Classifying land cover from satellite images using time series analytics

Patrick Schäfer
Computer Science Department
Humboldt-Universität zu Berlin
Germany
patrick.schaefer@hu-berlin.de

Dirk Pflugmacher
Patrick Hostert
Geography Department
Humboldt-Universität zu Berlin
[dirk.pflugmacher,patrick.hostert]
@geo.hu-berlin.de

Ulf Leser
Computer Science Department
Humboldt-Universität zu Berlin
leser@informatik.hu-berlin.de

ABSTRACT

The Earth's surface is continuously observed by satellites, leading to large multi-spectral image data sets of increasing spatial resolution and temporal density. One important application of satellite data is the mapping of land cover and land use changes such as urbanization, deforestation, and desertification. This information should be obtained automatically, with high accuracy, and at the pixel level, which implies the need to classify millions of pixels even when only small regions are studied. Balancing runtime and accuracy for this task becomes even more challenging with the recent availability of multiple time points per pixel, created by periodically performed satellite scans. In this paper we describe a novel approach to classify land cover from series of multi-spectral satellite images based on multivariate time series analytics. The main advantage of our method is that it inherently models the periodic changes (seasons, agriculture etc.) underlying many types of land covers and that it is comparably robust to noise. Compared to a classical feature-based classifier, our new method shows a slightly superior overall accuracy, with an increase of up to 20% in accuracy for rare land cover classes, though at the cost of notably increased runtime. The highest accuracy is achieved by combining both approaches.

1 INTRODUCTION

Monitoring changes in land usage is an important area of research as land cover is a key variable driving the Earth's energy balance, hydrological and carbon cycle, and the provisioning of natural resources and habitat [22]. Over the last three decades, satellite-based *Earth Observation (EO)* programs have made tremendous progress in acquiring medium-resolution ($10-100$m) images around the globe systematically and with increasing frequency (revisit time). As a result, large volumes of medium-resolution satellite images are now available free of charge, enabling automatic approaches to the identification of land usage and the detection of land surface changes over large areas. For example, the American Landsat 8 sensor images the Earth at 30-m spatial resolution every 16 days, and two European Sentinel-2 sensors acquire images with a revisit time of 5 days and a spatial resolution of $10-20$ meters, which amounts to roughly 60 measurements for more than 300 Billions pixels (excluding oceans) in a year.

The free availability of medium-resolution satellite image time series has spawned new possibilities for mapping land cover [9]. In the past, classification approaches operated on single images or stacks of images (i.e. composite classification). Multi-date classification approaches exploit the notion that land cover can vary over time, e.g. because of vegetation senescence or harvesting [27]. The task can be approached in different ways. In a typical baseline setting, the different measurements per pixel are used as independent features for a classical machine learning-based classifier, such as Naive Bayes or Decision Trees [20]. In this approach, the temporal order of the measurements is ignored as all features are treated as orthogonal dimensions of the feature space. An alternative method is to include the consideration of the order of measurements by using methods from time series analytics [2, 24]. Here, every pixel is considered as a temporally ordered (and aligned) series of measurements, and the specific changes (increasing or decreasing slopes, periodic changes etc.) of the measurements over time are analyzed to find commonalities and to derive classification models. Previous works (see Section 3) have shown that this can be advantageous as land cover are temporally variable and often follow characteristic temporal patterns, such as those imposed by seasons. However, given the enormous scale of the data to be classified, not only the accuracy of an approach is important, but also the runtime performance has a critical role in any practical application.

In this work, we evaluate the recently proposed multivariate time series classification algorithm *WEASEL+MUSE* [26] for land cover classification using temporally dense, medium-resolution satellite images. WEASEL+MUSE models multivariate time series using the truncated Fourier transformation and discretizes measurements, both of which to reduce noise, builds a rich feature space to capture also subsequences in the time series, is able to exploit similar temporal subsequences even when appearing at very different offsets within a time series, and uses aggressive feature selection to remove irrelevant features and thereby speed-up classification. Although WEASEL+MUSE was not developed specifically for land cover classification, many of its aspects fit nicely to the specificities of this domain, such as the inherent noise reduction and the exploitation of repetitive behaviour.

We compare the prediction performance of WEASEL+MUSE with the performance of an established and popular machine learning-based approach, Random Forests, using the same input features on 23 Landsat 8 images collected in 16 day-intervals over Reunion Island. The study region covers an area of $2866x2633$ pixels at 30~m spatial resolution. As reference dataset for model training and validation, we used a sample of 81714 pixels that had been manually classified into 9 land cover classes.

Our results indicate that time-series-based algorithms improve land cover classification accuracy compared to non-temporal algorithms. Our time series algorithm WEASEL+MUSE achieved higher classification accuracies than Random Forests. The improvement in accuracy was most notable with rare and/or difficult classes. Here, class-wise accuracies increased by 8 and 3 percentage points, respectively. Overall accuracy improved by 1%-point owing to the fact that the dominant classes were less affected by the choice of algorithm. Interestingly, Random Forests captures different signals in the data than the time-series-based approach, as a simple Ensemble of both approaches further improved classification accuracy. However, this increased accuracy comes at the cost of an increased runtime. Thus, we will focus future work on improving the runtime of our method without loosing the advancements in prediction performance we observed.

The rest of this paper is organized as follows: Section 2 provides background on land cover classification using satellite images. Section 3 presents the current state of the art in land cover mapping and time series classification. Section 4 describes the test dataset and the tested classification methods: Random Forests and WEASEL+MUSE. Section 5 presents the experiments.

2 LAND COVER CLASSIFICATION FROM SATELLITE IMAGES

The classification of satellite images to extract information on land cover and land use has a long history. A significant turning point in terrestrial Earth Observation was the launch of Landsat-1 in 1972 (then called Earth Resource Technology Satellite). For the first time, Landsat delivered systematic observations of the Earth land surface for land monitoring, leading to new ways of machine-assisted approaches for mapping land cover from space [7]. In the 1990s and early 2000s, international and national agencies started to adopt operational land cover mapping programs with Landsat and Landsat-like data, e.g. in Europe (CORINE), USA (NLCD), Canada (EOSD), and Australia (NCAS-LCCP). Even decades later, the Landsat program is still active today. Since February 2013, Landsat 8 is taking images of the Earth at 30 m spatial resolution in 8 spectral bands every 16 days. The radiometric quality and spectral resolution has greatly improved since the early satellites, and because of new global acquisition strategies and on-board storage and download capabilities, so has the sheer number of available images. The number of medium spatial resolution (10 – 100 m) sensors has been increasing [4], thus dense time series of medium resolution are available for many parts of the globe.

Multi-spectral sensors like Landsat record the sun's energy reflected by a surface in a few distinct spectral wavelengths (bands), e.g. blue, green, red in the visible spectrum (400 nm to 700 nm), near infrared (700 to 1100 nm), and short-wave infrared (1100 to 3000 nm). Since land surfaces with different chemical and structural properties often absorb and reflect sunlight differently and wavelength-dependent, information on land cover can be derived from these spectral bands. For example, water absorbs much of the near-infrared radiation, so these wavelengths are useful for discerning land-water boundaries that are not obvious in visible light. Similarly, green vegetation absorbs much of the incoming radiation in the red spectrum while reflecting about 50% of the radiation in the near-infrared spectrum.

Much of the past research on classification algorithms has focused on exploiting the spectral and spatial properties of land covers, including artificial neural networks [1], decision trees [11], support vector machines [15], and spatial segmentation algorithms [10]. Each algorithm has its strength and weakness with respect to: the distributional assumptions made about the data, training requirements, computational complexity, and robustness to overfitting, data noise, and errors in training data. Also common to all algorithms is that the work is supervised, i.e., they need an independent reference data set (i.e., land cover information collected in the field or from air-photos) for training a model. It is fair to say, that no single algorithm works best for all applications and reference data. However, there has been a trend away from parametric statistical models to machine learning to deal with the complexity of input data and class legends.

It has only recently become feasible to build land cover mapping algorithms that exploit the temporal domain of entire pixel time series with medium resolution data [12]. To this end, satellite images typically first undergo a series of pre-processing steps, including the correction of atmospheric effects, geometric alignment and cloud and cloud-shadow masking. Once these steps are finished, spectral values of pixels can be traced over time to identify and detect land surface changes such as deforestation [8], or urbanization [19].

3 RELATED WORK

Although, time-series based classification (TSC) is a relatively new area in remote sensing, the topic itself has a long tradition and dozens of approaches exist (see [2, 24], for instance). Time-series-based classifiers can broadly be categorized into two classes: Similarity-based methods use a similarity measure over sequences, such as Dynamic Time Warping (DTW), to perform a point-wise comparison of two time series. In contrast, feature-based TSC rely on comparing features extracted from the different time series, typically generated from their substructures.

These two approaches are also the basis for methods in multivariate time series classification (MTSC), where a time series is not made of a single stream of values but by multiple streams, each one usually synchronized in time. A popular similarity-based MTSC method is dynamic time warping (DTW) [21]. Feature-based MTSC can be grouped into those methods that build feature from so-called shapelets [30], which are short and maximally discriminative subsequences of the time series, and methods using the bag-of-patterns (BOP) approach [3]. The standard BOP model [17] break up a time series into windows, represent the corresponding subsequences within each window as discrete features (or words), and finally derives a classifier from the frequencies of the words. Different approaches to TSC (and MTSC) differ not only in their accuracy on different data sets, but also in their runtime for classification, which is a critical issue when it comes to very large data sets as is the case of satellite images. For instance, we recently evaluated the runtime of 12 different state-of-the-art methods for (univariate) TSC and found differences of up to three orders of magnitude [24, 25].

Most approaches to land cover classification rely on traditional machine-learning methods (see previous section), and there have been only a few prior studies on using time series information. For example, [32] fitted harmonic functions to each satellite band and used the fitted parameters as features in subsequent classification (which implies that it falls into the class of feature-based MTSC methods). [14] found that including temporal information into a model can have a bigger impact on classification accuracy than the choice of the particular classification algorithm. New time-series-based algorithms are needed to leverage the predictive potential of satellite time series images [6].

4 METHODS

4.1 Description of Data

For our analysis, we used a public dataset taken from the TiSeLaC (Time Series Land Cover Classification Challenge) [28]. This dataset consists of time series of 23 Landsat 8 images collected in 16 day-intervals over Reunion Island in 2014. Landsat data were provided by the French *Pôle Thématique Surfaces Continentales*

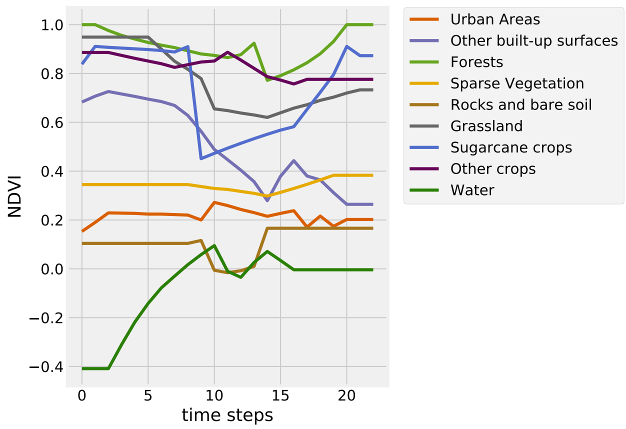

Figure 1: Normalized Difference Vegetation Index (NDVI) of 9 different land cover classes for our reference dataset [28].

ID	Land cover class	Samples
1	Urban Areas	16000
2	Other built-up surfaces	3236
3	Forests	16000
4	Sparse Vegetation	16000
5	Rocks and bare soil	12942
6	Grassland	5681
7	Sugarcane crops	7656
8	Other crops	1600
9	Water	2599

Table 1: Land cover classes and distribution of reference samples [28].

(THEIA), and atmospherically corrected, geometrically corrected, and cloud-masked with the *Multi-sensor Atmospheric Correction and Cloud Screening (MACCS)* level 2A processor developed at the *French National Space Agency (CNES)*. Data pre-processing and temporal gap filling was performed using the iota2[1] Land Cover processor developed by CESBIO[2]. For each time step and pixel, ten spectral features were extracted, i.e., the seven reflectance bands and three vegetation indices: the *Normalized Difference Vegetation Index (NDVI)*, the *Normalized Difference Water Index (NDWI)*, and the *Brightness Index (BI)*. Figure 1 shows examples of NDVI time series for different land cover classes.

Reference land cover data were derived from two publicly available dataset: the 2012 *CORINE Land Cover (CLC)* map and the 2014 farmers' graphical land parcel registration (*Régistre Parcellaire Graphique - RPG*). The most significant classes for the study area were retained, and a spatial processing (aided by photo-interpretation) was performed to ensure consistency with image geometry. Finally, a pixel-based random sampling of this dataset was applied to provide an almost balanced ground truth. The final reference dataset consisted of a total of 81714 pixels distributed over 9 classes (Table1). We split this reference dataset randomly in half for model training and testing.

[1]http://tully.ups-tlse.fr/jordi/iota2.git
[2]http://www.cesbio.ups-tlse.fr

4.2 Time Series Analytics

In general, a time series dataset contains N time series. Each time series is associated with a class label y from a predefined set of labels Y. *Time series classification (TSC)* is the task of predicting a class label for a time series whose label is unknown. A classifier is a function that is learned from a set of labelled time series (the training data), takes an unlabelled time series as input and outputs a label. In this paper, each pixel should be labeled by one of the nine reference land cover classes.

The 23 Landsat 8 images contain 10 time series of spectral features each of length 23. At the pixel level, this data can be interpreted as a multivariate time series (MTS).

A *multivariate time series (MTS)* $T = \{t_1, \ldots, t_n\}$ is an ordered sequence of $n \in \mathbb{N}$ streams $t_i = (t_{i,1}, \ldots t_{i,m}) \in \mathbb{R}^m$, i.e., m recorded values at each fixed time stamp. As we address MTS generated from satellites with a fixed revisit time, we can safely ignore the concrete time stamps. MTS are typically produced by sensors recording data over time like motion captures, gestures, EEG signals, hand-written letters, sign language, or the multi-spectral image data sets captured by satellites.

4.3 Machine Learning Approach using Random Forests

Random forests (RF) are an ensemble learning method widely used in *Earth Observation (EO)* for classifying land cover and land use (e.g. [31]). RF build ensembles of decision trees wherein each tree is trained on randomly selected features of a bootstrapped training sample. Node splits are performed using a random subset of predictor variables. Because of these random components, the RF approach does not require tree pruning and is relatively insensitive to overfitting [5]. To predict a class label, a RF classifies the sample with all its decision trees and returns the mode of the predicted classes. A RF approach to satellite image / pixel classification uses the concatenated spectral features of each pixel as input (this is a vector of length 230 and the 2 coordinates, longitude and latitude). RF are a typical machine learning method which do not consider any order of the features. If time series are used as features, each point in time is considered as an independent feature and the order of measurements is ignored. This implies that such methods are unable to reproduce serial correlations or to detect temporal trends in the data.

4.4 Land Cover Classification with WEASEL+MUSE

WEASEL+MUSE (Word ExtrAction for time SEries cLassification + MUltivariate Symbols and dErivatives) [26] is a state-of-the-art MTS classifier that is composed of the building blocks depicted in Figure 2. It conceptually builds upon the univariate Bag-of-Patterns model applied to each dimension. In the BOP model [23], a time series is characterized by the frequency of occurrence of substructures. BOP-based algorithms build a classification model by (1) extracting subsequences from a time series, (2) transforming each subsequence (of real values, the measurements) into a discrete-valued word (a sequence of symbols over a fixed alphabet), (3) building a feature vector from word counts (histogram), and (4) finally using a classification model from the machine learning repertoire on these feature vectors.

Specifically, WEASEL+MUSE treats the pixel time series of the 10 spectral features as a MTS with 10 dimensions. For each spectral feature, subsequences using varying lengths are extracted,

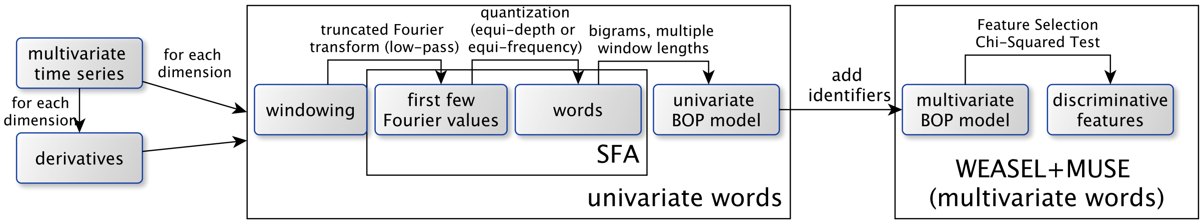

Figure 2: WEASEL+MUSE Pipeline: Feature extraction, univariate BOP models and the multivariate BOP model.

approximated using the truncated Fourier transform, and discretized into words using equi-depth or equi-frequency binning. A feature vector is built from the words (unigrams) and pairs of words (bigrams) to obtain order awareness. Finally, features are concatenated with the sensor id, to maintain a disjoint word space for each dimension. This high dimensional feature space is subsequently filtered using statistical feature selection (Chi-squared test); finally, a logistic regression classifier is trained, assigning weights to characteristic word in each spectral band.

Because WEASEL+MUSE is multivariate, the algorithm can leverage the multi-spectral information of satellite time series. This is an advantage over univariate time series models that operate on single indices (i.e. vegetation indices), as spectral-temporal patterns may differ from sensor band to sensor band.

The feature extraction and selection in WEASEL+MUSE make it interesting for land cover recognition:

- **Features extraction:** The words are derived from subsequences extracted at multiple window lengths in each spectral feature using the truncated Fourier transform and discretization. The Fourier transform reduces noise introduced by preprocessing, such as the cloud mask or geometric alignment, and the different window lengths capture the seasonal trends at different time granularity. Bigrams can capture seasonal trends, e.g., higher intensities in the spectrum in summer than in autumn. By extracting subsequences from the pixel stream, the classifier allows for small shifts in the time line, e.g. a delayed bloom of crops in some regions.
- **Feature selection:** The wide range of words considered also introduces many irrelevant features. Therefore, WEASEL+MUSE applies statistical feature selection to remove irrelevant words from each class. These may be a result of erroneous information introduced by the image capture or preprocessing.

The resulting feature vector is highly discriminative and contains words that are characteristic for each class, which allows the use of fast logistic regression classifier. To perform our analysis, we used the JAVA implementation available from [29].

4.5 Ensemble Approach

To understand whether there is value in combining the time-series-based approach and the feature-based approach, we build and tested a third model based on the ensemble of WEASEL+MUSE and RF. Both approaches output class probabilities for each pixel and select the class with the highest probability. Pixels belonging to a unique spectral class may be associated with high class probabilities, whereas pixels with less distinct class membership may have more equally distributed probabilities, e.g. 49% vs 51%. To combine the two sets of class probabilities ($2x9$ class probabilities), we trained a RF model using the 18 class probabilities from the training dataset as predictors and the corresponding land cover class labels as response.

5 EXPERIMENTS

5.1 Experimental setup

Datasets: We evaluated our competitors using the described Landsat dataset. The dataset was randomly split into 40857 training and 40857 test samples. The training dataset was used to train each classifier. All reported accuracies are based on the test dataset.

Competitors: We performed a series of experiments:

- Build RF on the training dataset and apply it to test dataset.
- Build WEASEL+MUSE on the training dataset and apply it to test dataset.
- Build the combined model on the RF- and MUSE-predicted class probabilities on the training dataset and apply it to the test dataset.

Training: For WEASEL+MUSE we performed 10-fold cross-validation on the training dataset to optimize the parameters for the SFA quantization method (equi-depth or equi-frequency binning). To perform the RF analysis, we used the R statistical language and the RF package from [16]. We set the algorithm to build 1000 trees and randomly sampled $\sqrt{p}$ variables as candidates at each split (where $p = 232$, and p is the total number of predictor variables).

5.2 Results

Table 2 presents the class-wise accuracies and overall accuracy (weighted and simple average F1-scores) on the test samples for each of the three methods. Overall, the combined model had the highest F1-score of 91.1% followed by WEASEL+MUSE with 89.6% and RF with 89.0% (weighted averages). Thus, the choice of the classifier had a relatively small effect on the overall accuracy. However, classes with high F1-scores were also better represented in the training and test samples. The effect on overall accuracy was therefore higher when sample sizes were ignored.

When looking at the results in detail, all classifiers showed high F1-scores of about 90% for all but two classes: "other built-up" and "other crops". For these classes, the accuracy was only ~60% and ~50%, respectively. While RF showed a high precision, WEASEL+MUSE had a higher recall and F1-score, indicating that the temporal profile improved the detection of challenging classes. However, the largest improvement in accuracy was obtained from combining the two classifiers, which further improved the F1-score by up to 20 percentage points for the "other crops" class.

	Random Forests (RF)			WEASEL+MUSE			Combined		
Land cover	**Precision**	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**
Urban	86.7%	91.1%	88.9%	87.6%	91.6%	89.6%	88.9%	91.6%	90.2%
Other built-up	75.0%	52.6%	61.9%	73.8%	57.7%	64.8%	71.9%	61.6%	66.3%
Forests	86.1%	93.1%	89.4%	87.9%	91.7%	89.8%	90.2%	92.5%	91.3%
Sparse Vegetation	92.3%	93.5%	92.9%	90.5%	93.3%	91.8%	93.9%	94.6%	94.2%
Barren	96.2%	95.3%	95.7%	96.0%	92.6%	94.3%	96.7%	95.9%	96.3%
Grassland	83.1%	84.5%	83.8%	85.1%	86.5%	85.8%	87.1%	89.1%	88.1%
Sugarcane	94.1%	94.1%	94.1%	94.1%	94.9%	94.5%	94.7%	95.3%	95.0%
Other crops	***95.2%***	***34.5%***	***50.6%***	***81.6%***	***45.3%***	***58.3%***	***85.3%***	***58.2%***	***69.2%***
Water	90.0%	78.1%	83.6%	91.3%	84.0%	87.5%	91.0%	86.6%	88.7%
Weight. Average	**89.4%**	**89.4%**	***89.0%***	**89.5%**	**89.6%**	***89.6%***	**91.1%**	**91.2%**	***91.1%***
Average	**88.7%**	**79.6%**	***82.3%***	**87.5%**	**82.0%**	***84.0%***	**88.9%**	**85.1%**	***86.6%***

Table 2: Class-wise accuracy and overall accuracy based on test sample for RF, WEASEL+MUSE, and our combined model.

	Urban	**Other built-up**	**Forests**	**Sparse Vegetation .**	**Barren**	**Grassland**	**Sugarcane**	**Other crops**	**Water**	**Users Accuracy**
Urban	7274	544	118	59	24	79	90	142	59	86.7%
Other built-up	163	840	12	41	4	8	16	5	31	75.0%
Forests	273	60	7448	146	30	282	73	237	106	86.1%
Sparse Vegetation	50	33	195	7494	226	33	8	9	72	92.3%
Barren	5	7	15	209	6164	2	0	3	4	96.2%
Grassland	123	60	137	48	2	2382	38	75	2	83.1%
Sugarcane	69	20	50	0	0	32	3669	52	6	94.1%
Other crops	1	1	11	0	0	0	1	276	0	95.2%
Water	26	31	14	17	18	1	2	2	998	90.0%
Producers Accuracy	91.1%	52.6%	93.1%	93.5%	95.3%	84.5%	94.2%	34.5%	78.1%	

Table 3: Confusion matrix from Random Forests classification.

The confusion matrix for the RF (Table 3) gives a detailed picture. “other build-up” is often confused with “urban”, and “other crops” is often confused with the “other forests”, “urban” or “grassland”. This might be a result of an under-representation of these land cover classes in the dataset, as these two classes are the ones with the lowest number of instances (Table 1). On average WEASEL+MUSE required 5.8 ms for a single pixel prediction, as a result of the feature extraction and selection phases prior to classification. The RF took 2.7 ms on average per pixel for classification. For the 7.5 Mio pixels of the study area, this translates into a total, single-CPU runtime of about 5.4 hours for RF compared to 12.2 hours with WEASEL+MUSE. WEASEL+MUSE obtained very promising accuracy for many classes. However we observed a limitation of WEASEL+MUSE and all Bag-of-Pattern approaches when applied in the context of land cover classification, namely the discretization step introduced for noise reduction and to obtain words from real-valued sequences. For discretization, the value range is divided into bins, and each one is associated with a label. However, only a limited number of symbols, typically between 4 to 8, can be used to discretize the value range without negatively impacting accuracy [17]. For the Landsat data the spectral range is between 0 and 1000 and the absolute difference between pixels is important for classification. However, after discretization using 8 symbols (i.e., $a = [000 - 125]$ and $b = [151 - 250], \ldots$) there can be a difference between 0 up to 125 between the values of the same symbol. This noise reduction is useful for applications like gesture recognition [25], but it seems to be too aggressive here.

5.3 Sensitivity Analysis

To better judge the performance of the two classification methods, we tested their sensitivity regarding the weighted F1-score to varying sizes of the training sample (Figure 3). Specifically, it was unclear whether the superior performance of the time series classifier could be replicated with small sample sizes. Starting with a random sample of 10% of the original training sample, we iteratively increased the training sample size up to 100% , and tested all models using the test samples. The results show that WEASEL+MUSE scored consistently higher than RF across all sample sizes. The absolute difference was constant, indicating that both approaches were equally sensitive to the sample size.

6 CONCLUSION

Our objective was to test a time-series-based classification approach (*WEASEL+MUSE*) to earth observation time series for land cover classification and conventional machine learning approach (*Random Forests*). We reported results of a series of experiments using 23 Landsat 8 images collected in 16 day-intervals over Reunion Island. The reference dataset consisted of a total of 81714 pixels distributed over 9 classes. Our key finding is that the use of temporal information improved the classification accuracy,

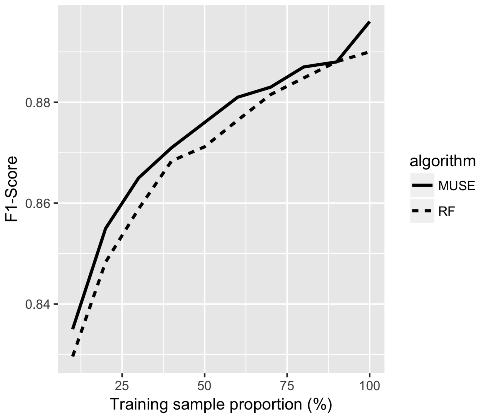

Figure 3: Overall accuracy of WEASEL+MUSE and Random Forests (RF) classification with varying train sample sizes, i.e., 10%-100% of the training samples to build the model and 100% to estimate accuracy.

but not for all land cover classes. The improvement in accuracy was highest for rare and difficult classes. For these, a combined classifier was able to improve the F1-score even further. We used Random Forests because they are widely used and robust, but the ensemble would also work with other classifiers that output class probabilities. Regarding classification times, the Random Forests approach was twice as fast as WEASEL+MUSE, as Random Forests do not require feature extraction or selection, as opposed to time-series-based approaches.

Further research is needed to understand how MUSE+WEASEL scales over larger areas, e.g. continental scale. The complexity of land cover processes and their spectral-temporal patterns will probably grow with increasing area size. From an application point of view, our presented method may be of particular interest for mapping agricultural land use patterns. Agricultural land can be highly dynamic and spectrally variable throughout the year [13]. This land use class is therefore likely to benefit from time-series based approaches. Future research could test the performance of our time-series based method for classifying a broader range of crop types and cropping cycles. The presented approaches essentially disregard the spatial neighbourhood of a pixel. Experiments have shown that including spatial information can improve classification accuracies, similar to the results reported in [18].

ACKNOWLEDGEMENTS

This work was partly supported by the German Federal Ministry of Education and Research through the GeoMultiSens project (grant no. 01IS14010B).

REFERENCES

[1] H. Bagan, Q. Wang, M. Watanabe, Y. Yang, and J. Ma. 2005. Land cover classification from MODIS EVI times-series data using SOM neural network. *International Journal of Remote Sensing* 26, 22 (2005), 4999–5012.

[2] A. Bagnall, J. Lines, A. Bostrom, J. Large, and E. Keogh. 2016. The Great Time Series Classification Bake Off: An Experimental Evaluation of Recently Proposed Algorithms. Extended Version. *Data Mining and Knowledge Discovery* (2016), 1–55.

[3] M. Baydogan and G. Runger. 2016. Time series representation and similarity based on local autopatterns. *Data Mining and Knowledge Discovery* 30, 2 (2016), 476–509.

[4] A. Belward and J. Skøien. 2015. Who launched what, when and why; trends in global land-cover observation capacity from civilian earth observation satellites. *ISPRS Journal of Photogrammetry and Remote Sensing* 103 (2015), 115–128.

[5] L. Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.

[6] L. Bruzzone and B. Demir. 2014. A review of modern approaches to classification of remote sensing data. In *Land Use and Land Cover Mapping in Europe*. Springer, 127–143.

[7] J.R.W. Ellefsen, L. Gaydos. 1974. Computer Aided Mapping of Land Use ERTS Multispectral Scanner Data. *First Pan American Congress on Photogrammetry*.

[8] K. Grogan, D. Pflugmacher, P. Hostert, R. Kennedy, and R. Fensholt. 2015. Cross-border forest disturbance and the role of natural rubber in mainland Southeast Asia using annual Landsat time series. *Remote Sensing of Environment* 169 (2015), 438–453.

[9] M. Hansen and T. Loveland. 2012. A review of large area monitoring of land cover change using Landsat data. *Remote sensing of Environment* 122 (2012), 66–74.

[10] G. Hay, G. Castilla, M. Wulder, and J. Ruiz. 2005. An automated object-based approach for the multiscale image segmentation of forest scenes. *International Journal of Applied Earth Observation and Geoinformation* 7, 4 (2005), 339–359.

[11] C. Homer, C. Huang, L. Yang, B. Wylie, and M. Coan. 2004. Development of a 2001 national land-cover database for the United States. *Photogrammetric Engineering & Remote Sensing* 70, 7 (2004), 829–840.

[12] P. Hostert, P. Griffiths, S. van der Linden, and D. Pflugmacher. 2015. Time series analyses in a new era of optical satellite data. In *Remote Sensing Time Series*. Springer, 25–41.

[13] B. Jakimow, P. Griffiths, S. van der Linden, and P. Hostert. 2017. Mapping pasture management in the Brazilian Amazon from dense Landsat time series. *Remote Sensing of Environment* (2017).

[14] K. Jia, X. Wei, X. Gu, Y. Yao, X. Xie, and B. Li. 2014. Land cover classification using Landsat 8 operational land imager data in Beijing, China. *Geocarto International* 29, 8 (2014), 941–951.

[15] T. Kuemmerle, O. Chaskovskyy, J. Knorn, V. Radeloff, I. Kruhlov, W. Keeton, and P. Hostert. 2009. Forest cover change and illegal logging in the Ukrainian Carpathians in the transition period from 1988 to 2007. *Remote Sensing of Environment* 113, 6 (2009), 1194–1207.

[16] A. Liaw, M. Wiener, et al. 2002. Classification and regression by randomForest. *R news* 2, 3 (2002), 18–22.

[17] J. Lin, R. Khade, and Y. Li. 2012. Rotation-invariant similarity in time series using bag-of-patterns representation. *Journal of Intelligent Information Systems* 39, 2 (2012), 287–315.

[18] Nicola Di Mauro, Antonio Vergari, Teresa Maria Altomare Basile, Fabrizio G. Ventola, and Floriana Esposito. 2017. End-to-end Learning of Deep Spatio-temporal Representations for Satellite Image Time Series Classification. In *Proceedings of the ECML/PKDD Discovery Challenges*.

[19] S. Powell, W. Cohen, Z. Yang, J. Pierce, and M. Alberti. 2008. Quantification of impervious surface in the Snohomish water resources inventory area of western Washington from 1972–2006. *Remote Sensing of Environment* 112, 4 (2008), 1895–1908.

[20] J. Ross Quinlan. 1986. Induction of decision trees. *Machine learning* 1, 1 (1986), 81–106.

[21] T. Rakthanmanon, B. Campana, A. Mueen, G. Batista, B. Westover, Q. Zhu, J. Zakaria, and E. Keogh. 2012. Searching and mining trillions of time series subsequences under dynamic time warping. In *Proceedings of the 2012 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 262–270.

[22] S. Running. 2008. Climate change - Ecosystem disturbance, carbon, and climate. *Science* 321, 5889 (2008), 652–653.

[23] P. Schäfer. 2015. The BOSS is concerned with time series classification in the presence of noise. *Data Mining and Knowledge Discovery* 29, 6 (2015), 1505–1530.

[24] P. Schäfer and U. Leser. 2017. Benchmarking Univariate Time Series Classifiers. In *BTW 2017*. 289–298.

[25] P. Schäfer and U. Leser. 2017. Fast and Accurate Time Series Classification with WEASEL. *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* (2017), 637–646.

[26] P. Schäfer and U. Leser. 2017. Multivariate Time Series Classification with WEASEL+MUSE. *ArXiv e-prints* (Nov. 2017). arXiv:1711.11343

[27] C. Senf, P. Leitão, D. Pflugmacher, S. van der Linden, and P. Hostert. 2015. Mapping land cover in complex Mediterranean landscapes using Landsat: Improved classification accuracies from integrating multi-seasonal and synthetic imagery. *Remote Sensing of Environment* 156 (2015), 527–536.

[28] TiSeLaC : Time Series Land Cover Classification Challenge. 2017. https://sites.google.com/site/dinoienco/tiselc. (2017).

[29] WEASEL+MUSE implementation. 2016. https://github.com/patrickzib/SFA/. (2016).

[30] M. Wistuba, J. Grabocka, and L. Schmidt-Thieme. 2015. Ultra-fast shapelets for time series classification. *arXiv preprint arXiv:1503.05018* (2015).

[31] H. Yin, D. Pflugmacher, R. Kennedy, D. Sulla-Menashe, and P. Hostert. 2014. Mapping annual land use and land cover changes using MODIS time series. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 7, 8 (2014), 3421–3427.

[32] Z. Zhu and C. Woodcock. 2014. Continuous change detection and classification of land cover using all available Landsat data. *Remote sensing of Environment* 144 (2014), 152–171.