did not use PCA decomposition, the best classification accuracy was still provided with the PCA composited components of bandwise median, median NDVI and median GNDVI. The PCA ratio 0.99 adopted in Chapter 5.2.3 used 50 generated PCA components for the MEDIAN+NDVI+GNDVI data. If the PCA components corresponding to all 192 variables $(16 * (10 + 1 + 1))$ were to be used, the explained variance ratio of the components is shown in Figure 24. The first few components are the most important ones and the components from the end of the graph could as well be removed as was done when using the explained ratio of 0.99. To find out the most meaningful original components, the mean and variance of the original component contribution was calculated and the values are presented in 25. The components with highest absolute mean were selected for comparison in an order of highest to lowest. The indices of the first 10 were: 145, 121, 177, 113, 119, 159, 118, 156, 14 and 123. The most meaningful variable was the variable at index 145, which was B2 median reflectance for interpolated date 5th August 2017.

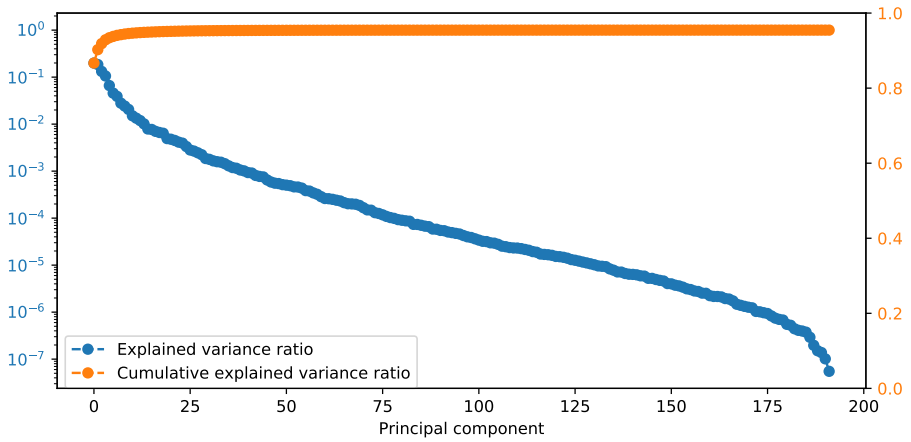


Figure 24: Explained ratios of the PCA components of parcel data including 10 reflectance values and 2 VI values

The comparisons between first 10, 20, 30, 40 and 50 most meaningful variables was calculated to find out whether the accuracies would be any better than with PCA decomposed components. The accuracies with the most meaningful original variables stayed below the 50 first PCA component with every number of original components below 50 (Table 16). This suggested that the PCA components manage to capture more dependencies than the original variables. The further experiments in this study were made with the PCA ratio of 0.99 after all.

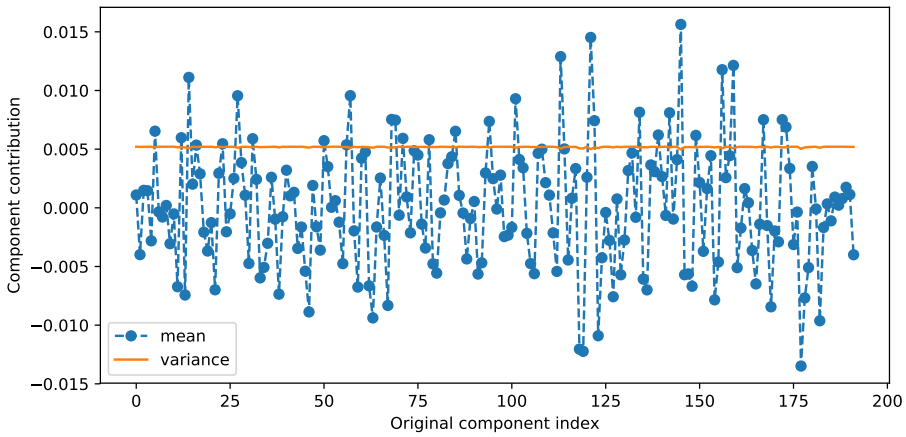


Figure 25: Mean variance of original variables of parcel data including 10 reflectance values and 2 VI values

Table 16: The accuracies of the original most meaningful values.

# of variables	OA	K	f1	PA	UA
10	72.3	48.8	25.2	52.0	23.3
20	86.0	74.7	43.4	62.5	38.5
30	87.9	78.4	51.8	67.2	46.1
40	88.5	79.4	53.7	67.9	47.9
50	**88.7**	**79.9**	**56.5**	**69.7**	**50.8**

5.4 Classification step

5.4.1 Imbalance of the crop classes

The use of over-sampling technique ROS with ratio of 1100 was found to be the best balance between UA and OA and was adopted for further experiments. The initial comparison of over-sampling and under-sampling revealed that out of the two over-sampling methods, ROS outperformed SMOTE (Table 17) and actually improved the accuracies when compared accuracies of the one without resampling (MEDIAN+NDVI+GNDVI in Table 15). For under-sampling, RUS improved UA values but with the cost of other accuracy metrics (Table 18).

Based on the resampling results, ROS and RUS were selected for further comparison with over-sampling values between 900 ans 15000. ROS value with 1100 (later ROS:1100) improved all metrics except PA when compared to classification without resampling (Table 19).

Yet another iteration was made with both ROS:1100 and different values of RUS ranging from 3000 to 13000. The best result ROS:1100+RUS:13000 of that comparison (Table 19) did not meet the accuracies of ROS:1100. The SVM balanced weights

Table 17: The results of a over-sampling comparison

Name	Value	OA	K	f1	PA	UA
ROS	600	**90.3**	**82.7**	63.9	73.8	60.2
	700	90.3	82.7	**64.9**	**73.9**	**61.8**
	13600	88.7	80.3	63.3	68.2	60.7
	18800	88.9	80.5	63.7	70.4	60.1
SMOTE	600	90.1	82.4	63.7	72.5	60.8
	700	90.1	82.3	64.1	71.7	61.7
	13600	88.4	79.7	62.1	67.0	58.8
	18800	88.5	79.8	61.6	68.3	57.2

Table 18: The results of a under-sampling comparison

Name	Value	OA	K	f1	PA	UA
AKNN	auto	88.9	80.1	55.3	70.3	52.0
CC	2000	79.4	67.4	51.9	45.0	68.2
	1000	66.0	51.6	43.8	38.7	67.5
	700	59.4	44.8	41.0	37.3	67.2
	600	57.4	42.9	40.2	37.0	67.8
	200	42.6	29.3	28.8	30.0	62.9
CNN	auto	81.3	67.8	46.0	45.4	51.9
IHT	auto	36.4	24.1	28.5	31.8	59.4
NCL	auto	88.4	79.2	51.9	69.1	48.1
OSS	auto	89.0	80.3	58.5	66.6	55.0
RUS	2000	85.5	75.7	58.0	52.1	67.6
	1000	78.7	66.4	52.2	45.5	68.9
	700	74.1	60.6	49.6	43.4	**69.3**
	600	72.8	59.1	48.4	42.3	68.8
	200	65.8	50.4	40.8	36.0	65.9
TL	auto	**90.0**	**82.1**	**62.4**	**74.7**	58.1

did not give good results either on any metric, as shown in Table 21. Therefore, ROS resampling was adopted with ratio of 1100 to deal with crop class imbalance in further experiment

5.4.2 Hyperparameter tuning

After the hyperparameter tuning, the following hyperparameters were obtained: SVM: $C = 3.1$ and $\gamma = 0.02$, RF: n_estimators=226, max_depth=608, min_samples_split=2, max_features=10 and MLP: hidden layers=(51,51,35). The tuning of the hyperparameters (Table 22) made SVM perform more poorly when compared to previous

Table 19: The results of another over-sampling comparison

Name	Value	OA	K	f1	PA	UA
ROS	1200	90.2	82.6	65.6	71.1	63.7
	1000	90.2	82.5	64.9	71.7	62.4
	1100	90.3	82.7	**66.2**	**72.8**	63.7
	1500	90.1	82.6	66.0	69.3	**64.9**
	900	**90.3**	**82.8**	65.4	71.9	62.9

Table 20: The results of combined over-sampling and under-sampling.

Name	Value	OA	K	f1	PA	UA
ROS+RUS	1100+13000	**90.1**	**82.5**	**65.3**	**69.7**	63.9
	1100+10000	89.8	82.2	65.1	66.8	65.0
	1100+8000	89.2	81.1	64.2	63.4	66.0
	1100+7000	89.0	80.8	64.1	62.4	66.7
	1100+6000	88.6	80.3	63.7	60.9	67.5
	1100+5000	87.9	79.2	63.2	59.4	68.4
	1100+4000	87.2	78.3	62.4	57.3	69.7
	1100+3000	85.6	75.9	60.3	54.6	**69.8**

best ROS:1100 in Table 19. This might suggest that the information about training set might have leaked to the testing set. Fortunately the validation dataset of X_2 was still left untouched, since training was done with test set of X_1.

When comparing the metrics between SVM, RF and MLP, every classifier was best with some metric. RF outperformed others with PA, but the low UA score suggested that the model had suffered from imbalance and had failed to perform accurate predictions on all classes. MLP got the highest UA, but the other metrics were lower than the metrics of SVM. These tuned models were used to predict the parcels from X_2 in next section.

Table 21: The results of SVM with balanced weights.

Name	Value	OA	K	f1	PA	UA
SVM weights	balanced	**88.4**	**79.3**	**57.0**	**64.0**	**54.6**

Table 22: The comparison between hypertuned classifiers. Best values for each metric are bolded.

Name	OA	K	f1	PA	UA
SVM	**90.0**	**82.0**	**61.5**	75.8	54.5
RF	87.3	76.8	45.3	**91.2**	37.2
MLP	88.6	80.2	60.2	58.4	**63.0**

5.5 Classification method comparison

5.5.1 Crop identification comparison for test area of whole Finland

In previous section the classification models SVM, RF and MLP were trained and optimized with X_1 data. Here the models were used to predict all the parcels from X_2 as a final evaluation of TA1. Based on confusion matrices presented in Figure 26, it was clear that all of the models were negatively affected from imbalance and assigned samples to the dominating two classes SCE and GRA. Also, all three models failed to distinguish FAL from GRA. RF performed worst and basically failed to predict any other class than SCE and GRA. According to overall metrics SVM and MLP both performed well, but MLP was able to predict classes more evenly (Table 23).

Table 23: Accuracy metrics of SVM, RF and MLP classification methods trained with X_1 and evaluated with X_2. Avg row represents macro average of the values

	SVM			RF			MLP			count
Class	PA	UA	f1	PA	UA	f1	PA	UA	f1	
BEA	65.2	27.2	38.4	**96.3**	7.5	13.9	39.3	46.3	42.5	6273
PEA	36.9	14.3	20.6	94.5	3.6	6.9	18.6	27.3	22.1	5285
BEE	63.7	38.1	47.7	78.9	15.8	26.3	40.8	63.4	49.6	3254
FAL	37.3	7.1	11.9	93.3	2.0	3.8	21.3	13.1	16.2	13478
SRA	72.0	40.9	52.1	73.3	28.2	40.8	60.2	55.0	57.5	8307
SCE	88.4	91.9	90.1	82.7	**94.1**	88.1	90.7	90.9	90.8	334153
GRA	**91.0**	**95.7**	**93.3**	91.6	94.1	**92.8**	**93.8**	**94.0**	**93.9**	462744
POT	65.1	37.3	47.5	79.6	21.8	34.3	52.7	56.1	54.3	10277
TRA	62.0	23.4	34.0	72.5	13.4	22.7	42.8	42.0	42.4	8276
WCE	84.3	62.4	71.7	95.9	27.9	43.3	74.6	69.2	71.8	18171
Avg.	66.6	43.8	50.7	85.9	30.9	37.3	53.5	55.7	54.1	
OA	89.1			87.6			88.8			
K	80.1			77.1			80.4			

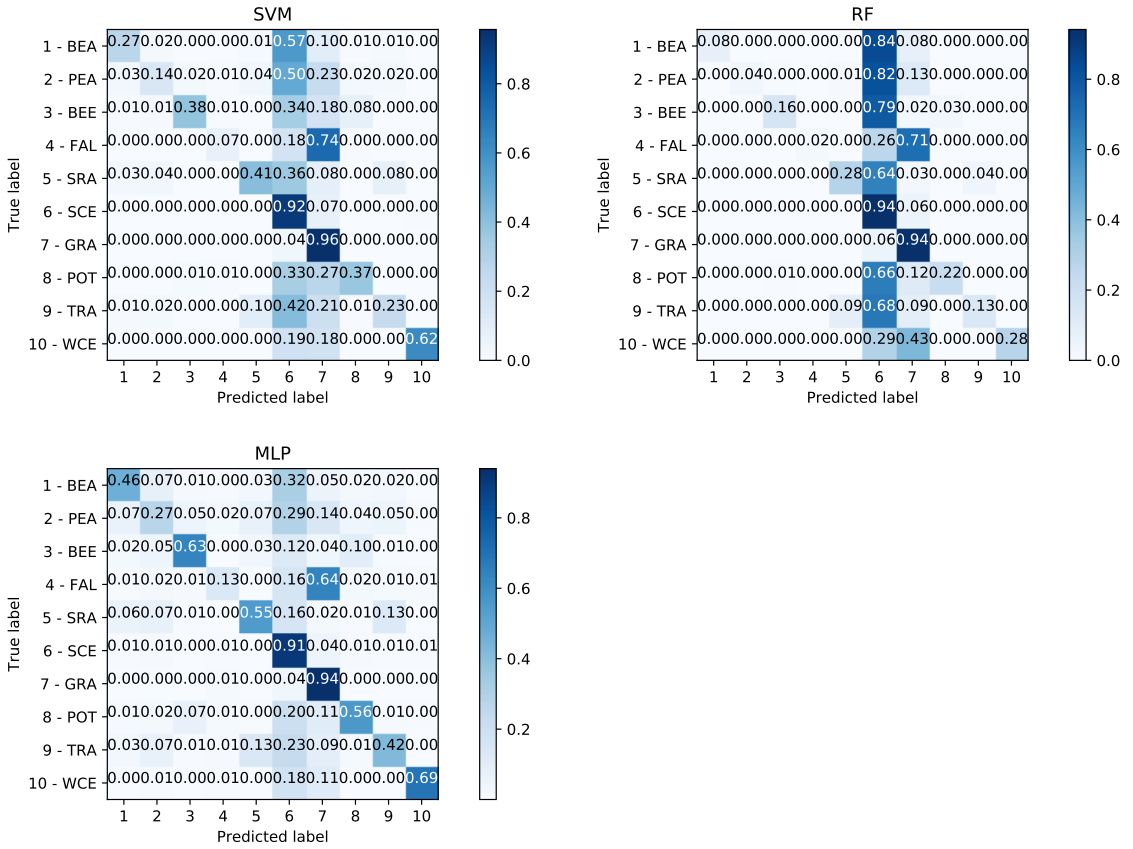


Figure 26: Confusion matrices of SVM, RF and MLP classification methods trained with X_1 and evaluated with X_2. Cell value and color correspond to the user's accuracy of the class.

5.5.2 Crop identification comparison for test area of one tile

Training of the ConvRNN took over 48 hours, while training and hyperparameter tuning of other methods took under 2 hours. Based on the OA and K metrics, MLP seemed to perform the best, but f1, PA and UA reveal that ConvRNN might perform better (Table 24). Based on the confusion matrices presented in Figure 27, it can be seen, that ConvRNN distributed the wrong predictions into much less cells than SVM or MLP. The reason of BEE row missing in ConvRNN confusion matrix, is that none of the parcels containing the class are assigned to the validation set. Other algorithms perform more poorly here than in previous comparison. One of the reason might be that there is no resampling in the classification data here and also ConvRNN had more images available, since Sentinel-2A and Sentinel-2B were both used in the training.

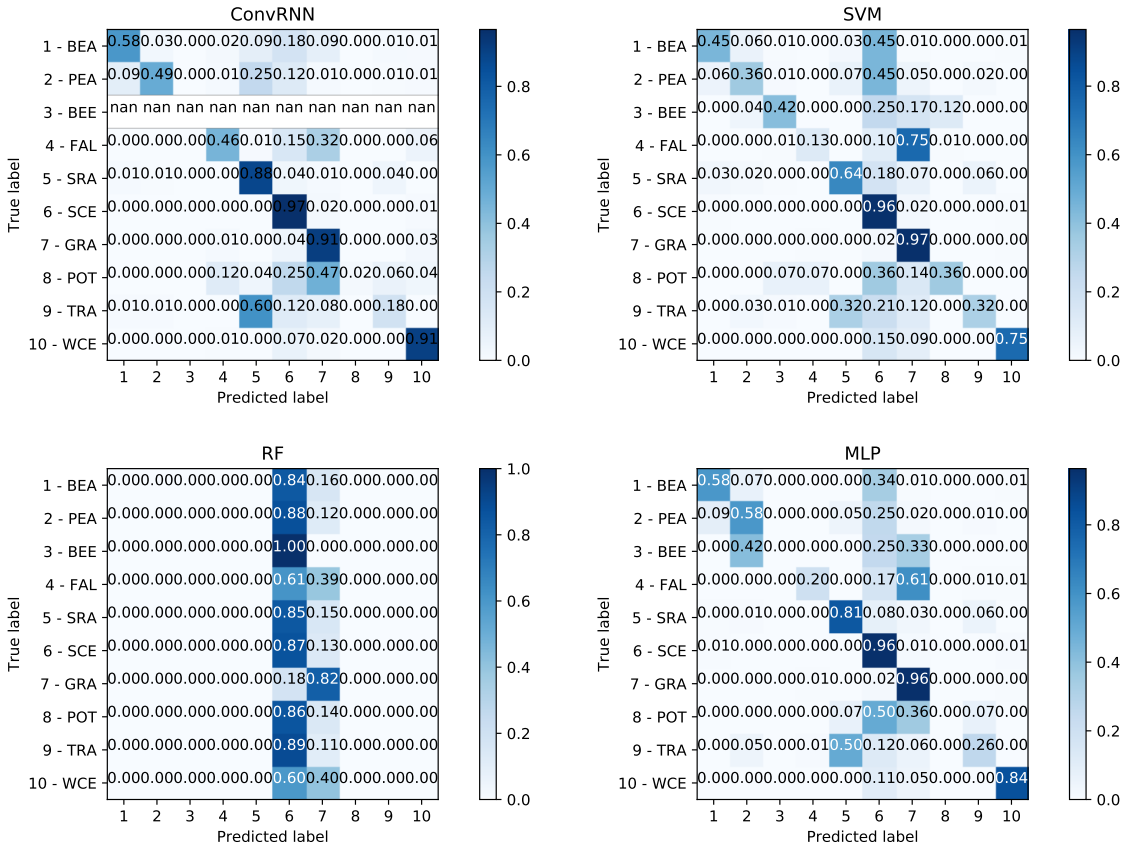


Figure 27: Confusion matrices of ConvRNN, SVM, RF and MLP classification methods trained and evaluated with X_2 parcels from tile 34VFN. Cell value and color correspond to the user's accuracy of the class.

Table 24: The ConvRNN result Best values for each metric are bolded.

Name	OA	K	f1	PA	UA
ConvRNN	91.3	85.1	**61.1**	**70.0**	**59.9**
SVM	89.9	83.3	58.2	67.2	53.3
RF	72.4	50.7	15.6	14.5	16.8
MLP	**91.5**	**86.1**	53.9	58.1	52.0

5.6 Classification calibration

For the classification calibration with reliability level λ, the MLP from 23 was chosen. The λ steps 0.5-1.00 were evaluated with the relationship between λ and overall ACP. The ACP got lower as λ increasesed, but very slowly until reliability level 0.9 was reached (Figure 28). Then after 0.95 ACP dropped dramatically to almost zero. Also ACP was evaluated for each class separately with different values of λ. Classes SCE and GRA had high ACP even at reliability level of 0.95, while other

classes started losing samples with earlier levels (Figure 29). Even the calibration did not help predicting FAL correctly and PEA vanished also at low levels. The relationship between classification accuracy and λ was also inspected for deciding the most suitable reliability level. UA, OA and K rise while increasing λ, but f1 and PA got worse before the end because calibration mainly affects UA (Figure 30). $\lambda = 0.7$ seemed to be like a balance between classwise ACP and classification accuracy. In the confusion matrix made for $\lambda = 0.7$ it is no surprise, that the most of the values in the matrix are 0.7, since that was set to be the lower limit of UA by the definition (Figure 31). Minimum confidence limits q_js for each crop class were calculated and for SCE and GRA, the limits were 0.22 and 0.20 respectively but for other classes limit was at least 0.43 (Table 25). these confidence values, the reliability level should be met without accessing the ground truth information of the parcels if the model was to be used with unseen data.

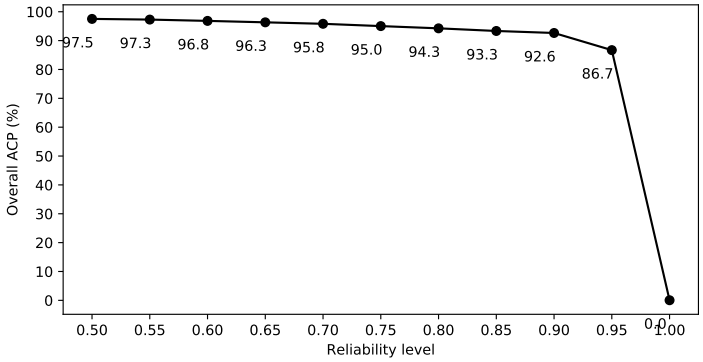


Figure 28: Automatic Classification Proportion as a function of reliability level λ

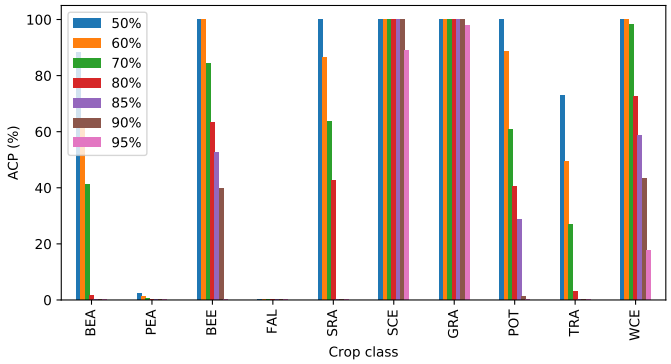


Figure 29: Automatic Classification Proportion as a function of reliability level in each crop class

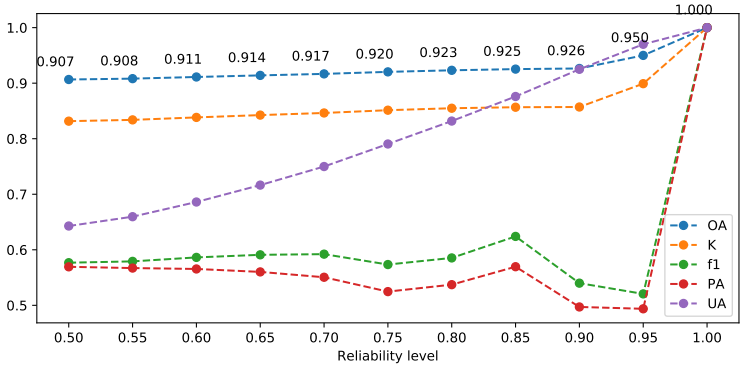


Figure 30: Classification metrics as function of reliability level

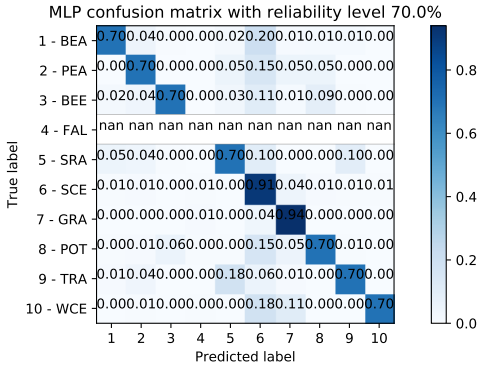


Figure 31: Confusion matrix of MLP model after calibration with reliability level $\lambda = 0.7$

Table 25: Estimations of q_j and ACP with reliability level 0.7. Cell color and value corrsepond to User's accuracy

	BEA	PEA	BEE	FAL	SRA	SCE	GRA	POT	TRA	WCE
q_j	0.980	1.000	0.630	NaN	0.930	0.22	0.2	0.82	0.990	0.430
ACP (%)	41.033	0.378	84.235	0.0	63.621	100.00	100.0	60.64	27.006	98.355

6 Discussion

In the performed case study of crop identification in Finland for the year 2017, four machine learning classification algorithms were tested. In 2017, the calculated average monthly cloud percentage was over 50% for all the 6 months of the thermal growth season, which made the season 2017 a cloudy one. On a three days average, none of the days showed cloud coverage under 20%. For the study area of whole Finland, SVM, RF and MLP classification methods were tested from which MLP method was chosen to be the best with original OA of 89%, K of 80% and f1 of 54%. After calibration with reliability level $\lambda = 0.7$, it was possible to rise OA to 92%, K to 85% and f1 to 59% with automatic classification proportion of 96%, i.e. by leaving 1618 parcels out from the total of 833856 parcels. Both the original OA and calibrated OA fell within the range of the best accuracies achieved in previous studies, indicating that the crop identification in Finland is possible even in cloudy years.

However, the relatively low f1 values indicated that the identification quality was not even with all the crop classes. Out of 10 crop classes, grass and spring cereal were the most dominating classes. This made both training and test classification datasets highly imbalanced. These two classes were predicted with UA of 91% and 94% respectively with uncalibrated ML model. The second best UA scores were 69%, 63% 56% and 55% for winter cereal, beet, potato and spring rapeseed respectively. The second best model SVM, was able to predict grass and spring cereal with UAs of 96% and 92%, but it failed to get as good results with all other classes. All three methods failed classify fallow, mixing it partly with spring cereal but mostly with grass. For MLP the other classes than the dominating two had at least 12% of the parcels classified to spring cereal. For SVM, the same ratio was 18%. RF, which performed most poorly, failed to predict any other class than spring cereal or grass well. This study found the imbalance to be one of challenging problem in classification, as has been found in previous studies.

The UA score ordering from largest to smallest of MLP follow the parcel count of the training dataset until beet. Thus, the imbalance seems to be related to the distribution of parcels in classes of the training of the models. Resampling of the classes was used to balance the distribution, but the resampling affected the other metrics in a negative way. One way to compensate the imbalance would be to divide the crops in crop classes in different way and to use different number of classes. A further research is needed for improving the results of the crop identification with different division of crops in crop classes and different crop class formation.

With the smaller are of interest, ConvRNN method was chosen to be the best with OA of 91%, K of 85% and f1 of 61%. MLP had better OA and K scores of 92% and 86%, but lower f1 score 54%. ConvRNN method did not suffer from imbalance that much, since it managed to classify smaller classes winter cereal and spring rapeseed of UAs 91% and 88% as well as larger classes spring cereal and grass with UAs 97% and 91%. MLP, SVM and especially RF had more significant differences between

the two dominating and the rest. Out of all methods, ConvRNN was the only one that could predict the most poorly predicted class within all experiments, fallow. ConvRNN was able to predict fallow with UA of 46%, while the second best UA score 20% was produced by MLP. ConvRNN and MLP methods failed to predict potato completely, by mixing it with grass and spring cereal, but SVM was able to perform little better there. Since it seems that none of the classifiers was able to perform best with all classes, a further research is needed for joining the classification methods in a manner that each classifier detects the ones that it is most suitable for.

The best classification methods were thus found to be MLP and ConvRNN while SVM performed also quite well. Unlike in previous studies, MLP performed better than SVM for identifying crops more equally. Also other studies have achieved good results with RF but this study found it to perform worst. It would be a subject of further study, if RF would work better with different implementation or processing steps and better hyperparameter optimization.

The study found that the optimal time period for identifying crops in 2017 was between period of 1st May 2017 and 1st September 2017 with 16 interpolated dates. Generally, for every tested starting date, the results got better when making the period longer until a certain point, where the results started decreasing. Typically including October and often also September, made the accuracy of the classification worse. As shown in Figure 23, broad bean, beet, spring cereal, grass and turnip rape all had similar NDVI values at the end of August. Making the period last longer would not help for recognizing the crops better. In the studies, also September was included, but here it was not found to make the accuracies better. Otherwise, the start of the optimal period was at the start of the thermal growing season 2017.

Inside the optimal time period, the optimal number for monthly interpolation steps was 4, with best OA and f1 values. By increasing monthly interpolation steps, the results went better until 4 was reached, and after that the results got worse. 4 monthly interpolation steps mean that for each month, the number of images were interpolated to be four. For period of four months this totals to 16 interpolated dates. For supervised parcels in X_1, for most of the months the number of images, where the parcel was visible, was less than four. In August, there were few parcels that had even three images and most of the parcels did not have any images available (Figure 21). By making the number of monthly images artificially above four by interpolation, would not improve the results, since only few parcels would even have that much original images. Although less than 3000 supervised parcels had at least 16 images for the whole period between 1st May 2017 and 1st September 2017, it still made sense to use 16 interpolated dates (Figure 19). For majority of the parcels, this lead to repeating same values during many interpolated dates, but nevertheless, it helped to make the identification more accurate.

Some of the spectral differences between different crop classes could be visually identified using NDVI time series in Figure 23. Two of the most different signatures from others are with winter cereal and potato. Winter cereal mean NDVI starts from 0.3 and achieves its peak 0.8 already at beginning of July and then it goes

down to belove 0.7 towards end of August. The shape is different from any other classes and no other class was found to mix with winter cereal in classifications and winter cereal was found to achieve the third best accuracy scores (Figures 26, 27 and Table 23). With potato, the mean NDVI values rose from 0.2 and achieved the peak 0.6 in August. Fallow had same shaped signature and it also started from 0.2 and had peak of 0.6 after steady rise, but it achieved the peak already at the beginning of July, where potato had the mean value of 0.4. From Figure 26 it can be seen than potato and fallow did not mix with each greatly with any classification method. Despite having different signatures from the dominating classes spring cereal and grass, winter cereal and potato still got mixed up with them with most of the classification algorithms.

Some of the signatures were quite similar with each other. For example, pea, spring rapeseed, spring cereal and turnip rape resemble each other. All of those classes were at least partly predicted to be spring cereal in area of whole Finland (Figure 26). As stated in previous studies, the similarities in crop signatures and phenology made distinguishing some of the crops difficult also in this study.

Training of ConvRNN was computationally expensive even with the use of GPUs. One way to make the training faster would be to use the same processing for Sentinel-2 images as was done with other classification methods. This way the temporal size of the training blocks would be smaller, and it would make the training faster. On the other hand, this would lose the ConvRNN's benefit to detect clouds automatically and to be relatively free of preprocessing. With more efficient algorithms and computers, the training could be also made faster. On its current state, it would have been extremely computationally expensive to train and evaluate ConvRNN model with samples of whole Finland. Therefore, further study is needed to optimize the speed of ConvRNN classification method to be able to compare it with other methods for the area of Finland.

In previous studies some of the best crop identification performances were achieved by using SAR sensor images in addition to the optical satellite imagery. Since this study focused only to use of optical imagery and more precisely to Sentinel-2, it is possible that the method would perform better with different set of imagery. Also, with just Sentinel-2, it would be possible to get better results when the both of the twin satellites Sentinel-2A and Sentinel-2B could be used.

Throughout the study a lot of choices and decisions had to be made based on previous studies and own experimental plans. All tested classification methods except ConvRNN used object-based classification, or more precisely PBC, instead of pixel-based for creating the data for the classification. Pixel-based classification would have allowed pixelwise interpolation of the values instead of interpolating the median and it would have been able to spot the anomalies inside the parcels while calculating prediction for each pixel individually. Also comparing the PBC methods with ConvRNN that used pixel-based classification, would have been more justified if all methods would have used that approach.

All potential classification methods, such as KELM, FNN and TWDTW, could not be tested because of tight schedule and resources, since they would have needed a different kind of workflow implementation compared to SVM, MLP and RF, which all used the same workflow. Therefore, further study is needed if one of the methods not used in this study, would actually perform better on crop identification in Finland.

The creation of VIs used statistical features instead of all the pixels in calculations. The VI values would have described the parcels better, if all the available pixel percentages would have been used in the calculations. The CV comparison done in few of the experiments would have been suitable for all comparisons before actual classification method comparison. It would have allowed to use all the supervised parcels instead of 80% of them for validating the right parameters and processing steps. This leads to one of the most suspicious decisions made in the study: to use solely SVM for determining the best processing steps.

After hyperparameter optimization SVM performed more poorly with the test data, so it could be possible that the hyperparameter tuning phase was not able to find the optimal parameters after all. Since SVM did not even produce the overall best results during classification method comparison, it could have been possible to produce even better results when all the three classification methods would have been utilized. However, the described processing comparison would be relatively easy to perform with other algorithms in further studies. By using Sentinel-2A and Sentinel-2B images, ConvRNN got much more images to analyse in training compared to other methods using Sentinel-2A images. This might have also affected the results and it could be tested to use just TOA reflectance or BOA reflectance with all the methods.

Implementing some of the processing steps and computations in this study were extremely challenging and some of the subjects were not familiar beforehand. That is why there might be some mistakes or bugs in the workflow even if it has been evaluated during the development. With peer code review the probability of bugs would have been smaller. Some of the results in this study even imply that there might have been some bugs in the workflow. For example, the average areas for parcels in X_2 in Table 8 are much larger than the average parcel areas for parcels in X_1. This suggests that there has been some mistake in rasterizing the parcels, so that also some valid parcels were not rasterized and were thus left out with no valid reason. However, the amount of parcels in X_2 was still sufficient for being a reliable test data for the classification methods.

After all, the results of this study were rather satisfactory, since beforehand the identifiction was not thought to work as well as it actually did. However, with the identification accuracies varying between crop classes, it is unclear whether the developed method could be used for replacing the previous methods used within IACS in Finland. In practice use of automated crop identification would reduce the costs of crop monitoring significantly. Using arbitrary reliability level set by decision makers and by accepting that some of the crops could not be automatically classified, it is possible to calibrate the method to be able to identify most parcels. Rest of the parcels could be identified using existing methods, but that number would be lot

less. Also, experimental confidence limits for each crops corresponding to the chosen reliability level found in this study, could be used when testing the suitability.

7 Conclusions

This thesis studied the crop identification with Sentinel-2 imagery using machine learning classification methods. The aim was to find the best method and time period for the task and to find out which crop classes could be identified best. The approach of the thesis was to investigate previous studies of crop identification to find out the recommended processing steps of satellite imagery and land parcels and to narrow down the alternatives for the most suitable methods. Afterwards case study was performed for the growing period of 2017 in Finland with the selected methods and developed processing and optimization steps and based on the results the best methods were identified.

The main conclusion of the study is that MLP classification method performed the best in whole Finland and ConvRNN classification method performed best in the area of one tile in Finland 2017. The optimal temporal period for the identification was found to be between 1st May 2017 and 1st September 2017. The crop classes that could be identified with greatest certainty were spring cereal and grass in all the experiments.

This study created a method to process the optical satellite imagery and parcel geometries to be able to perform crop identification in cloudy year in Finland. The method processed as many Sentinel-2 images as possible to get sufficient amount of images for every parcel to tackle the cloudy weather conditions. The images were masked with two cloud masks, snow mask and cloud shadow mask to get only clear pixels out of them. The statistical BOA reflectance quantities for each parcel were interpolated using each available image where the parcel was visible and met the filtering threshold values. The number of interpolated dates to use in the process was found to be 16 between period of 1st May 2017 and 1st September 2017. The statistical quantities to calculate were found to be bandwise median and vegetation indices NDVI and GNDVI. Method used over-sampling to balance the imbalanced crop class distribution.

The method that was developed was tested with four classification three classification methods SVM, RF and MLP for whole Finland and SVM, RF, MLP and ConvRNN for a one tile in Finland. For whole Finland SVM and MLP were able to perform relatively well according to the overall accuracy of 89% that the both were able to achieve. However, because of the macro averaged f1 scores of 51% and 54% and imbalanced per class metrics, the identification performance was not equally good for all crop classes. For the smaller area of interest, ConvRNN and MLP performed best with overall accuracies of 91% and 91%. Imbalanced crop distribution affected ConvRNN method less than the other methods.

The crop identification could work in other countries as well with the sufficient training data. Apart from land parcels, the methods could be applied to identify other spatial data in a similar manner as well. With the sufficient amount of ground truth data and balanced class distribution, the classification could work well.

Further study is needed to find out whether the crop identification method created in this study would perform as good or even better with other years with different weather conditions in Finland and especially, if the method could be utilized with different agro-regions and countries as well.

References

Abadi, Martín, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng (2015). *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from tensorflow.org. URL: https://www.tensorflow.org/ (visited on 07/24/2018).

Amazon Web Services (2018). URL: https://aws.amazon.com/ (visited on 02/10/2018).

Belgiu, M. and O. Csillik (2018). "Sentinel-2 cropland mapping using pixel-based and object-based time-weighted dynamic time warping analysis". In: *Remote Sensing of Environment* 204, pp. 509–523. DOI: 10.1016/j.rse.2017.10.005.

Bergstra, James and Yoshua Bengio (2012). "Random Search for Hyper-Parameter Optimization". In: *Journal of Machine Learning Research* 13, pp. 281–305. URL: http://www.jmlr.org/papers/volume13/bergstra12a/bergstra12a.pdf (visited on 07/10/2018).

Boudewyn, P.A., D. Seemann, M.A. Wulder, and S Magnussen (2000). "The Effects of Polygon Boundary Pixels on Image Classification Accuracy". In: p. 7. URL: https://cfs.nrcan.gc.ca/publications?id=5509.

Buitinck, Lars, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux (2013). "API design for machine learning software: experiences from the scikit-learn project". In: *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pp. 108–122. DOI: arXiv:1309.0238.

Camps-Valls, G. and L. Bruzzone (2009). *Kernel Methods for Remote Sensing Data Analysis*, pp. 1–403. DOI: 10.1002/9780470748992.

CNES (2018). *MAJA*. URL: https://logiciels.cnes.fr/en/content/maja (visited on 07/12/2018).

Duro, D. C., S. E. Franklin, and M. G. Dubé (2012). "A comparison of pixel-based and object-based image analysis with selected machine learning algorithms for the classification of agricultural landscapes using SPOT-5 HRG imagery". In: *Remote Sensing of Environment* 118, pp. 259–272. DOI: 10.1016/j.rse.2011.11.020.

ESA (Apr. 2017). *Pilot production of Sentinel-2 L2A products over Europe*. URL: https://sentinels.copernicus.eu/web/sentinel/news/-/asset_publisher/xR9e/content/pilot-production-of-sentinel-2-l2a-products-over-europe (visited on 07/27/2018).

ESA (2018a). *Sen2Cor*. URL: http://step.esa.int/main/third-party-plugins-2/sen2cor/ (visited on 03/13/2018).

ESA (2018b). *Sentinel-2 MSI Products and Algorithms*. URL: https://earth.esa.int/web/sentinel/technical-guides/sentinel-2-msi/products-algorithms (visited on 05/23/2018).

ESA (Jan. 2018c). *Sentinel-2B L2A pilot products available from 16 January 2018*. URL: https://earth.esa.int/web/sentinel/missions/sentinel-2/news/-/article/sentinel-2b-l2a-pilot-products-available-from-16-january-2018 (visited on 04/23/2018).

European Comission (2017a). *Agriculture in the European Union and the Member States - Statistical factsheets*. URL: https://ec.europa.eu/agriculture/sites/agriculture/files/statistics/factsheets/pdf/eu_en.pdf (visited on 04/24/2018).

European Comission (2017b). *CAP at a glance*. URL: https://ec.europa.eu/agriculture/cap-overview_en (visited on 04/24/2018).

European Comission (2014). *COMMISSION IMPLEMENTING REGULATION (EU) No 809/2014*. URL: http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=OJ:L:2014:227:TOC (visited on 01/18/2018).

European Comission (2017c). *Direct support*. URL: https://ec.europa.eu/agriculture/direct-support/iacs_en (visited on 04/24/2018).

European Comission (2018). *Glossary of terms related to the Common Agricultural Policy*. URL: https://ec.europa.eu/agriculture/glossary/producer-organisation_en (visited on 05/23/2018).

Finnish Meteorological Institute (2018a). *Climate elements*. URL: http://en.ilmatieteenlaitos.fi/climate-elements (visited on 07/26/2018).

Finnish Meteorological Institute (2018b). *Clouds are divided in 10 groups*. URL: http://ilmatieteenlaitos.fi/pilvikuvasto (visited on 07/26/2018).

Finnish Meteorological Institute (2018c). *Thermal growing season*. URL: http://ilmatieteenlaitos.fi/terminen-kasvukausi (visited on 03/20/2018).

Finnish Meteorological Institute (2017). *Thermal growing season 2017*. URL: http://ilmatieteenlaitos.fi/kasvukausi-2017 (visited on 03/20/2018).

Finnish Meteorological Institute (2018d). *Yerly statistics*. URL: http://ilmatieteenlaitos.fi/vuositilastot (visited on 07/26/2018).

Finnish Parliament (2017). *Valiokunnan lausunto MmVL 25/2017 vp- E 79/2017 vp*. URL: https://www.eduskunta.fi/FI/vaski/Lausunto/Sivut/MmVL_25+2017.aspx (visited on 01/18/2018).

Foerster, S., K. Kaden, M. Foerster, and S. Itzerott (2012). "Crop type mapping using spectral-temporal profiles and phenological information". In: *Computers and Electronics in Agriculture* 89, pp. 30–40. DOI: 10.1016/j.compag.2012.07.015.

Frampton, William James, Jadunandan Dash, Gary Watmough, and Edward James Milton (2013). "Evaluating the capabilities of Sentinel-2 for quantitative estimation of biophysical variables in vegetation". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 82, pp. 83–92. DOI: 10.1016/j.isprsjprs.2013.04.007.

Galbraith, A., J. Theiler, and S. Bender (2003). "Resampling Methods for the MTI Coregistration Product". In: *Proceedings of SPIE - The International Society for Optical Engineering*. Vol. 5093, pp. 283–293. DOI: 10.1117/12.485985.

Gardner, M.W and S.R Dorling (1998). "Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences". In: *Atmospheric Environment* 32.14, pp. 2627 –2636. ISSN: 1352-2310. DOI: 10.1016/S1352-2310(97)00447-0.

Gatti, A. and C. Naud (2017). *Sentinel-2 Products Specification Document*. URL: https://sentinel.esa.int/documents/247904/685211/Sentinel-2+Products+Specification+Document+%28PSD%29/0f7bedeb-9fbb-4b60-91aa-809162de456c; (visited on 01/18/2018).

GISAT s.r.o (Jan. 2017). *The Czech Agri national demonstration (CzechAgri) Final report*. Université Catholique de Louvain, GISAT s.r.o. URL: http://www.esa-sen2agri.org/wp-content/uploads/docs/CzechAgri%20Final%20Report%201.2.pdf (visited on 04/19/2018).

Haboudane, D., J.R. Miller, N. Tremblay, P.J. Zarco-Tejada, and L. Dextraze (2002). "Integrated narrow-band vegetation indices for prediction of crop chlorophyll content for application to precision agriculture". In: *Remote Sensing of Environment* 81.2-3, pp. 416–426. DOI: 10.1016/S0034-4257(02)00018-4.

Halko, Nathan, Per-Gunnar Martinsson, and Joel A. Tropp (2009). "Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions". In: *SIAM Rev., Survey and Review section* 53.2, pp. 217–288. DOI: arXiv:0909.4061.

Haralick, R.M., F. Caspall, and D.S. Simonett (1970). "Using radar imagery for crop discrimination: a statistical and conditional probability study". In: *Remote Sensing of Environment* 1.2, pp. 131–142. DOI: 10.1016/S0034-4257(70)80015-3.

Hart, K., D. Mottershead, H. Tucker, E. Underwood, A. Maréchal, L. Menet, I.Martin, C. Dayde, C. Bresson, E. Deniel, J. Sanders, N. Röber, B. Osterburg, and S. Klages (Nov. 2017). *Evaluation study of the payment for agricultural practices beneficial for the climate and the environment.* Thünen Institute. URL: https://ec.europa.eu/agriculture/sites/agriculture/files/fullrep_en.pdf (visited on 05/23/2018).

Head, Tim, MechCoder, Gilles Louppe, Iaroslav Shcherbatyi, fcharras, Vinícius Zé, and Alexander Fabisch (2018). "Scikit-Optimize". In: DOI: 10.5281/zenodo.1207017.

Honkavaara, E., J. Kaivosoja, P. Litkey, R. Näsi, and M. Pandzic (June 2017). *Novel enabling IT technologies boosting efficient utilization of open satellite data in precision agriculture (BoostSat) ESRIN/Contract No. 4000117401/16/I-NB.* URL: selvita.saatavuus.com (visited on 01/23/2018).

Huete, A., K. Didan, T. Miura, E.P. Rodriguez, X. Gao, and L.G. Ferreira (2002). "Overview of the radiometric and biophysical performance of the MODIS vegetation indices". In: *Remote Sensing of Environment* 83.1-2, pp. 195–213. DOI: 10.1016/S0034-4257(02)00096-2.

Hutson, Matthew (2017). "AI Glossary: Artificial intelligence, in so many words". In: *Science* 357, p. 19. DOI: 10.1126/science.357.6346.19.

Ienco, D., R. Gaetano, C. Dupaquier, and P. Maurel (2017). "Land Cover Classification via Multitemporal Spatial Data by Deep Recurrent Neural Networks". In: *IEEE Geoscience and Remote Sensing Letters* 14.10, pp. 1685–1689. DOI: 10.1109/LGRS.2017.2728698.

Iizumi, Toshichika and Navin Ramankutty (2015). "How do weather and climate influence cropping area and intensity?" In: *Global Food Security* 4, pp. 46 –50. ISSN: 2211-9124. DOI: https://doi.org/10.1016/j.gfs.2014.11.003.

Jacob, Cohen (1960). "A coefficient of agreeement for nominal scales". In: *Educational and Psychological Measurement* 20.1, pp. 37–46. DOI: 10.1177/001316446002000104.

Jolliffe, I.T. (2002). *Principal Component Analysis*. 2nd. New York: Springer. DOI: 10.1007/b98835.

Kingma, Diederik and Jimmy Ba (Dec. 2014). "Adam: A Method for Stochastic Optimization". In: p. 15. URL: https://arxiv.org/pdf/1412.6980.pdf.

Lemaître, Guillaume, Fernando Nogueira, and Christos K. Aridas (2017). "Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning". In: *Journal of Machine Learning Research* 18.17, pp. 1–5. URL: http://jmlr.org/papers/v18/16-365.

Liang, L., Z. Qin, S. Zhao, L. Di, C. Zhang, M. Deng, H. Lin, L. Zhang, L. Wang, and Z. Liu (2016). "Estimating crop chlorophyll content with hyperspectral vegetation indices and the hybrid inversion method". In: *International Journal of Remote Sensing* 37.13, pp. 2923–2949. DOI: 10.1080/01431161.2016.1186850.

Lussem, U., C. Hütt, and G. Waldhoff (2016). "Combined analysis of Sentinel-1 and RapidEye data for improved crop type classification: An early season approach for rapeseed and cereals". In: *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*. Vol. 41, pp. 959–963. DOI: 10.5194/isprsarchives-XLI-B8-959-2016.

Maaseutuvirasto (2018a). *Field monitoring*. URL: http://www.mavi.fi/fi/maksut-ja-valvonta/valvonta/viljelija/peltovalvonta/Sivut/peltovalvonta.aspx (visited on 01/18/2018).

Maaseutuvirasto (2018b). *Why are farmer and livestock subsidies controlled?* URL: http://www.mavi.fi/fi/maksut-ja-valvonta/valvonta/viljelija/Sivut/miksi-tukia-valvotaan.aspx (visited on 01/18/2018).

Marir, B., M. Kalla, F. Douak, and A. Daamouche (2018). "A Modular Support Vector Machine for Active Learning of Urban Remote Sensing Images Classification in Algeria". In: *Journal of the Indian Society of Remote Sensing* 46.4, pp. 515–529. DOI: 10.1007/s12524-017-0719-1.

Mountrakis, Giorgos, Jungho Im, and Caesar Ogole (2011). "Support vector machines in remote sensing: A review". In: *ISPRS Journal of Photogrammetry and Remote*

Sensing 66.3, pp. 247 –259. ISSN: 0924-2716. DOI: https://doi.org/10.1016/j.isprsjprs.2010.11.001.

Murphy, S., I. Housman, and D. Gennadii (2017). *Cloud masking with Sentinel 2*. URL: https://github.com/samsammurphy/cloud-masking-sentinel2/blob/master/cloud-masking-sentinel2.ipynb (visited on 04/10/2018).

Nair, V. and G.E. Hinton (2010). "Rectified linear units improve Restricted Boltzmann machines". In: pp. 807–814. URL: https://www.cs.toronto.edu/~hinton/absps/reluICML.pdf (visited on 07/26/2018).

NASA (June 1986). *Earth Observing System, Data and Information System, Data Panel Report, Vol. IIa, Technical Memorandum 87777*. NASA. URL: https://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/19860021622.pdf (visited on 05/23/2018).

NASA (2013). *Landsat 7 Science Data Users Handbook*. URL: https://landsat.gsfc.nasa.gov/wp-content/uploads/2016/08/Landsat7_Handbook.pdf.

National Geospatial-Intelligence Agency (2012). *Military Grid Reference System (MGRS) 1km Polygon Shapefile (.shp): Downloads*. URL: http://earth-info.nga.mil/GandG/coordsys/grids/mgrs_1km_polygon_dloads.html (visited on 02/04/2018).

National Land Survey of Finland (2018). *Maps and spatial data*. URL: https://www.maanmittauslaitos.fi/en/maps-and-spatial-data (visited on 07/24/2018).

NGA (2006). *DMA Technical Manual 8358.1*. URL: http://earth-info.nga.mil/GandG/publications/tm8358.1/tr83581b.html#ZZ26 (visited on 05/24/2018).

Niculescu, S., D. Ienco, and J. Hanganu (2018). "Application of deep learning of multi-temporal Sentinel-1 images for the classification of coastal vegetation zone of the danube delta". In: vol. 42. 3, pp. 1311–1318. DOI: 10.5194/isprs-archives-XLII-3-1311-2018.

Otterbach, Johannes (2016). *Principal Component Analysis (PCA) for Feature Selection and some of its Pitfalls*. Blog. URL: http://jotterbach.github.io/2016/03/24/Principal_Component_Analysis/ (visited on 06/23/2018).

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay (2011). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12,

pp. 2825–2830. URL: http://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf (visited on 02/02/2018).

Peña, J. M., P. A. Gutiérrez, C. Hervás-Martínez, J. Six, R. E. Plant, and F. López-Granados (2014). "Object-based image classification of summer crops with machine learning methods". In: *Remote Sensing* 6.6, pp. 5019–5041. DOI: 10.3390/rs6065019.

Ray, Sunil (2017). *Understanding Support Vector Machine algorithm from examples*. URL: https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/ (visited on 07/26/2018).

Roy, David P., Jian Li, Hankui K. Zhang, and Lin Yan (2016). "Best practices for the reprojection and resampling of Sentinel-2 Multi Spectral Instrument Level 1C data". In: *Remote Sensing Letters* 7.11, pp. 1023–1032. DOI: 10.1080/2150704X.2016.1212419.

Rußwurm, M. and M. Korner (2018). "Multi-temporal land cover classification with sequential recurrent encoders". In: *ISPRS International Journal of Geo-Information* 7.4. DOI: 10.3390/ijgi7040129.

Rußwurm, M. and M. Korner (2017). "Temporal Vegetation Modelling Using Long Short-Term Memory Networks for Crop Identification from Medium-Resolution Multi-spectral Satellite Images". In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops* 2017-July, pp. 1496–1504. DOI: 10.1109/CVPRW.2017.193.

Saarinen, N., J.C. White, M.A. Wulder, A. Kangas, S. Tuominen, V. Kankare, M. Holopainen, J. Hyyppä, and M. Vastaranta (2018). "Landsat archive holdings for Finland: opportunities for forest monitoring". In: *Silva Fennica* 52.3. DOI: 10.14214/sf.9986.

Schaepman-Strub, G., M. E. Schaepman, T. H. Painter, S. Dangel, and J. V. Martonchik (2006). "Reflectance quantities in optical remote sensing — definitions and case studies". In: *Remote Sensing of Environment* 13, pp. 27–42. DOI: 10.1016/j.rse.2006.03.002.

Schmedtmann, J. and M. L. Campagnolo (2015). "Reliable crop identification with satellite imagery in the context of Common Agriculture Policy subsidy control". In: *Remote Sensing* 7.7, pp. 9325–9346. DOI: 10.3390/rs70709325.

Sinergise (2018). *Sentinel on AWS*. URL: http://sentinel-pds.s3-website.eu-central-1.amazonaws.com/ (visited on 07/10/2018).

Sollich, P. (2002). "Bayesian methods for support vector machines: Evidence and predictive class probabilities". In: *Machine Learning* 46.1-3, pp. 21–52. DOI: 10.1023/A:1012489924661.

Sonobe, R., Y. Yamaya, H. Tani, X. Wang, N. Kobayashi, and K. I Mochizuki (2017). "Assessing the suitability of data from Sentinel-1A and 2A for crop classification". In: *GIScience and Remote Sensing* 54.6, pp. 918–938. DOI: 10.1080/15481603.2017.1351149.

Statistics Finland (2017). *Map datasets*. URL: https://www.stat.fi/tup/karttaaineistot/index.html (visited on 07/26/2018).

SUHET (2015). *Sentinel-2 User Handbook*. URL: https://sentinel.esa.int/documents/247904/685211/Sentinel-2_User_Handbook (visited on 05/23/2018).

Sun, X., X. Lin, S. Shen, and Z. Hu (2017). "High-resolution remote sensing data classification over urban areas using random forest ensemble and fully connected conditional random field". In: *ISPRS International Journal of Geo-Information* 6.8. DOI: 10.3390/ijgi6080245.

Zhu, Z., S. Wang, and C. E. Woodcock (2015). "Improvement and expansion of the Fmask algorithm: Cloud, cloud shadow, and snow detection for Landsats 4-7, 8, and Sentinel 2 images". In: *Remote Sensing of Environment* 159, pp. 269–277. DOI: 10.1016/j.rse.2014.12.014.

Zupanc, Anze (2017). *Improving Cloud Detection with Machine Learning*. URL: https://medium.com/sentinel-hub/improving-cloud-detection-with-machine-learning-c09dc5d7cf13 (visited on 03/13/2018).

A Crops in crop classes

Table A1: Crops that belong to the crop classes mentioned in Table 5.

Crop class	Crop name (fin)
BEA	Härkäpapu
PEA	Tarhaherne
PEA	Ruokaherne
PEA	Rehuherne
BEE	Sokerijuurikas, sokerintuotantoon
BEE	Punajuurikas ja keltajuurikas
FAL	Sänkikesanto
FAL	Avokesanto
SRA	Kevätrapsi
SCE	Vihantavilja (kaura)
SCE	Seoskasvusto (viljat)
SCE	Kevätruis
SCE	Vihantavilja (viljaseos)
SCE	Mallasohra
SCE	Vihantavilja (ohra)
SCE	Kevätruisvehnä
SCE	Vihantavilja (vehnä)
SCE	Kevätvehnä
SCE	Kaura
SCE	Rehuohra
GRA	Suojavyöhykenurmi (sopimukset ennen 2015)
GRA	Suojakaista
GRA	Apilan siemen, valvottu tuotanto
GRA	Ympäristösopimusala, pysyvä nurmi
GRA	Monivuotinen siemennurmi, yksilajinen
GRA	Pysyvä laidunnurmi (väh 5, alle 10 v)
GRA	Monivuotinen ympäristönurmi
GRA	1-vuotiset laidunnurmet
GRA	Monimuotoisuuspelto, maisema
GRA	1-vuot. kuivaheinä-, säilörehu-, tuorerehunurmet
GRA	Monivuotiset laidunnurmet
GRA	Apila
GRA	Monimuotoisuuspelto, niitty 1. ja 2. vuosi
GRA	Ruokohelpi (kuivike/rehu)
GRA	Pysyvä kuivah.,säilör., tuorer. (väh 5, alle10 v)
GRA	Viherkesanto (nurmi ja niitty)
GRA	Monivuotiset siemennurmet

Continued on next page

Table A1: Crops that belong to the crop classes mentioned in Table 5.

Crop class	Crop name (fin)
GRA	Luonnonlaidun ja -niitty
GRA	Ruokohelpi (energia)
GRA	Viherlannoitusnurmi (ei ympäristösitoumusta)
GRA	Luonnonhoitopelto (nurmikasvit, väh. 2 v.)
GRA	Moniv. kuivaheinä-, säilörehu- ja tuorerehunurmet
GRA	Suojavyöhyke (sitoumus alkaen 2015)
GRA	Viherlannoitusnurmi
POT	Ruokaperuna
POT	Siemenperuna (sertifioidun siemenen tuotantoon)
POT	Tärkkelysperunan oma siemenlisäys
POT	Varhaisperuna (katteenalainen)
POT	Ruokateollisuusperuna
POT	Tärkkelysperuna
TRA	Kevätrypsi
WCE	Syysvehnä
WCE	Syysspelttivehnä
WCE	Syysruis