Certificat Complémentaire en Géomatique 2022
Université de Genève

A feasibility study of Machine Learning methods applied to Sentinel-2 imagery for classification of Land Cover in Switzerland

Isabel NICHOLSON THOMAS
Mémoire

Sous la direction du Professeur Dr Gregory Giuliani
Affiliation UNIGE ISE/GRID

UNIVERSITÉ DE GENÈVE

Table of Contents

Table of Figures

Table of Tables

Table of Abbreviations

ARD	Analysis-Ready Data
CLC	CORINE Land Cover
CNN	Convolutional Neural Network
DEM	Digital Elevation Model
EO	Earth Observation
LC	Land Cover
LU	Land Use
ML	Machine Learning
MLR	Multinomial Linear Regression
NDBI	Normalised Differentiated Building Index
NDVI	Normalised Differentiated Vegetation Index
NDWI	Normalised Differentiated Water Index
NIR	Near Infra-Red
OFS	Office Fédérale de la Statistique (Federal Office of Statistics)
RF	Random Forests
SDG	Sustainable Development Goal
SITS	Satellite Image Time Series
SMOTE	Synthetic Minority Over-sampling Technique
SWIR	Short-Wave Infra-Red
SVM	Support Vector Machines

Summary

Land Cover (LC) is a key environmental data variable and the study of changes in LC requires accurate, regular data that represents the spatial and temporal scales of potential LC change. However, operational LC products for Switzerland are currently limited in their spatio-temporal resolution. Continued development of research in Earth Observation and Machine Learning has shown that classification of satellite imagery is a feasible option for the production of LC datasets, but there are still a number of approaches to choose from. The aim of this study was to provide insight into the feasibility of using of ML classification of satellite imagery to produce LC data for Switzerland, based on a subset of data from Western Switzerland. More specifically, the objective was to compare the accuracy of space-first and time-first approaches to classifying satellite imagery, as well as the impact of methods for improving model performance and inclusion of additional input variables. The accuracy of using a space-first approach was lower compared to that using the time-first Satellite Image Time Series (SITS) approach. In all iterations, model performance was lowest for LC classes representing Brush Vegetation. Methods to counter the underrepresentation of minority classes were ineffective, and the greatest model improvement came from the addition of DEM auxiliary variables. The overall performance of the various supervised classifications is comparable to current operational land cover datasets, however performance for minority classes may present an obstacle to their use in studies of LC change, considering the relatively subtle rates of change occurring in some LC classes in Switzerland. Continued development of the SITS methodology to test its transferability over time and expansion of testing to other bioregions in Switzerland will be necessary to produce an operational annual land cover dataset which can be used at the national scale.

1. Introduction

Land Cover (LC) is a key environmental data variable representing the (bio)physical features covering the Earth surface (Di Gregorio & Jansen, 2005), necessary for a wide variety of applications including urban planning, vegetation and agricultural monitoring, and modelling ecosystem services. LC and LC change play an important role in the assessment of multiple Sustainable Development Goals (SDGs). Changes in LC are often indicative of social, economic and political drivers at local, national and international scales, and can reflect conflicting states of stability or rapid transformation co-occurring across small spatial distances (Gómez et al., 2016). The availability of accurate, reliable, and timely LC data is therefore crucial for understanding and modelling environmental processes on policy relevant time-scales (Verde et al., 2020).

Currently, LC and Land Use (LU) data for Switzerland is produced by the Federal Office of Statistics (OFS), is available at a resolution of 100m, and covers the periods of 1979-1985, 1992-1997, 2004-2009, and 2013-2018 in 4 distinct datasets. These data, known as the *ArealStatistik,* are useful in terms of their high thematic resolution and their tailored categories which cover the specific features of Swiss landscapes. Even so, the low update frequency and relatively coarse spatial resolution are at odds with the data needs for quantifying variables which are known to be as dynamic and spatially variable as LC and LU (Ban et al., 2015). The ability to analyse LC maps on an annual, or sub-annual, basis would facilitate greater understanding of the processes and environmental pressures driving LC change (Kennedy et al., 2014; Teixeira et al., 2014), and allow for consideration of both subtle and rapid changes which can indicate diverse pressures on LC. However, the criteria for operational LC products are demanding. Such datasets need to be reproducible and automatable, meet requirements for spatial continuity and temporal coherence between map updates, and easily adapt to changes in nomenclature to ensure their continued relevance for policymakers and researchers (Inglada et al., 2017).

Within the context of increasing open access data, analysis ready data (ARD) formats, and capacity improvements for data processing, the state-of-the-art of LC mapping has greatly advanced in recent years (Wulder et al., 2018), keeping pace with increasing demands for LC data from research communities(Gómez et al., 2016) . Recent efforts to downscale LC data for Switzerland have improved the spatial resolution from 100m to 25 m using an algorithmic approach, however, the temporal resolution of datasets produced remains insufficient (Giuliani et al., 2022). The increasing availability of open-access high-resolution remote sensing data has vastly increased the potential for the development of land cover datasets from local to global scales (Ban et al., 2015). Satellite imagery provides a consistent dataset of earth observations which is spatially continuous and contains the temporal resolution necessary to identify classes with strong temporal dynamics (Inglada et al., 2017; Verde et al., 2020). Use of Machine Learning (ML) algorithms, with their ability to cope with high-dimensional data and map classes with complex characteristics, provides an automated approach to classification (Maxwell et al., 2018). Even so, ML is not yet widely implemented in the production of operational Land Cover data for Switzerland, with its use currently limited to partial automation of classification of aerial photography of the *ArealStatistik* surveys (OFS, 2022a).

The aim of this study is to provide insight into the feasibility of using of ML classification of satellite imagery to produce an annual LC dataset for Switzerland. More specifically, the objective is to compare the accuracy of space-first and time-first approaches to classifying satellite imagery, as well as the impact of methods for improving model performance and inclusion of additional input variables.

2. Theoretical concepts mobilised

2.1. Analysis-Ready Data and Data Cubes

Whilst availability of remote sensing data has greatly improved, the accessibility of its use and analysis remains a limitation in some cases. Obstacles to the use of remote sensing data include the pre-processing steps such as radiometric and atmospheric correction, as well as increasing data volumes which limit the feasibility of downloading EO data to desktop environments when the aim is to study large spatial, multi-temporal datasets (Szantoi et al., 2020). The standard of Analysis-Ready Data (ARD) has been developed to counter the limitations presented by complex pre-processing and provides data processed to a minimum level to facilitate its use. The Committee on Earth Observation Satellites (CEOS) defines ARD as “satellite data that have been processed to a minimum set of requirements and organized into a form that allows immediate analysis with a minimum of additional user effort and interoperability both through time and with other datasets” (CEOS, 2022). ARD can give users access to data which is already aggregated in space and time in analysis appropriate ways and allows the circumnavigation of typical problems such as cloud cover (Potapov et al., 2020) which can affect the use of these images in LC classification due to inconsistencies (Inglada et al., 2017).

The Swiss Data Cube (Chatenoux et al., 2021) provides ARD covering Switzerland since 1984, providing the spatial and temporal continuity necessary for national land cover mapping efforts with a full catalogue of Landsat and Sentinel ARD and additional derived products (The Swiss Data Cube, 2017). Data Cubes allow users to counteract the ‘Big Data’ challenges posed by large volumes of EO data, by providing “an architecture allowing a time-series multi-dimensional stack of spatially aligned pixels” (Giuliani et al., 2017, pg. 103) with which the user can directly start analysis. Switzerland therefore has a feasible, low-cost, data source for consistent land cover map production, provided an appropriate methodology can be applied.

2.2. Operational Land Cover maps

Beyond a single snapshot of LC, operational LC maps are data products which are accurate, reliable, and can be produced in accordance with a pre-defined schedule at regular time intervals (Inglada et al., 2017). Various operational LC classification systems and spatial data products have been developed by national and international agencies, and Table 1 gives an overview of such products which cover Switzerland. Most of these LC products are static, providing a discrete snapshot of LC for a particular year, however Google's Dynamic World dataset provides near real-time LC classification of Sentinel-2 imagery with an updated classification every 2-5 days (Brown et al., 2022). The production of high-resolution LC data is one of the central aims of the Sentinel-2 mission, and indeed it is now widely used for this in many regions (Phiri et al., 2020).

Whilst recent releases of global land cover datasets have improved spatial and temporal resolution, the implementation of these datasets at the local scale is often limited by their low thematic resolution, and the lack of training data on the global scale. The exact impact of this on classification accuracy remains a gap in research (Inglada et al., 2017), but development of LC data is increasingly focused on the production of improved LC data for small-scale applications. Standardised approaches to LC classification which can be linked to EO data cubes have become increasingly popular, such as *Living Earth* which aims to aid in coherent reporting towards SDG targets, but these can be restrictive in terms of the resources required to produce input data which has been determined to be useful for one country (Owers et al., 2021). Additionally, some datasets attempt to provide generic data variables for essential land cover attributes to counter the specificity of thematic classes, such as data on imperviousness, forest, grassland, wetland and water bodies generated by the European CORINE project (Gómez et al., 2016).

The *ArealStatistik* datasets are distinct from most operational LC maps, as the remote sensing data they are based on is in the form of aerial photographs rather than satellite imagery. These photographs are taken over a 6-year survey period and visually inspected to assign LC and LU classes to sample points on a 100m by 100m grid. The resulting point dataset covers Switzerland at a resolution of 100m, with a total of approximately 4 million points nationwide. Transformation to a gridded map involves assigning each cell assigned the value of the point at its lower-left corner, however the classification remains valid only for the point to which the label is assigned.

The size of estimation error for the *ArealStatistik* is based on the proportion of the dataset considered and decreases with the number of data points considered. At the national level the largest estimation error for a Basic Category is around 6.5% (OFS, 2022c), compared to up to 67% for the Copernicus CORINE Land Cover (CLC) 2012 dataset (Jaffrain, 2017). Recent developments to refine the interpretation of the source aerial photography using Deep Learning, which will be implemented for the next dataset to be released, present an overall accuracy of over 90% (OFS, 2022b).

Table 1 - Comparison of operational Land Cover datasets.

Name	Producing agency	Spatial coverage	Source imagery	Cell size	Temporal resolution	Thematic resolution
ArealStatistik	Swiss Federal Office of Statistics	Switzerland	Aerial photography	100 m	1979 - 1985, 1992 - 1997, 2004 - 2009, 2013 - 2018	27 classes

CORINE Land Cover (CLC)	EEA Copernicus Land Monitoring Service	Europe	Sentinel-2 gap filled with Landsat-8 (2018 map)	100 m	1990, 2000, 2006, 2012, 2018	44 classes
GLC2000	EC Joint Research Centre	Global and Regional datasets	SPOT	1 km	2000	19 classes
Dynamic World	Google	Global	Sentinel-2	10 m	Every 2-5 days	9 classes
ESRI Land Cover	ESRI	Global	Sentinel-2	10 m	Annual 2017-2021	10 classes

2.3. Machine Learning for Land Cover Classification

The use of pixel-based supervised ML methods to automate regular production of datasets has had success at spatial scales from the regional to global (Inglada et al., 2017). Comparisons show that supervised classification methods, in which the ML model is provided with a training dataset containing variables extracted from the images and the matching ground truth values for these variables, generally outperform unsupervised methods (Holloway & Mengersen, 2018; Szuster et al., 2011). However, these methods can sometimes be restrictive in their need for accurate reference data.

The choice of which ML model to use is based on several criteria, including resources for processing and availability of training data. For example, ensemble models, in which predictions of multiple individual models are combined to produce a result, in general have higher accuracy and can provide information on the uncertainty of classifications. However, these models have a trade off with reduced interpretability of the model and increased computational complexity (Gómez et al., 2016). Research on classification of satellite imagery has shown that Random Forests (RF) and Support Vector Machine (SVM) classifiers tend to show the best compromise between accuracy and complexity (Gómez et al., 2016; Khatami et al., 2016).

RF is an ensemble learning method based on decision tree algorithms which produces its final prediction for each observation through aggregating the predictions of multiple trees (Géron, 2019). Decision trees are simple and interpretable algorithms, in which the 'leaves' of the tree refer to the labels and the 'branches' refer to the unique combinations of input variables which produce those labels. Within RF, each decision tree is trained on a different random subset of the training data and predictions are then based on the majority vote from all of the individual trees (Breiman, 2001). The high accuracy and low propensity for overfitting has led to RF being widely adopted as a classifier for remote sensing data (Belgiu & Drăguţ, 2016; Talukdar et al., 2020). This method has been successfully applied for land cover classification in similar landscapes to Switzerland in France using data from Landsat (Pelletier et al., 2016; Inglada et al., 2017). Additionally, RF classifiers can be used to determine which input variables provide the most useful information, which helps to reduce the number of dimensions within the dataset and therefore reduce the computational complexity of the model (Belgiu & Drăguţ, 2016).

SVM classifiers work by determining a hyper plane, or optimal separation, between two classes within a dataset (Géron, 2019). SVMs are known for having a longer training time than RF (Pal, 2005), and are more sensitive to the choice of hyperparameters such as the kernel function which enables the hyper plane to be fitted. However, they frequently produce higher scores than RF in terms of overall accuracy, and their performance is less dependent on the size of training samples than other algorithms (Thanh Noi & Kappas, 2018).

Whilst the pixel-based supervised classification methods described above have shown good results, such approaches can be limited by intra-class variability of spectral signatures, similarity of spectral signatures between classes and classes which are highly discontinuous at small spatial scales (Stoian

et al., 2019). Other approaches include the use of object-based image analysis in which groups of spatially contiguous pixels which represent a geographical feature are defined, and the model trained to detect these as separate objects (Costa et al., 2018). By far the current state of the art for LC mapping are Deep Learning methodologies traditionally applied to image classification such as Convolutional Neural Networks (CNNs) (Pelletier et al., 2019). CNNs have been shown to out-perform pixel-based classification methods, with their success due in part to an explicit ability to encode spatial information contained within satellite imagery, as each prediction involves the values of neighbouring pixels (Carranza-García et al., 2019). However, these methods are subject to large requirements for computing power and storage space, as well as difficulty in obtaining sufficient reference data.

2.4. 'Space-first' versus 'time-first' approaches to LC classification

Two main approaches to classifying LC from satellite imagery can be identified based on the priority placed on the spatial or temporal dimension. Classification of Satellite Image Time Series (SITS) maximises the value of big data volumes of EO data, with the high revisit time of satellite imagery resulting in a dataset which effectively captures change in LC over time. This is known as a '*time-first, space-later*' approach, with pixels having a stronger temporal autocorrelation than spatial autocorrelation (Picoli et al., 2018). In comparison, the traditional '*space-first, time-later*' approach performs classification solely using spatial dimensions, and differences over time can incorporated at a later stage when assessing change detection between already classified images (Camara et al., 2014). *Space-first* approaches have produced widely used datasets based on annual composites (e.g. Hansen et al., 2013), however it has been argued that SITS classification enables to capture the dynamics of LC classes resulting in a more accurate classifications and greater potential for near real-time monitoring (Woodcock et al., 2020).

With SITS, processes of LC change become an integrated part of the generation of LC data itself (Wulder et al., 2018). Variables used in input data such as spectral indices can be highly variable over a year, and the incorporation of this temporal variation is a factor which can contribute to improved accuracy of classification. For example, variation in vegetation indices calculated from spectral bands over time can highlight important phenological differences between different vegetation classes (Defries & Townshend, 1994). Despite the advantages of this approach, time-series data can also contain a high degree of noise and inter-annual variability, posing methodological challenges as efforts to remove this noise to enable use in ML models can remove important information contained within the time series (Picoli et al., 2018).

3. Data

3.1. Study area

The study area in the western part of Switzerland between 46.0° to 46.8° latitude and between 5.95° and 6.98° longitude is shown in Figure 2. This region contains the urban centres of Geneva and Lausanne, as well as the surface of Lake Geneva which lies within the Swiss territory. The climate is temperate with a high degree of topographical variation between the mountains in the south-east of the study area and the plateau in the north-east.

3.2. Reference data

Reference data used is the ArealStatistik for the 2013-2018 survey period (OFS, 2022), which provides a classification for points spaced at 100m. The subset of reference data covering the study area includes around 410,000 features. The LC nomenclature used distinguishes between 6 'Principal domains' of Artificial Areas, Grass & Herb Vegetation, Brush Vegetation, Tree Vegetation, Bare Land, and Watery Areas, and between 27 "Basic Categories" which fall within these Principal Domains.

3.3. Input data

ARD was provided by the Swiss Data Cube which collates available satellite imagery for Switzerland (Chatenoux et al., 2021). The present study uses data from Sentinel-2 which has been pre-processed to produce top-of-atmosphere reflectance and summarised to give the median annual value. Location of the study area within Switzerland and the coverage of the study area by Sentinel-2 tile 31TGM are shown in Figure 2. Reference data were reprojected into the coordinate system of the Sentinel-2 imagery (EPSG:4326), and the surface reflectance values for each band extracted for each point to create the dataset used in training and validation. Median data for 2018 was used for training all models, with data from 2021 used for additional testing to assess transferability of the model over time.

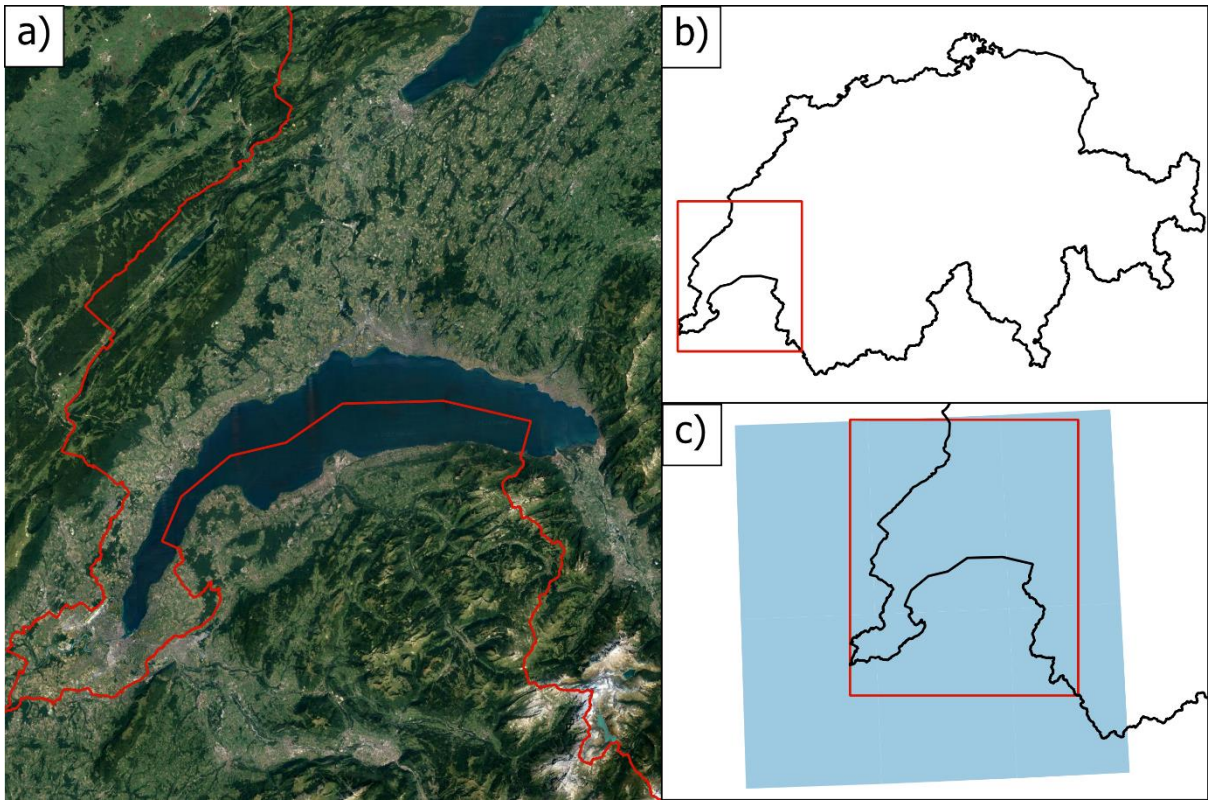

Figure 1 - Study area as a) shown in Google Satellite imagery, b) position within Switzerland, c) as covered by Sentinel-2 tile 31TGM.

The characteristics of the spectral bands of the Sentinel-2 images used in the classification models are provided in Table 2.

Table 2 - Characteristics of Sentinel-2 data.

Band	**Spectral Resolution (nm)**	**Spatial Resolution (m)**
Band 1 – Coastal aerosol	443	60
Band 2 – Blue	490	10
Band 3 – Green	560	10
Band 4 - Red	665	10
Band 5 – Red Edge 1	705	20
Band 6 – Red Edge 2	740	20
Band 7 – Red Edge 3	783	20

Band 8 – NIR	842	10
Band 8A – Red Edge 4	865	20
Band 11 – SWIR 1	1610	20

Elevation and DEM derivatives have been shown to be auxiliary datasets for LC classification which improve model performance, particularly for forest categories (Zhu et al., 2016), and were included as additional input data for one iteration of classification. The DEM used was the 25m grid Digital Height Model DHM25 (swisstopo, 2022). Slope and aspect were calculated from the elevation field. Elevation and derivatives were converted from EPSG:21781 to EPSG:4326 to enable interoperability with the other datasets, which led to a slight loss of extreme values.

4. Methods

4.1. Machine Learning models used

Classification algorithms were implemented using the *sklearn* Python module. Initially, classification on the median 2018 Sentinel-2 images was compared using RF and SVM with the default parameters. Multinomial Linear Regression (MLR) was also included in comparison as a simple baseline model to set a benchmark for comparison with more complex models. The trained RF model was used to classify the 2021 Sentinel-2 data, which was then compared to the 2013-2018 ArealStatistik labels. Results for the 2018 median model were then compared to those produced using RF through the R package *SITS* (Satellite Image Time Series Analysis for Earth Observation Data Cubes) (Gregory Giuliani, internal communication). The *SITS* package provides an interface to connect to data cubes, using a time-first, space-later approach to LC classification which optimises the use of time-series input data (Simoes et al., 2021). The time series for each instance of the dataset is incorporated in the classification.

4.2. Sampling strategy

For comparison with data produced using SITS, training and validation features were selected randomly at a ratio of 60:40, with equal class distributions between the training and validation sets. The resulting class distribution of pixels used for training and validation for LC Principal Domains is given in Figure 4.

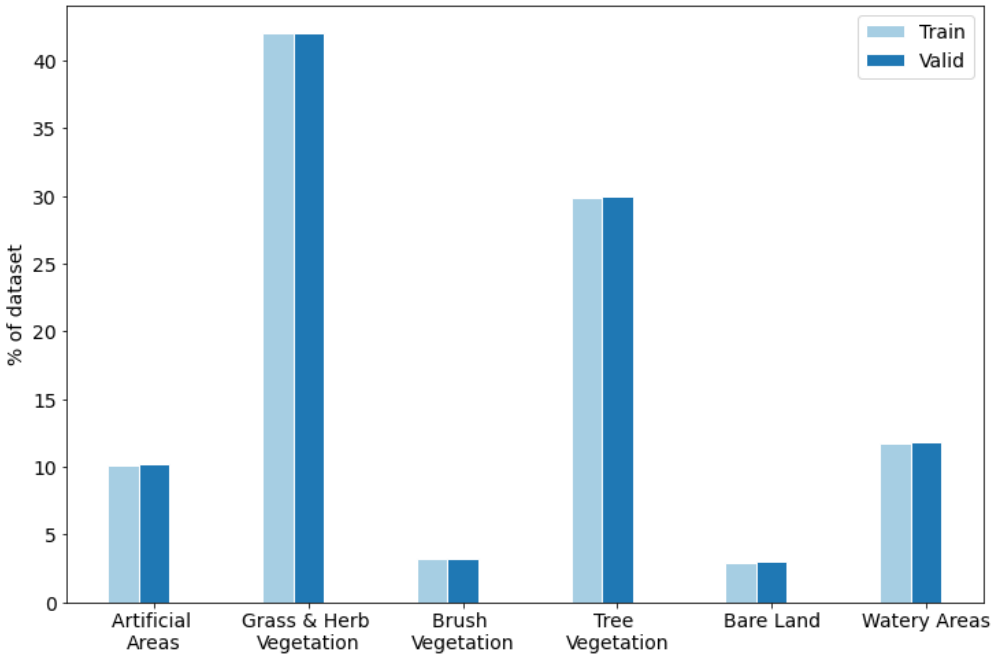

Figure 2 - Class distribution of samples for LC Principal Domains training and validation data sets.

The class distribution of pixels in the dataset used for training and validation for LC Basic Categories is given in Figure 4.

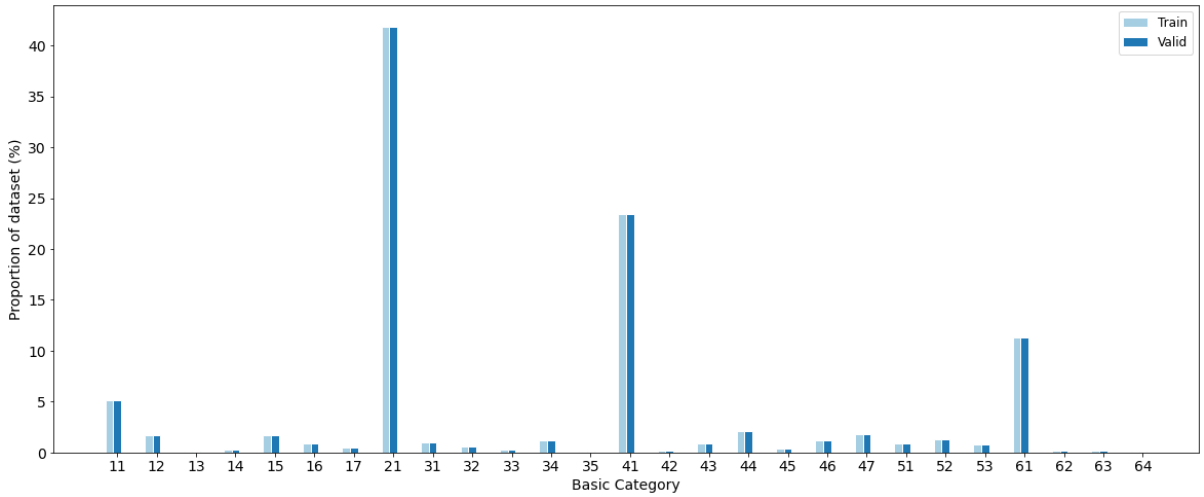

Figure 3 – Class distribution of samples for LC Basic Categories. Descriptions corresponding to the class codes are included in Annex 1.

The variation in land area coverage of different LC classes is a frequent problem in the supervised classification of LC (Douzas et al., 2019). To attempt to counteract the class imbalance evident above, RF-2018 was run once with balanced class weights, and once with oversampling using Synthetic Minority Over-sampling Technique (SMOTE). SMOTE is a data augmentation technique which trains the RF model with additional synthesised instances of minority classes which are duplicated from the actual minority data (Chawla et al., 2002).

4.3. Spectral indices

Inclusion of spectral indices has shown to increase classifier performance through providing information regarding the non-linear relationships which exist between the different spectral bands of a satellite image (Chaves et al., 2020; Inglada et al., 2017). 3 commonly used spectral indices were therefore included as input data to the model: Normalised Difference Vegetation Index (NDVI) for the identification of vegetation, Normalised Difference Building Index (NDBI) as a measure of built-up areas and Normalised Difference Water Index (NDWI) for the identification of water. The equations for these indices are given below.

Equation 1 - Normalised Difference Vegetation Index (NDVI):

$$NDVI = \frac{NIR - RED}{NIR + RED}$$

Equation 2 - Normalised Difference Building Index (NDBI):

$$NDBI = \frac{SWIR - NIR}{SWIR + NIR}$$

Equation 3 - Normalised Difference Water Index (NDWI):

$$NDWI = \frac{G - NIR}{G + NIR}$$

Figure 4 shows the mapped values of these spectral indices for the study area.

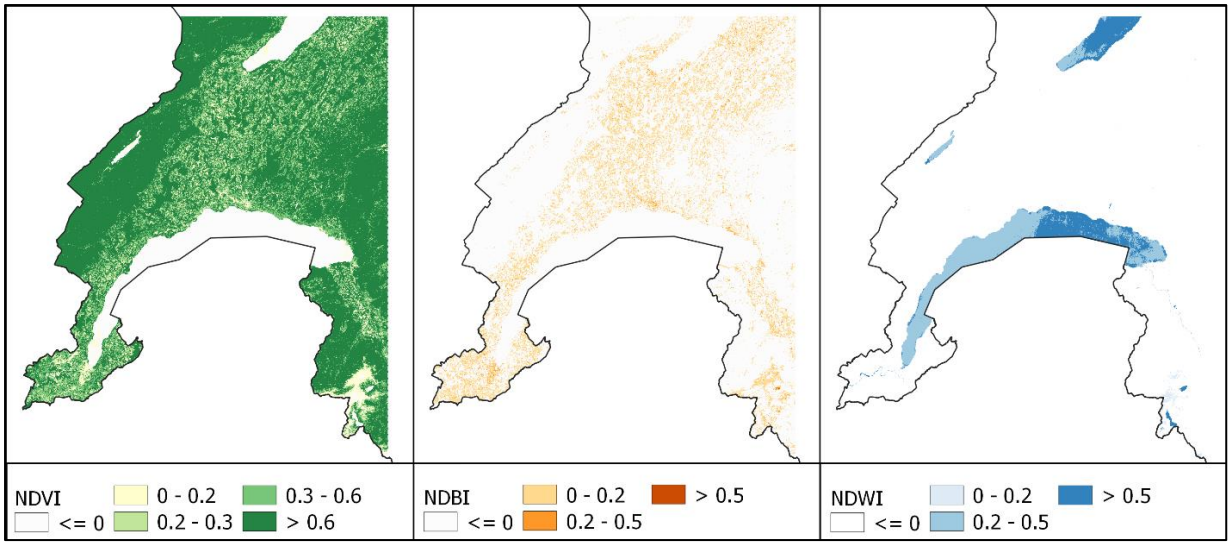

Figure 4 - Calculated spectral indices for 2018 median Sentinel-2 images (a) NDVI, (b) NDBI, (c) NDWI.

4.4. Feature importance

Permutation feature importance was calculated to determine the contribution of auxiliary input data. The permutation feature importance is the decrease in accuracy of the model that occurs when the values of an individual feature are randomly shuffled, and is therefore representative of the degree to which the model is dependent on the information provided by the feature.

4.5. Optimised RF model

RF were also used to run additional classification tests to assess the impact on classifier performance of a spatially sensitive selection of training and validation data, as well as the incorporation of hyperparameter tuning.

4.5.1. Sampling strategy considering spatial autocorrelation

For additional model testing, the dataset was divided into 3 sets (training, validation & test) to enable tuning of hyperparameters to optimise model performance.

Random selection of training and validation data at the pixel level can lead to model evaluation over-estimating the performance of a classifier, due to the potential for spatial autocorrelation between instances (Inglada et al., 2017; Tonini et al., 2020). A potential strategy to overcome this is to select training and validation sets from separate polygons (Pelletier et al., 2016), however this approach could introduce greater imbalance between LC classes. Hence, data was split using spatial k-fold cross validation, using an approach similar to the method described in Tonini et al. (2020).

For this method, the area of interest was overlaid with a grid of 900 10km^2 cells, and each data point was then assigned to its overlaying cell. Data for 54 cells, which corresponded to around 12% of the total data, was set aside to form the test set. The cell identifier for the remaining instances which was used as the 'group' input for a grouped k-fold split, which split data into two sets for training and validation, whilst ensuring that instances for each set originated from different cells. 7 folds were used in the grouped k-fold split, to ensure consistency in the size of the validation and test sets (the size of the validation set being 1/k of the total input to a k-fold split). The final ratio was 76:12:12 for training, validation, and testing.

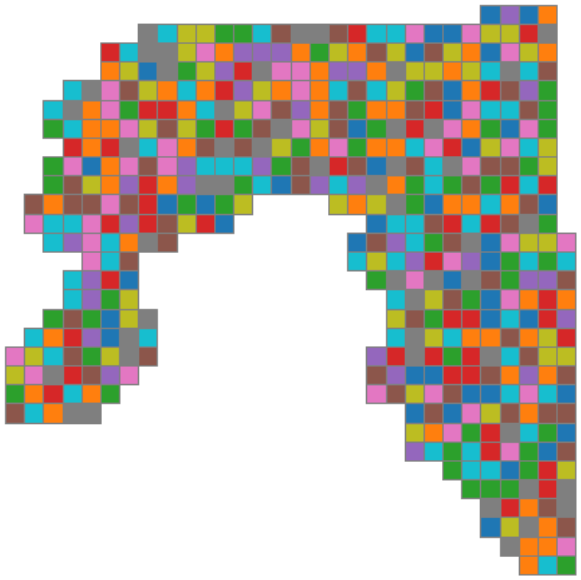

Figure 5 – Visualisation of randomly assigned spatial grid used for data split.

The class distribution of samples used for training, validation and test is shown in Figure 6, where the values for training and validation represent the average distribution over the 7 cross-validation folds.

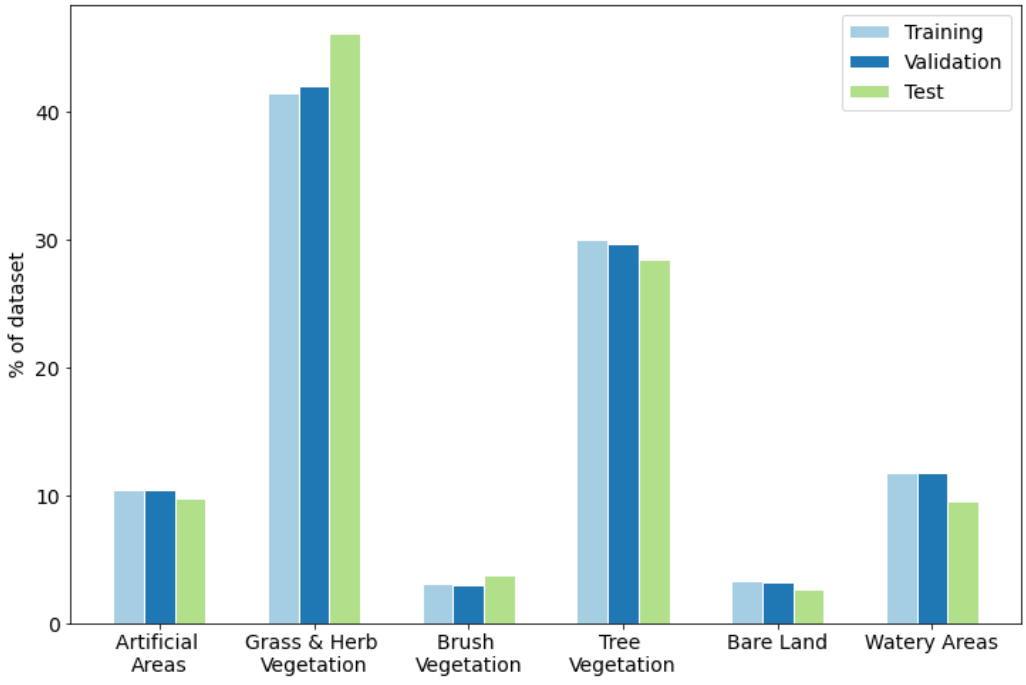

Figure 6 - Distribution of samples for LC Principal Domains using random selection of grids in Figure 5.

4.5.2. Optimisation of Random Forests using hyperparameter tuning

Hyperparameters are the parameters of the machine learning algorithm, which can have an important impact on the performance of a model. Evaluating the performance of different hyperparameter combinations on a subset of the training data allows for the selection of the optimal model to use for full training (Géron, 2021). The hyperparameters evaluated for the RF model were the number of trees

in the forest *('N estimators'*), the maximal depth of each tree (*'Max depth'*), the minimum samples required to split an internal node (*'Min samples split'*), and the minimum samples per node (*'Min samples leaf'*). These hyperparameters enable the model's approach to grouping similar features to be refined.

Successive halving grid search was implemented to perform a hyperparameter search using the training and validation sets. This method begins by evaluating model performance over all hyperparameter combinations provided using a small sample of data. Under successive iterations, the best candidates from the previous round are selected and the model is re-evaluated using a greater number of samples.

The range of hyperparameters used in grid search and the results of the hyperparameters leading to the best model performance, determined by highest overall accuracy, are displayed in Table 3. Whilst other work has included 'N estimators' of up to 400 in hyperparameter determination, this option was omitted due to the diminishing return in increased accuracy relative to the vastly increased computation time required (Pelletier et al., 2016).

Table 3 – Range of values used for RF hyperparameter search, and hyperparameters of the highest performing model.

Hyperparameter	Range	Best
N estimators	50, 100, 150, 200	200
Max depth	10, 25, 50	50
Min samples split	2, 5, 10	10
Min samples leaf	1, 10, 25, 50	1

5.5. Performance metrics

The main assessment of model performance uses the confusion matrix, a table in which the rows represent the actual classes as provided by the reference data, and the columns represent the predicted classes as produced by the classifier. The values on the diagonal of the table are therefore the results which have been correctly classified by the model, with all other values indicating misclassifications. The overall accuracy of classification is calculated as the sum of all diagonal values divided by the sum of all matrix values.

Additionally, several metrics were calculated based on the confusion matrix, to give further information on the model's performance:

- *Recall*, or producer's accuracy, refers to the proportion of predictions of a land cover class which are correct, relative to the total number of pixels within the ground truth class.
- *Precision*, or user's accuracy, refers to the proportion of on-the-ground land cover classes correctly predicted.
- The weighted *F1 score* represents a weighted average of the model's precision and recall, and is an effective summary of model performance on datasets with imbalanced class distributions.

6. Results

RF produced the highest overall accuracy score of all models compared, but in all cases the results of all models run using the 2018 median data were lower than those run using the SITS methodology. Whilst SVM had only produced a slightly lower accuracy, training time took over 6 hours compared to just two minutes for the RF runs. MLR produced the lowest scores, indicating a complexity in the

relationships between features and their class which is not effectively captured by simple models. The use of imbalanced training data is a known issue for ML algorithms such as RF, leading to poor performance for minority classes. However, oversampling and class balancing did not increase the model performance, and actually decreased accuracy, suggesting that the minority classes might actually have low discriminatory power and increasing their importance does not provide increased information to the model. Inclusion of DEM and its derivatives resulted in the greatest improvement in model performance, increasing accuracy to 87.1%. For all models, the weighted F1 score is slightly lower than the overall accuracy, showing a slightly detrimental effect of the imbalanced class distribution.

Table 4 - Performance metrics of ML models using default parameters to predict ArealStatistik Principal Domains.

	Overall accuracy	**F1**
RF (2018 median)	85%	84.7%
RF (SITS)	88.7%	88.2%
RF (with DEM)	87.1%	86.2%
RF (class balanced)	84.8%	83.4%
SVM (2018 median)	83.2%	81%
RF (with SMOTE)	82.9%	83.3%
MLR (2018 median)	82%	79.8%

The results of feature importance in Figure 7 indicate that the Green spectral band provides by far the greatest information to the model. The inclusion of spectral indices NDVI, NDBI and NDWI in the input features has a relatively low effect on the classification accuracy, although NDWI is the most important spectral index for this dataset. Elevation and slope provide greater information to the model as auxiliary variables.

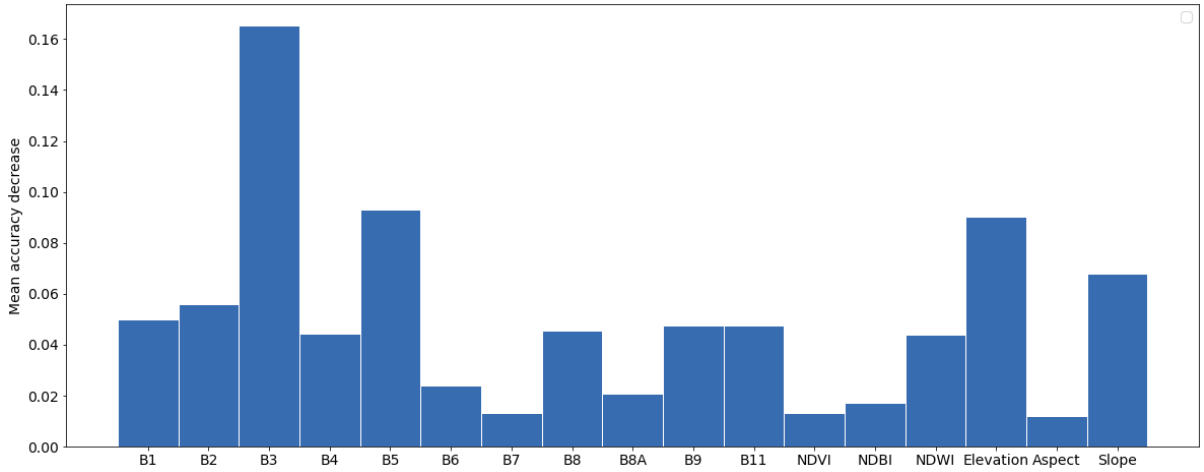

Figure 7 - Permutation feature importance for RF (with DEM), in the form of mean accuracy decrease resulting from feature omission.

Figure 8 shows the precision and recall scores for each Principal Domain. Performance for the individual classes of Watery Areas, Tree Vegetation and Grass & Herb Vegetation are comparable between the two approaches with low levels of error for each of these classes. The model performs poorly in terms of recall for Brush Vegetation and Bare Land, with over 56% of Brush Vegetation pixels mis-classified as Grass & Herb Vegetation. Precision for these categories is higher than recall,

indicating a lower proportion of other classes are incorrectly classified as Brush Vegetation or Bare Land. In general, the SITS model performs better for all LC classes. Irrespective of model used, the Brush Vegetation class shows very low recall compared to other classes.

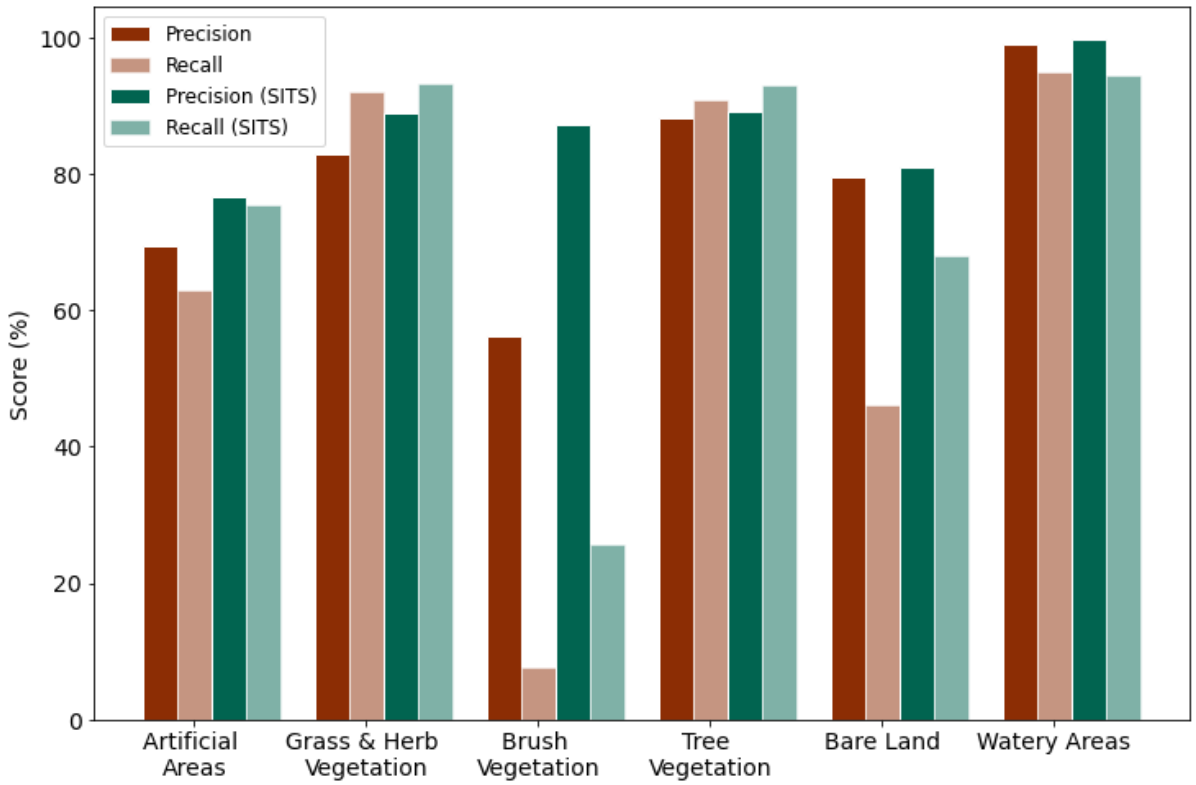

Figure 8 – Precision and recall for the highest performing RF-2018 model, and RF using SITS.

Training the RF median model using the 27 basic categories results in an overall accuracy of 77.6% and an F1 score of 72.6%, compared to an overall accuracy of 88% achieved using *SITS*. The model performs best for Grass & Herb Vegetation, Closed Forest, Water and Glacier/Perpetual Snow Basic Categories. As with the Principal Domains, the minority LC classes within the Basic Categories are not well classified, there are few other LC types which are incorrectly classified as members of these classes, but the precision score for many LC types is very low or zero.

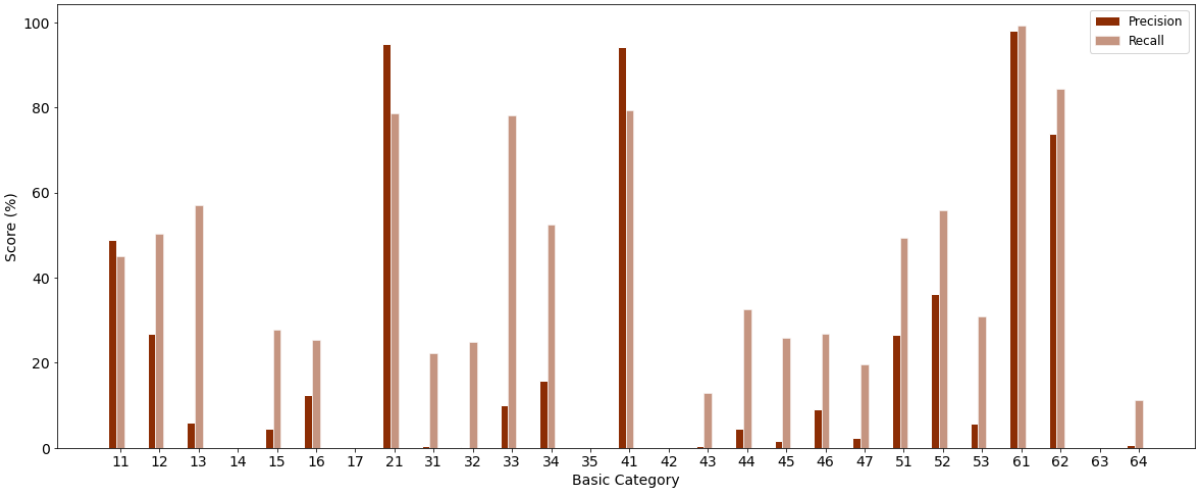

Figure 9 - Precision and recall by class for 2018 Basic Categories. Descriptions corresponding to the class codes are included in Annex 1.

Some decrease in performance when applying the trained RF model to Sentinel-2 data from 2021 could be attributed to the use of outdated reference data, as the average annual rate of change between the 2004-2009 and 2013-2018 ArealStatistik datasets was approximately 0.57%. However, the observed decrease was 6% and therefore much more than could be expected from normal processes of LC change, indicating low transferability of the model to other years. Performance per class on the 2021 test data shown in Figure 8 is consistent with performance on the validation data, with relatively weak scores for Brush Vegetation and Bare Land.

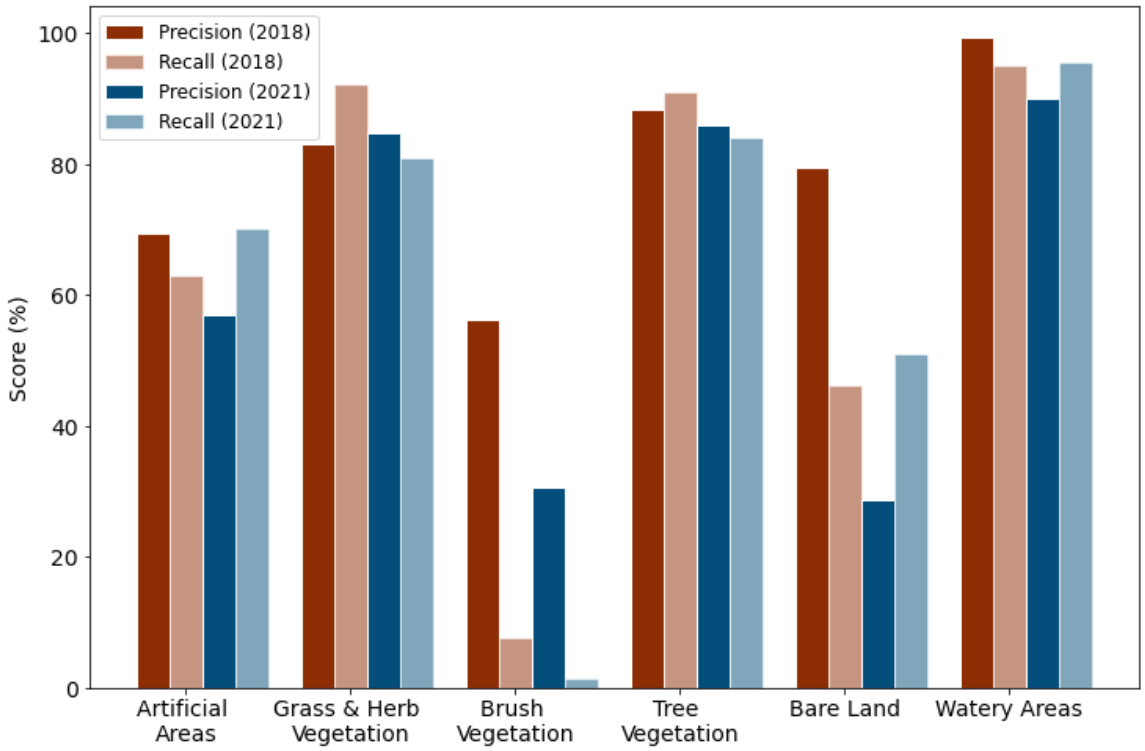

Figure 10 – Precision and recall by principal domain for RF median model applied to 2018 validation data and 2021 test data.

Some differences visible in Figure 9 indicate a weakness in the transferability of the model across time. For example, between 6.8° and 6.98° longitude, the model classifies large new areas of Bare Land and Watery Areas, but the scale of such changes does not reflect existing LC change processes in Switzerland (OFS, 2022). In the case of new Watery Areas, this could reflect an artefact of median values of the imagery being influenced by the presence of seasonal snow cover. Patches of vineyards around Lake Geneva, are not correctly classified, as would be expected from the model's poor ability to classify Brush Vegetation.

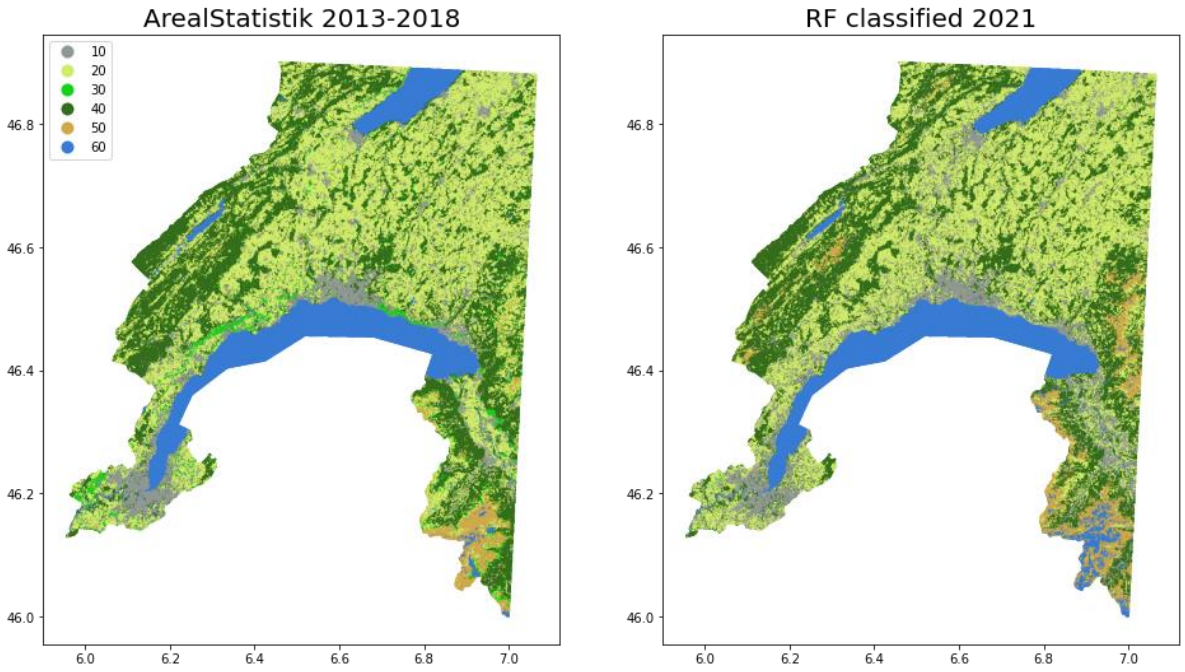

Figure 11 – Maps showing reference ArealStatistik data for 2013-2018 (right), and classification results for 2021 median Sentinel-2 data (left).

Running the RF model with hyperparameter optimisation and controlling for spatial autocorrelation resulted in only a minor improvement in model performance with an overall accuracy on the test set of 85.3%, however the total training time was increased to 3.5 hours. Relative performance of the model for minority classes was consistent with the other methods shown above. Compared to an initial study which used a similar optimised model applied to several individual dates of Landsat-5 images, use of the median annual Sentinel-2 data led to higher overall accuracy (85% versus 76%).

Table 5 – Performance metrics for hyperparameter optimised RF model.

	Overall accuracy
Training	93.8%
Validation	85.7%
Test	85.3%

7. Discussion

Comparison of time-first and space-first approaches

High overall accuracy and F1 scores of the classification results are encouraging but mask an important inter-class variation. In particular, in the Plateau region of Switzerland which features in the study area, the rate of change in Brush Vegetation categories has been highly dynamic over the last 33 years of *ArealStatistik* surveys (OFS, 2022b). Poor performance for these classes raises concerns over the analysis potential of these classified datasets to fully inspect such changes which could be necessary to understand pressures acting on LC over time. The high scores achieved for LC classes within the Watery Areas principal domain and Closed Forests could reflect the large contiguous lake and forest areas present in the dataset and the high accuracy often achieved in mapping areas of homogenous land cover (Tsutsumida & Comber, 2015). The results above indicate that spatial autocorrelation did not lead to significant overestimation of performance after training, so the high performance of these

classes may be because of similar spectral signatures and low proportion of edge cases found in these contiguous areas, rather than the effect of neighbourhood conditions.

Within the ML literature, the addition of temporal features to LC classification models has been shown to have varied effects, with some studies finding their inclusion results in only minor gains in accuracy which are outweighed by additional computational cost (e.g. Pelletier et al., 2016) and others finding clear accuracy gains with time-series metrics compared to single-date results (Franklin et al., 2015). Here, inclusion of the full time series of data through the *SITS* method outperforms the base RF model by 3.7% and remains the most accurate method of all model iterations considered despite efforts to improve performance through the addition of auxiliary input variables and efforts to balance class distribution. For this dataset, the use of a time-first approach therefore appears to be superior, and crucially key gains are presented in classifying the minority classes in Brush Vegetation.

The inclusion of spectral indices provided little additional information for the RF model using median data, and in the case of NDVI and NDWI this may be due to the median value not fully capturing the information which can be derived from these indices. Incorporation of the whole annual NDVI cycle has shown to be effective for classifying shrub vegetation (Evans & Geerken, 2006) with the full potential of information on phenological variation provided by vegetation indices helping to improve the separability of LC classes (Chaves et al., 2020). The space-first approach could be improved through the inclusion of additional spatial predictors such as distance-to metrics which take into account the likelihood of certain LC classes occurring next to each other (Hermosilla et al., 2022).

Quality of training samples

Relatively poor classification of Brush Vegetation categories is however a constant feature irrespective of model choice, and the inability of oversampling or class balancing techniques to resolve the issue of imbalanced data in this study suggests that the ground-truth data for the minority classes such as Brush Vegetation has low discriminatory power. This could arise from an inability to generalise the spectral signatures to the same extent as the training labels. Vineyards, for example, can be characterised by variations in interrow vegetation from one field to another, depending on the specific management practices applied (Fox et al., 2012).

Indeed, a consistent challenge in developing LC classification models is ensuring the existence of sufficient high-quality training data (Pandey et al., 2021), and furthermore in ensuring that the training data are representative of the classes assigned. This can be secured through adopting a sampling design which chooses features which accurately represent the range of spectral diversity within LC classes, and effectively include spectral sub-classes, or by removing mixed and outlier pixels from the training data (Kavzoglu, 2009). Oversampling methods tailored for spatial land cover classification such as Geometric-SMOTE may produce better results than the standard SMOTE method used here due to their ability to create an increased diversity of generated instances (Douzas et al., 2019)

The presence of 'noise', taking the form of non-systematic errors or confusion between the features of an instance and its class, in the labels of training samples can have an important detrimental impact on the accuracy of classifiers (Frenay & Verleysen, 2014). Whilst error of the *ArealStatistik* is generally accepted to be low, such large-area classifications can still be subject to bias and inconsistency brought in during interpretation. Additionally, subtle changes in ground-truth values may be missed by the fact that the *ArealStatistik* surveys are only accurate for one of the six years they cover. For an operational LC dataset to move past use of the *ArealStatistik* would require creation of a new ground-truthed training dataset which would likely be prohibitively resource-intensive. Choice of classifier can also have an effect, for example the RF classifier achieves high robustness to random and systematic label

noise for all the tested configurations (showing to be more robust than SVM for instance) (Pelletier et al., 2017). Uncertainty in the attribution of training labels is therefore unlikely to have a noticeable effect in this study.

Integrated model approaches and data

It is increasingly recognised that LC classification needs to harness the potential of big EO data, and several platforms for processing large volumes of SITS data exist including Google Earth Engine and Open Data Cube (Simoes et al., 2022). However, assessment of the feasibility of these methods for producing operational land cover products must consider ease of use and reproducibility. In addition to improved model performance, a further benefit of using software packages such as *SITS* is that default values and parameters have been selected based on expert assessment and so individual users do not require advanced skills in ML in order to run the classifications (Simoes et al., 2022). As shown above, optimisation of the model in the form of hyperparameter tuning can be considered as non-essential for RF classification. Additionally, a direct connection to a data cube can be made, which facilitates the automated and iterative classification of big data volumes.

Perspectives

Improvement of classification for vegetation categories is priority to be able to develop these datasets into operational land cover products. As has been shown for running RF with median data, the inclusion of auxiliary input variables, and refinement of the model, may lead to further gains in accuracy of the SITS approach. The benefit of adding DEM features, a static variable, to the space-first method is clear for a region with such varied topography and a flat plateau section mainly covered by agricultural, residential features and water bodies, and could add additional discriminatory power to time-first models as well. Additional auxiliary datasets such as runoff coefficients could help to distinguish between some categories such as vineyards and other low-growing vegetation (Fox et al., 2012).

An important development would need to be testing of the model's applicability over time, as the spatial distribution of accuracies can vary over time as evidenced here (Tsutsumida & Comber, 2015). Time-first approaches in the form of Temporal CNNs have been shown to further increase the accuracy of SITS classification (Pelletier et al., 2019), however this requires greater effort in the creation of ground truth maps to be able to train the model on patches (Carranza-García et al., 2019). The current point-based reference data of the *ArealStatistik* would be insufficient for this. The advantages of using Sentinel-2 data, with its increased spatial and temporal resolution, are unfortunately limited by the lack of sustained time-series available as the mission started in 2017. Methodologies combining Landsat-8 and Sentinel-2 data have proved to solve problems with gaps in time-series and also reduce uncertainties in Land Use/Land Cover studies, with combined use of the two systems providing repeat observations every 2 to 5 days (Chaves et al., 2020).

8. Conclusions

The methods explored produce overall accuracy scores which are comparable to the current standard of operational land cover products. The LC data produced by applying the SITS methodology to ARD from the Swiss Data Cube have the highest accuracy, and meet the criteria for a reproducible, automatable, spatially and temporally continuous dataset which can adapt to nomenclature changes (Inglada et al., 2017). However, whilst the overall accuracies achieved are acceptable for static LC maps, the use of these classified datasets for specific applications such as pixel-level analysis of LC change may be limited, particularly for classes where the rate of change is lower than the degree of error (Inglada et al., 2017). The assessment of LC trends and dynamics is therefore feasible using ML

classified datasets; however, any such study will need to consider the uncertainty the model error brings to any conclusions drawn, especially if these relate to minority LC classes. The feasibility moving from a producing a single well-classified image to a consistent, annual LC dataset is however questionable, due to the lower accuracy achieved for the 2021 Sentinel-2 data. Testing of the methodology using *SITS* may produce better results due to its inherent consideration of temporal autocorrelation.

As is a typical issue with studies involving ML methods, the realisation of the objectives for this study was limited by computational capacity and the time taken to run iterations of more complex model versions with more features. Increased amounts of input features are known contribute to improved performance, and the ability to test the use of multiple time-slices throughout the year may have added insight into the information added by time-series data. The extension of testing to the national level is a key step in assessing the feasibility of these methods to produce operational LC data products for Switzerland, but the expansion of the spatial scope of this study would require use of high-performance computing clusters. Generalisation from the regional scale to a nationally appropriate model can be difficult, and will certainly require further testing, but could be aided by a regionalised approach to the selection of training data (Hermosilla et al., 2022). Expansion of feasibility studies to the alpine regions is a crucial next step, as this would include increased representation of areas of forest and glaciers which are key priorities for LC monitoring.

9. Code and data availability

Code used is available at https://github.com/isabelntho/CGEOM. ArealStatistik data is available open source from https://www.bfs.admin.ch/bfs/fr/home/services/geostat/geodonnees-statistique-federale/sol-utilisation-couverture/statistique-suisse-superficie.html.

10. References

Ban, Y., Gong, P., & Giri, C. (2015). Global land cover mapping using Earth observation satellite data: Recent progresses and challenges. *ISPRS Journal of Photogrammetry and Remote Sensing*, *103*, 1–6. https://doi.org/10.1016/j.isprsjprs.2015.01.001

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

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Brown, C. F., Brumby, S. P., Guzder-Williams, B., Birch, T., Hyde, S. B., Mazzariello, J., Czerwinski, W., Pasquarella, V. J., Haertel, R., Ilyushchenko, S., Schwehr, K., Weisse, M., Stolle, F., Hanson, C., Guinan, O., Moore, R., & Tait, A. M. (2022). Dynamic World, Near real-time global 10 m land

use land cover mapping. *Scientific Data*, *9*(1), 251. https://doi.org/10.1038/s41597-022-01307-4

Carranza-García, M., García-Gutiérrez, J., & Riquelme, J. C. (2019). A Framework for Evaluating Land Use and Land Cover Classification Using Convolutional Neural Networks. *Remote Sensing*, *11*(3). https://doi.org/10.3390/rs11030274

CEOS (2022) CEOS Analysis Ready Data (http://ceos.org/ard/; accessed 14/09/2022)

Chatenoux, B., Richard, J.-P., Small, D., Roeoesli, C., Wingate, V., Poussin, C., Rodila, D., Peduzzi, P., Steinmeier, C., Ginzler, C., Psomas, A., Schaepman, M. E., & Giuliani, G. (2021). The Swiss data cube, analysis ready data archive using earth observations of Switzerland. *Scientific Data*, *8*(1), 295. https://doi.org/10.1038/s41597-021-01076-6

Chaves, M. E. D., C. A. Picoli, M., & D. Sanches, I. (2020). Recent Applications of Landsat 8/OLI and Sentinel-2/MSI for Land Use and Land Cover Mapping: A Systematic Review. *Remote Sensing*, *12*(18). https://doi.org/10.3390/rs12183062

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.

Costa, H., Foody, G. M., & Boyd, D. S. (2018). Supervised methods of image segmentation accuracy assessment in land cover mapping. *Remote Sensing of Environment*, *205*, 338–351. https://doi.org/10.1016/j.rse.2017.11.024

Defries, R. S., & Townshend, J. R. G. (1994). NDVI-derived land cover classifications at a global scale. *International Journal of Remote Sensing*, *15*(17), 3567–3586. https://doi.org/10.1080/01431169408954345

Di Gregorio, A., & Jansen, L. J. M. (2005). *Land cover classification system (LCCS): Classification concepts and user manual—Version 2*. Food and Agriculture Organization of the United Nations (FAO).

Douzas, G., Bacao, F., Fonseca, J., & Khudinyan, M. (2019). Imbalanced Learning in Land Cover Classification: Improving Minority Classes' Prediction Accuracy Using the Geometric SMOTE Algorithm. *Remote Sensing*, *11*(24). https://doi.org/10.3390/rs11243040

Evans, J. P., & Geerken, R. (2006). Classifying rangeland vegetation type and coverage using a Fourier component based similarity measure. *Remote Sensing of Environment*, *105*(1), 1–8. https://doi.org/10.1016/j.rse.2006.05.017

Fox, D. M., Witz, E., Blanc, V., Soulié, C., Penalver-Navarro, M., & Dervieux, A. (2012). A case study of land cover change (1950–2003) and runoff in a Mediterranean catchment. *Applied Geography*, *32*(2), 810–821. https://doi.org/10.1016/j.apgeog.2011.07.007

Franklin, S. E., Ahmed, O. S., Wulder, M. A., White, J. C., Hermosilla, T., & Coops, N. C. (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), 293–314. https://doi.org/10.1080/07038992.2015.1089401

Frenay, B., & Verleysen, M. (2014). Classification in the Presence of Label Noise: A Survey. *IEEE Transactions on Neural Networks and Learning Systems*, *25*(5), 845–869. https://doi.org/10.1109/TNNLS.2013.2292894

Géron, A. (2019). *Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow* (2nd ed.). O'Reilly.

Giuliani, G., Chatenoux, B., Bono, A. D., Rodila, D., Richard, J.-P., Allenbach, K., Dao, H., & Peduzzi, P. (2017). Building an Earth Observations Data Cube: Lessons learned from the Swiss Data Cube (SDC) on generating Analysis Ready Data (ARD). *Big Earth Data*, *1*(1–2), 100–117. https://doi.org/10.1080/20964471.2017.1398903

Giuliani, G., Rodila, D., Külling, N., Maggini, R., & Lehmann, A. (2022). Downscaling Switzerland Land Use/Land Cover Data Using Nearest Neighbors and an Expert System. *Land*, *11*(5). https://doi.org/10.3390/land11050615

Gómez, C., White, J. C., & Wulder, M. A. (2016). Optical remotely sensed time series data for land cover classification: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, *116*, 55–72. https://doi.org/10.1016/j.isprsjprs.2016.03.008

Hermosilla, T., Wulder, M. A., White, J. C., & Coops, N. C. (2022). Land cover classification in an era of big and open data: Optimizing localized implementation and training data selection to improve mapping outcomes. *Remote Sensing of Environment*, *268*, 112780. https://doi.org/10.1016/j.rse.2021.112780

Holloway, J., & Mengersen, K. (2018). Statistical Machine Learning Methods and Remote Sensing for Sustainable Development Goals: A Review. *Remote Sensing*, *10*(9). https://doi.org/10.3390/rs10091365

Inglada, J., Vincent, A., Arias, M., Tardy, B., Morin, D., & Rodes, I. (2017). Operational High Resolution Land Cover Map Production at the Country Scale Using Satellite Image Time Series. *Remote Sensing*, *9*(1). https://doi.org/10.3390/rs9010095

Jaffrain, G. (2017). *Corine Land Cover 2012 Final Validation Report*. https://land.copernicus.eu/user-corner/technical-library/clc-2012-validation-report-1

Kavzoglu, T. (2009). Increasing the accuracy of neural network classification using refined training data. *Environmental Modelling & Software*, *24*(7), 850–858. https://doi.org/10.1016/j.envsoft.2008.11.012

Kennedy, R. E., Andréfouët, S., Cohen, W. B., Gómez, C., Griffiths, P., Hais, M., Healey, S. P., Helmer, E. H., Hostert, P., Lyons, M. B., Meigs, G. W., Pflugmacher, D., Phinn, S. R., Powell, S. L., Scarth, P., Sen, S., Schroeder, T. A., Schneider, A., Sonnenschein, R., … Zhu, Z. (2014). Bringing an ecological view of change to Landsat-based remote sensing. *Frontiers in Ecology and the Environment*, *12*(6), 339–346. https://doi.org/10.1890/130066

Khatami, R., Mountrakis, G., & Stehman, S. V. (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*, 89–100. https://doi.org/10.1016/j.rse.2016.02.028

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

OFS (2022a) Transfert en production ArealStatistik Deep Learning (ADELE), OFS: Neuchatel, 109p.

OFS (2022b) Statistique suisse de la superficie (WWW) https://www.bfs.admin.ch/bfs/fr/home/services/geostat/geodonnees-statistique-federale/sol-utilisation-couverture/statistique-suisse-superficie.html; accessed 2nd November 2022.

OFS (2022c) Qualité des données, erreur aléatoire, (https://www.bfs.admin.ch/bfs/fr/home/statistiques/espace-environnement/enquetes/area/exploitation-donnees/qualite-donnees-erreur-aleatoire.html; accessed 15/10/2022).

Owers, C. J., Lucas, R. M., Clewley, D., Planque, C., Punalekar, S., Tissott, B., Chua, S. M. T., Bunting, P., Mueller, N., & Metternicht, G. (2021). Living Earth: Implementing national standardised land cover classification systems for Earth Observation in support of sustainable development. *Big Earth Data*, *5*(3), 368–390. https://doi.org/10.1080/20964471.2021.1948179

Pandey, P. C., Koutsias, N., Petropoulos, G. P., Srivastava, P. K., & Ben Dor, E. (2021). Land use/land cover in view of earth observation: Data sources, input dimensions, and classifiers—A review of the state of the art. *Geocarto International*, *36*(9), 957–988. https://doi.org/10.1080/10106049.2019.1629647

Pelletier, C., Valero, S., Inglada, J., Champion, N., & Dedieu, G. (2016). Assessing the robustness of Random Forests to map land cover with high resolution satellite image time series over large areas. *Remote Sensing of Environment*, *187*, 156–168. https://doi.org/10.1016/j.rse.2016.10.010

Pelletier, C., Valero, S., Inglada, J., Champion, N., Marais Sicre, C., & Dedieu, G. (2017). Effect of Training Class Label Noise on Classification Performances for Land Cover Mapping with Satellite Image Time Series. *Remote Sensing*, *9*(2). https://doi.org/10.3390/rs9020173

Pelletier, C., Webb, G. I., & Petitjean, F. (2019). Temporal Convolutional Neural Network for the Classification of Satellite Image Time Series. *Remote Sensing*, *11*(5). https://doi.org/10.3390/rs11050523

Phiri, D., Simwanda, M., Salekin, S., Nyirenda, V. R., Murayama, Y., & Ranagalage, M. (2020). Sentinel-2 Data for Land Cover/Use Mapping: A Review. *Remote Sensing*, *12*(14). https://doi.org/10.3390/rs12142291

Picoli, M. C. A., Camara, G., Sanches, I., Simões, R., Carvalho, A., Maciel, A., Coutinho, A., Esquerdo, J., Antunes, J., Begotti, R. A., Arvor, D., & Almeida, C. (2018). Big earth observation time series analysis for monitoring Brazilian agriculture. *ISPRS Journal of Photogrammetry and Remote Sensing*, *145*, 328–339. https://doi.org/10.1016/j.isprsjprs.2018.08.007

Simoes, R., Camara, G., Queiroz, G., Souza, F., Andrade, P. R., Santos, L., Carvalho, A., & Ferreira, K. (2021). Satellite Image Time Series Analysis for Big Earth Observation Data. *Remote Sensing*, *13*(13). https://doi.org/10.3390/rs13132428

Stoian, A., Poulain, V., Inglada, J., Poughon, V., & Derksen, D. (2019). Land Cover Maps Production with High Resolution Satellite Image Time Series and Convolutional Neural Networks: Adaptations and Limits for Operational Systems. *Remote Sensing*, *11*(17). https://doi.org/10.3390/rs11171986

swisstopo (2022) DHM25 (https://www.swisstopo.admin.ch/en/geodata/height/dhm25200.html; accessed 10/10/2022).

Szantoi, Z., Geller, G. N., Tsendbazar, N.-E., See, L., Griffiths, P., Fritz, S., Gong, P., Herold, M., Mora, B., & Obregón, A. (2020). Addressing the need for improved land cover map products for policy support. *Environmental Science & Policy*, *112*, 28–35. https://doi.org/10.1016/j.envsci.2020.04.005

Szuster, B. W., Chen, Q., & Borger, M. (2011). A comparison of classification techniques to support land cover and land use analysis in tropical coastal zones. *Applied Geography*, *31*(2), 525–532. https://doi.org/10.1016/j.apgeog.2010.11.007

Talukdar, S., Singha, P., Mahato, S., Shahfahad, Pal, S., Liou, Y.-A., & Rahman, A. (2020). Land-Use Land-Cover Classification by Machine Learning Classifiers for Satellite Observations—A Review. *Remote Sensing*, *12*(7). https://doi.org/10.3390/rs12071135

Teixeira, Z., Teixeira, H., & Marques, J. C. (2014). Systematic processes of land use/land cover change to identify relevant driving forces: Implications on water quality. *Science of The Total Environment*, *470–471*, 1320–1335. https://doi.org/10.1016/j.scitotenv.2013.10.098

Thanh Noi, P., & Kappas, M. (2018). Comparison of Random Forest, k-Nearest Neighbor, and Support Vector Machine Classifiers for Land Cover Classification Using Sentinel-2 Imagery. *Sensors*, *18*(1). https://doi.org/10.3390/s18010018

Tonini, M., D'Andrea, M., Biondi, G., Degli Esposti, S., Trucchia, A., & Fiorucci, P. (2020). A Machine Learning-Based Approach for Wildfire Susceptibility Mapping. The Case Study of the Liguria Region in Italy. *Geosciences*, *10*(3). https://doi.org/10.3390/geosciences10030105

Tsutsumida, N., & Comber, A. J. (2015). Measures of spatio-temporal accuracy for time series land cover data. *International Journal of Applied Earth Observation and Geoinformation*, *41*, 46–55. https://doi.org/10.1016/j.jag.2015.04.018

Verde, N., Kokkoris, I. P., Georgiadis, C., Kaimaris, D., Dimopoulos, P., Mitsopoulos, I., & Mallinis, G. (2020). National Scale Land Cover Classification for Ecosystem Services Mapping and Assessment, Using Multitemporal Copernicus EO Data and Google Earth Engine. *Remote Sensing*, *12*(20). https://doi.org/10.3390/rs12203303

Wulder, M. A., Coops, N. C., Roy, D. P., White, J. C., & Hermosilla, T. (2018). Land cover 2.0. *International Journal of Remote Sensing*, *39*(12), 4254–4284. https://doi.org/10.1080/01431161.2018.1452075

Zhu, Z., Gallant, A. L., Woodcock, C. E., Pengra, B., Olofsson, P., Loveland, T. R., Jin, S., Dahal, D., Yang, L., & Auch, R. F. (2016). Optimizing selection of training and auxiliary data for operational land cover classification for the LCMAP initiative. *ISPRS Journal of Photogrammetry and Remote Sensing*, *122*, 206–221. https://doi.org/10.1016/j.isprsjprs.2016.11.004

11. Annex

Annex 1 - Codes of Basic Categories

Code	Description
11	Consolidated surfaces
12	Buildings
13	Greenhouses
14	Gardens with border and patch structures
15	Lawns
16	Trees in artificial areas
17	Mix of small structures
21	Grass and herb vegetation
31	Shrubs
32	Brush meadows
33	Short-stem fruit trees
34	Vines
35	Permanent garden plants and brush crops
41	Closed forest
42	Forest edges
43	Forest strips
44	Open forest
45	Brush forest
46	Linear woods
47	Clusters of trees
51	Solid rock
52	Granular soil
53	Rocky areas
61	Water
62	Glacier, perpetual snow
63	Wetlands
64	Reedy marshes