University of Zurich^UZH

# Multitemporal land cover classification using synthetic optical satellite data in Switzerland

GEO 511 Master's Thesis

**Author**
Sebastian Hafner
14-719-215

**Supervised by**
Dr. Hendrik Wulf
Dr. Charlotte Steinmeier (charlotte.steinmeier@wsl.ch)

**Faculty representative**
Prof. Dr. Michael Schaepman

30.09.2019
Department of Geography, University of Zurich

GEO 511
Master Thesis
September 30, 2019

# Multitemporal land cover classification using synthetic optical satellite data in Switzerland

Sebastian HAFNER
14-719-215

Supervised by:
Dr. Hendrik WULF
Dr. Charlotte STEINMEIER

Faculty representative:
Prof. Dr. Michael SCHAEPMAN

Remote Sensing Laboratories
Department of Geography
University of Zurich

# Abstract

Multiple applications in land management and environmental monitoring require spatial information on land cover. The consistent mapping of land cover at various spatial scales is facilitated by remote sensing data. Previous studies successfully classified Landsat time series using conventional machine learning algorithms such as Random Forest (RF) or Support Vector Machine (SVM). In recent years, however, deep learning algorithms have gained ground in remote sensing coinciding with an unprecedented increase in freely available multispectral imagery from Sentinel-2A/B. To explore the added values of these new opportunities in land cover mapping, we compared performances of RF and SVM to a Deep Neural Network (DNN) under three input datasets: (1) an annual composite, (2) coefficients of an intra-annual time series model and (3) a spatial variation on input 2. We also investigated per-class separability and adequate training dataset size. Furthermore, the effect of augmenting Landsat time series with Sentinel-2 on classification results were explored. These case studies were run for six land cover classes within three equally sized study areas (345 km$^2$) contrasting in terrain across Switzerland. Averaged over the three input datasets, considerably higher Mean Accuracies (MAs) were recorded for RF (82.2 %) and SVM (79.3 %) as compared to DNNs (71.1 %). This difference can be largely attributed to the conventional machine learning algorithms' resistance to overfitting. The inclusion of temporal and spatial information increased classifier performances by 6.0 % and 2.1 % in MA, respectively (compared to the composite input). Moreover, all land cover classes were well separable from one another (user's and producer's accuracies > 77.0 %). Classification performances increased logarithmically with the number of training samples per class (1 to 1,000), levelling at about 1,000. Finally, augmenting Landsat time series with Sentinel-2 yielded a marginal improvement in classification results. This finding indicates that more sophisticated time-series modeling approaches are required to fully exploit the existing wealth in data for land cover mapping.

# Contents

**Abstract** III

**Abbreviations** VII

**List of Figures** IX

**List of Tables** XI

**1 Introduction** 1

**2 Study area** 5

**3 Data** 7

3.1 Land cover data 7

3.2 Satellite data 8

**4 Methods** 11

4.1 Overview 11

4.2 Composite input 12

4.3 Time series input 12

4.3.1 Preprocessing satellite data 12

4.3.2 Temporal feature extraction 13

4.4 Preprocessing land cover data and labeling inputs 15

4.5 Classifier selection and parameterization 16

4.6 Classification accuracy assessment 18

4.7 Experimental setup 19

**5 Results** 21

5.1 Time series method 21

5.1.1 Qualitative assessment 22

5.1.2 Quantitative assessment 25

5.2 Reference data 28

5.3 Land cover classification 30

5.3.1 Classifier and input comparison 30

5.3.2 Per-class separability 35

5.3.3 Training dataset size 36

5.3.4 Time series augmentation . . . . . . . . . . 37

**6 Discussion** **41**

6.1 The benefits of adding temporal and spatial information . . . . . . . . . . 41

6.1.1 Temporal information . . . . . . . . . . 41

6.1.2 Spatial information . . . . . . . . . . 42

6.2 Classifier comparison . . . . . . . . . . 43

6.3 Per-class separability . . . . . . . . . . 45

6.4 Trade-off between training dataset size and classification accuracy . . . . . . . . . . 46

6.5 The benefits of denser time series . . . . . . . . . . 47

**7 Conclusion** **49**

**Bibliography** **53**

**Appendices** **59**

A.1 Figures . . . . . . . . . . 59

A.2 Tables . . . . . . . . . . 62

A.3 Scripts . . . . . . . . . . 64

**Acknowledgements** **65**

**Personal Declaration** **67**

# Abbreviations

| | |
|---|---|
| AA | Average Accuracy |
| ANN | Artificial Neural Network |
| CRS | Coordinate Reference System |
| DNN | Dense Neural Network |
| DSM | Digital Surface Model |
| EO | Earth Observation |
| ESA | European Space Agency |
| ETM+ | Enhanced Thematic Mapper Plus |
| GEE | Google Earth Engine |
| LaSRC | Land Surface Reflectance Code |
| LEDAPS | Landsat Ecosystem Disturbance Adaptive Processing System |
| MA | Mean Accuracy |
| NASA | National Aeronautics and Space Administration |
| NDSI | Normalized Difference Snow Index |
| NIR | Near Infrared |
| NOLC04 | Swiss land use statistics′ nomenclature of land cover |
| RF | Random Forest |
| ROI | Region of Interest |
| SAR | Synthetic Aperture Radar |
| SVC | C-Support Vector Classification |
| SVM | Support Vector Machine |
| SWIR | Shortwave Infrared |
| USGS | U.S. Geological Survey |
| TOA | Top-Of-Atmosphere |
| UA | User′s Accuracy |
| PA | Producer′s Accuracy |
| OA | Overall Accuracy |
| RMSE | Root Mean Squared Error |

# List of Figures

2.1 Overview of Regions of Interest . . . 5

3.1 Land cover of the Arealstatisitk 2004/09. . . . 7
3.2 Temporal availability of multispectral satellite imagery. . . . 9

4.1 Overview of workflow. . . . 11
4.2 Preprocessing of satellite observations. . . . 13
4.3 Estimated fit of time series model to preprocessed satellite observations. . . . 14
4.4 Architecture of Deep Neural Network. . . . 18

5.1 Number of clear observations and model complexity for the Arealstatistik 2004/09. 22
5.2 Qualitative assessment of time series model in the lowland region. . . . 23
5.3 Qualitative assessment of time series model in the pre-alpine region. . . . 24
5.4 Qualitative assessment of time series model in the alpine region. . . . 25
5.5 Quantitative assessment of time series model (error scatter plots). . . . 26
5.6 Quantitative assessment of time series model (error time series). . . . 27
5.7 Summary statistics of the land cover reference data for the Arealstatistik 2004/09. . . 28
5.8 Class-wise medians of the time series coefficients for the Arealstatistik 2004/09. . . . 29
5.9 Reference data quality. . . . 30
5.10 Classifier and input comparison. . . . 31
5.11 Land cover maps from different classifiers with the time series-spatial input for the Arealstatisitk 2004/09. . . . 34
5.12 Relationship between number of training samples and class-wise accuracies for Random Forest. . . . 36
5.13 Number of clear observations and model complexity for the Arealstatistik 2013/18. 37
5.14 Land cover maps for the Arealstatisitk 2013/18. . . . 38
5.15 Class-wise accuracy comparison between the Arealstatistik 2004/09 and Arealstatistik 2013/18. . . . 39

A.1 Summary statistics of the land cover reference data for the Arealstatistik 2013/18. . . 59
A.2 Class-wise medians of the time series coefficients for the Arealstatistik 2013/18. . . . 59
A.3 Land cover maps from different classifiers with the composite input for the Arealstatisitk 2004/09. . . . 60
A.4 Land cover maps from different classifiers with the time series input for the Arealstatisitk 2004/09. . . . 61

A.5 Relationship between number of training samples and class-wise accuracies for Support Vector Machine. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 62

# List of Tables

3.1 Data collection years in the Regions of Interest. . . . . . . . . . . . . . . . . . . . . . . . 8
3.2 Spectral band designations for Landsat 5, 7 and 8 and Sentinel-2A and 2B. . . . . . . 9

4.1 Model complexity related to clear observation count. . . . . . . . . . . . . . . . . . . . 15
4.2 Classification scheme. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 16
4.3 Interpretation of Cohen's kappa coefficient. . . . . . . . . . . . . . . . . . . . . . . . . 19

5.1 Classifier and input comparison. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 32
5.2 Confusion matrix for Random Forest with the time series-spatial input. . . . . . . . . . 35

6.1 Strengths and weaknesses of classifiers. . . . . . . . . . . . . . . . . . . . . . . . . . . 43

A.1 Confusion matrix for Support Vector Machine with the time series-spatial input. . . 62
A.2 Confusion matrix for Deep Neural Network with the time series-spatial input. . . . 62
A.3 Land cover nomenclature of principal domains and basic categories used by the Swiss land use statistics. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 63

# 1 Introduction

Land cover mapping acquires spatial information on the biophysical cover of the Earth's surface. This information is essential for many applications in land management and environmental monitoring (Khatami et al., 2016). For example, land cover maps have been successfully applied for natural hazard assessment (Lee and Pradhan, 2007), land cover change analysis (Yuan et al., 2005) and biodiversity monitoring (Duro et al., 2007). Characterizing and mapping land cover consistently over large areas is made possible by satellite-based Earth Observation (EO). Since different land cover types exhibit distinct spectral signatures, they are particularly well separable with multispectral imagery (Adams et al., 1995). Such imagery is systematically collected and archived by space agencies. The Landsat program (1972), a joint program operated by the National Aeronautics and Space Administration (NASA) in collaboration with the the U.S. Geological Survey (USGS), has acquired multispectral imagery at a 30 m spatial resolution for over four decades. Due to its open access policy, adopted in 2008 (Woodcock et al., 2008), Landsat imagery is available free off charge. Similar medium resolution (10–60 m) imagery is collected by the European Space Agency (ESA) with the Copernicus Sentinel-2 mission (2015). However, land cover mapping with Sentinel-2 is still a relatively new topic, particularly contrasted with Landsat's rich history in that field (Phiri and Morgenroth, 2017).

The combined endeavor of NASA and ESA to observe the Earth's surface has increased the availability of satellites images to an unprecedented level. Today, the combined constellations of Landsat and Sentinel-2 have the capability to provide global observations with a daily to weekly frequency, depending on the latitude (Li and Roy, 2017; Wulder et al., 2015). Dense time series of satellite observations offer new opportunities to improve the characterization of land cover types by providing intra-annual information on land cover dynamics and phenological differences (Gómez et al., 2016). However, the irregular availability of observations (i.e., irregular temporal sampling), due to meteorological phenomena such as snow or clouds, make the derivation of spectrotemporal features for land cover classification challenging. Gómez et al. (2016) distinguished between three different approaches to incorporate time series information into land cover mapping, namely statistical metrics, change metrics and pattern components. Spectrotemporal statistical metrics (e.g. average, minimum, maximum) are derived from one or more time series segments corresponding to predefined temporal periods. They inform on seasonality and phenology and have been successfully incorporated into land cover classification (Petitjean et al., 2012; Gebhardt et al., 2014; Azzari and Lobell, 2017). However, different land cover types may be represented by similar values dependent on the choice of metrics (Gómez et al., 2016), and partitioning the time series into temporal segments requires location-specific modifications (Azzari and Lobell, 2017). Change metrics (e.g. magnitude, slope, duration), on the other hand, describe the evolution of time series segments over time. Change metrics were also successfully used for land cover classification (Franklin et al., 2015), but similar to statistical metrics, they require application specific tuning (Gómez et al., 2016). Pattern components, contrary to time series metrics, attempt to capture the shape of the complete time series. Their big potential to fully incorporate the temporal dimension into land cover classification was demonstrated by Zhu and Woodcock (2014) who classified coefficients of a periodic time series model fitted to all available Landsat observations. However, up to the author's knowledge this approach has not been tested yet with the combined use of Landsat and Sentinel-2 data and, moreover, variations in data availability are not addressed, which may be detrimental in areas where large parts of the year lack observations due to persistent snow coverage (Liu et al., 2016).

While the derivation of complex spectrotemporal features from time series has gained popularity in recent years, little effort has been made to pair them with novel classification algorithms. In most cases, spectrotemporal features are in fact classified with the same conventional machine learning classifiers also used for single-date land cover classification (Gómez et al., 2016). Deep learning, a subset of machine learning using multi-layered artificial networks, has the potential to be applied in a wide array of applications in remote sensing (Zhang et al., 2016). Deep Neural Networks (DNNs), for example, already proved their worth in the classification of hyperspectral imagery (Zhu et al., 2017). Consequently, deep learning may also offer advances in the classification of complex spectrotemporal features, despite recent studies indicating that DNNs offer limited benefits in pixel-based land cover classification with Landsat imagery in comparison to conventional machine learning classifiers (Heydari and Mountrakis, 2018),

An important characteristics of machine learning is that it requires training data (supervised classification). Reference data for training and validation is critical for the accuracy of classification results (Shao and Lunetta, 2012). Acquiring reference data for land cover classification is commonly done in a tedious and time consuming interpretation of high resolution remote imagery, e.g., Google Earth imagery (Schneider, 2012; Jia et al., 2014). In Switzerland, on the other hand, land cover data is systematically collected for the entire country by the Swiss Federal Statistical Office as part of their land use/land cover inventory called Arealstatistik (Swiss Federal Statistical Office, 2016). The Arealstatistik therefore offers unique access to a plethora of high quality land cover samples (over 4 million samples per Arealstatistik update).

In addition to large volumes of data and advanced classification algorithms, a key facet of modern land cover mapping is high performance computing (Wulder et al., 2018). Google Earth Engine (GEE) is a cloud-based platform that offers high computational powers to run geospatial analysis on its vast pool of Earth science raster datasets (Gorelick et al., 2017). A variety of EO applications, including land cover mapping (Azzari and Lobell, 2017), have already successfully used GEE

(Pekel et al., 2016; Hansen et al., 2013). In this work, GEE is used as a highly valuable tool for the time-efficient processing of large volumes of satellite imagery.

In the context of modern land cover mapping, we identified promising trends that have not been adequately addressed yet. In particular, time series from the virtual Landsat-Sentinel-2 constellation and deep learning offer potential to improve existing classification methodologies. This thesis therefore explores the capability of recent land cover mapping trends in a comparative analysis. In doing this, the following five research questions will be addressed:

1. What are the benefits of additional temporal and spatial context information to land cover classification?

2. How do popular machine learning classifiers perform under varying data input scenarios?

3. Which land cover classes can be sufficiently separated from others?

4. What is an adequate number of training samples for each land cover class?

5. What is the added value of denser satellite time series on the land cover classification?

The remainder of this thesis is organized in six main chapters. Chapter 2 gives on overview of the study area. Chapter 3 then summarizes all land cover and satellite data used in this thesis, followed by a description of the applied methods in Chapter 4. Chapter 5 and Chapter 6 present the results of this thesis and their discussion, respectively. Finally, we draw a conclusion and give advice for future research in Chapter 7.

# 2 Study area

The overall study area for this thesis is Switzerland. The country is located in central Europe and has an area of 41,285 km$^2$. Within Switzerland, three Regions or Interest (ROIs) representing distinct landscapes at different elevations were selected. The ROIs are of identical shape with an east west and north south extent of 23 km and 15 km, respectively. Consequently, each of them has an area of 345 km$^2$. Google Earth images of the three ROIs as well as a Digital Surface Model (DSM) of Switzerland are shown below (Figure 2.1).

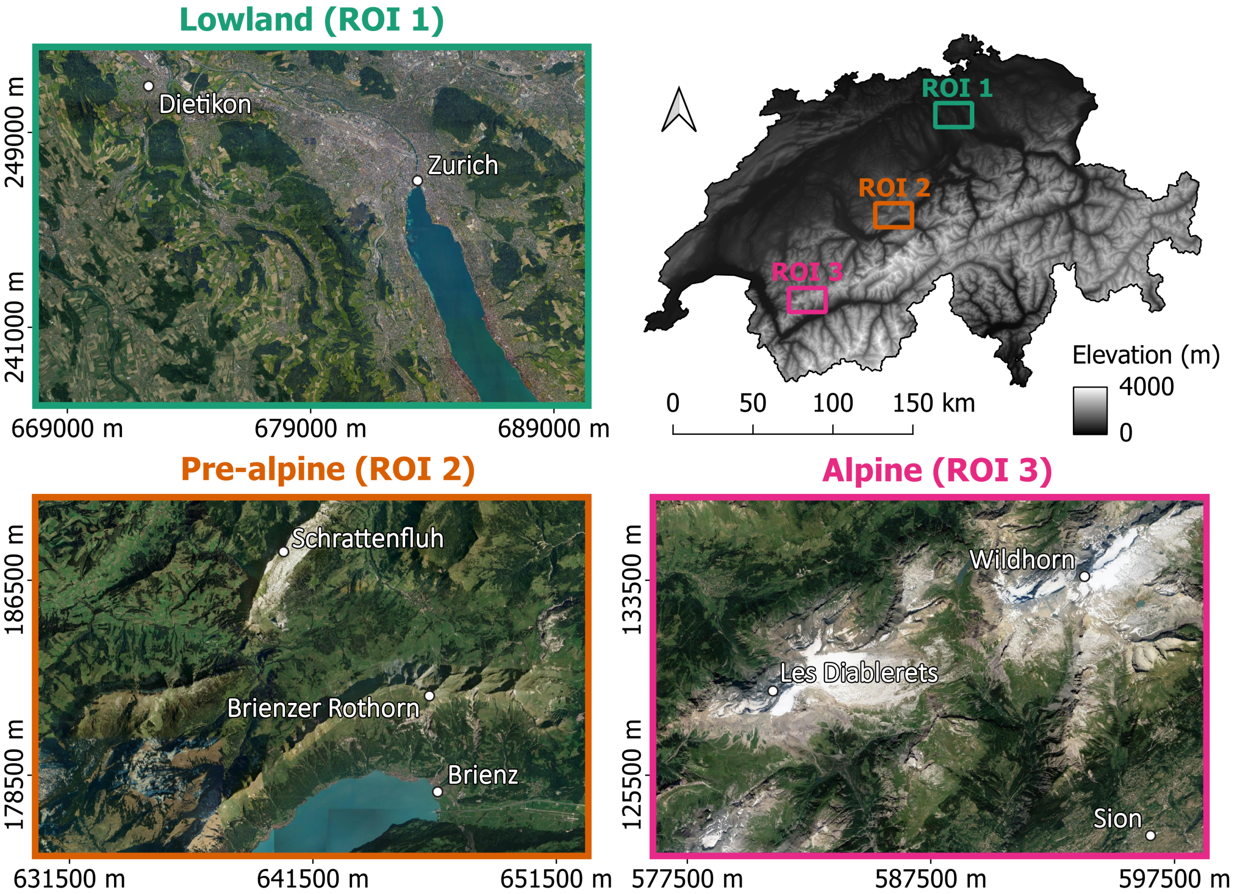

FIGURE 2.1: Digital Surface Model of Switzerland (top right) and Google Earth images of Region of Interest (ROI) 1 (top left), 2 (bottom left) and 3 (bottom right).

The DSM and the following elevation information are based on the ALOS global 30 m DSM (Tadono et al., 2014). All ROIs are displayed in the Swiss Coordinate Reference System (CRS) CH1903/LV03 (EPSG: 21781) which is used throughout this work for the processing and visualization of geospatial data. The first ROI is located in the northeast of Switzerland. With 512 m, its average elevation is the lowest among the ROIs; therefore, we will hereinafter refer to it as lowland region. In terms of landscape, it encompasses a large built-up area showing the city center of Zurich, the northern part of lake Zurich, agricultural areas and multiple forest patches. The second ROI is located in the Bernese Highlands in central Switzerland and represents a pre-alpine (hereinafter referred to as such) area due to its proximity to the alps as well as its average elevation of 1,316 m. The mostly rural landscape in the pre-alpine region encompasses widespread grasslands and forests, multiple mountains and the city of Brienz bordering lake Brienz on its eastern shore. Its highest point is the summit of the Brienzer Rothorn (2,343 m). The third ROI is located in the Swiss Alps in the southwest of Switzerland. We will hereinafter refer to it as alpine region. The landscape is located at an average elevation of 1,943 m (maximum elevation 3,238 m) and encompasses multiple mountains (e.g. Les Diablerets and Wildhorn) and the northwester part of the city Sion. It should be noted that at such elevations snow can be present throughout the year.

# 3 Data

## 3.1 Land cover data

Information about Switzerland's land cover and land use is collected by the Federal Statistical Office (Swiss Federal Statistical Office, 2016). This inventory called Arealstatistik (English: Swiss land use statistics) was initiated in 1979/85 and updated in 1992/97 and 2004/09. The most recent update for 2013/18 is currently being finished. Data for the Arealstatistik is systematically collected at the intersections of a 100 m grid across Switzerland. The over four million resulting intersections are then labeled with a land cover, land use and land use/land cover class based on the interpretation of high resolution aerial photographs, collected by the Federal Office of Topography (`swisstopo.admin.ch`), as well as in-field inspections. In this thesis, we are solely interested in the land cover product. Its most detailed classification scheme distinguishes 27 classes named basic categories. Those are aggregated into six principal domains which are visualized for the update 2004/09 in Figure 3.1. Additionally, the land cover nomenclature (NOLC04) including the basic categories is listed in the Appendices (see Table A.3).

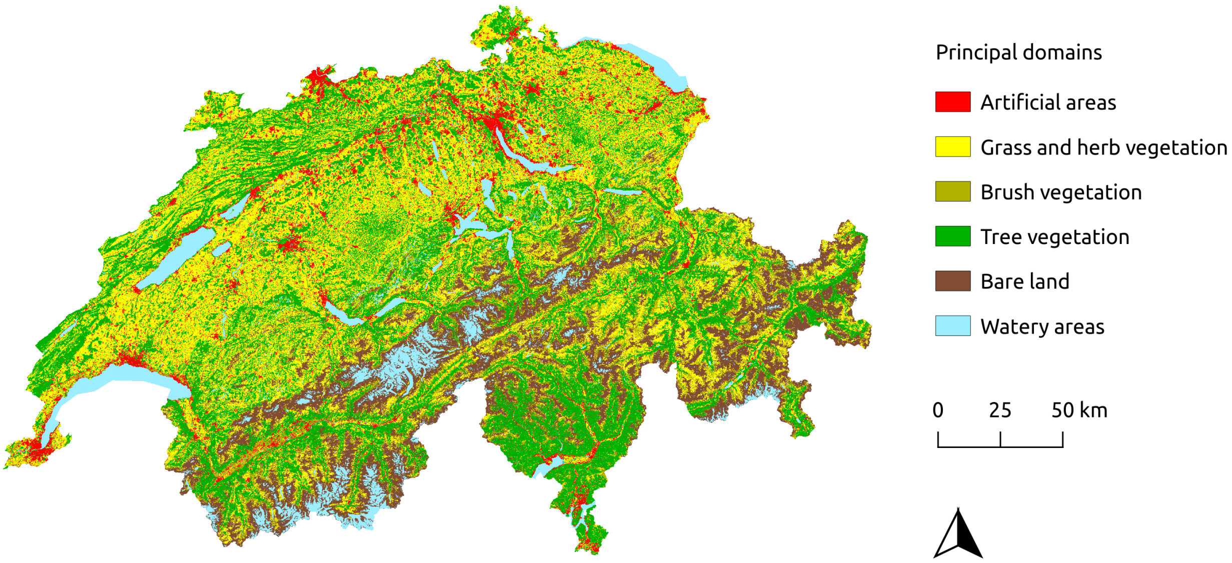

FIGURE 3.1: Arealstatistik 2004/09 - Rasterization (100 m spatial resolution) of the land cover principal domains.

Noteworthy is that the Arealstatistik does not differentiate between agricultural areas but rather maps all of them to the principal domain grass and herb vegetation. An update of the Arealstatistik is done over multiple years, whereby Switzerland is divided into sub-regions each corresponding to a year. As a result, the year in which the Arealstatistik data is collected varies between ROIs. Table 3.1 lists those data collection years for the two most recent Arealstatistik updates. All three ROIs fall withing different sub-regions. Data for the alpine region is collected at the beginning of an Arealstatistik update; in contrast, data for the pre-alpine and lowland regions two and three years later, respectively. In regard to future updates of the Arealstatistik, there is ongoing research to partially or entirely automate the time intensive labeling, since the inventory is foreseen to be updated in shorter time intervals of six years. In that context, Picterra (2017) presented a feasibility study that reviews different machine learning methods to reduce the labelling workload. They concluded that supervised deep learning is the most promising method due to the available abundance of reference data.

TABLE 3.1: Data collection years in the Regions of Interest (ROIs) for the two most recent Arealstatistik updates.

| Arealstatistik update | Lowland | Pre-alpine | Alpine |
|---|---|---|---|
| 2004/09 | 2007 | 2006 | 2004 |
| 2013/18 | 2016 | 2015 | 2013 |

## 3.2 Satellite data

The Landsat and Copernicus program collect optical EO data at a medium spatial (10–60 m) and a high temporal ($\leq$ 16 days) resolution. Landsat is a joint program of NASA in collaboration with the USGS and has been observing the Earth continuously since 1972. On the other hand, Copernicus, an EO program headed by the European Commission in partnership with ESA, has been collecting optical data since the launch of Sentinel-2A in 2015. An overview of the temporal availability of satellite data for the study area is shown in Figure 3.2.

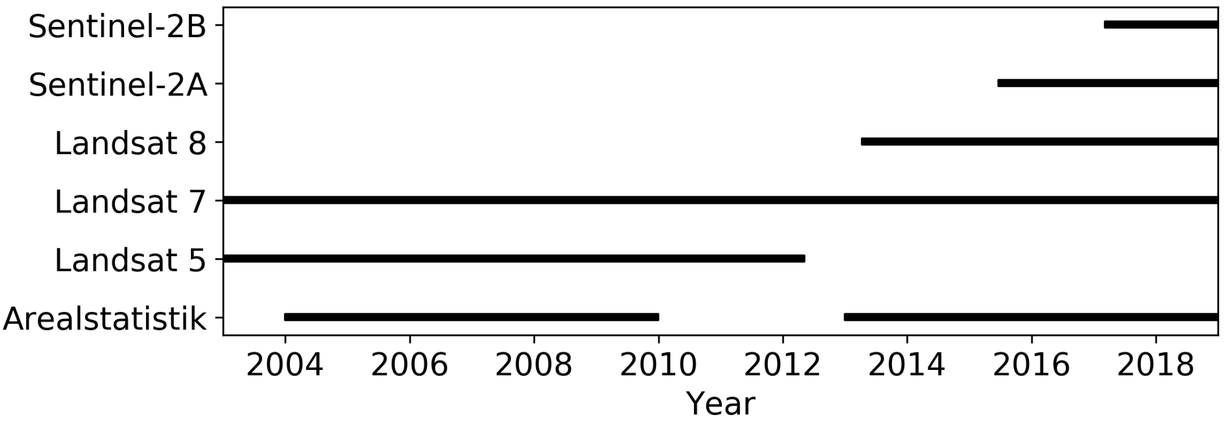

FIGURE 3.2: Temporal availability of multispectral satellite imagery from the Landsat (Landsat 5, 7 and 8) and Copernicus (Sentinel-2A and Sentinel-2B) program, as well as the temporal availability of land cover data (Arealstatistik) for Switzerland since 2003.

Satellite data from Landsat 5 and Landsat 7 are available for the Arealstatistik 2004/09. In contrast, not only Landsat data (Landsat 7 and Landsat 8) but also Copernicus data (Sentinel-2A and Sentinel-2B) are available for the most recent Arealstatistik (2013/18). All of these satellites acquire similar multispectral imagery at the wavelengths: blue, green, red, Near Infrared (NIR), Shortwave Infrared (SWIR) 1 and SWIR 2. The exact band designations and spatial resolutions for each satellite are listed in Table 3.2. Although the radiometric characteristics slightly differ between the satellites, tests regarding the combined use of Sentinel-2 and Landsat 8 data demonstrated very good correlations between corresponding spectral bands (Pearson coefficients generally higher than 0.98) (Mandanici and Bitelli, 2016). Moreover, transformation functions to minimize differences between Sentinel-2A and Landsat 8 (Zhang et al., 2018) and Landsat 7 and Landsat 8 (Roy et al., 2016) were developed. The similarity between spectral bands should be therefore sufficient for the extraction of general seasonal reflectance dynamics. It is also noteworthy that characterization requirements are always application dependent (Mandanici and Bitelli, 2016).

TABLE 3.2: Spectral band designations for Landsat 5, 7 and 8 and Sentinel-2A and 2B.

| | **Landsat** | | | | **Sentinel** | | |
|---|---|---|---|---|---|---|---|
| | Wavelength (*nm*) | | | Res. (*m*) | Wavelength (*nm*) | | Res. (*m*) |
| | Landsat 5 | Landsat 7 | Landsat 8 | | Sentinel-2A | Sentinel-2B | |
| Blue | 450-520 | 450-520 | 452-512 | 30 | 448-546 | 443-541 | 10 |
| Green | 520-600 | 520-600 | 533-590 | 30 | 538-583 | 536-582 | 10 |
| Red | 630-690 | 630-690 | 636-673 | 30 | 646-684 | 646-685 | 10 |
| NIR | 760-900 | 770-900 | 851-879 | 30 | 763-908 | 767-900 | 10 |
| SWIR 1 | 1550-1750 | 1550-1750 | 1566-1651 | 30 | 1542-1685 | 1540-1681 | 20 |
| SWIR 2 | 2080-2350 | 2090-2350 | 2107-2294 | 30 | 2081-2323 | 2067-2305 | 20 |

All of the listed satellite data is retrieved from GEE. GEE's data catalog (`developers.google.com/earth-engine/datasets/catalog/`) provides an extensive list of publicly available Earth science raster datasets. Landsat data is available in GEE at its native 30 m resolution grouped into two tiers and at three processing levels, namely at-sensor radiance (raw), Top-Of-Atmosphere (TOA) reflectance and surface reflectance. The tier of a scene denotes its data quality in terms of georegistration, radiometry, terrain precision and inter-calibration across the different Landsat sensors (see more information in the USGS docs at `landsat.usgs.gov/landsat-collections`). In this work, we are only considering Landsat scenes meeting tier 1 with the highest processing level (surface reflectance). Landsat 5 and Landsat 7 data were processed to surface reflectance using the Landsat Ecosystem Disturbance Adaptive Processing System (LEDAPS) algorithm (Schmidt et al., 2013). For Landsat 8, the processing to surface reflectance was done with the Land Surface Reflectance Code (LaSRC) (Guide, 2018). Landsat data from all three satellites was processed with CFMask, a C translation of the popular cloud and cloud shadow detection algorithm FMask (Zhu and Woodcock, 2012). Further information about CFMask was also published by Foga et al. (2017) in the context of a cloud detection comparison for Landsat data products. In terms of temporal resolution, Landsat 5, 7 and 8 all have a revisit time of 16 days, but the orbits of Landsat 7 and 8 are offset to create an eight-day revisit time between them. It should be noted that since 31 May 2003, Landsat 7 scenes are missing approximately 22 % of their normal area due to the failure of the scan line corrector (Maxwell et al., 2007). Contrary to Landsat imagery, Sentinel-2A and Sentinel-2B imagery are available in GEE only at two processing levels, namely TOA reflectance (Level-1C) and surface reflectance (Level-2A). The latter dataset, Level-2A, was added to GEE's catalog on 27 March 2019 and consists of scenes acquired not earlier than 28 March 2017. Surface reflectance was computed from TOA reflectance with Sen2Cor, an atmospheric correction processor develped by Main-Knorn et al. (2017) on behalf of ESA. In addition to the atmospheric correction, Sen2Cor ouputs a scene classification map containing cloud and snow/ice information. The TOA reflectance dataset, on the other hand, only contains information about clouds but not about snow/ice. Nevertheless, we retrieve TOA reflectance scenes until 28 March 2017 when surface reflectance scenes become available. In terms of temporal and spatial resolution, the constellation of Sentinel-2A and Sentinel-2B visits the same area every five days or less, and all six spectral bands of interest are acquired at either a 10 m (blue, green, red and NIR) or 20 m (SWIR 1 and 2) spatial resolution. Moreover, the virtual constellation of Sentinel-2 in combination with Landsat 8 provides a global revisit time of 2.9 days (Li and Roy, 2017).

# 4 Methods

## 4.1 Overview

An overview of the workflow is illustrated in Figure 4.1. We first applied two different methods to combine multiple optical satellite scenes into a single image of classifiable information. The applied composite method and time series method (including preprocessing) are presented in Section 4.2 and 4.3, respectively. The preprocessed land cover data from the Arealstatistik were then overlaid onto the images to assign a class label to the corresponding pixels and kernels, as described in Section 4.4. The former dataset was used for the pixel-based classification without spatial information and the latter dataset for the kernel-based classification with spatial information. This section is consecutively followed by the classifier selection and parameterization (Section 4.5) and the classification accuracy assessment (Section 4.6). Finally, the classifiers are trained and validated with labeled data from the composite and time series method. In this context, we set up different experiments in Section 4.7 to address the earlier raised research questions.

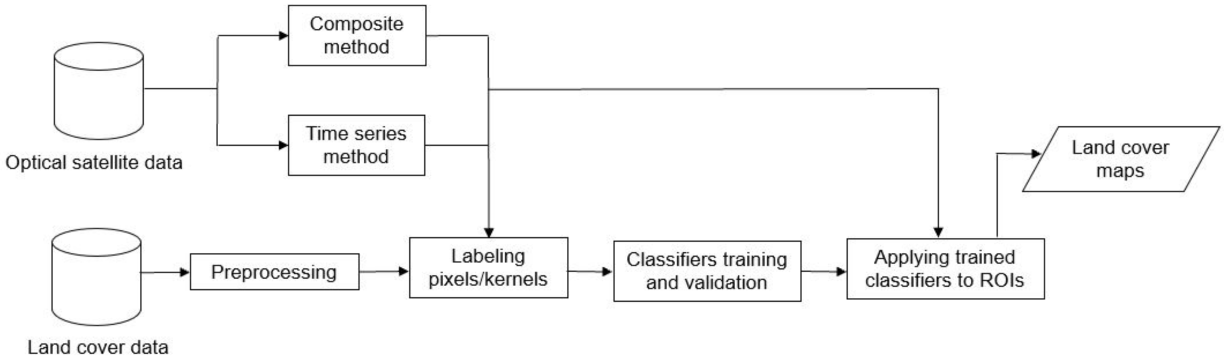

FIGURE 4.1: Overview of workflow.

## 4.2 Composite input

The objective of compositing is to generate cloud-free images by combining multiple satellite scenes according to user-defined rules (White et al., 2014). Composites have been produced for many applications in remote sensing, including land cover classification (Griffiths et al., 2013; Gómez et al., 2016). Among the various available compositing rules, computing a low percentile of the set of observations for each pixel in the image is an effective way to remove clouds and snow (Poortinga et al., 2019). This is the case because cloud and snow observations generally have high reflectance values which fall within high percentile ranks. We implemented percentile compositing in GEE by first retrieving all available Landsat 5, 7 and 8 scenes for the ROIs and temporally filtered them to the the Arealstatistik data collection years. The Landsat 5 and 7 scenes were then characterized with the Landsat 8 scenes by applying transformation functions (Roy et al., 2016). Finally, the scenes were combined by computing the 10$^{\text{th}}$ percentile of the preprocessed scenes.

## 4.3 Time series input

The time series method is a new method to obtain classifiable information from optical satellite scenes. The aim of this method is to minimize the loss in temporal information compared to annual composites. First, the preprocessing of the satellite data is presented, then the extraction of the temporal information.

### 4.3.1 Preprocessing satellite data

We first retrieved all available scenes from the aforementioned Landsat and Sentinel datasets in GEE. Sentinel TOA reflectance sences were enriched with a NDSI-based snow mask (Normalized Difference Snow Index: Salomonson and Appel, 2004). Consequently, all retrieved scenes now contain pixel-wise information about snow, clouds and cloud shadows (in addition to the six spectral bands). We then applied transformation functions to the Landsat 5 and 7 datasets (Roy et al., 2016) and the Sentinel 2 dataset (Zhang et al., 2018) to minimize the differences between corresponding spectral bands. Thereafter, the scenes were sorted by acquisition date and temporally filtered to the Arealstatistik data collection year (±1 year) for each ROI. Scenes from adjacent years were included to obtain longer time series, and consequently robuster temporal information. After temporally and spatially filtering the satellite images to the ROIs, all pixels affected by clouds, snow or ice were masked. We then identified time series outliers based on the blue spectral band. More specifically, for a time series of observations the mean and standard deviation of the blue spectral band were calculated. In a consecutive step, each observation lying outside the mean ±2 standard deviations was masked. The blue spectral band was used because its seasonal variability is generally the lowest among the six spectral bands. As a result, the remaining observations are masked for most clouds, cloud shadows, snow, ice and outliers. Henceforth, they are referred to as clear observations. The described preprocessing of satellite data is exemplified for a time series of a forest pixel in the NIR band in Figure 4.2. In this example, 34 out of the 80 observations were flagged as unclear, and consequently removed from the dataset. The remaining 46 clear observations are used for the temporal feature extraction described in the following subsection (Subsection 4.3.2).

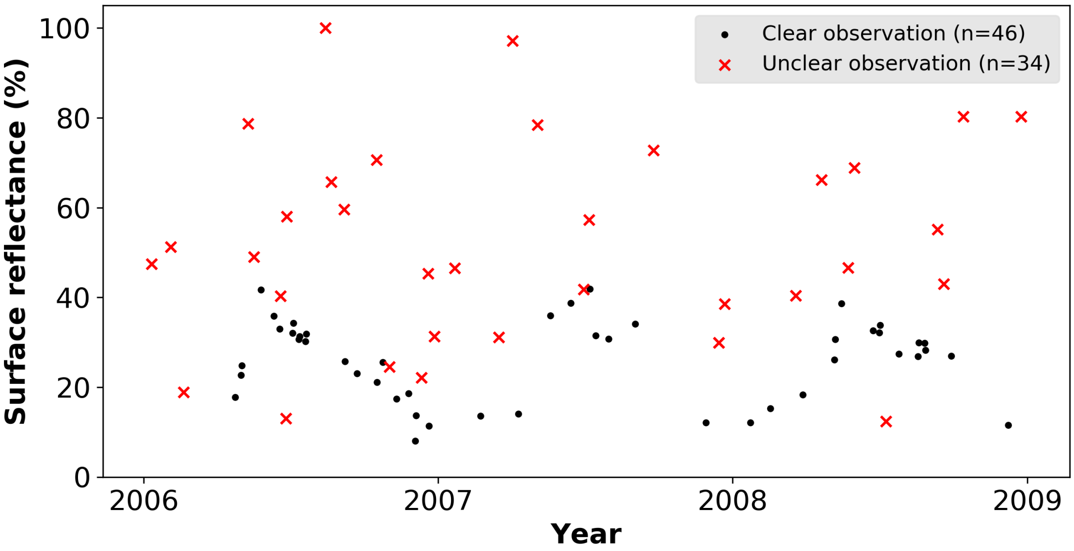

FIGURE 4.2: All available near infrared observations for a forest pixel over a three years period. Observations were flagged by the preprocessing as either clear or unclear.

### 4.3.2 Temporal feature extraction

We estimated a separate time series model for each pixel and spectral band. A time series model is composed of two parts: a constant and a harmonic (Fourier) model (Davis, 1986; Rayner, 1971). A mathematical description of the time series model is given below (Equation 4.1).

$$X_t = c + A\cos(t + p) \tag{4.1}$$

Where:

$t$: date.

$c$: coefficient for overall value.

$A, p$: coefficients for intra-annual change.

The constant ($c$) is used to estimate the overall value of the time series and is invariable over time. On the other hand, the harmonic model composed of two coefficients, namely an amplitude ($A$) and a phase shift ($p$), adds a periodic component to the time series model. A period of the harmonic model corresponds to a calendar year, and therefore is used to estimate intra-annual changes that are recurring on a yearly basis. In practice, these intra-annual surface reflectance dynamics are caused by phenology and sun angle differences (Zhu et al., 2015). Similar time series models were successfully applied in remote sensing to time series analysis by Zhu et al. (2012), Zhu and Woodcock (2014), Zhu et al. (2015), Liu et al. (2016) and Wilson et al. (2018). Most of these authors, however, applied more sophisticated models with up to three harmonic frequencies to capture not only unimodal change but also bimodal and trimodal changes. Furthermore, they added long term trends and breaks to model inter-annual differences and abrupt surface changes, respectively. For simplicity, we decided that a time series model composed of only a constant and a harmonic model with one frequency is sufficient to extract general annual surface reflectance dynamics for most land cover types. Figure 4.3 exemplifies an estimated model for a time series of clear observations of a forest pixel in the near infrared band.

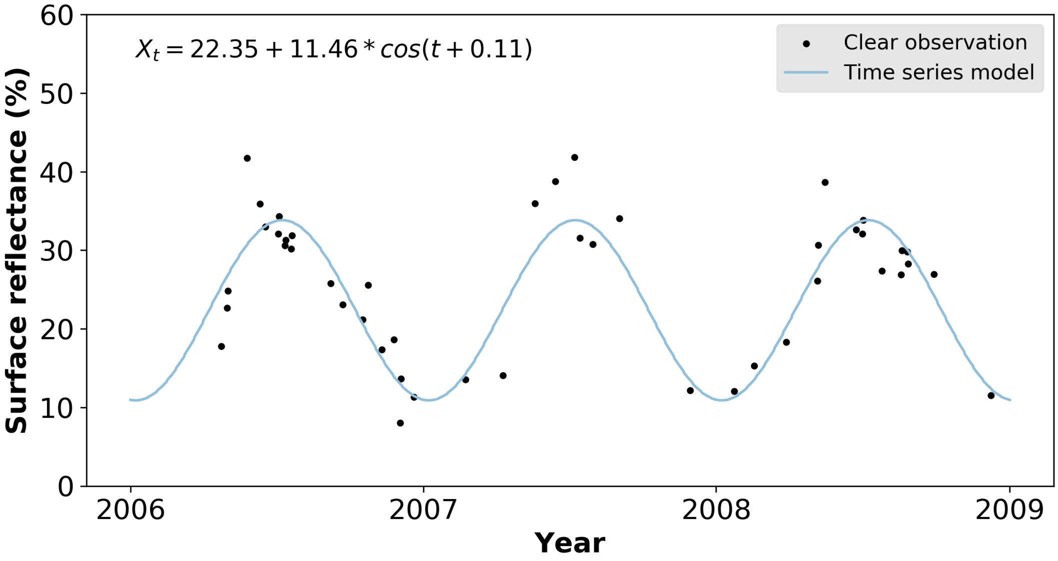

FIGURE 4.3: Clear near infrared observations for a forest pixel over a three years period and the corresponding estimated fit by the time series model.

An important characteristics of the preprocessed satellite data is that the number of clear observations varies not only across elevation (i.e., between the ROIs) due to snow, but also on a local scale (i.e., within the ROIs) due to sensor artefacts or clouds. Moreover, the number of clear observations for many locations is in fact less than desirable (Roy et al., 2010). We addressed this spatial variability by implementing a threshold for the harmonic model. If the number of clear observations in a given three-year-period is lower than or equal to 12, the overall value ($c$) of the time series model is set to the 10$^{\text{th}}$ percentile of all observations (including unclear ones) and the coefficients for intra-annual change ($A$ and $p$) are set to zero. The low percentile method ensures that if there are any clear observations available, a value close to them is set as constant because cloud or snow observations typically have high surface reflectances (see Section 4.2). Not using a harmonic model for sparse time series of clear observations also addresses overfitting which occurs when a complex model is fit too closely to a limited time series of observations. The two described scenarios are summarized in Table 4.1.

TABLE 4.1: Model complexity related to clear observation count.

| Model complexity | Clear observations count (n) |
|---|---|
| Constant + harmonic model | $n > 12$ |
| Constant (10th percentile of all observations) | $n \leq 12$ |

Postprocessing of the coefficients is divided into two steps: First, the time series model was validated based on the visible spectrum (blue, green and red) for all pixels modeled with a temporal component (i.e., with more than 12 clear observations). The model was identified as invalid if any of the visual spectrum coefficients resulted in surface reflectance predictions outside the interval [0 %, 50 %]. The complexity of invalid models was reduced to the simple model composed of only a constant, which was done for all spectral bands of the pixel. The upper bound was set because surface reflectance values in the visual spectrum exceeding 50 % are highly unlikely. Second, independent of the model complexity, negative overall values and overall values greater than 100 % were set to 0 % and 100 %, respectively.

## 4.4 Preprocessing land cover data and labeling inputs

We first adapted a new classification scheme from the Arealstatistik land cover nomenclature (NOLC04) (NOLC04: Table A.3). The principal domain watery areas was split into its basic categories water, glacier, perpetual snow, wetlands and reedy marshes. The two basic categories water and glacier, perpetual snow were each mapped to a separate class, and the latter was renamed to snow & ice. This further distinction is necessary because the spectral signatures of snow and ice differ significantly from that of water. In fact, snow is considerably brighter in the blue and green band than any other surface (Dozier, 1989). From a remote sensing point of view, it is desirable to differentiate classes of distinct spectral signatures. Therefore, we decided on a single water class containing only lakes and rivers. The basic categories wetlands and reedy marshes, on the other hand, were added to the principal domain grass and herb vegetation. Before these additions, grass and herb vegetation had already summarized a large variety of spectral signatures because all agricultural areas fall within it. This made it the most suitable principal domain for the expansion. Furthermore, we merged the expanded grass and herb vegetation class with the principal domain brush vegetation. This new class named non-forest vegetation summarizes all vegetation but forest, hence its name. The other classes in the adapted classification scheme remained equivalent to the NOLC04 principal domains (artificial areas, tree vegetation and bare land) with the only difference that tree vegetation was renamed to forest vegetation. The correspondence between the basic categories of the NOLC04 and the adaptation used in this thesis is visualized in Table 4.2.

TABLE 4.2: Adapted classification scheme and corresponding basic categories of the Swiss land use statistics nomenclature of land cover (NOLC04).

| Land cover class | NOLC04 basic categories |
|---|---|
| Artificial areas | Consolidated surfaces, Buildings, Greenhouses, Gardens with border and patch structures, Lawns, Trees in artificial areas, Mix of small structures |
| Non-forest vegetation | Grass and herb vegetation, Wetlands, Reedy marshes, Shrubs, Brush meadows, Short-stem fruit trees, Vines, Permanent garden plants and brush crops |
| Forest vegetation | Closed forest, Forest edges, Forest strips, Open forest, Brush forest, Linear woods, Clusters of trees |
| Bare land | Solid rock, Granular soil, Rocky areas |
| Water | Water |
| Snow & Ice | Glacier, perpetual snow |

Using the preprocessed land cover points, we then labeled the composite input (see Subsection 4.2) and the time series input (see Subsection 4.3). This was done by finding for each land cover point the pixel in the input image corresponding to the point's coordinates. Thereafter, values of that pixel were labeled according to the point's land cover class. The number of values corresponds to the number of features for each input. It is 6 for the composite input (one for each spectral band) and 18 for the time series input (3 coefficients per spectral band). In addition to these pixel-based datasets, we also created a 3x3 kernel-based dataset for the time series input. Its features consist not only of the values of the pixel corresponding to the point's coordinates, but also of those of the pixel adjacent to it (8-adjacency). In doing that, spatial information was added and the number of features expanded from 18 to 162 (18 for each of the 9 pixels). This enhanced time series input will be hereinafter referred to as time series-spatial input.

## 4.5 Classifier selection and parameterization

The described data inputs were tested with three popular machine learning classifiers, namely Random Forest (RF), Support Vector Machine (SVM) and Artificial Neural Network (ANN). All three of them are supervised classifiers, meaning they learn based on training data. The first classifier, RF (Breiman, 2001), has been used for a multitude of classification (and regression) tasks, including land cover classification with remotely sensed data. A review of the classifier in the context of remote sensing is given in Rodriguez-Galiano et al. (2012). RF is built of a multitude of decision trees, whereby each tree is trained on a random subset of the training data (hence the name random forest). A trained RF classifier outputs the class based on majority voting, meaning the mode of the individual decision tree predictions is used. The combination of a multitude of decision trees corrects for overfitting which single tree instances are prone to do. The effect of this is exemplified for land cover classification in Rodriguez-Galiano et al. (2012) where the RF classifier significantly outperformed a single decision tree. In terms of parameterization, we set the number of decision trees to 100 because Thanh Noi and Kappas (2018) showed that a RF with ntree = 100 produces accurate land cover classification results. We also set class weights according to their frequency in the training data to avoid overfitting the classifier to overrepresented classes. For other parameters, the default settings of the Python machine learning library `scikit-learn` were used. The documentation for their implementation of the RF classifier including default parameters can be accessed online (see `scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier`).

The second tested classifier, SVM, is based on the concept of finding a decision boundary that best separates the data into two classes (SVM: Hearst, 1998). The decision boundary is placed in space such that it maximizing the distance between itself and the nearest points from both classes (support vectors). In order to support the classification of non-linearly separable data, a kernel trick can be applied to map the data into a higher dimensional space. The data may then be linearly separable. In this thesis, we used the C-Support Vector Classification (SVC) SVM because it supports multiclass classification. The `scikit-learn` library implements a SVC based on the popular SVM library `libsvm` (Chang and Lin, 2011). The documentation is accessible under `scikit-learn.org/stable/modules/generated/sklearn.svm.SVC`. In terms of parameterization, the default parameters from `scikit-learn` were used, except for the class weights parameter (same parameterization as for RF) and the C parameter. The C parameter decides to what degree misclassifications should be allowed when separating the training data. It was increased from 1 (default setting) to 1,000 to increase the number of correctly classified samples in the training dataset. Noteworthy is also that the time complexity of training the SVC classifier scales at least quadratically with the number of training samples. Therefore, this classifier might not be the optimal choice for large datasets due to the long training time.

The third tested classifier, ANN, was originally designed to mimic the human brain, but then started to be adapted for specific regression and classification tasks for example in computer vision and speech recognition. ANNs are composed of a set of layers each containing nodes. The number of nodes for the first layer (input layer) and the last layer (output layer) must correspond to the number of features and the number of classes, respectively. In between the input and output layer, are the so called hidden layers. Networks with multiple hidden layers are also referred to as Deep Neural Networks (DNNs: LeCun et al., 2015) and fall withing the subcategory deep learning of machine learning. When an input vector is passed from layer to layer forward through a network, at each layer, the layer's activation function is element-wise applied to the vector and then multiplied with weights (linear transformation) represented by the layer's nodes. For classification tasks, the last layer applies the softmax activation function that normalizes the vector into a probability distribution. The node with the highest probability is then used as prediction (argmax). ANNs are trained with the backpropagation algorithm which modifies the weights of all nodes according to an error term computed by passing a sample forward through the network and then comparing the ANN's prediction to the sample's label (backpropagation: Werbos, 1990). To build an ANN classifier, we used `TensorFlow`'s implementation of the deep learning library keras. The network architecture is visualized in Figure 4.4. In addition to an input and output layer, the architecture is composed of three hidden layers with 400, 200 and 100 consecutive nodes. All hidden layers use the tanh activation function. In contrast, the output layer applies the softmax activation function. Prior to training our DNN, the labels were one hot encoded which is a common technique used in machine learning for categorical data. The network was trained in 120 epochs with a doubling batchsize every 20 epochs starting from 256 samples. The number of epochs defines how many times the network sees the complete training dataset during its learning phase, and the size of a batch determines how many samples the network sees before updating its weights. An increasing batchsize has the same effect as a decreasing learning rate but is computationally more efficient (Smith et al., 2017).

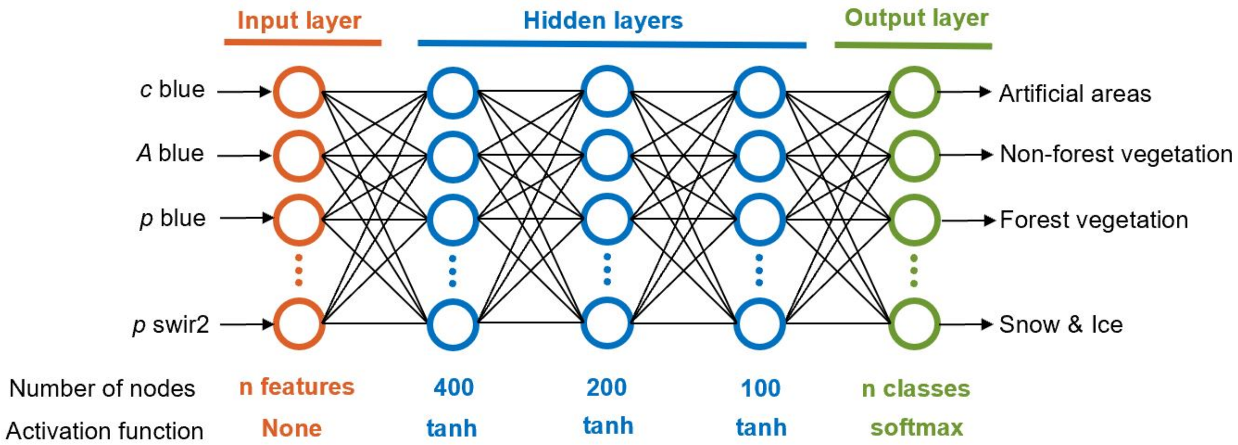

FIGURE 4.4: Architecture of the Deep Neural Network exemplified for the time series input.

## 4.6 Classification accuracy assessment

To assess the performance of a classifier in combinations with an input dataset, we used a confusion matrix, also known as error matrix (Stehman, 1997). This matrix summarizes how the classifier is confused when it makes predictions by listing class-wise count values of the classifier predictions in relation to the corresponding ground truths. User's Accuracy (UA) and Producer's Accuracy (PA) for each class are also commonly added to confusion matrices. The former (UA) denotes the probability of a classified pixel on the thematic map to be correct, and the latter (PA) denotes the percentage of pixels of a class that were correctly classified (see formulas in Box 4.2). Methods have also been developed to summarize a confusion matrix in one single accuracy metric. Overall Accuracy (OA) describes the ratio between correctly classified samples and total number of samples. It is, however, biased towards in the reference data over-represented classes. Average Accuracy (AA), on the other hand, accounts for this bias by using the mean UA of all classes, thereby giving a lot of weight to small classes. The combination of these two metrics (OA and AA) represents the Mean Accuracy (MA). Below are the Formulas for all three accuracy metrics (see Box 4.3).

$$UA_b = \frac{x_{bb}}{\sum_{a=1}^{n} x_{ba}} * 100 \qquad PA_b = \frac{x_{bb}}{\sum_{a=1}^{n} x_{ab}} * 100 \tag{4.2}$$

$$OA = \frac{\sum_{a=1}^{n} x_{aa}}{n} \qquad AA = \frac{\sum_{a=1}^{n} UA_a}{n} \qquad AM = \frac{OA + AA}{2} \tag{4.3}$$

Where:

$b$: class.

$n$: number of classes.

$x_{ij}$: value in error matrix at row i and column j.

Cohen's kappa coefficient (hereinafter referred to as Kappa) eliminates the chance agreement of OA by determining how much better a classification is than random (Cohen, 1960). Criticism questioning its usefulness for the assessment of classification maps has emerged in recent years, though (Pontius and Millones, 2011). Nevertheless, it is still widely used in remote sensing, making it an important metric when comparing classification results to other studies. In this thesis, we reported classification results in all four metrics (OA, AA, MA, Kappa) in order to ensure an extensive accuracy assessment.

TABLE 4.3: Interpretation of Cohen's kappa coefficient according to Landis and Koch (1977).

| Kappa coefficient | Strength of agreement |
|---|---|
| <0.00 | Poor |
| 0.00–0.20 | Slight |
| 0.21–0.40 | Fair |
| 0.41–0.60 | Moderate |
| 0.61–0.80 | Substantial |
| 0.81–1.00 | Almost perfect |

## 4.7 Experimental setup

In this section we describe the experiments performed to address our research questions. We first addressed the classifier comparison (RF vs. SVM vs. DNN) and the effects of adding temporal information (composite input vs. time series input) and spatial information (time series input vs. time-series-spatial input). Consequently, this experiment involved nine different classifier-input combinations. Each combination was validated with k-fold cross-validation (Kohavi, 1995). K-fold cross-validation involves randomly splitting the reference data into k equally sized sub-datasets and then training and validating the classifier k times, whereas a different validation set is used for each run (the other k-1 subsets are used for training). For all of our experiments we used 5-fold cross-validation which is a common choice for k (Kuhn and Johnson, 2013). In 5-fold cross-validation the training and validation subsets comprise about 80 % and 20 % of the reference data, respectively. The classifier-input combinations were then compared in terms of OA, AA, MA and Kappa. In this experiment we also produced land cover maps with all three classifiers trained on the complete time series-spatial input to compare their performances visually. Moreover, we rasterized the reference data (derived from the Arealstatistik 2004/09) to create a ground truth map. Importantly, this map has a coarser resolution than the classification maps (100 m vs. 30 m).

The highest performing classifier-input combination was then used to analyse the separability between classes, thereby addressing the second research question. For this class-wise analysis we examined the aggregated confusion matrix from the 5-fold cross-validation. The aggregation represents the mean value for each field.

The third experiment was set up by running the highest performing classifier-input combination with training datasets of exponentially increasing size ($10^n$). In contrast to the previous two experiments, we now used balanced training datasets for the 5-fold cross validation, meaning they are comprised of the same number of samples for each class. Four training dataset sizes (1, 10, 100 and 1,000 samples per class) were tested. Performances for the sample sizes were compared in terms of UA and PA.

For the final experiment we tested the highest performing classifier-input combination for the most recent Arealstatistik update from 2013/18 (5-fold cross-validation). This allowed us to augment the Landsat-based time series with observations from Sentinel-2 in the lowland and prealpine region. We are particularly interested in whether classifier performance can be improved by using features extracted from denser time series; therefore, a class-wise comparison between using Landsat time series (Arealstatistik 2004/09) and using Landsat-Sentinel-2 time series (Arealstatistik 2013/18) was set up.

# 5 Results

## 5.1 Time series method

In this section, we assess the capability of our time series model by analysing extracted coefficients for the three ROIs. The analysis is done for the year when the Arealstatistik 2004/09 data was collected in the respective ROI (see Table 3.1). Due to the Arealstatistik collection time span the coefficients are exclusively derived from Landsat 5 and 7 imagery. We first assess the number of available observations and the related model complexity, both visualized in Figure 5.1. The clear observation count in the lowland region is spatially homogeneous (standard deviation of 3 observations). There are, however, some scan line error artefacts in the form of stripes with lower clear observation counts than their surrounding areas. On average, a time series in the lowland region consists of 45 observations, which is sufficient to apply a time series model composed of a constant and a harmonic model. Consequently, the coefficients for the vast majority of pixels in the lowland region contain temporal information. The mean clear observation count in the pre-alpine region is similar to that in the lowland region (46 and 45 observations, respectively). In the pre-alpine region, the spatial distribution of clear observations is heterogeneous (standard deviation of 19 observations). A large part of this heterogeneity is a consequence of the high counts present on lake Brienz due to overlapping satellite orbits in combination with the low counts present on the mountain ridges of the Schrattenfluh and the Brienzer Rothorn due to many cloud or snow observations. Furthermore, there are multiple scan line error artefacts. Nevertheless, the time series for most pixels in the pre-alpine region were dense enough to extract valid temporal information. Model complexity only had to be reduced for pixels located at the aforementioned mountain ridges with low clear observation counts. For the alpine region, evidently fewer observations are available than for the other two regions (mean: 26, standard deviation: 11). In fact, at high elevations such as the areas centered around the summits of Les Diablerets and Wildhorn clear observation counts are predominantly below 10. Patches with higher counts are only present at lower elevations, for example close to Sion or in the northwest. As a direct consequence, no temporal information was extracted for the majority of pixels located in mountainous areas.

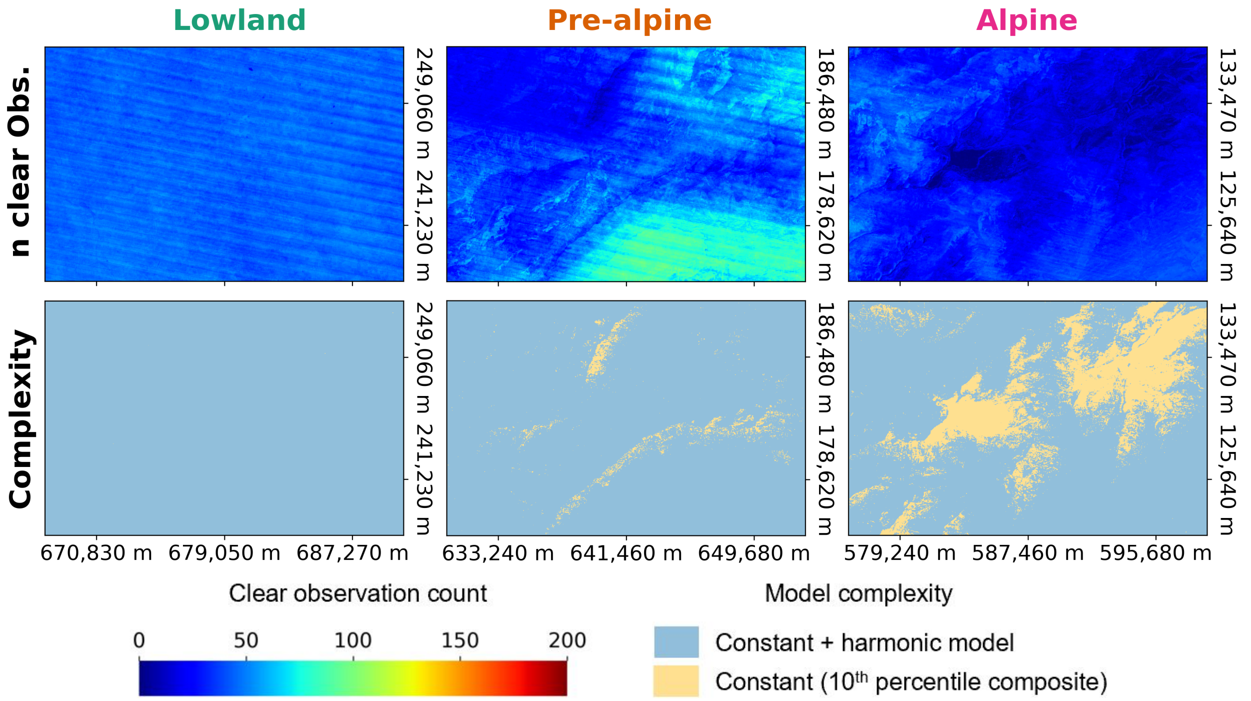

FIGURE 5.1: Number of clear observations available over a period of three years in the lowland (2006–2008), pre-alpine (2005–2007) and alpine (2003–2005) region (top row) and the respective model complexities (bottom row).

In the following two subsections we further assess the time series model's capability by comparing the coefficient-based predictions to observed data. This assessment is twofold - a qualitative, to show a visual comparison between observed Landsat scenes and predicted scenes; and a quantitative, to analyse model predictions with statistical metrics.

### 5.1.1 Qualitative assessment

For the qualitative assessment, Landsat scenes acquired in three seasons of the input year (spring, summer and autumn) were removed from the input image stack. This reduced input was then used to compute the time series model. It should be noted that the dates of the removed scenes differ between the ROIs because they are covered by different orbits and cloud cover is very high in some scenes making them unsuitable for a visual comparison. Predominantly cloudy scenes were only considered if there was a lack of alternatives. After computing the time series model, we generated images at the acquisition dates of the removed scenes. Images generated based on time series model coefficients are hereinafter referred to as synthetic images. In the following, we compare false color visualizations (red: NIR, green: red, blue: green) of the synthetic images (predicted) to those of the unprocessed Landsat scenes (observed). Similarity between predicted and observed images across the three seasons indicates that the time series model coefficients contain accurate temporal information.

The qualitative results for the lowland region are visualized in Figure 5.2. The most notable difference between the unprocessed Landsat scenes (top row) and the synthetic scenes (bottom row) is that the latter are completely free of clouds and scan line error artefacts. Contrary to that, there are wide stripes of missing data in the spring and autumn Landsat 7 images due to the scan line error of that satellite. Moreover, most of the surface in the autumn Landsat 5 image is obscured by clouds, and there are also several clouds in the summer image. The series of synthetic images reveals that the time series model coefficients contain a seasonal dynamic in surface reflectance for most land covers but urban and water. This is particularly apparent for vegetation. For example, the forest patches east of Zurich are considerably brighter during summer than during autumn or spring. Although the general seasonal dynamic of the synthetic scenes matches that of the observed scenes, there are also notable visual differences when it comes to color and brightness. In spring, for instance, the forest patches in the synthetic scenes are brighter than they are in the observed scenes. Furthermore, the observed scenes look crisper, which is particularly apparent for the agricultural fields in the southwest.

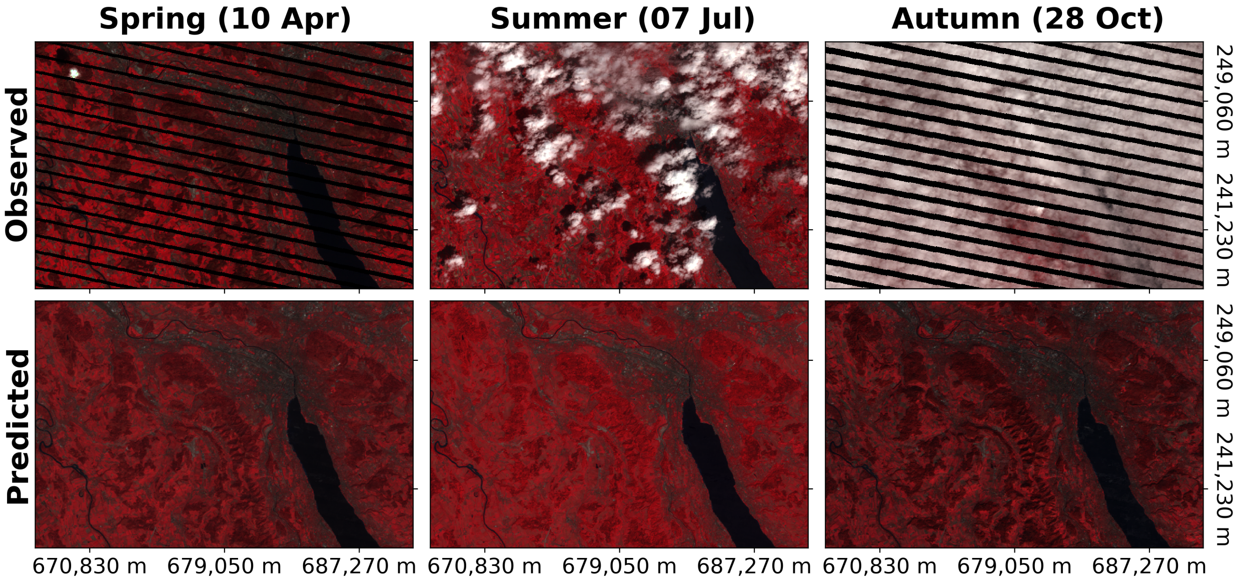

FIGURE 5.2: False color visualization (red: near infrared, green: red, blue: green) of observed Landsat 5 and 7 images (top row) and synthetic images (bottom row) for a date in spring (left column), summer (middle column) and autumn (right column) for the Lowland region in 2007.

Results for the pre-alpine region are visualilzed in Figure 5.3. In spring, a lot of snow was observed at high elevations by Landsat 5. For example, the Schrattenfluh and Brienzer Rothorn are mostly white. On the other hand, the synthetic spring image is completely snow free. As for the previous ROI, clouds and cloud shadows were also removed, which is apparent in the summer image pair. Interestingly, the northern mountain faces in the autumn Landsat 5 scene are very dark due to the mountains' shadows. This effect is slightly reduced in the synthetic scenes. In terms of surface reflectance dynamic, there are evident differences between the seasons in the Landsat scenes. The time series model is replicating those dynamics, indicated by the visual similarity between the seasonal image pairs. However, no assessment is possible for some areas in spring due to the presence of snow. Those high elevation areas are particularly interesting because parts of them are modeled with a lower complexity (see Figure 5.1), and consequently do not contain any temporal information. This is noticeable at the south face of the Schrattenfluh and Brienzer Rothorn where the synthetic spring image shows isolated bright red patches surrounded by grey or matt red areas. These bright red patches correspond to summer observations of vegetation modeled with the 10th percentile method; in contrast, the adjacent grey or matt red areas were modeled with a harmonic model.

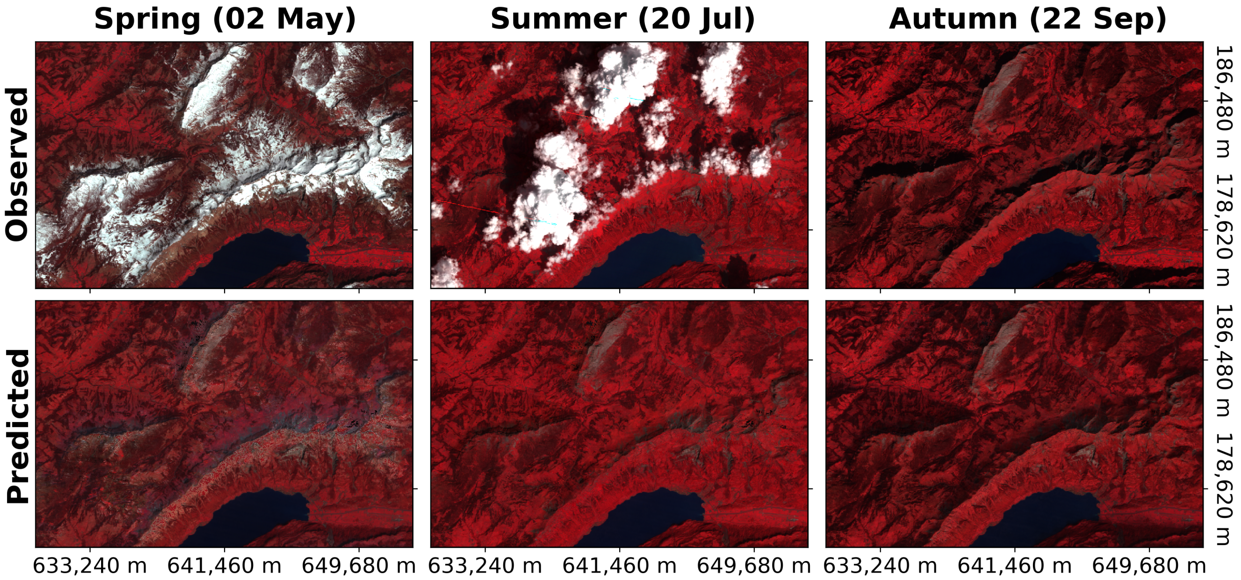

FIGURE 5.3: False color visualization (red: near infrared, green: red, blue: green) of observed Landsat 5 and 7 images (top row) and synthetic images (bottom row) for a date in spring (left column), summer (middle column) and autumn (right column) for the pre-alpine region in 2006.

Figure 5.4 visualizes the qualitative assessment for the alpine region. A great extend of the unprocessed Landsat 7 spring scene is covered by snow and its western part is affected by scan line artefacts. The corresponding synthetic image, on the other hand, is snow-free apart from the peaks of Les Diablerets and Wildhorn. Like the mountain ridges in the pre-alpine region, the snow patches as well as the surrounding rocky areas were modeled with the 10th percentile method (see Figure 5.1). At the border between the areas of different model complexities, there is a sharp drop in brightness from high to low elevation. The time series model generally captures the seasonal surface reflectance dynamic at lower elevations (e.g. in the southeast). At high elevations, the complete surface in the spring Landsat scene is obscured by snow; therefore it cannot be used as reference. Likewise, the summer Landsat scene due to clouds. In the autumn Landsat scene, the spatial extend of the snow is similar to the predicted one. However, the glacier east of Les Diablerets (Tsanfleuron Glacier) is larger in the synthetic image, and south of the Wildhorn there are small snow patches which are not present in the autumn Landsat scene.

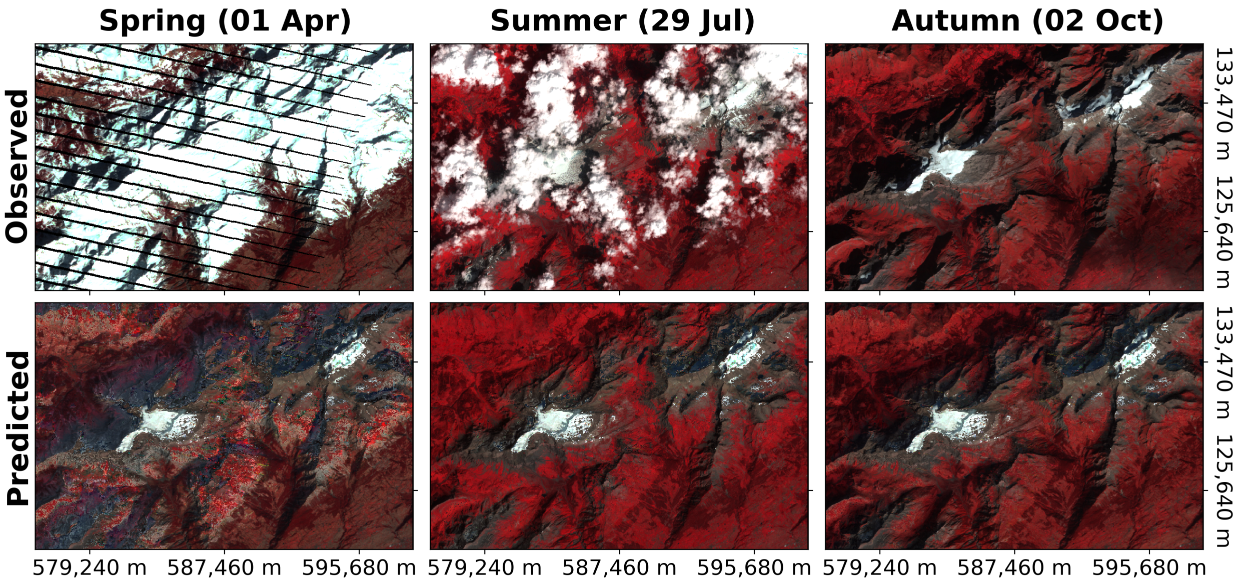

FIGURE 5.4: False color visualization (red: near infrared, green: red, blue: green) of observed Landsat 5 and 7 images (top row) and synthetic images (bottom row) for a date in spring (left column), summer (middle column) and autumn (right column) for the alpine region in 2004.

### 5.1.2 Quantitative assessment

For the quantitative assessment, we removed 0.01 % of the pixels in each Landsat scene of the input year. This reduced data set, in addition to the complete scenes from the two adjacent years, were used to calibrate the time series model. Thereafter, synthetic images were generated with the model for any day of the input year on which a Landsat scene was acquired. The masked surface reflectance values for all six spectral bands were compared to the corresponding model predictions.

Figure 5.5 visualizes the comparison by plotting the predicted against observed surface reflectance values for each spectral band and ROI. The color is used to denote the density of the point cloud, ranging from blue (low density) to red (high density). The dashed line in the scatter plots represents the 1:1 line. Proximity to this line indicates a good prediction. The solid line represents the robust linear regression of the point cloud (Huber regression: Owen, 2007). The more similar it is to the 1:1 line, the better the time series model works. The model's performance is also summarized in the R-squared ($R^2$) term, which is a statistical measurement for the proportion of the variance that is explained by the model. The lowest R-squared values, ranging from 0.15 to 0.68, are present for the visual bands (blue, green and red). For those bands, predicted and observed surface reflectance is generally very low (sub 0.1), resulting in hot spots of small size in the bottom left corner of the scatter plots. There is, however, a considerable amount of points distributed along the bottom of the scatter plots due to observations exceeding predictions. Furthermore, maximum observation values (1) largely underestimated by the time series model exist. Both phenomena are the most evident in the alpine region. As a consequence of this discrepancy, the robust linear regressions of the point clouds deviate a lot from the 1:1 lines with the former being much flatter. In comparison to the visual spectrum, R-squared values for longer wavelengths (NIR, SWIR 1 and SWIR 2) are higher, ranging from 0.69 to 0.93. In addition, the range of predicted values is larger (particularly for the NIR band), and the dense areas in the point cloud follow the 1:1 line. Therefore, the time series model works better for longer wavelengths. Nevertheless, some point clouds also contain a considerably amount of predictions that are clearly lower than the corresponding observations.

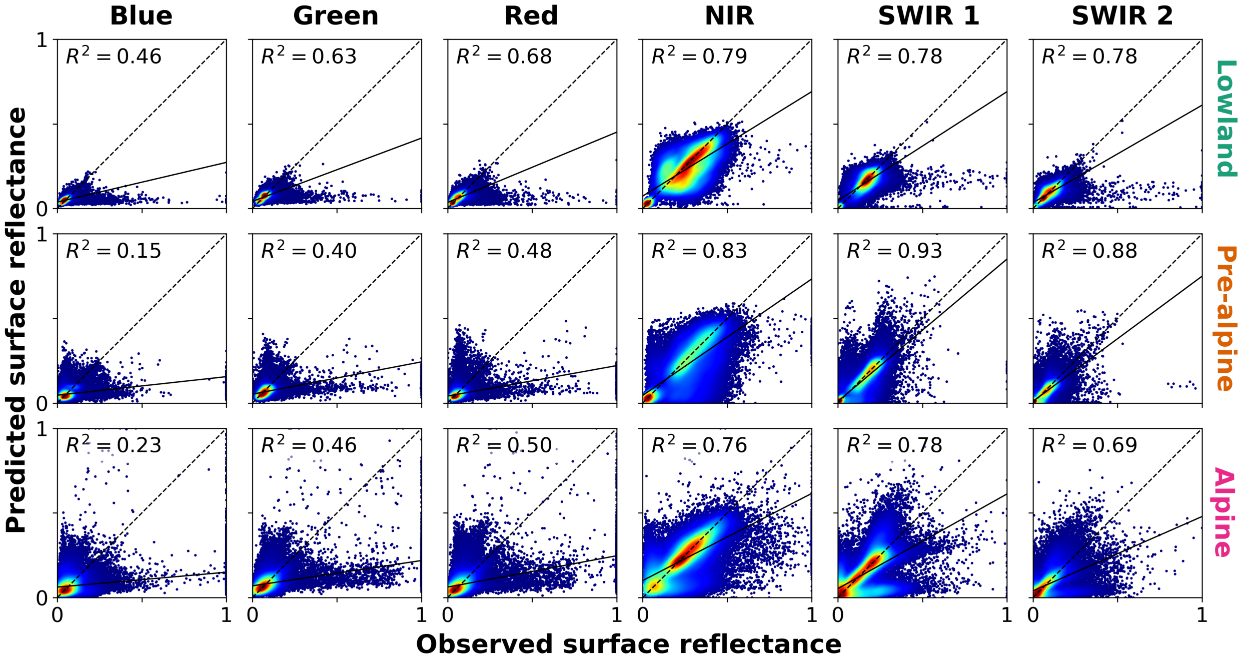

FIGURE 5.5: Accuracy of model predictions in the lowland (top row), pre-alpine (middle row) and alpine (bottom row) region for the six Landsat spectral bands (columns). Scatter plots show predicted vs. observed surface reflectance values. Red and blue denote regions with high and low point density, respectively. Solid line represents a robust linear regression of the point cloud and dashed line is the 1:1 line. The number of evaluation points across all bands is approximately 39,000, 38,000 and 35,000 for the lowland, pre-alpine and alpine region, respectively.

In addition to the scatter plots, time series showing the Root Mean Squared Errors (RMSEs) for each scene over the respective input years were visualized in Figure 5.6. The RMSE for each scene was calculated according to the formula below (5.1).

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}} \tag{5.1}$$

Where:

$y$: observed value.

$\hat{y}$: predicted value.

$n$: number of points.

The RMSE represents the average difference between observed values and model predictions. Hence, scenes with a low RMSE indicate a good working model. Scenes in the visual spectrum in the lowland region show the smallest RMSEs. Almost all of the 23 scenes have a RMSE below 0.1, except for a few outliers scenes in spring and one towards the end of summer. Slightly higher RMSEs exist for the SWIR 1 and SWIR 2 bands in the lowland region with several scenes exceeding 0.1. The highest RMSEs, however, exist for the NIR band, where multiple scenes are located only just below the 0.2 mark. The time series for the pre-alpine region are with 82 scenes noticeably denser than those of the other two ROIs (the alpine region has 40 scenes). The scenes acquired over the pre-alpine regions generally have a RMSE of approximately 0.1 and there are a few more outliers compared to the lowland region. The NIR band once again shows the highest RMSE with many spring and autumn scenes being close to, or slightly exceeding, a RMSE of 0.2. Higher errors at the start of the year and towards the end of the year are a recurring pattern for most spectral bands of the three ROIs. In addition, the time series are sparser in winter and spring than in the other two seasons. The highest RMSEs are found in the alpine region. All spectral bands except for SWIR 1 and 2 have several scenes with error terms higher than 0.2. Compared to the other two regions, there also is no apparent series of consecutive scenes with low errors.

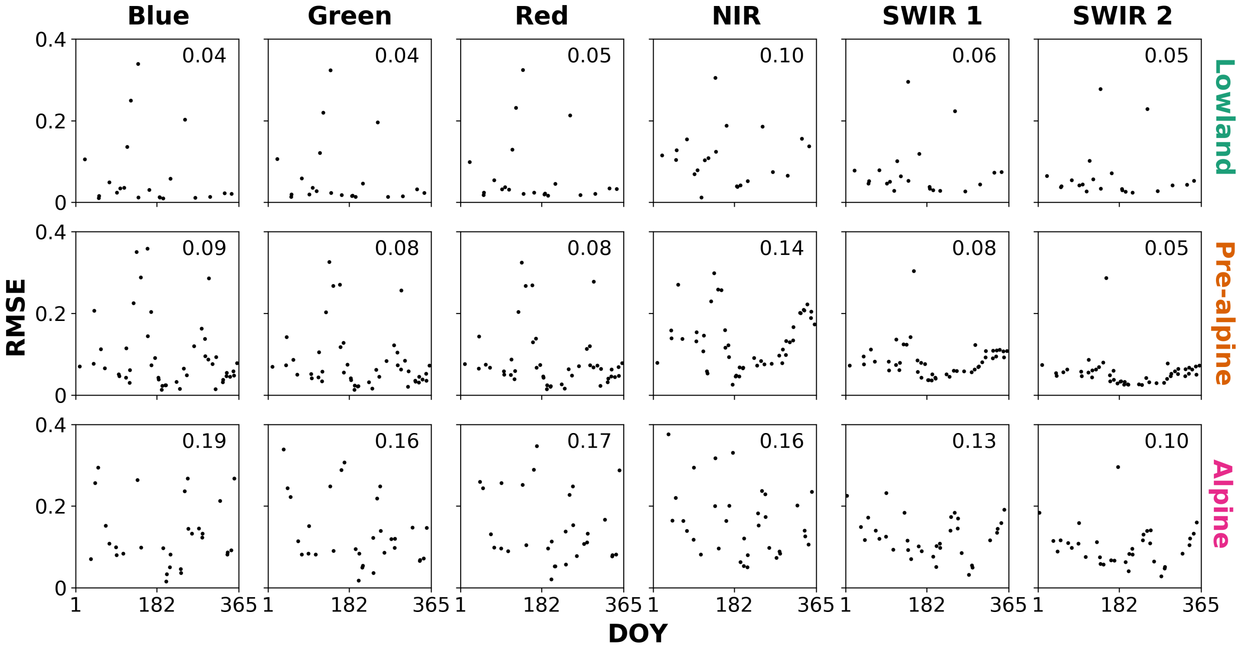

FIGURE 5.6: Accuracy of model predictions in the lowland (top row), pre-alpine (middle row) and alpine (bottom row) region for the six Landsat spectral bands (columns). Time series show the Root Mean Square Error (RMSE) for each Landsat scene over a one year period (Arealstatistik 2004/09 data collection years). The number of scenes for the lowland, pre-alpine and alpine region is 23, 82 and 40, respectively. The average RMSE of the scenes is denoted in the top right corner.

## 5.2 Reference data

The summary statistics of the land cover reference data derived from the Arealstatistik 2004/09 are visualized in Figure 5.7. In total, the dataset is comprised of 103,500 samples. The two most frequent classes are non-forest (36,275 samples) and forest vegetation (29,689 samples). Both vegetation classes are evenly represented among the three ROIs. On the other hand, the 18,907 bare land samples are mostly located in the alpine region except of some occurrences in the pre-alpine region. The distribution of artificial areas (13,593 samples), predominantly found in the lowland region, is even more one-sided. The only two classes with fewer than 10,000 samples are water (3,892 samples) and snow & ice (1,144 samples). Water is almost evenly distributed between the lowland and pre-alpine region. There is, however, almost no water in the alpine region. Snow & ice, on the other hand, is exclusively present in the alpine region and has the lowest sample count.

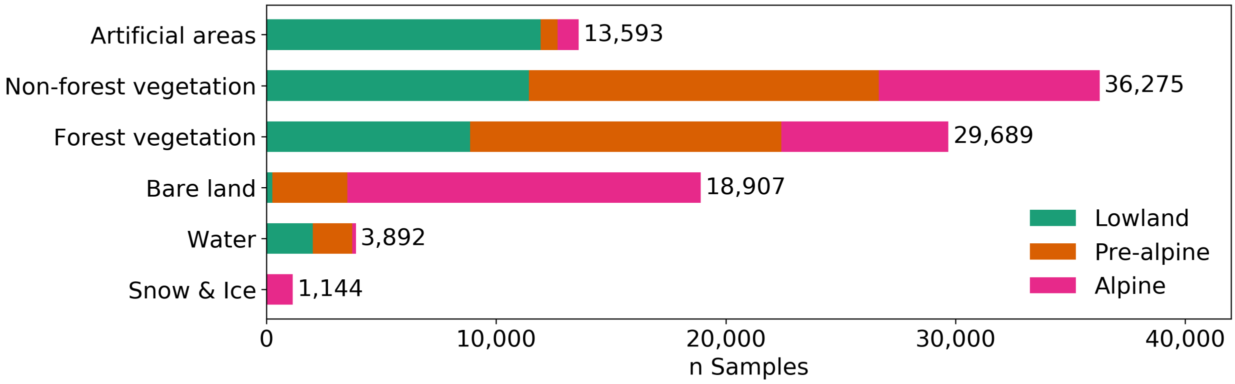

FIGURE 5.7: Summary statistics of the land cover reference data derived from the Arealstatistik 2004/09 dataset.

The land cover reference data was used to label the extracted time series features according to Section 4.4. Figure 5.8 visualizes the constants (left subplot) and amplitudes (right subplot) for each spectral band grouped into land cover classes (medians). In the visual spectrum, snow & ice have the highest surface reflectance constants. Constants for all other classes apart from bare land are substantially lower. In the NIR band, the constants of almost all classes strongly increase compared to shorter wavelengths. This increase is particularly strong for forest and non-forest vegetation. Only the median constant of water and snow & ice remain unchanged. In the SWIR 1 band, constants for all classes are decreasing again. Snow & ice show the steepest drop. In the SWIR 1 and 2 bands, water and snow & ice are clearly distinguishable from two groups of classes with higher constants. The first group is composed of forest vegetation and bare land and the second one of artificial areas and non-forest vegetation.

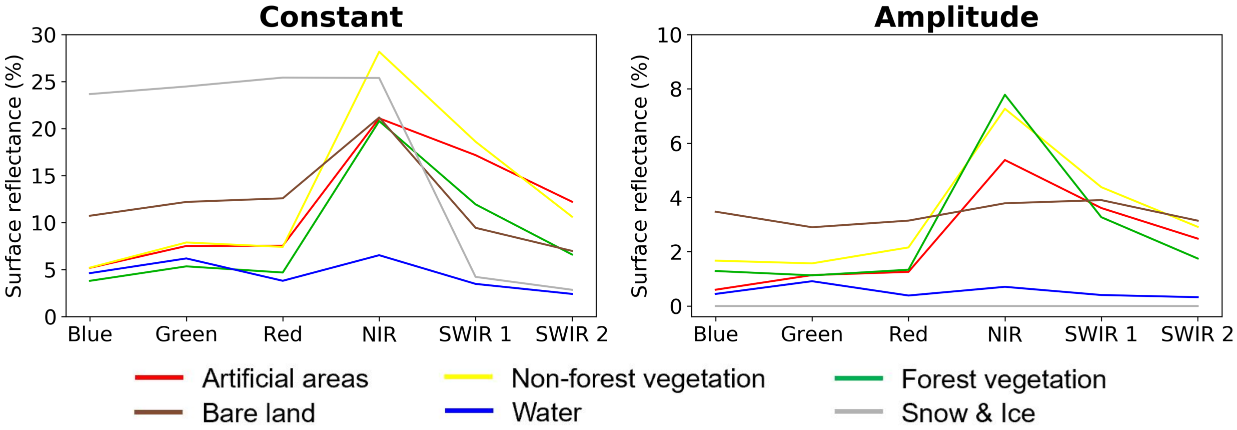

FIGURE 5.8: Class-wise medians of the time series constants (left) and amplitudes (right) for the six spectral bands (Arealstatistik 2004/09).

Amplitudes are generally much lower than the corresponding constants. In the visual spectrum, all classes but bare land have a median surface reflectance amplitude below 2 %. In particular, water and snow & ice are for all spectral bands only subject to very low changes in surface reflectance. Amplitude values for the latter class are actually zero for all wavelengths, indicating no temporal surface reflectance dynamics at all. Contrary to water and snow & ice, the other four classes show high amplitudes in the NIR and SWIR bands. The highest amplitude is shown by forest vegetation, closely followed by non-forest vegetation. Artificial areas also have high amplitudes in the NIR band, but slightly lower than those for the previous two classes. Bare land is the only class that has a high dynamic for all spectral bands.

As part of this section, we are also presenting the results of an analysis on reference data quality. Our reference data is based on the Arealstatistik which has a high quality due to its multistep data collection method including in-field verifications if areal photographs and ancillary data are insufficient for a clear interpretation (Swiss Federal Statistical Office, 2016). An important characteristic of these interpretations is that in some cases they are based on the distinction of elements in the sub-meter neighborhood of a point (Picterra, 2017). Consequently, the land cover class of some Arealstatistik points may not be distinguishable in medium resolution satellite scenes. Figure 5.9 exemplifies this for trees located along a river in an urban environment (left column) and for a country road surrounded by vegetation (center column). While the high resolution Google Earth images confirm the correctness of the Arealstatistik 2013/18 interpretations, the decisive objects are not observable on the medium resolution Landsat 8 scenes. Another problem affecting the quality of the reference data is shown in the right column of Figure 5.9. Snow & ice is mapped in the Arealstatistik based on a single summer aerial photograph; however, its extend varies throughout the year. Consequently, bare land in a Landsat scene may be labeled as snow & ice in the Arealstatistik and vice versa. The vast majority of reference samples, however, are presumably not affected by either of those problems considering that they mainly occur for edge-cases while the six land cover classes are predominantly present as homogeneous areas (Figure 3.1).

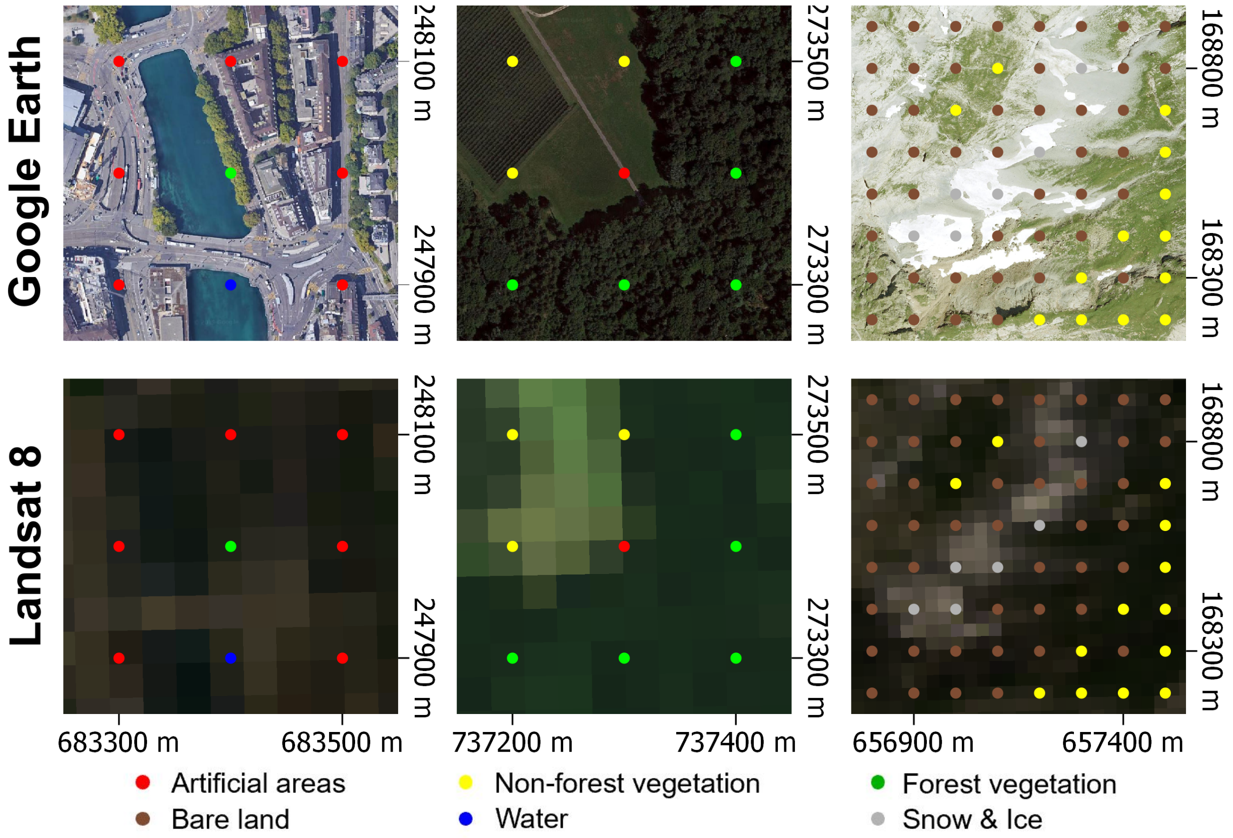

FIGURE 5.9: Arealstatistik 2013/18-derived land cover points overlaid onto Google Earth images (top row) and corresponding true color Landsat 8 summer scenes (bottom row).

## 5.3 Land cover classification

In what follows we present the results for the land cover classification. Each subsection corresponds to one of the previously set up experiments (see Section 4.7).

### 5.3.1 Classifier and input comparison

The results for the classifier and input comparison are visualized in Figure 5.10. The left subplot summarizes the classification results for the nine tested classifier-input combinations in terms of AA and Kappa coefficient. The right subplot shows the respective training times. The experiment's numeric accuracy assessment results (OA, AA, MA and Kappa coefficient) are listen in Table 5.1. Generally, all but the DNN-composite combinations achieved accuracies (OA, AA and MA) greater than 70% and Kappa coefficients greater than 0.6. Standard deviations for all classifier-input combinations denoted by the solid black lines are negligibly small. The highest standard deviation in accuracies was recorded for the DNN-time series-spatial approach (approximately 1 %). The same classifier-input combination also has the highest standard definition for the Kappa coefficient (approximately 0.01). All others combinations are well below that threshold.

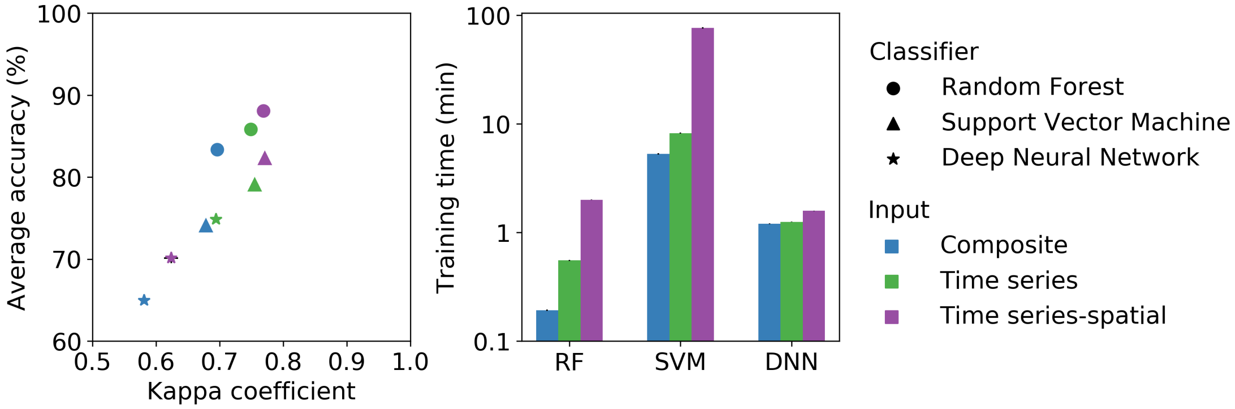

FIGURE 5.10: Classifier and input comparison. Left subplot shows average accuracy vs. Kappa coefficient for the different classifier-input combination and right subplot shows the respective runtimes. Results were obtained using 5-fold cross-validation.

Regarding classifier inputs, the worst results in terms of all metrics were recorded with the composite input: RF achieved accuracies (OA, AA and MA) between 77.6 and 83.3 %; SVM between 74.1 and 75.8 % and DNN between 65 and 68.3 %. Adding temporal information (time series input) improved performances of all classifiers. Most notable, accuracies obtained with the DNN increased by at least 8.7 % and Kappa by 0.114. Although to a smaller extend, accuracies of the non-deep learning classifiers also increased by several percent compared to the composite input (OA, AA and MA by at least 2.5 % and 5 % for RF and SVM, respectively). The time series input itself was outperformed by the time series-spatial input in combination with RF (+1.5 % at least) and SVM (+1.3 % at least). The performance improvement for the addition of spatial information, however, was smaller than that for the addition of temporal information, and the performance of the DNN classifier even decreased by 6.3 % (OA), 4.7 % (AA), 5 % (MA) and 0.071 (Kappa) with the addition of spatial information.

In the classifier comparison, RF obtained the overall best mapping accuracies of the experiment with an OA, AA and MA of 83 %, 88.1 % and 85.5 %, respectively (time series-spatial input). Only in terms of Kappa SVM in combination with the same input achieved a slightly higher value (0.771 compared to 0.769). RF also generally achieved the highest performances for the composite and time series inputs, especially in terms of AA. The performance of RF in terms of AA was in fact at least 6 % better than SVM for all inputs, while the two classifiers performed similarly in terms of OA and Kappa. DNNs, on the other hand, performed worse than the conventional shallow machine learning classifiers (RF and SVM), which is clearly supported by the fact that the lowest accuracy for each input was recorded with deep learning. Moreover (and as previously mentioned), the DNN is also the only classifier that performed worse with the time series-spatial input than with the time series input.

Among the classifiers, the shortest training times were generally recorded for RF with an average time of approximately 12 s, 33 s and 120 s for the composite (6 features), time series (18 features) and time series-spatial (162 features) input, respectively. The sole exception was the DNN-time series-spatial combination which was trained about 15 s faster than RF. The DNN is in fact also the only classifier that is not heavily affected by the number of features: Training time for the DNN only increased by about 22 s when comparing the composite and time series-spatial input. The longest training times independent of the number of features were recorded with SVM, approximately ranging from 5 min (composite input) to 77 min (time series-spatial input).

TABLE 5.1: Classifier (RF: Random Forest, SVM: Support Vector Machine and DNN: Deep Neural Network) and input comparison. The performance is compared in terms of Overall Accuracy (OA), Average Accuracy (AA), Mean Accuracy (MA) and Kappa coefficient. Results were obtained using 5-fold cross-validation.

| Input | Classifier | OA (%) | AA (%) | MA (%) | Kappa coefficient |
|---|---|---|---|---|---|
| Composite | RF | **77.6** | **83.3** | **80.5** | **0.696** |
| | SVM | 75.8 | 74.1 | 74.9 | 0.678 |
| | DNN | 68.3 | 65.0 | 66.6 | 0.581 |
| Time series | RF | 81.5 | **85.8** | **83.6** | 0.749 |
| | SVM | **81.6** | 79.1 | 80.4 | **0.755** |
| | DNN | 77.0 | 74.8 | 75.9 | 0.695 |
| Time series-spatial | RF | **83.0** | **88.1** | **85.5** | 0.769 |
| | SVM | 82.9 | 82.3 | 82.6 | **0.771** |
| | DNN | 71.7 | 70.1 | 70.9 | 0.624 |

A visual comparison of land cover maps for the three different classifiers with the time series-spatial input is shown in Figure 5.11. The Arealstatistik 2004/09-derived reference data, serving as ground truth, is also shown in the first row. Visual inspection indicates a high similarity between the classification maps and the ground truth. Major landscape features of all land cover classes such as lakes (water), cities (artificial areas), mountains (bare land and snow & ice), forests (forest vegetation) and agricultural fields (non-forest vegetation) apparent in the ground truth are also present in the classification maps. Furthermore, these landscape features are mostly represented as a homogeneous land cover patch with only very little salt-and-pepper noise. There are, however, also classifier-specific disagreements between predicted maps and ground truth. A selection of sub-regions showing different types of disagreement are marked in the predicted maps with black rectangles having unique letters (A to G) assigned to them for identification. Sub-region A, located in the alpine region, shows small patches of artificial areas in the RF result. In contrast, there are no artificial areas south of Les Diablerets in the ground truth. This difference is even more apparent for the results of the other two classifiers. Sub-region B marks another incorrectly classified patch of artificial areas south of the Schrattenfluh in the pre-alpine region. While these patches are even wider spread in the DNN map, they are not present in the RF map. Sub-region C is located southeast of the Wildhorn. In the ground truth and RF map this sub-region is homogeneously classified as bare land. SVM and DNN maps, on the contrary, contain several snow & ice patches. The extend of snow & ice is generally overestimated by these two classifiers (low UA and high PA). The D rectangle marks one of the few areas in the lowland region with apparent deviations from the ground truth. The river Limmat is partly classified as artificial area where Lake

Zurich drains into it. Although Limmat is also obscured by a bridge in the ground truth, in the DNN map the water line of the river is disrupted multiple times by artificial areas. The D square marks more misclassification at the southern end of Lake Zurich, namely there are multiple small patches of bare land in the lake. Another area where the DNN shows deviations from the ground truth as well as the other classifiers is in the pre-alpine region east of Brienz (sub-region E) where a large area of non-forest vegetation was classified as the incorrect vegetation type (forest). In comparison, RF and SVM classified the extend of this non-forest vegetation patch correctly; however, they disagree on the extend of the artificial patches scattered in that area. Sup-region F marks extra artificial areas and water in the alpine DNN map. The landscape in this region actually consists primarily of homogeneous bare land (see ground truth). The DNN, however, detected small patches of artificial areas and multiple isolated water pixels. The RF and SVM classifications of this area, on the contrary, match the ground truth. The final sub-region, G, is located in the same ROI but in the very southeast. The urban areas in Sion are surrounded by non-forest vegetation. A large part of them, however, is classified as bare land in the DNN map. Land cover maps of the other two classifiers match the ground truth in this sub-region.

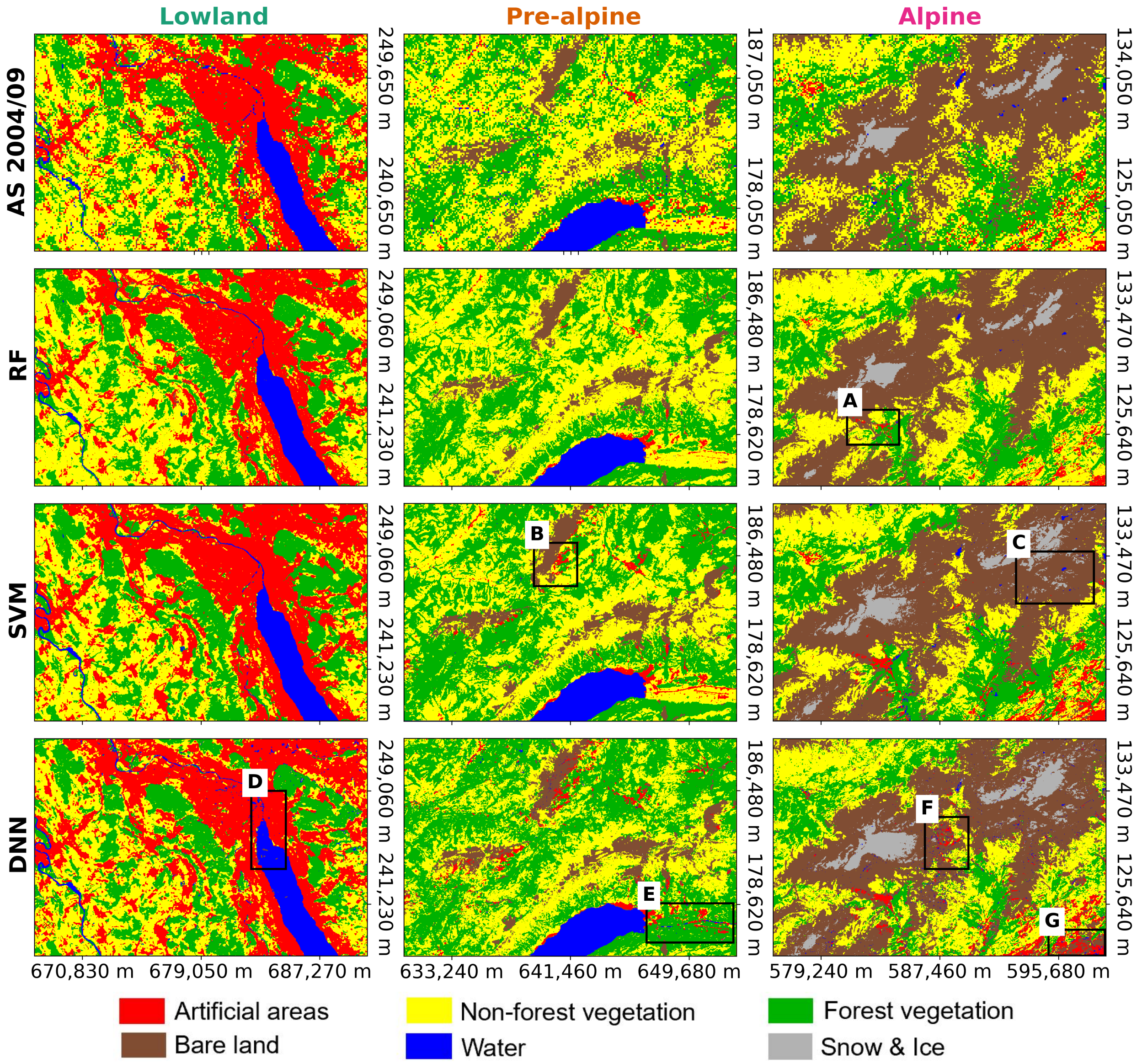

FIGURE 5.11: Visual comparison of land cover maps for the lowland (left column), pre-alpine (mid column) and alpine (right column) region. The first row shows the ground truth in the form of Arealstatistik (AS) 2004/09-derived land cover maps, and the following three rows show the classification results for Random Forest (RF), Support Vector Machine (SVM) and Deep Neural Network (DNN) with the time series-spatial input.

### 5.3.2 Per-class separability

We also examined the confusion matrix of the highest performing classifier-input combination from the previous section (RF-time series-spatial). The aggregated confusion matrix (5-fold cross-validation) visualized in Table 5.2 gives insight into the separability between individual classes by comparing classifier predictions to their true labels. The MA of the classification result listed in the confusion matrix is 86.0 % (OA and AA are 83.3 % and 88.6 %, respectively) and Kappa is 0.773. UAs and PAs for all classes generally exceeded the 80 % mark. A considerably large confusion exists between forest and non-forest vegetation. In fact, 14.8 % of the 5,938 forest samples were labeled as non-forest vegetation and 9.7 % of the 7,255 non-forest vegetation samples as forest vegetation. Non-forest vegetation was also misclassified, but to a lesser degree, as bare land (269 samples) and artificial areas (184 samples). This also applies to forest vegetation. Artificial areas were often mistaken for non-forest vegetation (16.1 %), but their identification was reliable (UA 87.1 %). Bare land was well distinguishable from other land cover classes (UA 87.9 % and PA 83.1 %). Noteworthy is only the frequent misclassification of bare land (3,781 samples) as non-forest vegetation (482 samples), accounting for 75.5 % of the misclassified bare land samples. Most water samples were correctly classified (PA 87.9 %) and, moreover, water was identified with a very high reliability (UA 98.1 %), meaning only very few dry land cover samples were mistaken for water. The identification of snow & ice was also very reliable (UA 94.5 %), but snow & ice was frequently misclassified as bare land (PA 74.7 %).

TABLE 5.2: Aggregated confusion matrix (5-fold cross-validation) including User's Accuracies (UAs) and Producer's Accuracies (PAs) for Random Forest with the time series-spatial input. Correct classifications (diagonal) are boldfaced.

| Classified as (pixels) | | Reference data (pixels) | | | | | | | UA (%) | PA (%) |
|---|---|---|---|---|---|---|---|---|---|---|
| | | AA | NFV | FV | BL | W | SI | Total | | |
| Artificial areas | AA | **2,162** | 184 | 82 | 39 | 15 | 0 | 2,482 | 87.1 | 79.5 |
| Non-forest veg. | NFV | 437 | **6,099** | 876 | 482 | 23 | 0 | 7,917 | 77.0 | 84.1 |
| Forest veg. | FV | 89 | 701 | **4,916** | 107 | 35 | 0 | 5,848 | 84.1 | 82.8 |
| Bare land | BL | 24 | 269 | 60 | **3,143** | 21 | 58 | 3,575 | 87.9 | 83.1 |
| Water | W | 7 | 2 | 4 | 0 | **685** | 0 | 698 | 98.1 | 87.9 |
| Snow & Ice | SI | 0 | 0 | 0 | 10 | 0 | **171** | 181 | 94.5 | 74.7 |
| | Total | 2,719 | 7,255 | 5,938 | 3,781 | 779 | 229 | | | |

### 5.3.3 Training dataset size

Having previously trained the classifiers on the entire training dataset, we ran RF on a subsets of the time series-spatial input with different training dataset sizes (1, 10, 100 and 1,000 samples per class) to determine an adequate number of samples for each class. Figure 5.12 visualizes UAs (left subplot) and PAs (right subplot) for the six classes for an increasing number of training samples with balanced datasets (i.e., all classes have the same number of samples). Results were obtained using 5-fold cross-validation. The lowest accuracies were recorded with 1 training sample per class (all UAs < 50 % and all PAs < 75 %). Low PAs were particularly recorded for the classes non-forest vegetation (16.4 %), bare land (17.4 %) and forest vegetation (37.0 %). Moreover, the extend of snow & ice was largely overestimated (UA 8.2 % and PA 70.8 %). Increasing the number of training samples per class from 1 to 10 resulted in a strong increase in average accuracy (UA +20.2 % and PA +23.5 %); most notably, the UA of snow & ice increased by 37.7 % and the PA of bare land by 50.7 %. The step from 10 to 100 samples per class further increased accuracies for all classes (on average by 10.0 % and 8.2 % in UA and PA, respectively), but to a lesser extent than the first step. The same trend continued for the last step from 100 to 1,000 (UA +6.9 % and PA +4.0 %). Consequently, the highest results were recorded with 1,000 training samples per class (UA 78.8 % and PA 84.4 %). Noteworthy is that all six land cover classes show a similar pattern characterized by high accuracy increases for small training data sizes (<100 samples per class) and low increases for large training dataset sizes (>100 samples per class).

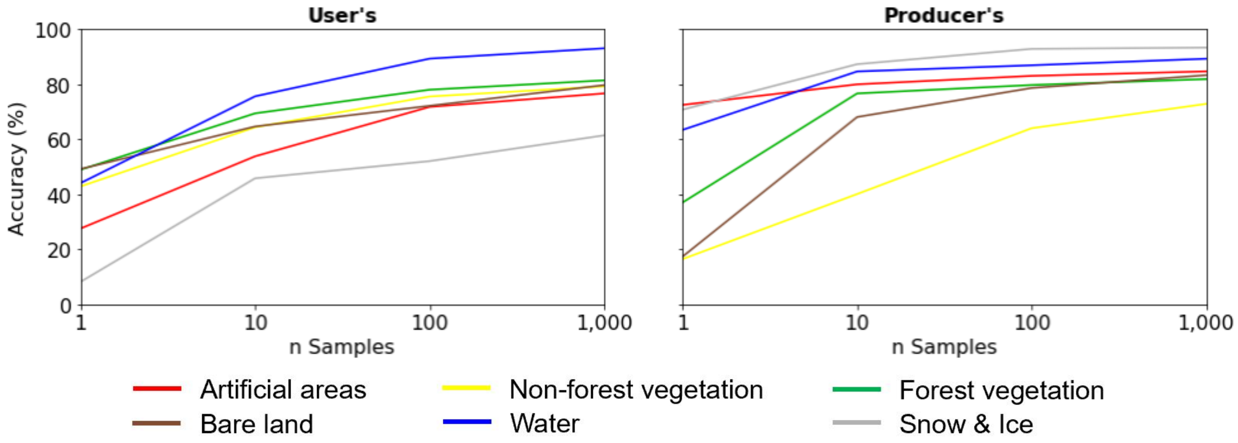

FIGURE 5.12: Relationship between the number of training samples (x-axis) and the user's (left) and producer's (right) accuracies (y-axis) for different land cover classes. Results were obtained using Random Forest with the time series-spatial input and 5-fold cross-validation.

### 5.3.4 Time series augmentation

In the final experiment we also tested the effect of augmenting Landsat time series with observations from Sentinel-2. To do so, we computed the time series-spatial input for the most recent update of the Arealstatistik (2013/18). Figure 5.13 shows the number of available clear observations and the related model complexities. Interesting is also the comparison between the time series results for 2013/18 to the ones for 2004/09 (see Figure 5.1). Therefore, we used colorbars with identical ranges for the clear observation count visualizations. In the lowland region, there are on average 148 clear observations available per pixel. Consequently, the mean count tripled with the addition of Sentinel-2 data. There are, however, no differences in model complexities. The mean clear observation count for the pre-alpine region (131 observations) also strongly increased compared to 2004/09 (46 observations). However, time series for pixel located on the mountain ridges remained sparse. Consequently, the percentage of pixels modeled with high complexity only increased by 1 % (97.3 % for 2004/09 to 98.3 % for 2013/18). For the alpine region only Landsat data (Landsat 7 and Landsat 8) is available because the Sentinel-2 constellation was launched after the Arealstatistik was updated in 2013 in that region (see Figure 3.2). Therefore, the mean number of clear observations remained largely unchanged (26 for 2004/09 vs. 25 for 2013/18), and the percentage of pixels modeled with a lower complexity actually increased from 27.8 % to 38.7 %.

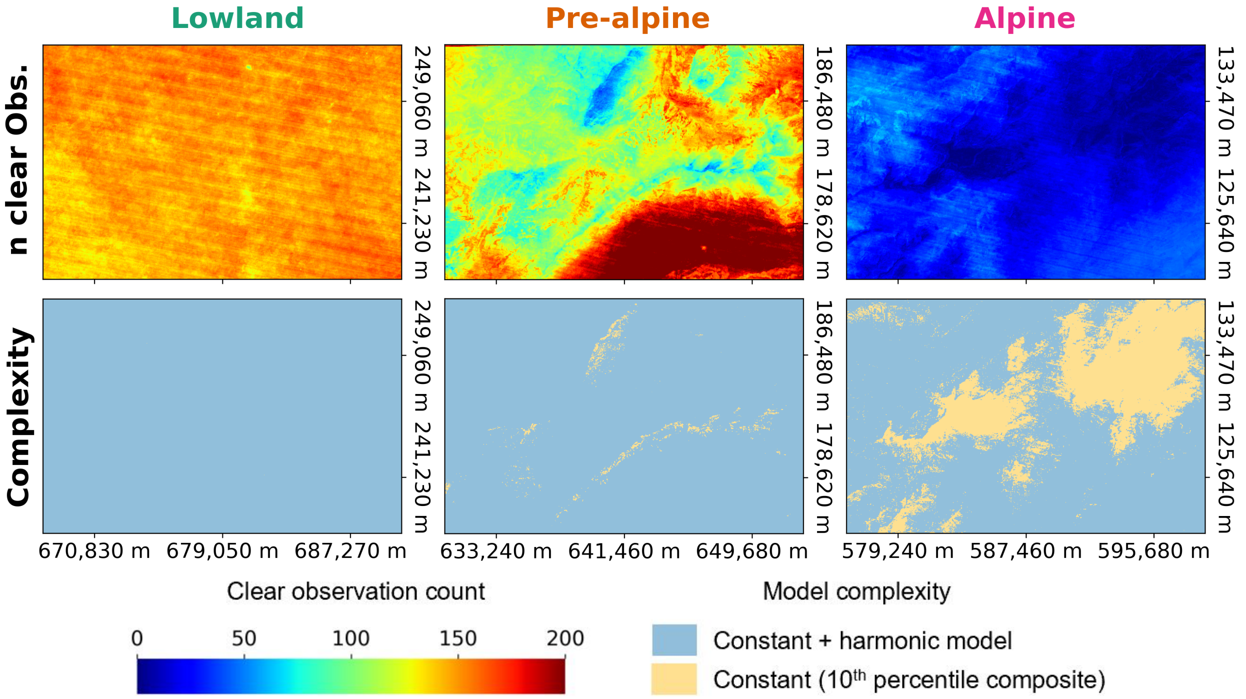

FIGURE 5.13: Number of clear observations available over a period of three years in the lowland (2015–2017), pre-alpine (2014–2016) and alpine (2012–2014) region (top row) and the respective model complexities (bottom row).

We then used the coefficients extracted from the denser time series in conjunction with the highest performing classifier-input combination (see Subsection 5.3.1). The classification results are visualized in Figure 5.14. Visual inspection of the land cover maps showed that there is a high level of agreement between the ground truth and the RF maps. Like for the Arealstatistik 2004/09 (see bottom row in Figure 5.11), all major features were correctly mapped and there is little salt-and-pepper noise. However, minor disagreements between the ground truth and the RF maps are still present on closer inspection. For example, the extend of snow & ice was slightly underestimated, and some of the artificial areas east of lake Brienz were not detected by RF.

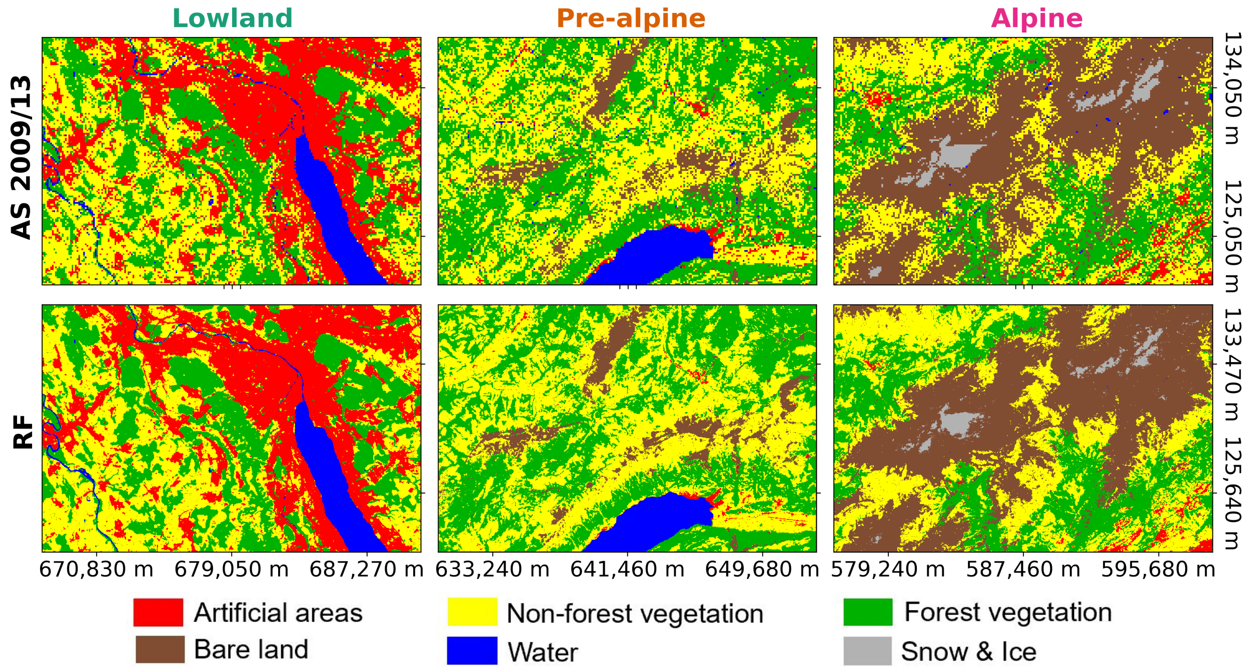

FIGURE 5.14: Visual comparison of land cover maps for the lowland (left column), pre-alpine (mid column) and alpine (right column) region. Top row shows the ground truth in the form of Arealstatistik (AS) 2013/18-derived land cover maps, and the bottom row shows the classification results for Random Forest with the time series-spatial input.

The good visual impression was also confirmed by the classification's quantitative results (OA 84.0 %, AA 86.9 %, MA 85.5 % and Kappa 0.784). Figure 5.15 shows class-wise performance differences between including temporal information from Landsat time series (Arealstatistik 2004/09) and temporal information from Landsat time series augmented with Sentinel-2 observations (Arealstatistik 2013/18). Higher UAs and PAs were recorded with the augmented time series (compared to the Landsat time series) for the classes: artificial areas (UA +1.2 % and PA +1.0 %), non-forest vegetation (UA +1.2 % and PA +0.5 %), forest vegetation (UA +2.1 % and PA +3.0 %) and water (UA +0.2 % and PA +0.3 %). While the PA of bare land also increased (+0.5 %), its UA decreased (-1.1 %). In contrast to all other land cover classes, the performance for snow & ice decreased by 10.3 % and 20.1 % in UA and PA, respectively. Snow & ice, however, is only present in the alpine region where no Sentinel-2 imagery is available for the data collection year of the Arealstatistik 2013/18. This decrease in accuracy is therefore unrelated to the density of time series.

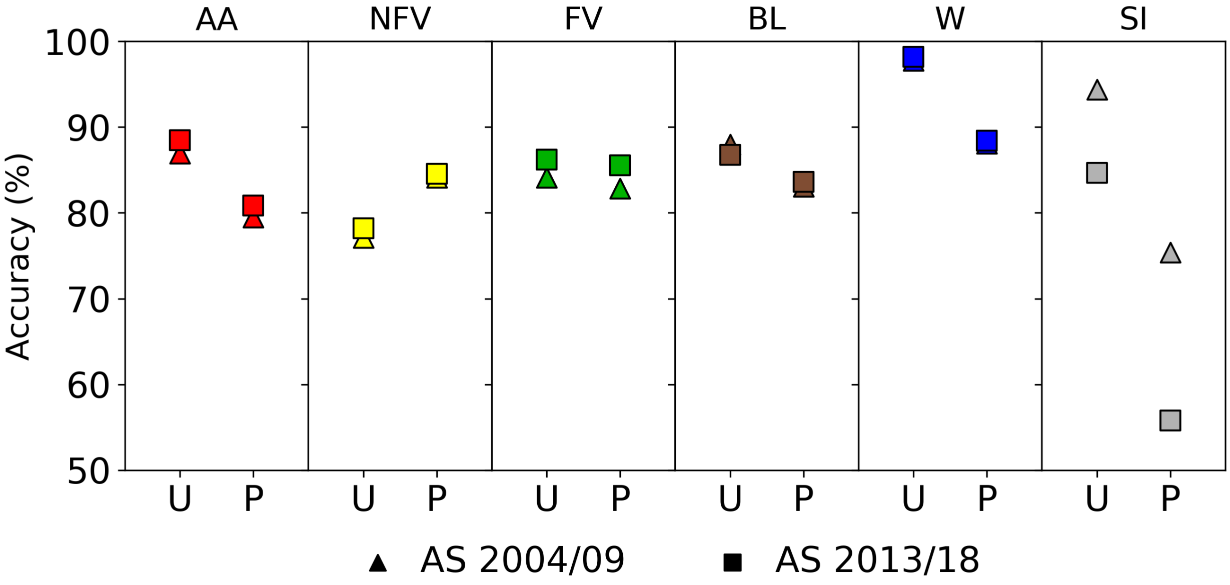

FIGURE 5.15: Comparison between classifying (Random Forest) the time series-spatial input derived from Landsat 5 and Landsat 7 data for the Arealstatistik (AS) 2004/09 and the time series-spatial input derived from Landsat 7, Landsat 8 and Sentinel-2 data for the Arealstatistik 2013/18. Performance is compared class-wise in terms of User's (U) and Producer's (P) accuracy. Results were obtained using 5-fold cross-validation.

# 6 Discussion

## 6.1 The benefits of adding temporal and spatial information

### 6.1.1 Temporal information

As part of our first research question, we investigated the benefits of adding temporal information to land cover classification. The corresponding experiment compared classification results from an annual composite input having no temporal information to classification results with a time series input. Results with the composite input averaged 74.0 % in MA (Kappa = 0.652) across the classifiers; in contrast, results averaged 80.0 % in MA (Kappa = 0.733) with the time series input (Figure 5.10 and Table 5.1). Consequently, the addition of temporal information increased classification results by 6.0 % in MA (+0.081 in Kappa). This finding is in agreement with those of a meta-analysis of remote sensing research on land cover classification (Khatami et al., 2016). Averaged over 16 analysed articles, Khatami et al. (2016) reported an increase of 6.9 % in OA (from 73.3 % to 80.2 %) due to the inclusion of multi-time images. Likewise, average OA increased by 6.1 % (73.9 % with composite input and 80.0 % with time series input) in our experiment, despite using balanced class weights to prevent the classifiers from achieving a high OA at the expense of small classes. The addition of temporal information via the developed time series method thus achieves improvements of similar magnitude to those published in articles on this topic. We attributed these improvements to our time series model's capability to capture distinct seasonal spectral dynamics of land covers (time series model assessment: Figures 5.2–5.6). The temporal information in turn facilitates the discrimination between land cover classes (Gebhardt et al., 2014; Petitjean et al., 2012). Vegetation in particular is better separable with temporal information from other land cover types due to high phenology-based seasonal dynamics (amplitude in Figure 5.8).

Land cover was also successfully classified with coefficients from a harmonic time series model in Burkina Faso (Liu et al., 2016). Liu et al. (2016), however, expressed concerns regarding the appropriateness of harmonic models for areas where large parts of the year lack observations. Regarding the spatial variation in the availability of clear observations in Switzerland (Figure 5.1 and 5.13), we had to ensure that our time series model extracts temporal information in a robust manner. Consequently, we reduced model complexity for sparse time series having 12 or fewer observations. Visual classification results in mountainous areas, where data availability is particularly sparse (Figure 5.1), show good agreement with the ground truth (Figure 5.11). The visual agreement is also supported by the accuracies of the predominant classes bare land (UA 87.9 % and PA 83.1 %) and snow & ice (94.5 % and 74.7 %) (Table 5.2). Our twofold model therefore extracts valuable information from time series independent of elevation and landscape. Furthermore, it can be applied globally due to GEE's worldwide data availability.

### 6.1.2 Spatial information

The second part of the first research question focused on the benefits of adding spatial context information to land cover classification. A third input including coefficients from adjacent pixels was thus compared to the time series input. Classification results with this time series-spatial input averaged 84.1 % in MA (Kappa = 0.770) across RF and SVM. Compared to the time series input, performances of these classifiers increased by 2.1 % in MA (+0.018 in Kappa) (Figure 5.10 and Table 5.1). Our results therefore reveal that adding spatial information information on top of temporal information further improves classification results. An exception to this finding is the DNN's performance which decreased by 5 % in MA (-0.071 in Kappa). We treated this anomaly as a classifier-specific issue, and consequently excluded it from this analysis. It is discussed as part of the classifier comparison in Section 6.2. The found benefits of adding spatial information to land cover classification are in agreement with the previously cited meta-analysis on land cover classification (Khatami et al., 2016). Khatami et al. (2016), however, reported that the inclusion of textural information yields an average increase of 12.1 % in OA, while we only recorded an average increase of 1.4 % (Table 5.1). We identified two potential reasons for this difference: (1) the 31 analysed articles by Khatami et al. (2016) also included articles using high resolution imagery which contains more textural information than medium resolution imagery and (2) the baseline OA in these articles is more than 10 % lower than ours (71.2 % vs. 81.6 %). Consequently, we are limited to the conclusion that adding features from adjacent pixels is beneficial for land cover classification. The benefits particularly include the improvement of the visual classification results. Compared to the classification maps with the time series input (Figure A.4), landscape features are evidently more homogeneous with the time series-spatial input (Figure 5.11). Therefore, spatial information is essential for reducing salt-and-pepper effects in pixel-based classifications. Alternatively, land cover maps can also be post processed to remove salt-and-pepper effects by reevaluating classifier predictions on the basis of predictions from adjacent pixels. A review of popular postprocessing methods in the context of remote sensing is given in Huang et al. (2014).

## 6.2 Classifier comparison

The second research question aimed to evaluate performances of popular machine learning classifiers (RF, SVM and DNN) under different spatiotemporal input scenarios (composite input, time series input and time series-spatial input). In the corresponding experiment, nine different classifier-input combinations were tested. All classifier generally achieved good results (MA > 70 % and Kappa > 0.6) under the different input scenarios (Figure 5.10 and Table 5.1). In the following paragraphs of this section, we are discussing performance differences between the classifiers, also taking into account classifier-specific strength and weaknesses. Table 6.1 lists the strength and weaknesses of RF, SVM and ANNs. The table was adapted from Gómez et al. (2016) who synthesized literature on large-area land cover mapping using time series optical satellite data.

TABLE 6.1: Strengths and weaknesses of Random Forest (RF), Support Vector Machine (SVM) and Artificial Neural Networks (ANNs) used for large-area land cover characterization with time series optical data adapted from Gómez et al. (2016).

| Algorithm | Strengths/characteristics | Weaknesses |
|---|---|---|
| RF | Does not overfit<br>Robust to data reduction<br>Capacity to determine variable importance | Needs input parameters<br>Computationally intense<br>Black box (rules are unknown) |
| SVM | Does not overfit<br>Works well with small training dataset<br>Manages well large feature space | Needs input parameters<br>Computationally intense<br>Poor performance with small feature space |
| ANN | Manage well large feature space<br>Generally high classification accuracy<br>Resistant to training data deficiencies<br>Indicate strength of class membership | Needs parameters for network design<br>Computationally intense<br>Black box (rules are unknown)<br>Tends to overfit data<br>Slow training |

DNNs averaged a MA of 71.1 % (Kappa = 0.633) over the three inputs. In contrast, RF and SVM averaged 12.1 % (+0.105 in Kappa) and 8.2 % (+0.102 in Kappa) higher MAs, respectively (Table 5.1). Therefore, deep learning performed considerably worse compared to shallow learning in our experiment. We compared this finding to those of a recent study that investigated classifier performance under different sample size, reference class distribution and scene complexity scenarios (Heydari and Mountrakis, 2018). Heydari and Mountrakis (2018) reported that compared to other fields, neural networks (ANNs and DNNs) do not offer considerable accuracy advantages over conventional classifiers (including RF and SVM) in the field of Landsat-based land cover classification. They mainly attributed this finding to the lack of rich contextual information in a pixel-based Landsat input space (6 spectral bands per-pixel), which may lead to data overfitting in combination with neural networks. Furthermore, overfitting was outlined by Gómez et al., 2016 as a key weakness of ANNs (Table 6.1). In our experiment, the two conventional machine learning classifiers (RF and SVM) outperformed DNNs considerably more for the composite input (MA +11.1 % and Kappa + 0.106) than for the time series input (MA +6.1 % and Kappa +0.057). Performance differences between deep and shallow learning therefore decreased with the addition of contextual information via the temporal dimension. This is in agreement with the two aforementioned studies (Gómez et al., 2016; Heydari and Mountrakis, 2018), thereby emphasizing deep learning's susceptibility to overfitting. The only contradicting result was the particularly bad performance of DNNs with the time series-spatial input which has the most contextual information among the tested inputs. In fact, the performance of DNNs considerably decreased with the addition of spatial information on top of temporal information (75.9 % vs. 70.9 % in MA and 0.695 vs. 0.624 in Kappa). Considering the benefits of the time series-spatial input in combination with RF and SVM (Subsection 6.1.2) and that ANNs manage large feature spaces well (Table 6.1), this phenomenon was regarded as a classifier malfunction for the lack of a more plausible explanation. Consequently, overfitting remains a probable reason for the neural network's underperformance. This may be particularly bad considering spatial resolution inconsistencies introduced erroneous samples into our reference data (Figure 5.9). During the training, the DNN (given enough complexity) may have learnt to correctly label these samples based on information that is not related to their land cover class. This could have led to discrepancies in classifier performance between the training and validation set, and consequently to bad classification results.

RF and SVM, on the other hand, produced promising results averaging a MA over the three inputs of 83.2 % (Kappa = 0.738) and 79.3 % (Kappa = 0.735), respectively. This can be attributed to the conventional machine learning classifiers' resistance to overfitting (Table 6.1), while still offering enough model complexity to utilize the contextual information in the spatiotemporal input scenarios. The comparison between RF and SVM also revealed that the former (RF) slightly outperformed the latter (SVM) averaged over the three inputs (+3.9 % MA and + 0.003 Kappa). Contrary to that, other similar studies (i.e. classifying land cover using Sentinel-2 or Landsat images) reported better performances with SVM than with RF (Heydari and Mountrakis, 2018; Thanh Noi and Kappas, 2018; He et al., 2015). Thanh Noi and Kappas (2018), for example, compared the classifiers under different training sample sizes in a land use/land cover classification (6 classes) using Sentinel-2 images within the Red River Delta of Vietnam. They reported that SVM produced higher OAs than RF with small training datasets. The classifiers performed similarly, though, when training sample sizes were large enough (greater than 750 samples per class). Our large training dataset with an average of 17,250 samples per class (Figure 5.7) may therefore have contributed to RF outperforming SVM. Another strength of SVM is that it manages well large feature spaces (Table 6.1). Our experiment confirmed this: The performance in MA of SVM was more similar to that of RF for larger feature spaces (-5.6 % for 6 features, -3.2 % for 18 features and -2.9 % for 162 features). The conventional machine learning classifier's promising accuracy results are supported by the respective land cover maps showing a high level of agreement with the ground truth (Figure 5.11). The fewest differences between classification maps and ground truth were found for RF, consecutively followed by SVM and DNN. The DNN was in fact the only classifier that misclassified major areas (e.g. Subregion E and G in Figure 5.11) what we also attribute to its previously discussed weakness to overfitting.

Our classifier comparison experiment also included a training time component. Training times ranging from 5 min (composite input) to 77 min (time series-spatial input) were recorded for SVM; in contrast, those of RF and DNNs did not exceed 2 min for any input (Figure 5.10). Consequently, SVM required considerably more training time in our experiment compared to RF and DNNs, even though all three classifiers are characterized as computationally intense (Table 6.1) It is important to take into account, however, that the recorded training times are strongly dependent on the computer hardware the training is run on. Nevertheless, they provide an approximate estimation to guide user decisions. Based on the relatively short training time coupled with the good performances, we recommend RF for moderately-detailed (about six classes) land cover classification with Sentinel-2 or Landsat imagery, given more than 1,000 training samples per class are available. However, literature on classifier selection indicates that with sparse training datasets SVM potentially produces better results than RF. DNNs produced inferior results compared to both conventional machine learning classifiers, which we attributed to overfitting in combination with erroneous training samples. A recent study, however, showed that deep learning (convolutional neural networks) outperforms RF in a land cover and crop type classification using optical Landsat 8 images in combination with Synthetic Aperture Radar (SAR) Sentinel-1 images (Kussul et al., 2017). Therefore, while limited benefits are offered for input spaces exclusively based on Sentinel-2 and Landsat imagery (even if temporal and spatial information are added), deep learning may offer advances compared to conventional machine learning classifiers when feature space dimensionality increases through the inclusion of ancillary data. Further studies are required to investigate this topic.

## 6.3 Per-class separability

This thesis also investigated how well the six land cover classes can be separated from one another. The results of the corresponding experiment composed of a classification with RF and the time series-spatial input (Table 5.2) showed that the identification of water is very reliable (UA 98.1 %) and only few water pixels were omitted (PA 87.9 %). Other studies confirmed that water is well separable from dry land surfaces based on Landsat imagery (Yamazaki et al., 2015). Problems are mainly limited to mixed pixels (Fisher et al., 2016) and to dark surfaces and areas that include shadow (Feyisa et al., 2014). The identification of snow & ice was also reliable (UA 94.5 %); however, only three-quarters of the reference snow & ice pixels were correctly classified (PA 74.7 %). The other quarter was classified as bare land. In contrast, Dozier (1989) stated that snow & ice is well distinguishable in Landsat imagery from other surfaces due to its distinct spectral signature. A considerable amount of misclassified snow & ice reference samples therefore presumably showed bare land, which is supported by the results of our quality analysis (Section 5.2). Bare land itself showed good separability (UA and PA > 80 %). Similarly well separable was forest vegetation from other land cover classes, even though a noticeable amount of confusion existed between forest and non-forest vegetation (15.8 % of their samples were mistaken for the wrong vegetation type). Other Landsat-based land cover classification studies reported separability issues between forest and grass caused by their coexistence (Jia et al., 2014). Since inspecting the ground truth confirmed that non-forest vegetation often exists along forest boundaries or in forest gaps (Figure 5.11), this is also a plausible explanation in our case. Artificial areas were well distinguishable from the other land cover classes (UA 87.1 % and PA 79.5 %), and we would also like to emphasize that due to the spatial heterogeneity in urban ecosystems (Cadenasso et al., 2007), artificial area samples are particularly vulnerable to resolution-related quality deficiencies (Section 5.2). Overall, we showed that all six land cover classes are well separable from one another based on spatiotemporal optical satellite imagery.

## 6.4 Trade-off between training dataset size and classification accuracy

An additional experiment was set up to find an adequate number of training samples for our land cover classification problem. We therefore investigated RF classification accuracies under different training dataset sizes of the time series-spatial input. Highest to lowest average accuracies were achieved using 1,000 (UA 78.8 % and PA 84.4 %), 100 (UA 71.8 % and 80.4 %), 10 (UA 61.8 % and 72.2 %), and 1 (UA 41.6 % and PA 48.7 %) sample(s) per class (Figure 5.12). Consequently, adding training samples increased classification accuracy. This applied to all classes, even though there were differences in the degree of accuracy increase between them. Moreover, while the step from 1 to 10 samples per class increased UAs and PAs on average by 20.2 % and 23.48 %, consecutive additions of samples led to smaller increases. The lowest increase was recorded for the step from 100 to 1,000 samples with +6.9 % in UA and +4.0 % in PA. This indicates that the benefits of additional training samples is strongly decreasing with larger training datasets. In addition to that, one has to take into consideration that acquiring training samples is expensive and time consuming (Shao and Lunetta, 2012). Based on the small increase in accuracy in relation to the number of additional training samples , we argue that collecting hundreds of training samples per class for the given setup (i.e., classifier, input and classes) is not a worthwhile use of time. We would also like to emphasize that sensitivity to training dataset size is a classifier-specific characteristic (Li et al., 2014; Thanh Noi and Kappas, 2018), and consequently also the adequate number of training samples. In our experiment, for example, SVM performed better on small training datasets than RF (+2.1 % in UA and +1.4 % in PA averaged over results for 1 and 10 samples per class) but worse on large datasets (-5.6 % in UA and -2.4 % in PA averaged over results for 100 and 1,000 samples per class) (Figure A.5). According to Li et al. (2014), however, classifier-specific characteristics are overshadowed by the effect of training dataset size. We therefore generally recommend to use approximately 100 training samples per class for a good trade-off between training dataset size and classification accuracy.

## 6.5 The benefits of denser time series

The final research question aimed to investigate the benefits of denser satellite time series to land cover classification. In the corresponding experiment, we first fused Landsat 7 and 8 with Sentinel-2 images and then ran the time series model on the augmented image stack. Compared to using only Landsat images (Figure 5.1), including Sentinel-2 images approximately tripled the number of clear observations per pixel in the lowland (148 observations vs. 45 observations) and pre-alpine (131 observations vs. 46 observations) region (Figure 5.13). The time series was therefore able to extract robuster temporal information, and the number of pixels with lower modeling complexity decreased by 1 % in the pre-alpine region. After running RF on this enhanced time series-spatial input, we compared the classification results to the results from the corresponding classifier-input combination from the first experiment (Landsat 5 and 7 images). Averaged over all classes but snow & ice, accuracies increased by 0.7 % in UA and by 1.0 % in PA with the inclusion of Sentinel images (Figure 5.15). Therefore, denser time series improved land cover classification. Snow & ice (decrease of 10.3 % and 20.1 % in UA and PA, respectively) was excluded from this analysis because it is only present in the alpine region (Figure 5.7) where time series were not augmented with Sentinel images (Figure 5.13).

Although we showed that land cover classification benefits from denser time series, the improvements were marginal. This indicates that the extraction of complex seasonal dynamics from denser time series is limited by the unimodality of our time series model, which calls for more model complexity. Interesting in the context of our work is the increase in complexity via additional harmonic frequencies allowing the modeling of multimodal dynamics. Bimodal and trimodal time series model were successfully used to monitor forest disturbances at the border between Georgia and Sourth Caronlina, USA (Zhu et al., 2012) and to detected drought-related vegetation disturbances in Somalia (Verbesselt et al., 2012), respectively. The difficulty in choosing an adequate number of harmonic frequency, however, is that it is largely dependent on data availability (Landmann et al., 2019). Zhu and Woodcock (2014) thus suggest that the number of clear observations in a time series should be more than three times the number of model coefficients in order for the estimated fit to be accurate and robust. Consequently, an integrated model on the basis of the number of available clear observations should be used to address the spatial variability in clear observation counts (Figure 5.1). The effectiveness of integrating harmonic models with different complexities was demonstrated by Zhu et al. (2015) who generated synthetic images based on all available Landsat data. The challenge in using such an integrated model for classification is, however, that the number of coefficients varies between the individual models, while classifiers require equally sized input vectors. Consequently, model coefficients first have to be combined into the same number of features, unless a separate classifier is trained for each model complexity. Combining the coefficients can be achieved by either not populating features for less complex models (i.e., setting coefficients only extracted by complex models to zero), or using the integrated model to predict synthetic images at predefined dates and then using those as classifier input. The latter can also be used in combination with non-harmonic models, and consequently opens up the possibilities to use more sophisticated interpolation techniques. Exploring the effectiveness of this approach is subject to further research.

# 7 Conclusion

Our main goal was to investigate performances of machine learning classifiers under different spatiotemporal input scenarios. We therefore developed a method to extract robust temporal information from time series of multispectral satellite observations and compared this temporal input, as well as a spatial variation on it, to a baseline classification using a composite input. Our study included a per-class separability analysis. Furthermore, we investigated the adequate number of training samples for each land cover class; and finally, we explored the class-wise accuracy effects of augmenting the Landsat-based time series for the feature extraction with Sentinel-2 observations. The main findings of these experiments can be summarized with the following points:

- Robust temporal information from time series of satellite observations was effectively incorporated into land cover classification with our twofold model. The temporal input considerably increased land cover mapping accuracy compared to a non-temporal composite input (+6.0 % in MA and +0.081 in Kappa). This benefit is attributed to the model's capability to inform on intra-annual land cover dynamics and phenological differences. In general, the benefits of using the time series model include but are not limited to: (1) full automation, (2) fast computation and (3) worldwide applicability.

- Adding spatial information from adjacent pixels on top of temporal information further improved land cover classification results (+2.1 % in MA and +0.018 in Kappa). In particular, the visual appearance of maps was improved by reducing salt-and-pepper effects. Spatial information is therefore essential for pixel-based classifications, given that no postprocessing is applied.

- All machine learning classifier generally achieved satisfactory results under the varying input scenarios (MA > 70 % and Kappa > 0.6). RF averaged the best performance (MA 83.2 % and Kappa 0.738), closely followed by SVM (MA 79.3 % and Kappa 0.735); in contrast, the DNN classifier performed considerably worse (MA 71.1 % and 0.644 Kappa). However, its susceptibility to overfitting in combination with quality deficiencies in the training data may have had a detrimental effect on the performance of deep learning. Moreover, the conventional machine learning classifiers offered enough model complexity for the tested inputs, while DNNs presumably offer advances for more complex input spaces.

- The six land cover classes, artificial areas, non-forest vegetation, forest vegetation, bare land, water and snow & ice are well separable from each other with Landsat-based spatiotemporal information (UAs and PAs > 77 %). Confusions may occur, however, between classes that often coexist (e.g. forest and non-forest vegetation). Caution should also be exercised in the context of medium resolution imagery such as Landsat (30 m) or Sentinel-2 (10–60 m) when classifying land covers that are characterized by sub-pixel information.

- Adding training samples increased land cover classification accuracy particularly for small training datasets (< 100 samples per class). We therefore recommend to acquire at least 100 training samples for broad land cover classes. Importantly, though, the benefit of additional training samples decreases considerably for large training datasets (> 100 samples per class). Considering acquiring training samples is expensive and time consuming, we therefore argue that collecting considerably more than 100 training samples per class is not a worthwhile endeavor. Finally, we would also like to stress that determining the adequate number of training samples may vary since its dependent on many factors including classifier.

- Landsat data was successfully augmented with Sentinel-2 data to obtain denser time series of observations. Classification results, however, improved only marginally by using extracted features from these augmented time series instead of time series consisting exclusively of Landsat data. We assume that lack of complexity in the time series model limited the extraction of temporal information from the augmented time series.

In spite of the encouraging classification results achieved in this thesis, there are many more opportunities to further improve land cover mapping. The three most obvious opportunities for follow-up research are listed below:

- Regarding the extraction of classifiable information from time series, the lack of complexity of our time series model is a limiting factor for temporal feature extraction from dense time series. Future research should therefore investigate extracting features using sophisticated time series reconstruction (or gap-filling) techniques in order to minimize the loss in temporal information. Examples of state-of-the-art reconstruction techniques are presented in Julien and Sobrino (2019). Advances in this field are particularly relevant to leverage time series from the synthetic constellation of Landsat 8 and Sentinel-2A and 2B.

- Future studies should also use temporal features from the developed time series model in conjunction with more detailed land cover data. Particularly interesting is the differentiation between vegetation types based on differences in phenological traits. In order to facilitate this endeavour, we provide a ready-to-use implementation of the time series model (see Section A.3).

- Current trends in land cover mapping also appear to improve classifier performance by adding ancillary data (Khatami et al., 2016). For example, Sentinel-2 optical data was fused with Sentinel-1 synthetic aperture radar data to increase mapping accuracy (Clerici et al., 2017). Since Sentinel-1 is also available in GEE, it could be easily integrated in the existing workflow. The augmented feature space would also be of interest in combination with deep learning, since input space complexity may have been a limiting factor in this work. Recent classification results based on both remote sensing technologies are promising (Kussul et al., 2017).

# Bibliography

Adams, J. B., D. E. Sabol, V. Kapos, R. Almeida Filho, D. A. Roberts, M. O. Smith, and A. R. Gillespie (May 1995). "Classification of multispectral images based on fractions of endmembers: Application to land-cover change in the Brazilian Amazon." *Remote Sensing of Environment* 52.2, pp. 137–154.

Azzari, G. and D. Lobell (Dec. 2017). "Landsat-based classification in the cloud: An opportunity for a paradigm shift in land cover monitoring." *Remote Sensing of Environment* 202, pp. 64–74.

Breiman, L. (Oct. 2001). "Random Forests." *Machine Learning* 45.1, pp. 5–32.

Cadenasso, M. L., S. T. A. Pickett, and K. Schwarz (2007). "Spatial heterogeneity in urban ecosystems: reconceptualizing land cover and a framework for classification." *Frontiers in Ecology and the Environment* 5.2, pp. 80–88.

Chang, C.-C. and C.-J. Lin (May 2011). "LIBSVM: A Library for Support Vector Machines." *ACM Trans. Intell. Syst. Technol.* 2.3, 27:1–27:27.

Clerici, N., C. A. V. Calderón, and J. M. Posada (Nov. 2017). "Fusion of Sentinel-1A and Sentinel-2A data for land cover mapping: a case study in the lower Magdalena region, Colombia." *Journal of Maps* 13.2, pp. 718–726.

Cohen, J. (Apr. 1960). "A Coefficient of Agreement for Nominal Scales." *Educational and Psychological Measurement* 20.1, pp. 37–46.

Dozier, J. (Apr. 1989). "Spectral signature of alpine snow cover from the landsat thematic mapper." *Remote Sensing of Environment* 28, pp. 9–22.

Duro, D. C., N. C. Coops, M. A. Wulder, and T. Han (June 2007). "Development of a large area biodiversity monitoring system driven by remote sensing." *Progress in Physical Geography: Earth and Environment* 31.3, pp. 235–260.

Feyisa, G. L., H. Meilby, R. Fensholt, and S. R. Proud (Jan. 2014). "Automated Water Extraction Index: A new technique for surface water mapping using Landsat imagery." *Remote Sensing of Environment* 140, pp. 23–35.

Fisher, A., N. Flood, and T. Danaher (Mar. 2016). "Comparing Landsat water index methods for automated water classification in eastern Australia." *Remote Sensing of Environment* 175, pp. 167–182.

Foga, S., P. L. Scaramuzza, S. Guo, Z. Zhu, R. D. Dilley, T. Beckmann, G. L. Schmidt, J. L. Dwyer, M. Joseph Hughes, and B. Laue (June 2017). "Cloud detection algorithm comparison and validation for operational Landsat data products." *Remote Sensing of Environment* 194, pp. 379–390.

Franklin, S. E., O. S. Ahmed, M. A. Wulder, J. C. White, T. Hermosilla, and N. C. Coops (July 2015). "Large Area Mapping of Annual Land Cover Dynamics Using Multitemporal Change

Detection and Classification of Landsat Time Series Data." *Canadian Journal of Remote Sensing* 41.4, pp. 293–314.

Gebhardt, S., T. Wehrmann, M. A. M. Ruiz, P. Maeda, J. Bishop, M. Schramm, R. Kopeinig, O. Cartus, J. Kellndorfer, R. Ressl, L. A. Santos, and M. Schmidt (May 2014). "MAD-MEX: Automatic Wall-to-Wall Land Cover Monitoring for the Mexican REDD-MRV Program Using All Landsat Data." *Remote Sensing* 6.5, pp. 3923–3943.

Gorelick, N., M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore (Dec. 2017). "Google Earth Engine: Planetary-scale geospatial analysis for everyone." *Remote Sensing of Environment* 202, pp. 18–27.

Griffiths, P., S. v. d. Linden, T. Kuemmerle, and P. Hostert (Oct. 2013). "A Pixel-Based Landsat Compositing Algorithm for Large Area Land Cover Mapping." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 6.5, pp. 2088–2101.

Guide, P (2018). *Landsat 8 surface reflectance code (LaSRC) product*. URL: `https://landsat.usgs.gov/sites/default/files/documents/lasrc_product_guide.pdf` (visited on 09/23/2019).

Gómez, C., J. C. White, and M. A. Wulder (June 2016). "Optical remotely sensed time series data for land cover classification: A review." *ISPRS Journal of Photogrammetry and Remote Sensing* 116, pp. 55–72.

Hansen, M. C., P. V. Potapov, R. Moore, M. Hancher, S. A. Turubanova, A. Tyukavina, D. Thau, S. V. Stehman, S. J. Goetz, T. R. Loveland, A. Kommareddy, A. Egorov, L. Chini, C. O. Justice, and J. R. G. Townshend (Nov. 2013). "High-Resolution Global Maps of 21st-Century Forest Cover Change." *Science* 342.6160, pp. 850–853.

He, J., J. R. Harris, M. Sawada, and P. Behnia (Apr. 2015). "A comparison of classification algorithms using Landsat-7 and Landsat-8 data for mapping lithology in Canada's Arctic." *International Journal of Remote Sensing* 36.8, pp. 2252–2276.

Hearst, M. A. (July 1998). "Support Vector Machines." *IEEE Intelligent Systems* 13.4, pp. 18–28.

Heydari, S. S. and G. Mountrakis (Jan. 2018). "Effect of classifier selection, reference sample size, reference class distribution and scene heterogeneity in per-pixel classification accuracy using 26 Landsat sites." *Remote Sensing of Environment* 204, pp. 648–658.

Huang, X., Q. Lu, L. Zhang, and A. Plaza (Nov. 2014). "New Postprocessing Methods for Remote Sensing Image Classification: A Systematic Study." *IEEE Transactions on Geoscience and Remote Sensing* 52.11, pp. 7140–7159.

Jia, K., X. Wei, X. Gu, Y. Yao, X. Xie, and B. Li (Nov. 2014). "Land cover classification using Landsat 8 Operational Land Imager data in Beijing, China." *Geocarto International* 29.8, pp. 941–951.

Julien, Y. and J. A. Sobrino (Apr. 2019). "Optimizing and comparing gap-filling techniques using simulated NDVI time series from remotely sensed global data." *International Journal of Applied Earth Observation and Geoinformation* 76, pp. 93–111.

Khatami, R., G. Mountrakis, and S. V. Stehman (May 2016). "A meta-analysis of remote sensing research on supervised pixel-based land-cover image classification processes: General guidelines for practitioners and future research." *Remote Sensing of Environment* 177, pp. 89–100.

Kohavi, R. (Jan. 1995). "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection." Vol. 14.

Kuhn, M. and K. Johnson (2013). "Over-Fitting and Model Tuning." *Applied Predictive Modeling*. Ed. by M. Kuhn and K. Johnson. New York, NY: Springer New York, pp. 61–92.

Kussul, N., M. Lavreniuk, S. Skakun, and A. Shelestov (May 2017). "Deep Learning Classification of Land Cover and Crop Types Using Remote Sensing Data." *IEEE Geoscience and Remote Sensing Letters* 14.5, pp. 778–782.

Landis, J. R. and G. G. Koch (Mar. 1977). "The measurement of observer agreement for categorical data." *Biometrics* 33.1, pp. 159–174.

Landmann, T., D. Eidmann, N. Cornish, J. Franke, and S. Siebert (Nov. 2019). "Optimizing harmonics from Landsat time series data: the case of mapping rainfed and irrigated agriculture in Zimbabwe." *Remote Sensing Letters* 10.11, pp. 1038–1046.

LeCun, Y., Y. Bengio, and G. Hinton (May 2015). "Deep learning." *Nature* 521.7553, pp. 436–444.

Lee, S. and B. Pradhan (Mar. 2007). "Landslide hazard mapping at Selangor, Malaysia using frequency ratio and logistic regression models." *Landslides* 4.1, pp. 33–41.

Li, C., J. Wang, L. Wang, L. Hu, and P. Gong (Feb. 2014). "Comparison of Classification Algorithms and Training Sample Sizes in Urban Land Classification with Landsat Thematic Mapper Imagery." *Remote Sensing* 6.2, pp. 964–983.

Li, J. and D. P. Roy (Sept. 2017). "A Global Analysis of Sentinel-2A, Sentinel-2B and Landsat-8 Data Revisit Intervals and Implications for Terrestrial Monitoring." *Remote Sensing* 9.9, p. 902.

Liu, J., J. Heiskanen, E. Aynekulu, E. E. Maeda, and P. K. E. Pellikka (May 2016). "Land Cover Characterization in West Sudanian Savannas Using Seasonal Features from Annual Landsat Time Series." *Remote Sensing* 8.5, p. 365.

Main-Knorn, M., B. Pflug, J. Louis, V. Debaecker, U. Müller-Wilm, and F. Gascon (Oct. 2017). "Sen2Cor for Sentinel-2." *Image and Signal Processing for Remote Sensing XXIII*. Vol. 10427. International Society for Optics and Photonics, p. 1042704.

Mandanici, E. and G. Bitelli (Dec. 2016). "Preliminary Comparison of Sentinel-2 and Landsat 8 Imagery for a Combined Use." *Remote Sensing* 8.12, p. 1014.

Maxwell, S. K., G. L. Schmidt, and J. C. Storey (Dec. 2007). "A multi-scale segmentation approach to filling gaps in Landsat ETM+ SLC-off images." *International Journal of Remote Sensing* 28.23, pp. 5339–5356.

Owen, A. B. (2007). "A robust hybrid of lasso and ridge regression." *Contemporary Mathematics*. Ed. by J. S. Verducci, X. Shen, and J. Lafferty. Vol. 443. Providence, Rhode Island: American Mathematical Society, pp. 59–71.

Pekel, J.-F., A. Cottam, N. Gorelick, and A. S. Belward (Dec. 2016). "High-resolution mapping of global surface water and its long-term changes." *Nature* 540.7633, pp. 418–422.

Petitjean, F., J. Inglada, and P. Gancarski (Aug. 2012). "Satellite Image Time Series Analysis Under Time Warping." *IEEE Transactions on Geoscience and Remote Sensing* 50.8, pp. 3081–3095.

Phiri, D. and J. Morgenroth (Sept. 2017). "Developments in Landsat Land Cover Classification Methods: A Review." *Remote Sensing* 9.9, p. 967.

Picterra, S. (2017). *Feasibility study Arealstatistik 2020: Review of state-of-the-art, tests on data & recommendations*. URL: `https://www.bfs.admin.ch/bfs/de/home/statistiken/raum-umwelt/erhebungen/area.assetdetail.5687737.html` (visited on 09/19/2019).

Pontius, R. G. and M. Millones (Aug. 2011). "Death to Kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment." *International Journal of Remote Sensing* 32.15, pp. 4407–4429.

Poortinga, A., K. Tenneson, A. Shapiro, Q. Nquyen, K. San Aung, F. Chishtie, and D. Saah (Jan. 2019). "Mapping Plantations in Myanmar by Fusing Landsat-8, Sentinel-2 and Sentinel-1 Data along with Systematic Error Quantification." *Remote Sensing* 11.7, p. 831.

Rodriguez-Galiano, V. F., B. Ghimire, J. Rogan, M. Chica-Olmo, and J. P. Rigol-Sanchez (Jan. 2012). "An assessment of the effectiveness of a random forest classifier for land-cover classification." *ISPRS Journal of Photogrammetry and Remote Sensing* 67, pp. 93–104.

Roy, D. P., V. Kovalskyy, H. K. Zhang, E. F. Vermote, L. Yan, S. S. Kumar, and A. Egorov (Nov. 2016). "Characterization of Landsat-7 to Landsat-8 reflective wavelength and normalized difference vegetation index continuity." *Remote Sensing of Environment*. Landsat 8 Science Results 185, pp. 57–70.

Roy, D. P., J. Ju, C. Mbow, P. Frost, and T. Loveland (June 2010). "Accessing free Landsat data via the Internet: Africa's challenge." *Remote Sensing Letters* 1.2, pp. 111–117.

Salomonson, V. V and I Appel (Feb. 2004). "Estimating fractional snow cover from MODIS using the normalized difference snow index." *Remote Sensing of Environment* 89.3, pp. 351–360.

Schmidt, G., C. B. Jenkerson, J. Masek, E. Vermote, and F. Gao (2013). *Landsat ecosystem disturbance adaptive processing system (LEDAPS) algorithm description*. USGS Numbered Series 2013-1057. Reston, VA: U.S. Geological Survey, p. 27.

Schneider, A. (Sept. 2012). "Monitoring land cover change in urban and peri-urban areas using dense time stacks of Landsat satellite data and a data mining approach." *Remote Sensing of Environment* 124, pp. 689–704.

Shao, Y. and R. S. Lunetta (June 2012). "Comparison of support vector machine, neural network, and CART algorithms for the land-cover classification using limited training data points." *ISPRS Journal of Photogrammetry and Remote Sensing* 70, pp. 78–87.

Smith, S. L., P.-J. Kindermans, C. Ying, and Q. V. Le (Nov. 2017). "Don't Decay the Learning Rate, Increase the Batch Size." *arXiv:1711.00489 [cs, stat]*. arXiv: 1711.00489.

Stehman, S. V. (Oct. 1997). "Selecting and interpreting measures of thematic classification accuracy." *Remote Sensing of Environment* 62.1, pp. 77–89.

Swiss Federal Statistical Office (2016). *Arealstatistik der Schweiz*. URL: `https://www.bfs.admin.ch/bfs/de/home/statistiken/raum-umwelt/erhebungen/area.assetdetail.6813.html` (visited on 09/19/2019).

Tadono, T., H. Ishida, F. Oda, S. Naito, K. Minakawa, and H. Iwamoto (2014). "Precise global DEM generation by ALOS PRISM." *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 2.4, p. 71.

Thanh Noi, P. and M. Kappas (Jan. 2018). "Comparison of Random Forest, k-Nearest Neighbor, and Support Vector Machine Classifiers for Land Cover Classification Using Sentinel-2 Imagery." *Sensors* 18.1, p. 18.

Verbesselt, J., A. Zeileis, and M. Herold (Aug. 2012). "Near real-time disturbance detection using satellite image time series." *Remote Sensing of Environment* 123, pp. 98–108.

Werbos, P. J. (Oct. 1990). “Backpropagation through time: what it does and how to do it.” *Proceedings of the IEEE* 78.10, pp. 1550–1560.

White, J. C., M. A. Wulder, G. W. Hobart, J. E. Luther, T. Hermosilla, P. Griffiths, N. C. Coops, R. J. Hall, P. Hostert, A. Dyk, and L. Guindon (May 2014). “Pixel-Based Image Compositing for Large-Area Dense Time Series Applications and Science.” *Canadian Journal of Remote Sensing* 40.3, pp. 192–212.

Wilson, B. T., J. F. Knight, and R. E. McRoberts (Mar. 2018). “Harmonic regression of Landsat time series for modeling attributes from national forest inventory data.” *ISPRS Journal of Photogrammetry and Remote Sensing* 137, pp. 29–46.

Woodcock, C. E., R. Allen, M. Anderson, A. Belward, R. Bindschadler, W. Cohen, F. Gao, S. N. Goward, D. Helder, E. Helmer, R. Nemani, L. Oreopoulos, J. Schott, P. S. Thenkabail, E. F. Vermote, J. Vogelmann, and M. a. W. Wulder (2008). “Free Access to Landsat Imagery.” *SCIENCE VOL 320 :1011*.

Wulder, M. A., T. Hilker, J. C. White, N. C. Coops, J. G. Masek, D. Pflugmacher, and Y. Crevier (Dec. 2015). “Virtual constellations for global terrestrial monitoring.” *Remote Sensing of Environment* 170, pp. 62–76.

Wulder, M. A., N. C. Coops, D. P. Roy, J. C. White, and T. Hermosilla (June 2018). “Land cover 2.0.” *International Journal of Remote Sensing* 39.12, pp. 4254–4284.

Yamazaki, D., M. A. Trigg, and D. Ikeshima (Dec. 2015). “Development of a global ~90m water body map using multi-temporal Landsat images.” *Remote Sensing of Environment* 171, pp. 337–351.

Yuan, F., K. E. Sawaya, B. C. Loeffelholz, and M. E. Bauer (Oct. 2005). “Land cover classification and change analysis of the Twin Cities (Minnesota) Metropolitan Area by multitemporal Landsat remote sensing.” *Remote Sensing of Environment* 98.2, pp. 317–328.

Zhang, H. K., D. P. Roy, L. Yan, Z. Li, H. Huang, E. Vermote, S. Skakun, and J.-C. Roger (Sept. 2018). “Characterization of Sentinel-2A and Landsat-8 top of atmosphere, surface, and nadir BRDF adjusted reflectance and NDVI differences.” *Remote Sensing of Environment* 215, pp. 482–494.

Zhang, L., L. Zhang, and B. Du (June 2016). “Deep Learning for Remote Sensing Data: A Technical Tutorial on the State of the Art.” *IEEE Geoscience and Remote Sensing Magazine* 4.2, pp. 22–40.

Zhu, X. X., D. Tuia, L. Mou, G. Xia, L. Zhang, F. Xu, and F. Fraundorfer (Dec. 2017). “Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources.” *IEEE Geoscience and Remote Sensing Magazine* 5.4, pp. 8–36.

Zhu, Z. and C. E. Woodcock (Mar. 2012). “Object-based cloud and cloud shadow detection in Landsat imagery.” *Remote Sensing of Environment* 118, pp. 83–94.

– (Mar. 2014). “Continuous change detection and classification of land cover using all available Landsat data.” *Remote Sensing of Environment* 144, pp. 152–171.

Zhu, Z., C. E. Woodcock, and P. Olofsson (July 2012). “Continuous monitoring of forest disturbance using all available Landsat imagery.” *Remote Sensing of Environment*. Landsat Legacy Special Issue 122, pp. 75–91.

Zhu, Z., C. E. Woodcock, C. Holden, and Z. Yang (June 2015). “Generating synthetic Landsat images based on all available Landsat data: Predicting Landsat surface reflectance at any given time.” *Remote Sensing of Environment* 162, pp. 67–83.

# Appendices

## A.1 Figures

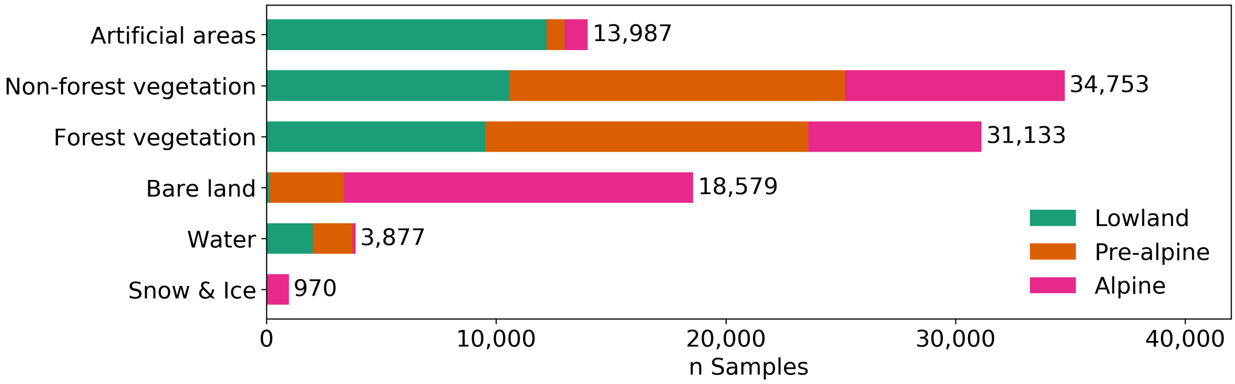

FIGURE A.1: Summary statistics of the land cover reference data derived from the Arealstatistik 2013/18 dataset.

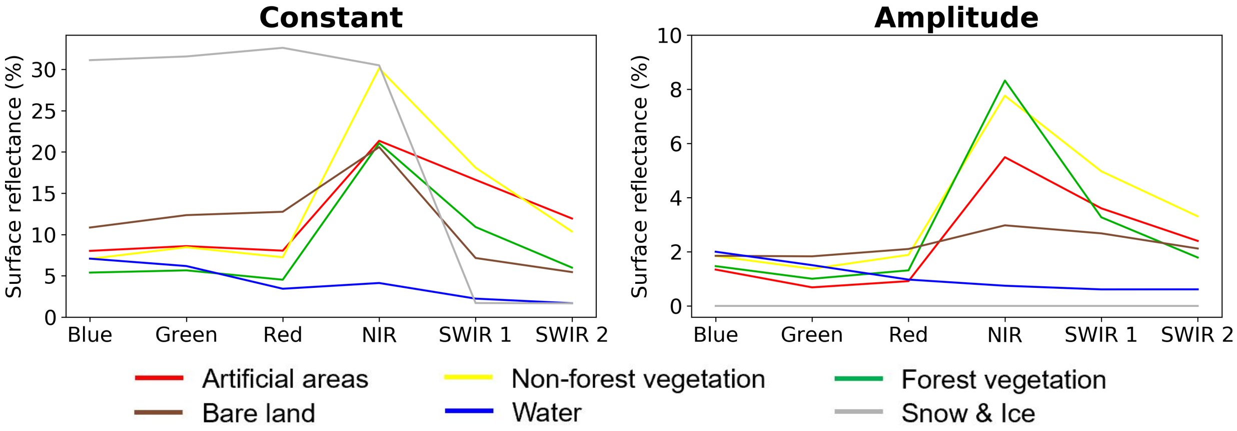

FIGURE A.2: Class-wise medians of the time series constants (left) and amplitudes (right) for the six spectral bands (Arealstatistik 2013/18).

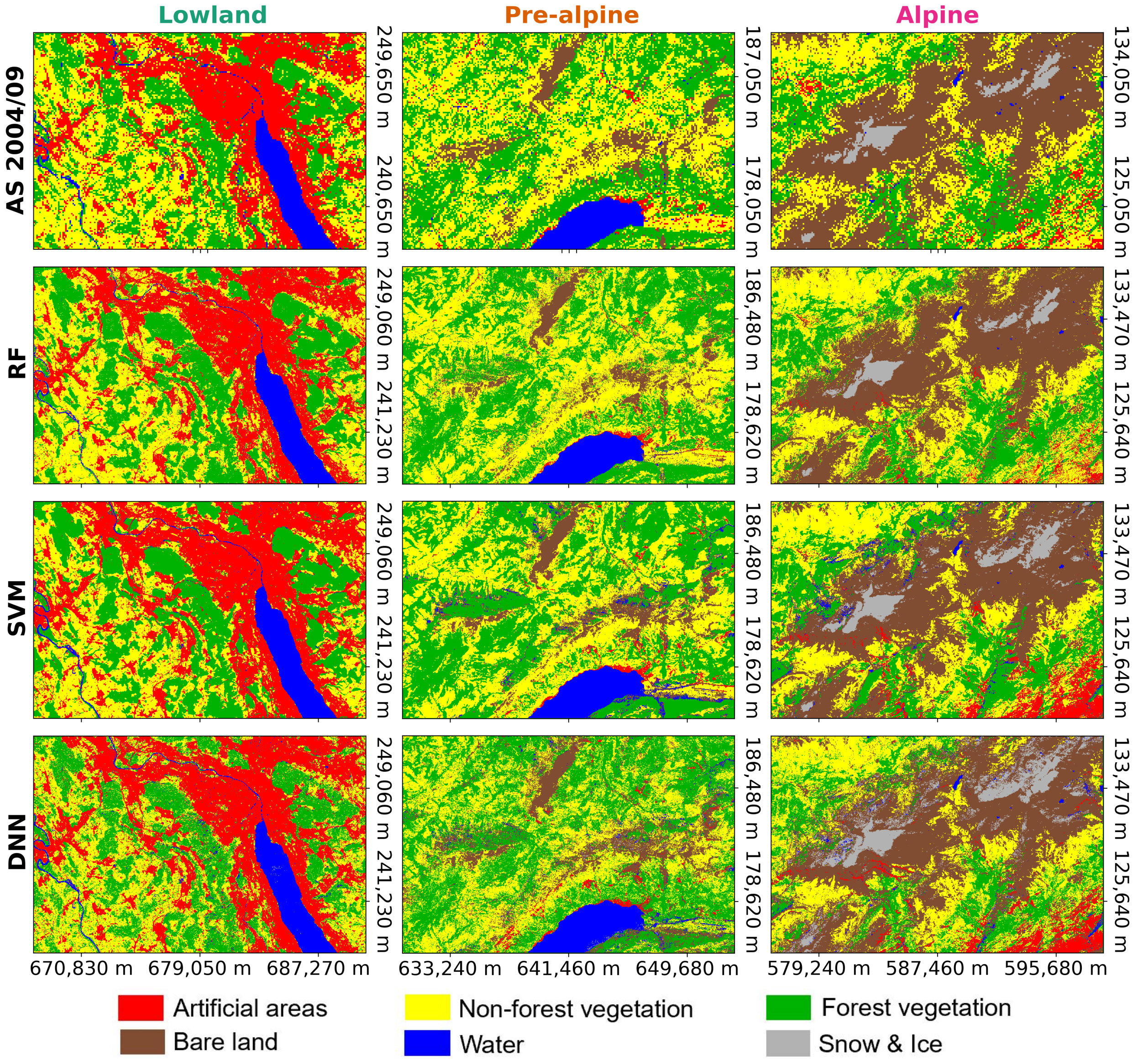

FIGURE A.3: Visual comparison of land cover maps for the lowland (left column), pre-alpine (mid column) and alpine (right column) region. The first row shows the ground truth in the form of Arealstatistik (AS) 2004/09-derived land cover maps, and the following three rows show the classification results for Random Forest (RF), Support Vector Machine (SVM) and Deep Neural Network (DNN) with the composite input.

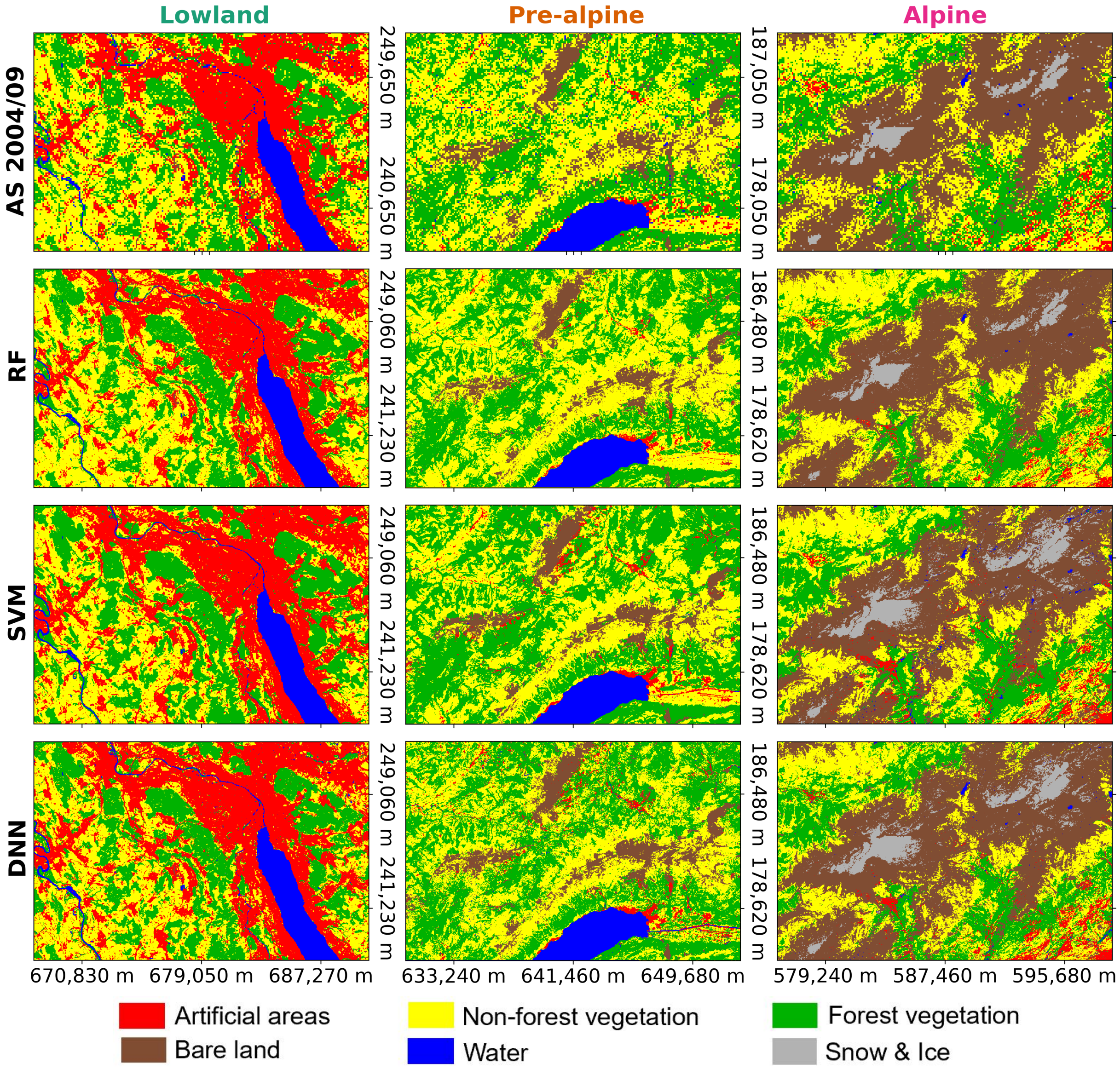

FIGURE A.4: Visual comparison of land cover maps for the lowland (left column), pre-alpine (mid column) and alpine (right column) region. The first row shows the ground truth in the form of Arealstatistik (AS) 2004/09-derived land cover maps, and the following three rows show the classification results for Random Forest (RF), Support Vector Machine (SVM) and Deep Neural Network (DNN) with the time series input.

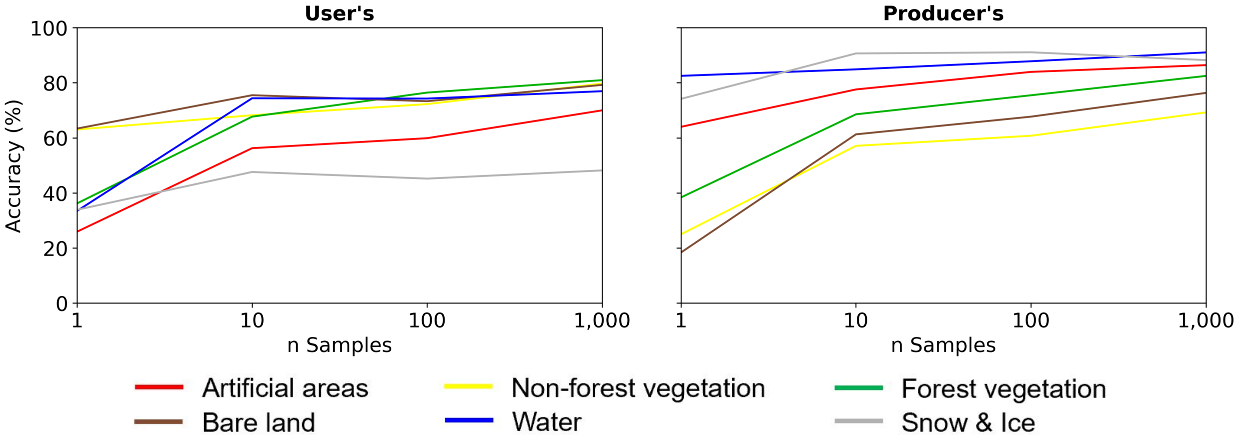

FIGURE A.5: Relationship between the number of training samples (x-axis) and the user's (left) and producer's (right) accuracies (y-axis) for different land cover classes. Results were obtained using Support Vector Machine with the time series-spatial input and 5-fold cross-validation.

## A.2 Tables

TABLE A.1: Aggregated confusion matrix (5-fold cross-validation) including User's Accuracies (UAs) and Producer's Accuracies (PAs) for Support Vector Machine with the time series-spatial input. Correct classifications (diagonal) are boldfaced.

| Classified as (pixels) | | Reference data (pixels) | | | | | | | UA (%) | PA (%) |
|---|---|---|---|---|---|---|---|---|---|---|
| | | AA | NFV | FV | BL | W | SI | Total | | |
| Artificial areas | AA | **2,405** | 422 | 164 | 62 | 17 | 0 | 3,070 | 78.3 | 88.5 |
| Non-forest veg. | NFV | 197 | **5,605** | 514 | 400 | 14 | 15 | 6,745 | 83.1 | 77.3 |
| Forest veg. | FV | 85 | 814 | **5,146** | 111 | 20 | 0 | 6,176 | 83.3 | 86.7 |
| Bare land | BL | 16 | 396 | 85 | **3,136** | 20 | 49 | 3,702 | 84.7 | 83.0 |
| Water | W | 16 | 18 | 29 | 15 | **708** | 0 | 786 | 90.1 | 90.9 |
| Snow & Ice | SI | 0 | 0 | 0 | 56 | 0 | **165** | 221 | 74.7 | 72.1 |
| | Total | 2,719 | 7,255 | 5,938 | 3,780 | 779 | 229 | | | |

TABLE A.2: Aggregated confusion matrix (5-fold cross-validation) including User's Accuracies (UAs) and Producer's Accuracies (PAs) for Deep Neural Network with the time series-spatial input. Correct classifications (diagonal) are boldfaced.

| Classified as (pixels) | | Reference data (pixels) | | | | | | | UA (%) | PA (%) |
|---|---|---|---|---|---|---|---|---|---|---|
| | | AA | NFV | FV | BL | W | SI | Total | | |
| Artificial areas | AA | **2,245** | 519 | 209 | 135 | 27 | 0 | 3,135 | 71.6 | 82.6 |
| Non-forest veg. | NFV | 308 | **4,362** | 1,118 | 308 | 15 | 0 | 6,111 | 71.4 | 60.1 |
| Forest veg. | FV | 109 | 1,716 | **4,391** | 101 | 27 | 0 | 6,344 | 69.2 | 74.0 |
| Bare land | BL | 36 | 634 | 187 | **2,936** | 20 | 13 | 3,826 | 76.7 | 77.7 |
| Water | W | 20 | 23 | 32 | 28 | **685** | 1 | 789 | 86.8 | 88.0 |
| Snow & Ice | SI | 0 | 0 | 0 | 273 | 4 | **214** | 491 | 43.6 | 93.9 |
| | Total | 2,718 | 7,254 | 5,937 | 3,781 | 778 | 228 | | | |

TABLE A.3: Land cover nomenclature of principal domains and basic categories used by the Swiss land use statistics.

| Principal domains | Basic categories |
|---|---|
| Artificial areas | Consolidated surfaces<br>Buildings<br>Greenhouses<br>Gardens with border and patch structures<br>Lawns<br>Trees in artificial areas<br>Mix of small structures |
| Grass and herb vegetation | Grass and herb vegetation |
| Brush vegetation | Shrubs<br>Brush meadows<br>Short-stem fruit trees<br>Vines<br>Permanent garden plants and brush crops |
| Tree vegetation | Closed forest<br>Forest edges<br>Forest strips<br>Open forest<br>Brush forest<br>Linear woods<br>Clusters of trees |
| Bare land | Solid rock<br>Granular soil<br>Rocky areas |
| Watery areas | Water<br>Glacier, perpetual snow<br>Wetlands<br>Reedy marshes |

## A.3 Scripts

The GEE implementation of the time series model can be found at https://code.earthengine.google.com/83137f976916b2d28b4871461ba9247a. The time series model's qualitative assessment (https://code.earthengine.google.com/e69d703927c355af89419cb9aac6639f) and quantitative assessment (https://code.earthengine.google.com/85bf6f51877e7092bfaec033f088ee82) are also available in GEE. All other non-GEE code can be publicly accessed on the following *GitHub* repository: https://github.com/SebastianHafner/masterthesis.

# Acknowledgements

First, I would like to express my gratitude to Prof. Dr. Michael Schaepman for giving me the opportunity to carry out this research. My sincere thanks go to Dr. Hendrik Wulf who supervised this thesis, as well as to Dr. Charlotte Steinmeier for her co-supervision. My thanks also go to all my colleagues and friends. Last but not least I would like to express my special gratitude to my parents.

# Personal Declaration

I hereby declare that the submitted thesis is the result of my own, independent work. All external sources are explicitly acknowledged in the thesis.

Date
September 30, 2019

Signature
Sebastian Hafner S. Hafner