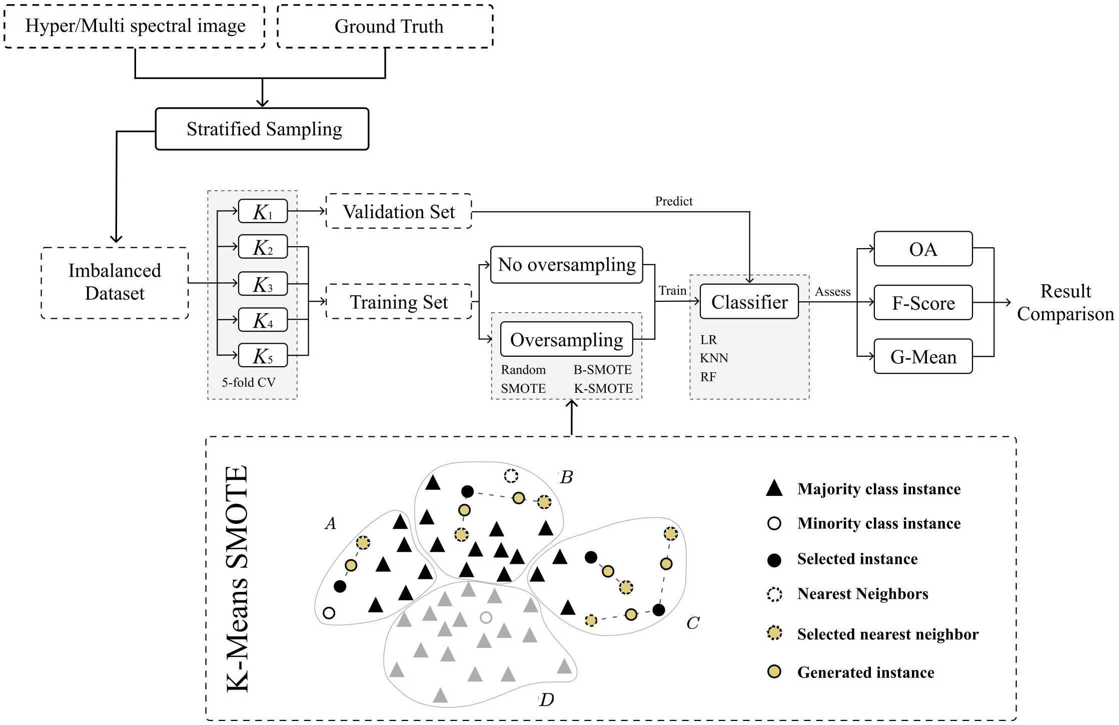
Graphical Abstract

Improving Imbalanced Land Cover Classification with K-means SMOTE: Detecting and Oversampling Distinctive Minority Spectral Signatures

Joao Fonseca, Georgios Douzas, Fernando Bacao



Highlights

Improving Imbalanced Land Cover Classification with K-means SMOTE: Detecting and Oversampling Distinctive Minority Spectral Signatures

Joao Fonseca, Georgios Douzas, Fernando Bacao

- We propose the application of a recent oversampler for LULC classification;
- The method implemented is shown using a variety of classifiers, oversamplers and performance metrics on seven different hyperspectral scenes;
- The proposed method outperformed the remaining configurations in the seven hyperspectral scenes used;

Improving Imbalanced Land Cover Classification with K-means SMOTE: Detecting and Oversampling Distinctive Minority Spectral Signatures

Joao Fonseca[a], Georgios Douzas[a], Fernando Bacao[a]

[a]*NOVA Information Management School, Universidade Nova de Lisboa, Campus de Campolide, 1070-312 Lisboa, Portugal*

Abstract

Land cover maps are a critical tool to support informed policy development, planning, and resource management decisions. The ability to automatically produce Land Use/Land Cover (LULC) maps, through the use of machine learning methods, will improve the accuracy and timeliness of decision-making. With significant upsides, the automatic production of Land Use/Land Cover maps has been a topic of interest for the remote sensing community for several years, but it is still fraught with technical challenges. One such challenge is the imbalanced nature of most remotely sensed data, where the number of samples of a few classes is significantly greater than the number of samples of the remaining classes. This asymmetric class distribution impacts negatively the performance of classifiers and adds a new source of error to the production of these maps. In this paper, we address the imbalanced learning problem, by using K-means and the Synthetic Minority Oversampling TEchnique (SMOTE) as an improved oversampling algorithm. K-Means SMOTE improves the quality of newly created artificial data by addressing both the between-class imbalance, as traditional oversamplers do, but also the within-class imbalance, avoiding the generation of noisy data while effectively overcoming data imbalance. The performance of K-means SMOTE is compared to other popular oversampling methods using seven remote sensing benchmark datasets and a variety of classifiers and evaluation metrics. The statistical analysis of the results show that the proposed method consistently outperforms the remaining oversamplers producing higher quality land cover classifications. These results suggest that LULC data can benefit significantly from the use of more sophisticated oversamplers as spectral signatures for the same class can vary according to geographical distribution.

Keywords: LULC Classification, Imbalanced Learning, Oversampling, Data Generation, Clustering

1. Introduction

The increasing amount of remote sensing missions granted the access to dense time series (TS) data at a global level and provides up-to-date, accurate land cover information (Drusch et al., 2012). This information is often materialized through Land Use and Land Cover (LULC) maps. While Land Cover maps define the biophysical cover found on the surface of the earth, Land Use maps define how it is used by humans (Fritz et al., 2017). Both Land Use and Land Cover maps constitute an essential asset for various purposes, such as land cover change detection, urban planning, environmental monitoring and natural hazard assessment (Khatami et al., 2016). However, the timely production of accurate and updated LULC maps is still a challenge within the remote sensing community (Wulder et al., 2018). LULC maps are produced based on two main approaches: photo-interpreted by the human eye, or automatic mapping using remotely sensed data and classification algorithms.

While photo-interpreted LULC maps rely on human operators and can be more reliable, they also present some significant disadvantages. The most important disadvantage is the cost of production, in fact photo-interpretation consumes significant resources, both in terms of money and time. Because of that, they are not frequently updated and not suitable for operational mapping over large areas. Finally, there is also the issue of overlooking rare or small-area classes, due to factors such as the minimum mapping unit being used.

Automatic mapping with classification algorithms based on machine-learning (ML) have been extensively researched and used to speed up and reduce the costs of the production process (Khatami et al., 2016; Gavade and Rajpurohit, 2019; Kaur et al., 2019). Improvements in classification algorithms are sure to have significant impact in the efficiency with which remote sensing imagery is used. Several challenges have been identified in order to improve automatic classification:

1. Improve the ability to handle high-dimensional datasets, in cases such as Multispectral TS composites high-dimensionality increases the complexity of the problem and creates a strain on computational power (Stromann et al., 2020).

2. Improve class separability, as the production of an accurate LULC map can be hindered by the existence of classes with similar spectral signatures, making these classes difficult to distinguish (Alonso-Sarria et al., 2019).

3. Resilience to mislabelled LULC patches, as the use of photo-interpreted training data poses a threat to the quality of any LULC map produced with this

strategy, since factors such as the minimum mapping unit tend to cause the overlooking of small-area LULC patches and generates noisy training data that may reduce the prediction accuracy of a classifier (Pelletier et al., 2017).

4. Dealing with rare land cover classes, due to the varying levels of area coverage for each class. In this case using a purely random sampling strategy will amount to a dataset with a roughly proportional class distribution as the one on the multi/hyperspectral image. On the other hand, the acquisition of training datasets containing balanced class frequencies is often unfeasible. This causes an asymmetry in class distribution, where some classes are frequent in the training dataset, while others have little expression (Wang et al., 2019; Feng et al., 2019).

The latter challenge is known, in machine learning, as the imbalanced learning problem (Chawla et al., 2004). It is defined as a skewed distribution of instances found in a dataset among classes in both binary and multi-class problems (Abdi and Hashemi, 2016). This asymmetry in class distribution negatively impacts the performance of classifiers, especially in multi-class problems. The problem comes from the fact that during the learning phase, classifiers are optimized to maximize an objective function, with overall accuracy being the most common one (Maxwell et al., 2018). This means that instances belonging to minority classes contribute less to the optimization process, translating into a bias towards majority classes. As an example, a trivial classifier can achieve 99% overall accuracy on a binary dataset where 1% of the instances belong to the minority class if it classifies all instances as belonging to the majority class. This is an especially significant issue in the automatic classification of LULC maps, as the distribution of the different land-use classes tends to be highly imbalanced. Therefore, improvements in the ability to deal with imbalanced datasets will translate into important progress in the automatic classification of LULC maps.

There are three different types of approaches to deal with the class imbalance problem (Fernández et al., 2013; Kaur et al., 2019):

1. Cost-sensitive solutions. Introduces a cost matrix to the learning phase with misclassification costs attributed to each class. Minority classes will have a higher cost than majority classes, forcing the algorithm to be more flexible and adapt better to predict minority classes.

2. Algorithmic level solutions. Specific classifiers are modified to reinforce the learning on minority classes. Consists on the creation or adaptation of classifiers.

3. Resampling solutions. Rebalances the dataset's class distribution by removing majority class instances and/or generating artificial minority instances. This can be seen as an external approach, where the intervention occurs before the learning phase, benefitting from versatility and independency from the classifier used.

Since resampling strategies represent a set of methods that are detached from classifiers by operating at the data level, they allow the use of any off the shelf algorithm, without the need for any type of changes or adaptions to the algorithm. Specifically, in the case of oversampling (defined below), the user is able to balance the dataset's class distribution by without the loss of information, which is not the case with undersampling techniques. This is a significant advantage especially considering that most users in remote sensing are not expert machine learning engineers.

Within resampling approaches there are three subgroups of approaches (Fernández et al., 2013; Kaur et al., 2019; Luengo et al., 2020):

1. Undersampling methods, which rebalance class distribution by removing instances from the majority classes.

2. Oversampling methods, which rebalance datasets by generating new artificial instances belonging to the minority classes.

3. Hybrid methods, which are a combination of both oversampling and undersampling, resulting in the removal of instances in the majority classes and the generation of artificial instances in the minority classes.

Resampling methods can be further distinguished between non-informed and heuristic (i.e., informed) resampling techniques (Fernández et al., 2013; Luengo et al., 2020; García et al., 2016). The former consist of methods that duplicate/remove a random selection of data points to set class distributions to user-specified levels, and are therefore a simpler approach to the problem. The latter consists of more sophisticated approaches that aim to perform over/undersampling based on the points' contextual information within their data space.

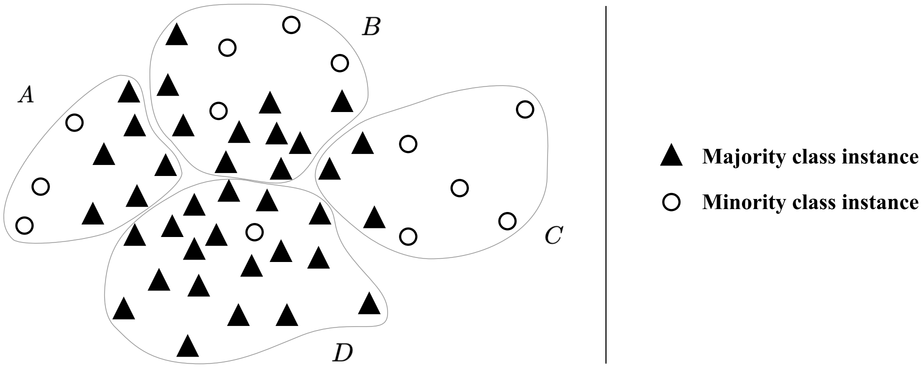
The imbalanced learning problem is not new in machine learning but its relevancy has been growing, as attested by (Haixiang et al., 2017). The problem has also been addressed in the context of remote sensing (Douzas et al., 2019). In this paper, we propose the application of a recent oversampler based on SMOTE (Chawla et al., 2002), the K-means SMOTE (Douzas et al., 2018) oversampler, to address the imbalanced learning problem in a multiclass context for LULC classification using various remote sensing datasets. Specifically, we use seven land use datasets commonly used in research literature, that vary among agricultural and urban land use. The K-means SMOTE algorithm couples two different procedures in the generation of artificial data. The algorithm starts by grouping the instances into clusters by using the K-means algorithm; next, the generation of the artificial data is done using the smote algorithm, taking into consideration the distribution of majority/minority cases in each individual cluster. The idea of starting with a clustering procedure before the data generation phase is important in remote sensing because the spectral signature of the different classes can change significantly based on the geographical area in which it is represented. In other words, the spectral signature of a specific class can vary greatly depending on the geography, meaning that often we will be facing within-class imbalance (Japkowicz, 2001).

In fact, we can decompose class imbalance into two different types: between-class imbalance and within-class imbalance (Douzas et al., 2018; Jo and Japkowicz, 2004). While the first refers to the overall asymmetry between majority and minority classes, the second results from the fact that in different areas of the input space there might be different levels of imbalance. Depending on the complexity of the input space, different subclusters of minority and majority instances may be present. In order to achieve a balance between minority and majority instances, these subclusters should be treated separately. Assuming that the role of a classifier is to create rules in such a way that it is able to isolate the different relevant sub-concepts that represent both the majority and minority classes, the classifier will create multiple disjunct rules that describe these concepts. If the input space is simple and the classes' instances are grouped together in a unique cluster, the classifier will only need to create (general) rules that comprise large portions of instances belonging to the same class. To the contrary, if the input space is complex and scatters through multiple small clusters, the classifier will need to learn a more complex set of (specific) rules, which can be seen in Figure 1. It is important to note that small clusters can happen both in the minority and majority class, although they will tend to be more frequent in the minority class due to its underrepresentation.

The efficacy of K-means SMOTE is tested using different types of classifiers. To do so, we employ both commonly used and/or state-of-the-art oversamplers as benchmarking methods: Random oversampling (ROS), SMOTE and Borderline-SMOTE (B-SMOTE) (Han et al., 2005). Also as a baseline score we include classification results without the use of any resampling method.

This paper is organized in 5 sections: section 2 provides an overview of the state-of-art, section 3 describes the proposed methodology, section 4 covers the results and discussion and section 5 presents the conclusions taken from this study.

Figure 1: Example of a complex input space. In this example, a classifier would need to separate the minority class' samples across 4 distinguishable clusters (A, B, C and D).



2. Imbalanced Learning Approaches

Imbalanced learning has been addressed in three different ways: over/undersampling, cost-sensitive training and changes/adaptations in the learning algorithms (Kaur et al., 2019). These approaches impact different phases of the learning process, while over/undersampling can be seen as a pre-processing step, cost-sensitive and changes in the algorithm imply a more customized and complex intervention in the algorithms. In this section, we focus on previous work related with resampling methods, while providing a brief explanation of cost-sensitive and algorithmic level solutions.

All of the most common classifiers used for LULC classification tasks (Khatami et al., 2016; Gavade and Rajpurohit, 2019) are sensitive to class imbalance (Blagus and Lusa, 2010). Algorithm-based approaches typically focus on adaptations based on ensemble classification methods (Mellor et al., 2015) or common non-ensemble based classifiers such as Support Vector Machines (Shao et al., 2014). In (Lee et al., 2016), the reported results show that algorithm-based methods have comparable performance to resampling methods.

Cost-sensitive solutions refer to changes in the importance attributed to each instance through a cost matrix (Huang et al., 2016; Cui et al., 2019; Dong et al., 2017). A relevant cost sensitive solution (Huang et al., 2016) uses the inverse class frequency (i.e., $1/|C_i|$, where C_i refers to the frequency of class i) to give higher weight to minority classes. Cui et al. (Cui et al., 2019) extended this method by adding a hyperparameter β to class weights as $(1-\beta)/(1-\beta^{|C_i|})$. When $\beta = 0$, no re-weighting is done. When $\beta \to 1$, weights are the inverse of the frequency class matrix. Another method (Dong et al., 2017) explores adaptations of Cross-entropy classification loss by adding different formulations of class rectification loss.

Resampling (over/undersampling) is the most common approach to imbalanced learning addressing the problem through data resampling in machine learning in general and remote sensing in particular (Feng et al., 2019). The generation of artificial instances (i.e., augmenting the dataset), based on rare instances, is done independently of any other step in the learning process. Once the procedure is applied, any standard machine learning algorithm can be used. Its simplicity makes resampling strategies particularly appealing for any user (especially the non-sophisticated user) interested in applying several classifiers, while maintaining a simple approach. It is also important to notice that over/undersampling methods can also be easily applied to multiclass problems, common in LULC classification tasks.

2.1. Non-informed resampling methods

There are two main non-informed resampling methods. Random Oversampling (ROS) generates artificial instances through random duplication of minority class instances. This method is used in remote sensing for its simplicity (Sharififar et al., 2019; Hounkpatin et al., 2018), even though its mechanism makes the classifier prone to overfitting (Krawczyk, 2016). Hounkpatin et al. (2018) found that using ROS returned worse results than keeping the original imbalance in their dataset.

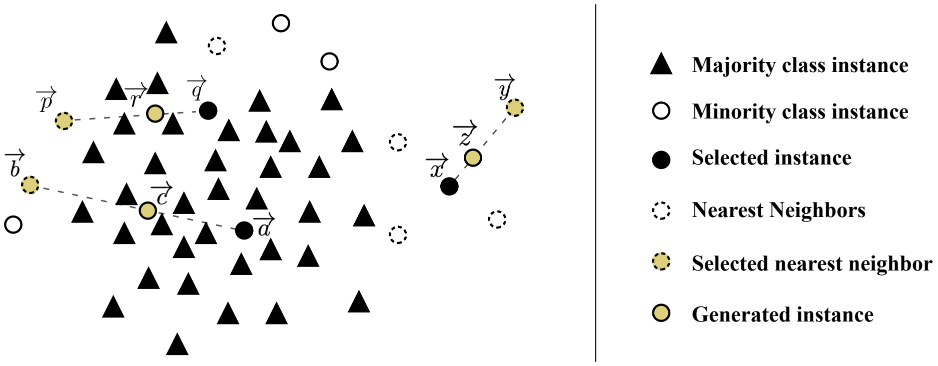
A few of the recent remote sensing studies employed Random Undersampling (RUS) (Ferreira et al., 2019), which randomly removes instances belonging to majority classes. Although it's not as prone to overfitting as ROS, it incurs into information loss by eliminating instances from the majority class (Feng et al., 2019), which can be detrimental to the quality of the results.

Another disadvantage of non-informed resampling methods is their performance-wise inconsistency across classifiers. ROS' impact on the Indian Pines dataset was found inconsistent between Random Forest Classifiers (RFC) and Support Vector Machines (SVM) and lowered the predictive power of an artificial neural network (ANN) (Maxwell et al., 2018). Similarly, RUS is found to generally lead to a lower overall accuracy due to the associated information loss (Maxwell et al., 2018).

2.2. Heuristic methods

The methods presented in this section appear as a means to overcome the insufficiencies found in non-informed resampling. They use either local or global information to generate new, relevant, non-duplicated instances to populate the minority classes and/or remove irrelevant instances from majority classes. In a comparative analysis between over- and undersamplers' performance for LULC classification (Feng et al., 2018) using the rotation forest ensemble classifier, authors found that

Figure 2: Example of SMOTE's data generation process. SMOTE randomly selects instance $\overrightarrow{x}$ and randomly selects one of its k-nearest neighbors $\overrightarrow{y}$ to produce $\overrightarrow{z}$. Noisy instance $\overrightarrow{r}$ was generated by randomly selecting $\overrightarrow{q}$ and randomly selecting its nearest neighbor $\overrightarrow{p}$ from a different minority class cluster. Noisy instance $\overrightarrow{c}$ was generated by randomly selecting the noisy minority class instance $\overrightarrow{a}$ and one of its nearest neighbors $\overrightarrow{b}$.



oversampling methods consistently outperform undersampling methods. This result led us to exclude undersampling from our study.

SMOTE (Chawla et al., 2002) was the first heuristic oversampling algorithm to be proposed and has been the most popular one since then, likely due to its fair degree of simplicity and quality of generated data. It takes a random minority class sample and introduces synthetic instances along the line segment that join a random k minority class nearest neighbor to the selected sample. Specifically, a single synthetic sample $\overrightarrow{z}$ is generated within the line segment of a randomly selected minority class instance $\overrightarrow{x}$ and one of its k nearest neighbors $\overrightarrow{y}$ such that $\overrightarrow{z} = \alpha\overrightarrow{x} + (1-\alpha)\overrightarrow{y}$, where α is a random real number between 0 and 1, as shown in Figure 2.

A number of studies implement SMOTE within the LULC classification context and reported improvements on the quality of the trained predictors (Jozdani et al., 2019; Bogner et al., 2018). Another study proposes an adaptation of SMOTE on an algorithmic level for deep learning applications (Zhu et al., 2020). This method combines both typical computer vision data augmentation techniques, such as image rotation, scaling and flipping on the generated instances to populate minority classes. Another algorithmic implementation is the variational semi-supervised learning model (Cenggoro et al., 2018). It consists of a generative model that allows learning from both labeled and unlabeled instances while using SMOTE to balance the data.

Despite SMOTE's popularity, its limitations have motivated the development of more sophisticated oversampling algorithms (Douzas and Bacao, 2019; Han et al., 2005; Ma and Fan, 2017; Douzas and Bacao, 2017; Douzas et al., 2018; Haibo He et al., 2008). Douzas and Bacao (2019) identify four major weaknesses of the SMOTE algorithm, which can be summarized as:

1. Generation of noisy instances due to random selection of a minority instance to oversample. The random selection of a minority instance makes SMOTE oversampling prone to the amplification of existing noisy data. This has been addressed by variants such as B-SMOTE (Han et al., 2005) and ADASYN (Haibo He et al., 2008).

2. Generation of noisy instances due to the selection of the k nearest neighbors. In the event an instance (or a small number thereof) is not noisy but is isolated from the remaining clusters, known as the "small disjuncts problem" (Holte et al., 1989), much like sample $\overrightarrow{b}$ from Figure 2, the selection of any nearest neighbor of the same class will have a high likelihood of producing a noisy sample.

3. Generation of nearly duplicated instances. Whenever the linear interpolation is done between two instances that are close to each other, the generated instance becomes very similar to its parents and increases the risk of overfitting. G-SMOTE (Douzas and Bacao, 2019) attempts to address both the k nearest neighbor selection mechanism problem as well as the generation of nearly duplicated instances problem.

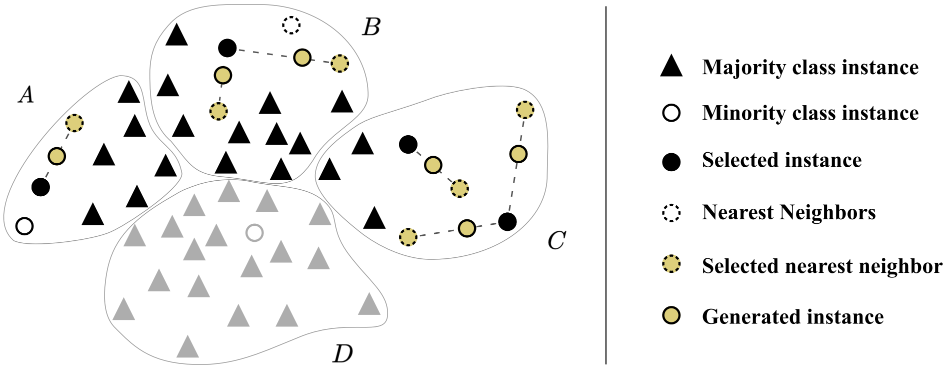
4. Generation of noisy instances due to the use of instances from two different minority class clusters. Although an increased k could potentially avoid the previous problem, it can also lead to the generation of artificial data between different minority clusters, as depicted Figure 2 with the generation of point $\overrightarrow{r}$ using minority class instances $\overrightarrow{p}$ and $\overrightarrow{q}$. Cluster-based oversampling methods attempt to address this problem.

This last issue, the generation of noisy instances due to the existence of several minority class clusters, is particularly relevant in remote sensing. It is frequent that instances belonging to the same minority class can have different spectral signatures, meaning that they will be clustered in different parts of the input space. For example, in the classification of a hyperspectral scene dominated by agricultural activities, patches relating to urban areas may constitute a minority class. These patches frequently refer to different types of land use, such as housing regions, small gardens, asphalt roads, etc., all these containing different spectral signatures. In this context, the use of SMOTE will lead to the generation of noisy instances of the minority class. This problem can be efficiently mitigated through the use of a cluster-based oversampling method. According to our literature review cluster-based oversampling approaches have never been applied in the context of remote sensing. On the other hand, while there are references of the application of cluster-based oversampling in the context of machine learning (Santos et al., 2015; Douzas and Bacao, 2017; Ma and Fan, 2017; Douzas et al., 2018), the multiclass case is rarely addressed, which is a fundamental requirement for the application of oversampling in the context of LULC.

Cluster-based oversampling approaches introduce an additional layer to SMOTE's selection mechanism, which is done through the inclusion of a clustering process. This ensures that both between-class data balance and within-class balance is preserved. The self-organizing map oversampling (SOMO) (Douzas and Bacao, 2017) algorithm transforms the dataset into a 2-dimensional input, where the areas with the highest density of minority samples are identified. SMOTE is then used to oversample each of the identified areas separately. ClUstered REsampling SMOTE (CURE-SMOTE) (Ma and Fan, 2017) applies a hierarchical clustering algorithm to discard isolated minority instances before applying SMOTE. Although it avoids noise generation problems, it ignores within-class data distribution. Another method (Santos et al., 2015) uses K-means to cluster the entire input space and applies SMOTE to clusters with the fewest instances, regardless of their class label. The label of the generated instance is copied from one of its parents. This method cannot ensure a balanced dataset since class imbalance is not specifically addressed, but rather dataset imbalance.

K-means SMOTE (Douzas et al., 2018) avoids noisy data generation by modifying the data selection mechanism. It employs k-means clustering to identify safe areas using cluster-specific Imbalance Ratio (IR, defined by $\frac{count(C_{majority})}{count(C_{minority})}$) and determine the quantity of generated samples per cluster based on a density measure. These samples are finally generated using the SMOTE algorithm. The K-means SMOTE's data generation process is depicted in Figure 3. Note that the number of samples generated for each cluster varies according to the sparsity of each cluster (the sparser the cluster is, the more samples will be generated) and a cluster is rejected if the cluster's IR surpasses the threshold. Therefore, this method can be combined with any data generation mechanism, such as G-SMOTE. Also K-means SMOTE includes the SMOTE algorithm as a special case when the number of clusters is set to one. Consequently, K-means SMOTE returns results as good as or better than SMOTE.

Figure 3: Example of K-means SMOTE's data generation process. Clusters A, B and C are selected for oversampling, whereas cluster D was rejected due to its high imbalance ratio. The oversampling is done using the SMOTE algorithm and the k nearest neighbors selection only considers instances within the same cluster.



Although no other study was found to implement cluster-based oversampling, another study (Douzas et al., 2019) compared the performance of SMOTE, ROS, ADASYN, B-SMOTE and G-SMOTE in a highly imbalanced LULC classification dataset. The authors found that G-SMOTE consistently outperformed the remaining oversampling algorithms regardless of the classifier used.

This paper's main contributions are:

- Propose a cluster-based multiclass oversampling method appropriate for LULC classification and compare its performance with the remaining oversamplers in a multiclass context with seven benchmark LULC classification datasets. Allows us to check the oversamplers' performance across benchmark LULC datasets.

- Introducing a cluster-based oversampling algorithm within the remote sensing domain, as well as comparing its performance with the remaining oversamplers in a multiclass context.

- Make available to the remote sensing community the implementation of the algorithm in a Python library and the experiment's source code.

3. Methodology

The purpose of this work is to understand the performance of K-means SMOTE as opposed to other popular and/or state-of-the-art oversamplers for LULC classification. This is done using 7 datasets with predominantly land use information, along with 3 evaluation metrics and 3 classifiers to evaluate the performance of oversamplers. In this section we describe the datasets, evaluation metrics, oversamplers, classifiers and software used as well as the procedure developed.

3.1. Datasets

The datasets used were extracted from publicly available hyperspectral scenes. Information regarding each of these scenes is provided in this subsection. The data collection and preprocessing pipeline is shown in Figure 4 and is common to all hyperspectral scenes:

1. Data collection of publicly available hyperspectral scenes. The original hyperspectral scenes and ground truth data were collected from a single publicly available data repository available here.

2. Conversion of each hyperspectral scene to a structured dataset and removal of instances with no associated LULC class. This done to reshape the dataset from $(h, w, b + gt)$ into a conventional dataframe of shape $(h * w, b + gt)$, where gt, h, w and b represents the ground truth, height, width and number of bands in the scene, respectively. The pixels without ground truth information are discarded from further analysis.

3. Stratified random sampling to maintain similar class proportions on a sample of 10% of each dataset. This is done by computing the relative class frequencies in the original hyperspectral scene (minus the class representing no ground truth availability) and retrieving a sample that ensures the original relative class frequencies remain unchanged.

4. Removal of instances belonging to a class with frequency lower than 20 or higher than 1000. This is done to maintain the datasets to a practicable size due to computational constraints, while conserving the relative LULC class frequencies and data distribution.

5. Data normalization using the MinMax scaler. This ensures all features (i.e., bands) are in the same scale. In this case, the data was rescaled between 0 and 1.

Table 1 provides a description of the final datasets used for this work, sorted according to its IR. Figure 5 shows the original hyperspectral scene out of which the dataset used in this experiment were extracted. In the representation of the ground truth of these scenes, the blue regions in the ground truth of each hyperspectral

Figure 4: Data collection and preprocessing pipeline.

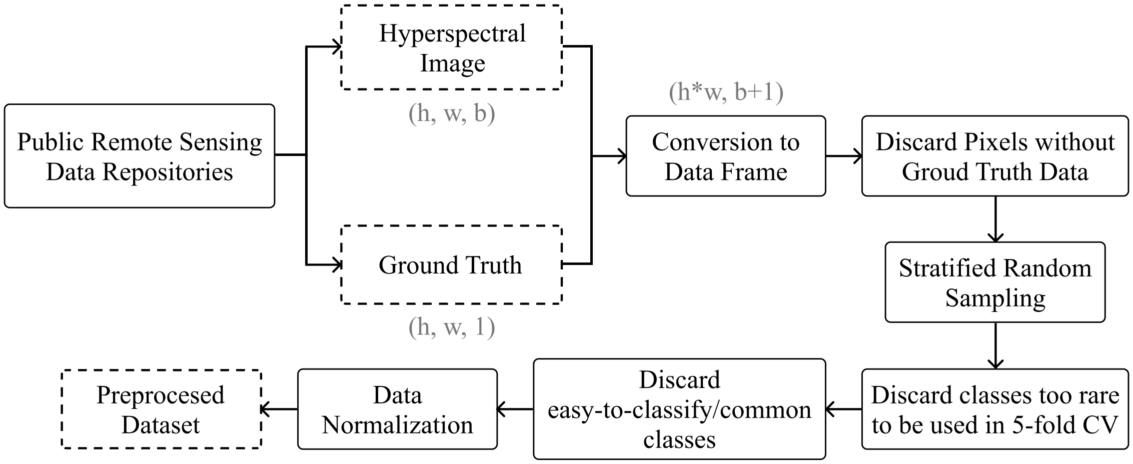

Public Remote Sensing Data Repositories
Hyperspectral Image
(h, w, b)
Ground Truth
(h, w, 1)
(h*w, b+1)
Conversion to Data Frame
Discard Pixels without Groud Truth Data
Stratified Random Sampling
Discard classes too rare to be used in 5-fold CV
Discard easy-to-classify/common classes
Data Normalization
Preprocesed Dataset

Table 1: Description of the datasets used for this experiment.

Dataset	Features	Instances	Min. Instances	Maj. Instances	IR	Classes
Botswana	145	288	20	41	2.05	11
Pavia Centre	102	3898	278	879	3.16	7
Kennedy Space Center	176	497	23	80	3.48	11
Salinas A	224	535	37	166	4.49	6
Pavia University	103	2392	89	679	7.63	8
Salinas	224	4236	91	719	7.9	15
Indian Pines	220	984	21	236	11.24	11

scene represent unlabeled regions (i.e., no ground truth is available). Particularly, in the Botswana and Kennedy Space Center scenes the truth was photointerpreted in more limited regions of the scene. However, the scenes are still represented as they are in order to maintain a standardized analysis over all datasets extracted for the experiment.

Botswana

The Botswana scene was acquired by the Hyperion sensor on the NASA EO-1 satellite over the Okavango Delta, Botswana in 2001-2004 at a 30m spatial resolution. Data preprocessing was performed by the UT Center for Space Research. The scene comprises a 1476 × 256 pixels with 145 bands and 14 classes regarding land cover types in seasonal and occasional swamps, as well as drier woodlands (see figure 5a).

Pavia Center and University

Both Pavia Center and University scenes were acquired by the ROSIS sensor. These scenes are located in Pavia, northern Italy. Pavia Center is a 1096×1096 pixels image with 102 spectral bands, whereas Pavia University is a 610×610 pixels image with 103 spectral bands. Both images have a geometrical resolution of 1.3m and their ground truths are composed of 9 classes each (see Figures 5b and 5c).

Kennedy Space Center

The Kennedy Space Center scene was acquired by the AVIRIS sensor over the Kennedy Space Center, Florida, on March 23, 1996. Out of the original 224 bands, water absorption and low SNR bands were removed and a total of 176 bands at a spatial resolution of 18m are used. The scene is a 512×614 pixel image and contains a total of 16 classes (see figure 5d).

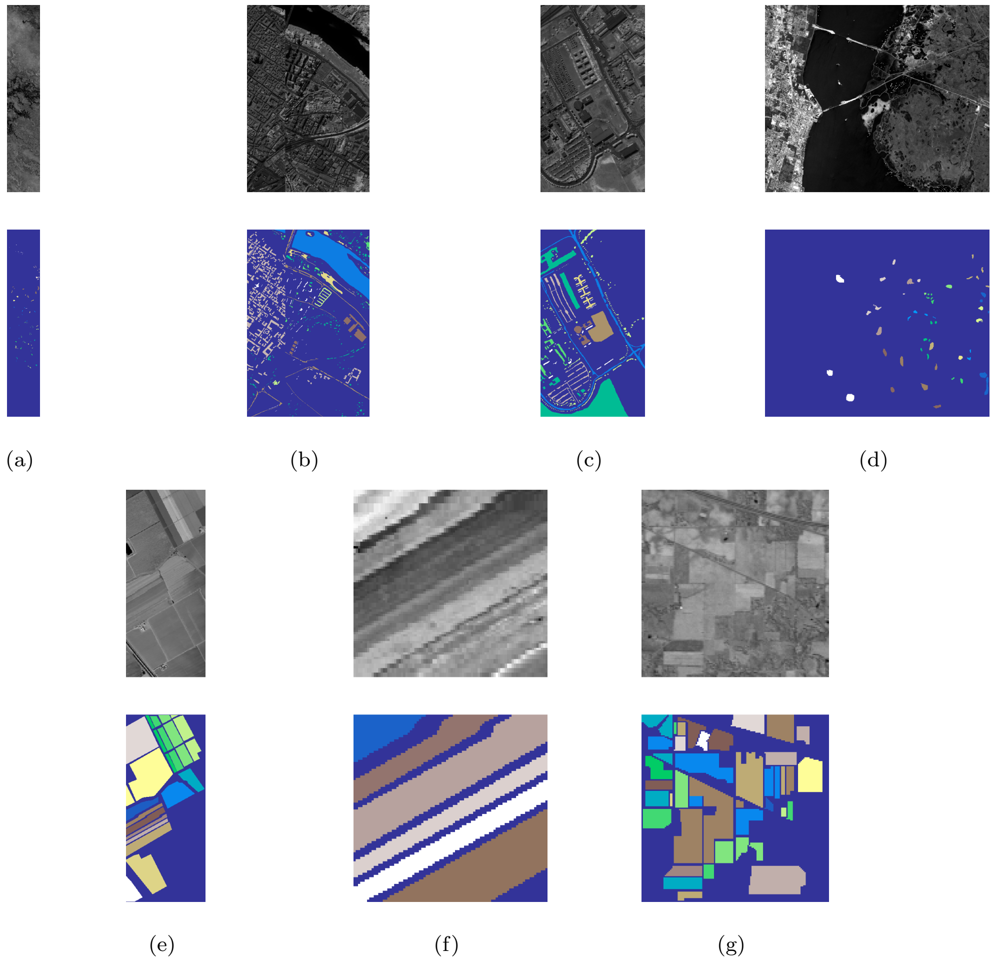
Salinas and Salinas-A

These scenes were collected by the AVIRIS sensor over Salinas Valley, California and contain at-sensor radiance data. Salinas is a 512×217 pixels image with 224 bands and 16 classes regarding vegetables, bare soil and vineyard fields (see Figure 5e). Salinas-A, a subscene of Salinas, comprises 86×83 pixels and contains 6 classes regarding vegetables (see Figure 5f). These scenes have a geometrical resolution of 3.7m.

Indian Pines

The Indian Pines scene (Baumgardner et al., 2015) was collected on June 12, 1992 and consists of AVIRIS hyperspectral image data covering the Indian Pine Test Site 3, located in North-western Indiana, USA. As a subset of a larger scene, it is composed of 145×145 pixels (see Figure 5g) and 220 spectral reflectance bands in the wavelength range 400 to 2500 nanometers at a spatial resolution of 20m. Approximately two thirds of this scene is composed by agriculture and the other third is composed of forest and other natural perennial vegetation. Additionally, the scene also contains low density buildup areas.

Figure 5: Gray scale visualization of a band (top row) and ground truth (bottom row) of each scene used in this study. (a) Botswana, (b) Pavia Center, (c) Pavia University, (d) Kennedy Space Center, (e) Salinas, (f) Salinas A, (g) Indian Pines.



3.2. Machine Learning Algorithms

To assess the quality of the K-means SMOTE algorithm, three other oversampling algorithms were used for benchmarking. ROS and SMOTE were chosen for their simplicity and popularity. B-SMOTE chosen as a popular variation of the SMOTE algorithm. We also include the classification results of no oversampling (NONE) as a baseline.

To assess the performance of each oversampler, we use the classifiers Logistic Regression (LR) (Nelder and Wedderburn, 1972), K-Nearest Neighbors (KNN) (Cover and Hart, 1967) and Random Forest (RF) (Liaw et al., 2002). This choice was based on the classifiers' popularity for LULC classification, learning type and training time (Maxwell et al., 2018; Gavade and Rajpurohit, 2019).

3.3. Evaluation Metrics

Most of the satellite-based LULC classification studies (nearly 80%) employ *Overall Accuracy* (OA) and the *Kappa Coefficient* (Gavade and Rajpurohit, 2019). Although, some authors argue that both evaluation metrics, even when used simultaneously, are insufficient to fully address the area estimation and uncertainty information needs (Olofsson et al., 2013; Pontius and Millones, 2011). Other metrics like User's Accuracy (or *Precision*) and Producer's Accuracy (or *Recall*) are also common metrics to evaluate per-class prediction power. These metrics consist of ratios employing the True and False Positives (*TP* and *FP*, number of correctly/incorrectly classified instances of a given class) and True and False Negatives (*TN* and *FN*, number of correctly/incorrectly classified instances as not belonging to a given class). These metrics are formulated as $Precision = \frac{TP}{TP+FP}$ and $Recall = \frac{TP}{TP+FN}$. While metrics like OA and *Kappa Coefficient* are significantly affected by imbalanced class distributions, *F-Score* is less sensitive to data imbalance and a more appropriate choice for performance evaluation (Jeni et al., 2013).

The datasets used present significantly high IRs (see Table 1). Therefore, it is especially important to attribute equal importance to the predictive power of all classes, which does not happen with OA and *Kappa Coefficient*. In this study, we employ 3 evaluation metrics: 1) *G-mean*, since it is not affected by skewed class distributions, 2) *F-Score*, as it proved to be a more appropriate metric for this problem when compared to other commonly used metrics (Jeni et al., 2013), and 3) *Overall Accuracy*, for discussion purposes.

- The *G-mean* consists of the geometric mean of $Specificity = \frac{TN}{TN+FP}$ and *Sensitivity* (also known as *Recall*). For multiclass problems, The *G-mean* is expressed as:

$$\textit{G-mean} = \sqrt{\overline{Sensitivity} \times \overline{Specificity}}$$

- *F-score* is the harmonic mean of *Precision* and *Recall*. The *F-score* for the multi-class case can be calculated using their average per class values (He and Garcia, 2009):

$$\mathit{F\text{-}score} = 2\frac{\overline{\mathit{Precision}} \times \overline{\mathit{Recall}}}{\overline{\mathit{Precision}} + \overline{\mathit{Recall}}}$$

- *Overall Accuracy* is the number of correctly classified instances divided by the total amount of instances. Having c as the label of the various classes, *Accuracy* is given by the following formula:

$$\mathit{Accuracy} = \frac{\sum\limits_{c} \mathrm{TP}_c}{\sum\limits_{c} (\mathrm{TP}_c + \mathrm{FP}_c)}$$

3.4. Experimental Procedure

The procedure for the experiment started with the definition of a hyperparameter search grid, where a list of possible values for each relevant hyperparameter in both classifiers and oversamplers is stored. Based on this search grid, all possible combinations of oversamplers, classifiers and hyperparameters are formed. Finally, for each dataset, hyperparameter combination and initialization we use the evaluation strategy shown in Figure 6: k-fold cross-validation strategy where $k = 5$ to train each model defined and save the averaged scores of each split.

In the 5-fold cross validation strategy, a combination of oversampler, classifier and hyperparameters vector is fit 5 times per dataset. Before the training phase, the training set (containing $\frac{4}{5}$ of the dataset) is oversampled using one of the methods described (except for the baseline method NONE), creating an augmented dataset with the exact same number of instances for each class. The newly formed training dataset is used to train the classifier and the test set ($\frac{1}{5}$ of the dataset) is used to evaluate the performance of the classifier. The evaluation scores are then averaged over the 5 times the process is repeated. The range of hyperparameters used are shown in table 2. The definition of hyperparameters for the K-means SMOTE oversampler is defined according to the recommendations discussed in the original K-means SMOTE paper (Douzas et al., 2018).

3.5. Software Implementation

The experiment was implemented using the Python programming language, using the Scikit-Learn (Pedregosa et al., 2011), Imbalanced-Learn (Lemaître et al., 2017), Geometric-SMOTE, Cluster-Over-Sampling and Research-Learn libraries. All functions, algorithms, experiments and results are provided at the GitHub repository of the project.

Figure 6: Experimental procedure. The performance metrics are averaged over the 5 folds across each of the 3 different initializations of this procedure for a given combination of oversampler, classifier and hyperparameter definition.

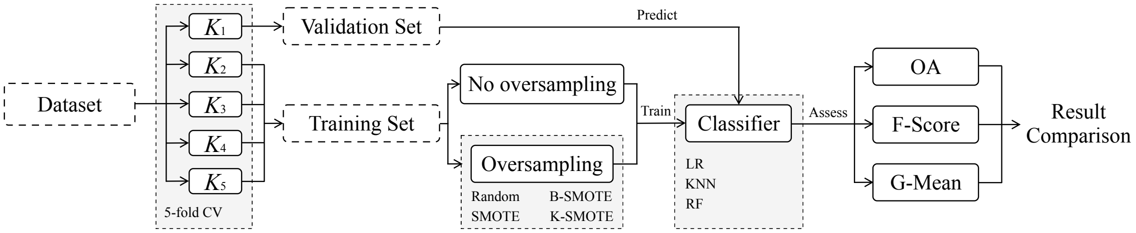


Table 2: Hyper-parameters grid. * One cluster is generated in total, a corner case that mimics the behavior of SMOTE

Classifier	Hyperparameters	Values
LR	maximum iterations	10000
KNN	# neighbors	3, 5, 8
RF	maximum depth	None, 3, 6
	# estimators	50, 100, 200
Oversampler		
K-means SMOTE	# neighbors	3, 5
	# clusters (as % of number of instances)	1*, 0.1, 0.3, 0.5, 0.7, 0.9
	Exponent of mean distance	auto, 2, 5, 7
	IR threshold	auto, 0.5, 0.75, 1.0
SMOTE	# neighbors	3, 5
BORDERLINE SMOTE	# neighbors	3, 5

Table 3: Results for mean ranking of oversamplers across datasets.

Classifier	Metric	NONE	ROS	SMOTE	B-SMOTE	K-SMOTE
LR	Accuracy	3.14 ± 0.47	3.00 ± 0.49	3.29 ± 0.42	4.29 ± 0.43	**1.29 ± 0.18**
LR	F-score	3.86 ± 0.26	2.71 ± 0.47	3.00 ± 0.44	4.29 ± 0.42	**1.14 ± 0.14**
LR	G-mean	4.29 ± 0.18	2.57 ± 0.48	2.57 ± 0.3	4.29 ± 0.42	**1.29 ± 0.18**
KNN	Accuracy	2.14 ± 0.55	4.14 ± 0.34	3.43 ± 0.43	3.86 ± 0.34	**1.43 ± 0.2**
KNN	F-score	2.86 ± 0.4	4.29 ± 0.29	2.71 ± 0.61	3.86 ± 0.34	**1.29 ± 0.18**
KNN	G-mean	3.57 ± 0.48	3.86 ± 0.4	1.86 ± 0.46	4.14 ± 0.26	**1.57 ± 0.2**
RF	Accuracy	3.29 ± 0.52	3.14 ± 0.4	3.14 ± 0.51	4.29 ± 0.29	**1.14 ± 0.14**
RF	F-score	3.86 ± 0.46	3.14 ± 0.51	3.00 ± 0.44	4.00 ± 0.22	**1.00 ± 0.0**
RF	G-mean	4.86 ± 0.14	3.00 ± 0.31	2.57 ± 0.3	3.57 ± 0.37	**1.00 ± 0.0**

4. Results & Discussion

When evaluating the performance of an algorithm across multiple datasets, it is generally recommended to avoid direct score comparisons and use classification rankings instead (Demšar, 2006a). This is done by assigning a ranking to oversamplers based on the different combinations of classifier, metric and dataset used. These rankings are also used for the statistical analyses presented in Section 4.2.

The rank values are assigned based on the mean validation scores resulting from the experiment described in Section 3. The averaged ranking results are computed over 3 different initialization seeds and a 5 fold cross validation scheme, returning a real number within the interval $[1, 5]$. The mean rankings are presented in Table 3.

4.1. Results

The mean ranking of oversamplers is presented in Table 3. This ranking was computed by averaging the ranks of the mean cross-validation scores per dataset, oversampler and classifier. K-means SMOTE achieves the best mean ranking across datasets with low standard deviation.

The mean cross-validation scores are shown in Table 4. As discussed previously in this section, the disparity of performance levels across datasets makes the analysis of these scores less informative.

The mean cross-validation scores for each dataset are presented in Table 5. This table allows the direct comparison of the performance metrics being analysed.

Table 4: Mean cross-validation scores of oversamplers.

Classifier	Metric	NONE	ROS	SMOTE	B-SMOTE	K-SMOTE
LR	Accuracy	0.906 ± 0.039	0.904 ± 0.04	0.904 ± 0.04	0.901 ± 0.04	**0.909 ± 0.038**
LR	F-score	0.891 ± 0.041	0.893 ± 0.042	0.893 ± 0.042	0.890 ± 0.042	**0.898 ± 0.04**
LR	G-mean	0.936 ± 0.025	0.940 ± 0.025	0.940 ± 0.025	0.937 ± 0.025	**0.943 ± 0.024**
KNN	Accuracy	0.879 ± 0.043	0.865 ± 0.048	0.867 ± 0.05	0.862 ± 0.054	**0.881 ± 0.045**
KNN	F-score	0.859 ± 0.05	0.853 ± 0.049	0.861 ± 0.047	0.851 ± 0.053	**0.866 ± 0.048**
KNN	G-mean	0.919 ± 0.03	0.920 ± 0.029	0.926 ± 0.027	0.918 ± 0.03	**0.927 ± 0.027**
RF	Accuracy	0.898 ± 0.032	0.901 ± 0.031	0.900 ± 0.031	0.898 ± 0.032	**0.905 ± 0.031**
RF	F-score	0.879 ± 0.041	0.885 ± 0.037	0.887 ± 0.036	0.883 ± 0.037	**0.891 ± 0.036**
RF	G-mean	0.930 ± 0.024	0.935 ± 0.022	0.937 ± 0.021	0.935 ± 0.021	**0.939 ± 0.02**

Table 5: Mean cross-validation scores for each dataset. Legend: IP - Indian Pines, KSC - Kennedy Space Center, PC - Pavia Center, PU - Pavia University, SA - Salinas A.

Dataset	Classifier	Metric	NONE	ROS	SMOTE	B-SMOTE	K-SMOTE
Botswana	LR	Accuracy	0.920	0.917	0.920	0.921	**0.927**
Botswana	RF	F-score	0.865	0.877	0.872	0.870	**0.883**
Botswana	RF	Accuracy	0.873	0.884	0.877	0.877	**0.890**
Botswana	KNN	G-mean	0.924	0.918	0.930	0.923	**0.933**
Botswana	RF	G-mean	0.925	0.933	0.929	0.928	**0.936**
Botswana	KNN	Accuracy	0.875	0.862	0.881	0.869	**0.889**
Botswana	LR	G-mean	0.952	0.950	0.952	0.952	**0.956**
Botswana	LR	F-score	0.913	0.909	0.913	0.914	**0.921**
Botswana	KNN	F-score	0.859	0.850	0.873	0.859	**0.879**
PC	LR	F-score	0.944	0.947	0.947	0.941	**0.948**
PC	RF	G-mean	0.959	0.964	**0.965**	0.961	**0.965**
PC	RF	F-score	0.928	0.932	0.931	0.928	**0.933**
PC	RF	Accuracy	0.938	0.941	0.940	0.938	**0.942**
PC	KNN	G-mean	0.953	0.955	**0.957**	0.954	**0.957**
PC	KNN	Accuracy	**0.926**	0.920	0.923	0.924	**0.926**
PC	LR	G-mean	0.968	0.972	0.972	0.966	**0.973**
PC	LR	Accuracy	0.954	0.955	0.955	0.950	**0.956**
PC	KNN	F-score	**0.915**	0.909	0.913	0.913	**0.915**
KSC	LR	Accuracy	0.904	0.905	0.905	0.899	**0.909**
KSC	LR	F-score	0.868	0.873	0.874	0.862	**0.877**

Table 5: Mean cross-validation scores for each dataset. Legend: IP - Indian Pines, KSC - Kennedy Space Center, PC - Pavia Center, PU - Pavia University, SA - Salinas A.

Dataset	Classifier	Metric	NONE	ROS	SMOTE	B-SMOTE	K-SMOTE
KSC	LR	G-mean	0.928	0.932	0.932	0.924	**0.934**
KSC	KNN	Accuracy	0.855	0.859	0.862	0.857	**0.865**
KSC	KNN	F-score	0.808	0.819	**0.827**	0.810	0.826
KSC	KNN	G-mean	0.893	0.901	**0.906**	0.895	0.905
KSC	RF	Accuracy	0.860	0.859	0.863	0.859	**0.868**
KSC	RF	F-score	0.817	0.815	0.826	0.816	**0.832**
KSC	RF	G-mean	0.898	0.899	0.905	0.898	**0.907**
SA	RF	Accuracy	0.980	0.983	0.984	0.979	**0.985**
SA	KNN	G-mean	0.992	0.989	0.990	0.991	**0.993**
SA	KNN	F-score	0.986	0.979	0.981	0.982	**0.987**
SA	KNN	Accuracy	0.987	0.979	0.982	0.983	**0.988**
SA	LR	G-mean	0.985	0.988	**0.990**	0.987	0.989
SA	LR	F-score	0.976	0.979	**0.982**	0.977	**0.982**
SA	LR	Accuracy	0.979	0.981	0.983	0.979	**0.984**
SA	RF	G-mean	0.987	0.988	0.989	0.986	**0.990**
SA	RF	F-score	0.979	0.982	0.983	0.978	**0.984**
PU	KNN	F-score	**0.891**	0.868	0.868	0.874	**0.891**
PU	LR	Accuracy	**0.905**	0.897	0.897	0.891	0.904
PU	LR	F-score	0.890	0.894	0.894	0.888	**0.898**
PU	LR	G-mean	0.932	0.947	0.947	0.942	**0.949**
PU	KNN	Accuracy	**0.895**	0.867	0.865	0.873	**0.895**
PU	KNN	G-mean	0.940	0.935	0.936	0.936	**0.941**
PU	RF	Accuracy	**0.912**	0.908	0.907	0.908	0.911
PU	RF	F-score	**0.909**	0.906	0.906	0.908	**0.909**
PU	RF	G-mean	0.946	0.946	0.948	0.948	**0.949**
Salinas	RF	G-mean	0.989	0.989	0.989	0.989	**0.990**
Salinas	RF	F-score	0.979	0.979	0.977	0.978	**0.980**
Salinas	RF	Accuracy	0.984	0.983	0.983	0.983	**0.985**
Salinas	KNN	G-mean	0.977	0.978	**0.981**	0.976	**0.981**
Salinas	KNN	Accuracy	**0.970**	0.967	0.969	0.967	**0.970**
Salinas	LR	G-mean	0.992	**0.993**	0.992	0.992	**0.993**
Salinas	LR	F-score	0.985	**0.986**	0.985	0.985	**0.986**
Salinas	LR	Accuracy	**0.990**	**0.990**	0.989	**0.990**	**0.990**
Salinas	KNN	F-score	0.959	0.957	**0.960**	0.957	**0.960**

Table 5: Mean cross-validation scores for each dataset. Legend: IP - Indian Pines, KSC - Kennedy Space Center, PC - Pavia Center, PU - Pavia University, SA - Salinas A.

Dataset	Classifier	Metric	NONE	ROS	SMOTE	B-SMOTE	K-SMOTE
IP	LR	F-score	0.662	0.663	0.659	0.659	**0.674**
IP	LR	G-mean	0.798	0.801	0.798	0.797	**0.807**
IP	KNN	Accuracy	**0.644**	0.602	0.589	0.557	0.632
IP	RF	G-mean	0.806	0.826	0.835	0.831	**0.838**
IP	KNN	F-score	0.593	0.591	0.603	0.560	**0.604**
IP	KNN	G-mean	0.757	0.764	**0.782**	0.751	0.781
IP	RF	Accuracy	0.742	0.747	0.747	0.740	**0.752**
IP	RF	F-score	0.673	0.704	0.713	0.701	**0.714**
IP	LR	Accuracy	0.687	0.681	0.680	0.678	**0.692**

4.2. Statistical Analysis

The experiment's multi-dataset context was used to perform a Friedman test (Friedman, 1937). Table 6 shows the results obtained in the Friedman test performed, where the null hypothesis is rejected in all cases. The rejection of the null hypothesis implies that the differences between the differences among the different oversamplers are not random, in other words, these differences are statistically significant.

Table 6: Results for Friedman test. Statistical significance is tested at a level of $\alpha = 0.05$. The null hypothesis is that there is no difference in the classification outcome across oversamplers.

Classifier	Metric	p-value	Significance
LR	Accuracy	9.8e-03	True
LR	F-score	2.3e-03	True
LR	G-mean	9.8e-04	True
KNN	Accuracy	4.3e-03	True
KNN	F-score	4.3e-03	True
KNN	G-mean	3.0e-03	True
RF	Accuracy	5.5e-03	True
RF	F-score	2.9e-03	True
RF	G-mean	1.8e-04	True

A Wilcoxon signed-rank test (Wilcoxon, 1992) was also performed to understand whether K-means SMOTE's superiority was statistically significant across datasets and oversamplers, as suggested in (Demšar, 2006b). This method is used as an alternative to the paired Student's t-test, since the distribution of the differences between the two samples cannot be assumed as normally distributed. The null hypothesis of the test is that K-means SMOTE's performance is similar to the compared oversampler (i.e., the oversamplers used follow a symmetric distribution around zero).

Table 8: *p-values* of the Wilcoxon signed-rank test. Boldface values are statistically significant at a significance level of $\alpha = 0.05$.

Dataset	NONE	ROS	SMOTE	B-SMOTE
Botswana	**3.1e-02**	**3.9e-03**	**3.9e-03**	**3.9e-03**
Pavia Centre	**3.1e-02**	**3.9e-03**	**1.2e-02**	**3.9e-03**
Kennedy Space Center	**3.1e-02**	**3.9e-03**	**2.7e-02**	**3.9e-03**
Salinas A	**3.1e-02**	**3.9e-03**	**1.2e-02**	**3.9e-03**
Pavia University	**3.1e-02**	**3.9e-03**	**3.9e-03**	**3.9e-03**
Salinas	**3.1e-02**	5.5e-02	**2.7e-02**	**3.9e-03**
Indian Pines	**3.1e-02**	**3.9e-03**	**7.8e-03**	**3.9e-03**

4.3. Discussion

The mean rankings presented in Table 3 show that on average, K-means SMOTE produced the best results for every classifier and performance metric used. This is due to the clustering phase and subsequent selection of data to be considered for oversampling. By successfully clustering and selecting the relevant areas in the data space to oversample, the generation of artificial instances is done only in the context of minority regions that represent well their spectral signature.

As previously discussed, the direct comparison of performance metrics averaged over various datasets is not recommended due to the varying levels of performance of classifiers across datasets (Demšar, 2006a). Nonetheless, these results are shown in Table 4 to provide a fuller picture of the results obtained in the experiment. We found that on average K-means SMOTE provides increased performance, regardless of the classifier and performance metric used. More importantly, K-means SMOTE guaranteed a more consistent performance across datasets and with less variability, which can be attested in Tables 3, 4 and 5.

As discussed in Subsection 3.3, Evaluation Metrics, our results are consistent with the findings in (Olofsson et al., 2013; Pontius and Millones, 2011). Particularly, we consider the results obtained in our experiment using Overall Accuracy to be less informative than the results obtained with the remaining performance metrics, since this metric is affected by imbalanced class distributions. The majority class bias in this metric can be observed in our experiment in Table 3 with the classifiers LR and KNN, where the control method (NONE) is only outperformed by K-means SMOTE. This effect is observed with more detail in Table 4, where the benchmark oversamplers are outperformed by the control method in 16 out of 63 tests (approximately 25%). Out of these, most refer to tests using overall accuracy among the four datasets with highest IR, showing the overall accuracy's class imbalance bias discussed in (Olofsson et al., 2013; Pontius and Millones, 2011). The K-means SMOTE oversampler is only outperformed by the control method in 3 of tests (all of them using overall accuracy). This is an improvement over the benchmark oversamplers, showing that generally K-means SMOTE is the best choice even when overall accuracy is used as the main performance metric.

In the majority of the cases, K-means SMOTE was able to generate higher quality data due to the non-random selection of data spaces to oversample. This can be seen in the performance of the classifiers trained on top of this data generation step, making it a more informed data generation method in the context of LULC.

The performance of both oversamplers and classifiers is generally dependent on the dataset being used. Although both absolute and relative scores between the different oversamplers are dependent on the choice of metric and classifier, K-means SMOTE's relative performance is consistent across datasets and generally outperforms the remaining oversampling methods in 56 of the 63 tests (approximately 89%). The mean cross-validation results found in Table 5 show that performance-wise, K-means SMOTE is always better than or as good as SMOTE, with the exception of 4 situations (representing 6% of the tests done), in which cases the percentage point difference is neglectable (≤ 0.1 percentage points).

The statistical tests showed that not only there is a statistically significant difference across the oversamplers used in this problem (found in the Friedman test presented in Table 6), but also that K-means SMOTE's superior performance is statistically significant at a level of 0.05 in 27 out of 28 tests in the Wilcoxon signed-rank test shown in Table 8 (approximately 96% of the tests performed). This shows that, in most cases, the usage of k-Means SMOTE improves the quality of LULC classification when compared to using SMOTE in its original format, which remains the most popular oversampler among the remote sensing community.

Although the usage of K-means SMOTE successfully captured the spectral signatures of the minority classes, it was done using K-means, a problem-agnostic clusterer. Consequently, the implementation of this method using a GIS-specific clusterer that considers the geographical traits of different regions (e.g., using the sampled pixels' geographical coordinates), may be a promising direction towards the development of more appropriate oversampling techniques in the remote sensing domain.

5. Conclusion

This research paper was motivated by the challenges faced when classifying rare classes for LULC mapping. Cluster-based oversampling is especially useful in this context because the spectral signature of a given class often varies, depending on its geographical distribution and the time period within which the image was acquired. This induces the representation of minority classes as small clusters in the input space. As a result, training a classifier capable of identifying LULC minority classes in the hyper/multi-spectral scene over different areas or periods becomes particularly challenging. The clustering procedure, performed before the data generation phase, allows for a more accurate generation of minority samples, as it identifies these minority clusters.

A number of existing methods to address the imbalanced learning problem were identified and their limitations discussed. Typically, algorithm-based approaches and cost-sensitive solutions are not only difficult to implement, but they are also context dependent. In this paper we focused on oversampling methods due to their widespread usage, easy implementation and flexibility. Specifically, this paper demonstrated the efficacy of a recent oversampler, K-Means SMOTE, applied in a multi-class context for Land Cover Classification tasks. This was done with sampled data from seven well known and naturally imbalanced benchmark datasets: Indian Pines, Pavia Center, Pavia University, Salinas, Salinas A, Botswana and Kennedy Space Center. For each combination of dataset, oversampler and classifier, the results of every classification task was averaged across a 5 fold stratification strategy with 3 different initialization seeds, resulting in a mean validation score of 15 classification tasks. The mean validation score of each combination was then used to perform the analyses presented in this report.

In 56 out of 63 classification tasks (approximately 89%), K-means SMOTE led to better results than ROS, SMOTE, B-SMOTE and no oversampling. More importantly, we found that K-Means SMOTE is always better or equal than the second best oversampling method. K-means SMOTE's performance was independent from both the classifier and performance metric under analysis. In general, K-means SMOTE shows a higher performance among the non tree-based classifiers employed (LR and KNN) when compared with the remaining oversamplers, where these oversamplers generally failed to improve the quality of classification. Although these findings are case dependent, they are consistent with the results presented in (Douzas et al., 2018). The proposed method also had the most consistent results across datasets, since it produced the lowest standard deviations across datasets in 7 out of 9 cases for both analyses, either based on ranking or mean cross-validation scores.

The proposed algorithm is a generalization of the original SMOTE algorithm. In fact, the SMOTE algorithm represents a corner case of K-means SMOTE i.e., when the number of clusters equals to 1. Its data selection phase differs from the one used in SMOTE and Borderline SMOTE, providing artificially augmented datasets with less noisy data than the commonly used methods. This allows the training of classifiers with better defined decision boundaries, especially in the most important regions of the data space (the ones populated by a higher percentage of minority class instances).

As stated previously, the usage of this oversampler is technically simple. It can be applied to any classification problem relying on an imbalanced dataset, alongside any classifier. K-means SMOTE is available as an open source implementation for the Python programming language (see Subsection 3.5). Consequently, it can be a useful tool for both remote sensing researchers and practitioners.

6. Funding

This research was funded by ”Fundação para a Ciência e a Tecnologia” (Portugal), grants' number PCIF/SSI/0102/2017 - foRESTER and DSAIPA/AI/0100/2018 - IPSTERS.

7. Data Availability

The data reported in this study is publicly available. It can be retrieved and preprocessed using the Python source code provided at https://github.com/joaopfonseca/research. Alternatively, the original data is available at http://www.ehu.eus/ccwintco/index.php?title=Hyperspectral_Remote_Sensing_Scenes.

8. Conflicts of Interest

The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

Abdi, L. and S. Hashemi (2016, jan). To Combat Multi-Class Imbalanced Problems by Means of Over-Sampling Techniques. *IEEE Transactions on Knowledge and Data Engineering 28*(1), 238–251.

Alonso-Sarria, F., C. Valdivieso-Ros, and F. Gomariz-Castillo (2019, dec). Isolation Forests to Evaluate Class Separability and the Representativeness of Training and Validation Areas in Land Cover Classification. *Remote Sensing 11*(24), 3000.

Baumgardner, M. F., L. L. Biehl, and D. A. Landgrebe (2015, sep). 220 Band AVIRIS Hyperspectral Image Data Set: June 12, 1992 Indian Pine Test Site 3. *Purdue University Research Repository*.

Blagus, R. and L. Lusa (2010, oct). Class prediction for high-dimensional class-imbalanced data. *BMC bioinformatics 11*(1), 523.

Bogner, C., B. Seo, D. Rohner, and B. Reineking (2018, jan). Classification of rare land cover types: Distinguishing annual and perennial crops in an agricultural catchment in South Korea. *PLoS ONE 13*(1).

Cenggoro, T. W., S. M. Isa, G. P. Kusuma, and B. Pardamean (2018, nov). Classification of imbalanced land-use/land-cover data using variational semi-supervised learning. In *Proceedings - 2017 International Conference on Innovative and Creative Information Technology: Computational Intelligence and IoT, ICITech 2017*, Volume 2018-Janua, pp. 1–6. IEEE.

Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer (2002, jun). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research 16*, 321–357.

Chawla, N. V., N. Japkowicz, and A. Kotcz (2004, jun). Editorial: special issue on learning from imbalanced data sets. *ACM SIGKDD Explorations Newsletter 6*(1), 1.

Cover, T. and P. Hart (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory 13*(1), 21–27.

Cui, Y., M. Jia, T. Y. Lin, Y. Song, and S. Belongie (2019). Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Volume 2019-June, pp. 9260–9269.

Demšar, J. (2006a). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research 7*(Jan), 1–30.

Demšar, J. (2006b). Statistical Comparisons of Classifiers over Multiple Data Sets. *Journal of Machine Learning Research 7*, 1–30.

Dong, Q., S. Gong, and X. Zhu (2017). Class Rectification Hard Mining for Imbalanced Deep Learning. In *Proceedings of the IEEE International Conference on Computer Vision*, Volume 2017-Octob, pp. 1869–1878.

Douzas, G. and F. Bacao (2017, oct). Self-Organizing Map Oversampling (SOMO) for imbalanced data set learning. *Expert Systems with Applications 82*, 40–52.

Douzas, G. and F. Bacao (2019, oct). Geometric SMOTE a geometrically enhanced drop-in replacement for SMOTE. *Information Sciences 501*, 118–135.

Douzas, G., F. Bacao, J. Fonseca, and M. Khudinyan (2019, dec). Imbalanced learning in land cover classification: Improving minority classes' prediction accuracy using the geometric SMOTE algorithm. *Remote Sensing 11*(24), 3040.

Douzas, G., F. Bacao, and F. Last (2018, oct). Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE. *Information Sciences 465*, 1–20.

Drusch, M., U. Del Bello, S. Carlier, O. Colin, V. Fernandez, F. Gascon, B. Hoersch, C. Isola, P. Laberinti, P. Martimort, A. Meygret, F. Spoto, O. Sy, F. Marchese, and P. Bargellini (2012, may). Sentinel-2: ESA's Optical High-Resolution Mission for GMES Operational Services. *Remote Sensing of Environment 120*, 25–36.

Feng, W., W. Huang, and W. Bao (2019). Imbalanced Hyperspectral Image Classification With an Adaptive Ensemble Method Based on SMOTE and Rotation Forest With Differentiated Sampling Rates. *IEEE Geoscience and Remote Sensing Letters*, 1–5.

Feng, W., W. Huang, H. Ye, and L. Zhao (2018, oct). Synthetic minority oversampling technique based rotation forest for the classification of unbalanced hyperspectral data. In *International Geoscience and Remote Sensing Symposium (IGARSS)*, Volume 2018-July, pp. 2651–2654. Institute of Electrical and Electronics Engineers Inc.

Fernández, A., V. López, M. Galar, M. J. del Jesus, and F. Herrera (2013, apr). Analysing the classification of imbalanced data-sets with multiple classes: Binarization techniques and ad-hoc approaches. *Knowledge-Based Systems 42*, 97–110.

Ferreira, M. P., F. H. Wagner, L. E. Aragão, Y. E. Shimabukuro, and C. R. de Souza Filho (2019, mar). Tree species classification in tropical forests using visible to shortwave infrared WorldView-3 images and texture analysis. *ISPRS Journal of Photogrammetry and Remote Sensing 149*, 119–131.

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the american statistical association 32*(200), 675–701.

Fritz, S., L. See, C. Perger, I. McCallum, C. Schill, D. Schepaschenko, M. Duerauer, M. Karner, C. Dresel, J. C. Laso-Bayas, M. Lesiv, I. Moorthy, C. F. Salk, O. Danylo, T. Sturn, F. Albrecht, L. You, F. Kraxner, and M. Obersteiner (2017, jun). A global dataset of crowdsourced land cover and land use reference data. *Scientific Data 4*(1), 1–8.

García, S., S. Ramírez-Gallego, J. Luengo, J. M. Benítez, and F. Herrera (2016, dec). Big data preprocessing: methods and prospects. *Big Data Analytics 1*(1), 9.

Gavade, A. B. and V. S. Rajpurohit (2019, feb). Systematic analysis of satellite image-based land cover classification techniques: literature review and challenges. *International Journal of Computers and Applications*, 1–10.

Haibo He, Yang Bai, E. A. Garcia, and Shutao Li (2008, jun). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, pp. 1322–1328. IEEE.

Haixiang, G., L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing (2017, May). Learning from class-imbalanced data. *Expert Syst. Appl. 73*(C), 220–239.

Han, H., W.-Y. Wang, and B.-H. Mao (2005). Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. In *International Conference on Intelligent Computing*, pp. 878–887. Springer, Berlin, Heidelberg.

He, H. and E. A. Garcia (2009). Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering 21*(9), 1263–1284.

Holte, R. C., L. Acker, B. W. Porter, et al. (1989). Concept learning and the problem of small disjuncts. In *IJCAI*, Volume 89, pp. 813–818. Citeseer.

Hounkpatin, K. O., K. Schmidt, F. Stumpf, G. Forkuor, T. Behrens, T. Scholten, W. Amelung, and G. Welp (2018, dec). Predicting reference soil groups using legacy data: A data pruning and Random Forest approach for tropical environment (Dano catchment, Burkina Faso). *Scientific Reports 8*(1), 1–16.

Huang, C., Y. Li, C. C. Loy, and X. Tang (2016). Learning deep representation for imbalanced classification. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Volume 2016-Decem, pp. 5375–5384.

Japkowicz, N. (2001). Concept-learning in the presence of between-class and within-class imbalances. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Volume 2056, pp. 67–77. Springer Verlag.

Jeni, L. A., J. F. Cohn, and F. De La Torre (2013). Facing imbalanced data - Recommendations for the use of performance metrics. In *Proceedings - 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, ACII 2013*, pp. 245–251.

Jo, T. and N. Japkowicz (2004, jun). Class imbalances versus small disjuncts. *ACM SIGKDD Explorations Newsletter 6*(1), 40–49.

Jozdani, S. E., B. A. Johnson, and D. Chen (2019, jul). Comparing Deep Neural Networks, Ensemble Classifiers, and Support Vector Machine Algorithms for Object-Based Urban Land Use/Land Cover Classification. *Remote Sensing 11*(14), 1713.

Kaur, H., H. S. Pannu, and A. K. Malhi (2019, August). A systematic review on imbalanced data challenges in machine learning: Applications and solutions. *ACM Comput. Surv. 52*(4).

Khatami, R., G. Mountrakis, and S. V. Stehman (2016, may). 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.

Krawczyk, B. (2016, nov). Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence 5*(4), 221–232.

Lee, T., K. B. Lee, and C. O. Kim (2016, nov). Performance of Machine Learning Algorithms for Class-Imbalanced Process Fault Detection Problems. *IEEE Transactions on Semiconductor Manufacturing 29*(4), 436–445.

Lemaître, G., F. Nogueira, and C. K. Aridas (2017). Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research 18*(17), 1–5.

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

Luengo, J., D. García-Gil, S. Ramírez-Gallego, S. García, and F. Herrera (2020). Imbalanced Data Preprocessing for Big Data. In *Big Data Preprocessing*, pp. 147–160. Cham: Springer International Publishing.

Ma, L. and S. Fan (2017, mar). CURE-SMOTE algorithm and hybrid algorithm for feature selection and parameter optimization based on random forests. *BMC Bioinformatics 18*(1), 169.

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

Mellor, A., S. Boukir, A. Haywood, and S. Jones (2015, jul). Exploring issues of training data imbalance and mislabelling on random forest performance for large area land cover classification using the ensemble margin. *ISPRS Journal of Photogrammetry and Remote Sensing 105*, 155–168.

Nelder, J. A. and R. W. Wedderburn (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General) 135*(3), 370–384.

Olofsson, P., G. M. Foody, S. V. Stehman, and C. E. Woodcock (2013, feb). Making better use of accuracy data in land change studies: Estimating accuracy and area and quantifying uncertainty using stratified estimation. *Remote Sensing of Environment 129*, 122–131.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research 12*(Oct), 2825–2830.

Pelletier, C., S. Valero, J. Inglada, N. Champion, C. Marais Sicre, and G. Dedieu (2017, feb). Effect of Training Class Label Noise on Classification Performances for Land Cover Mapping with Satellite Image Time Series. *Remote Sensing 9*(2), 173.

Pontius, R. G. and M. Millones (2011). Death to Kappa: Birth of quantity disagreement and allocation disagreement for accuracy assessment.

Santos, M. S., P. H. Abreu, P. J. García-Laencina, A. Simão, and A. Carvalho (2015, dec). A new cluster-based oversampling method for improving survival prediction of hepatocellular carcinoma patients. *Journal of Biomedical Informatics 58*, 49–59.

Shao, Y. H., W. J. Chen, J. J. Zhang, Z. Wang, and N. Y. Deng (2014, sep). An efficient weighted Lagrangian twin support vector machine for imbalanced data classification. *Pattern Recognition 47*(9), 3158–3167.

Sharififar, A., F. Sarmadian, and B. Minasny (2019, apr). Mapping imbalanced soil classes using Markov chain random fields models treated with data resampling technique. *Computers and Electronics in Agriculture 159*, 110–118.

Stromann, O., A. Nascetti, O. Yousif, and Y. Ban (2020). Dimensionality Reduction and Feature Selection for Object-Based Land Cover Classification based on Sentinel-1 and Sentinel-2 Time Series Using Google Earth Engine. *Remote Sensing 12*(1), 76.

Wang, R., J. Zhang, J.-W. Chen, L. Jiao, and M. Wang (2019, jun). Imbalanced Learning-based Automatic SAR Images Change Detection by Morphologically Supervised PCA-Net. *IEEE Geoscience and Remote Sensing Letters*.

Wilcoxon, F. (1992). Individual comparisons by ranking methods. In *Breakthroughs in statistics*, pp. 196–202. Springer.

Wulder, M. A., N. C. Coops, D. P. Roy, J. C. White, and T. Hermosilla (2018). Land cover 2.0. *International Journal of Remote Sensing 39*(12), 4254–4284.

Zhu, M., B. Wu, Y. N. He, and Y. Q. He (2020). Land Cover Classification Using High Resolution Satellite Image Based On Deep Learning. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XLII-3/W10*, 685–690.