

# Quality control and class noise reduction of satellite image time series

Lorena A. Santos<sup>\*</sup>, Karine R. Ferreira, Gilberto Camara, Michelle C.A. Picoli, Rolf E. Simoes

Earth Observation General-Coordination, National Institute for Space Research, INPE, Brazil

## ARTICLE INFO

### Keywords:

Self-organizing map  
Class noise reduction  
Bayesian inference  
Satellite image time series  
Land use and cover classification

## ABSTRACT

The extensive amount of Earth observation satellite images available brings opportunities and challenges for land mapping in global and regional scales. These large datasets have motivated the use of satellite image time series analysis coupled with machine learning techniques to produce land use and cover class maps. To be successful, these methods need good quality training samples, which are the most important factor for determining the accuracy of the results. For this reason, training samples need methods for quality control of class noise. In this paper, we propose a method to assess and improve the quality of satellite image time series training data. The method uses self-organizing maps (SOM) to produce clusters of time series and Bayesian inference to assess intra-cluster and inter-cluster similarity. Consistent samples of a class will be part of a neighborhood of clusters in the SOM map. Noisy samples will appear as outliers in the SOM. Using Bayesian inference in the SOM neighborhoods, we can infer which samples are noisy. To illustrate the methods, we present a case study in a large training set of land use and cover classes in the Cerrado biome, Brazil. The results prove that the method is efficient to reduce class noise and to assess the spatio-temporal variation of satellite image time series training samples.

## 1. Introduction

Humans are changing the Earth's environment at a fast pace. In the last decades, socio-economic and population growth in developing nations have increased the removal of natural lands for food and energy production. Such fast changes in land areas resulted in greater greenhouse gas emissions and biodiversity loss (Foley et al., 2005). In this context, mapping and monitoring of land use and cover change (LUCC) are essential for planning and managing natural resources (Gomez et al., 2016). Technologies and methods of remote sensing image processing play a crucial role in the identification, mapping, assessment, and monitoring LUCC.

In this paper, we deal with the problem of noise detection and quality improvement in satellite image time series (SITS). The new generation of open access remote sensing satellites has made petabytes of Earth observation (EO) data available online. From repeated orbits of remote sensing satellites, we obtain a sequence of images from the same area. After suitable calibrations, these images can be joined into a time series to measure change. Time series derived from EO satellite images allow us to detect complex underlying processes that would be difficult to identify using bi-temporal or other traditional change detection approaches (Pasquarella et al., 2016). Satellite image time series are being increasingly used land use and cover classification and change detection

with good results (Petitjean et al., 2012; Maus et al., 2016; Inglada et al., 2017; Picoli et al., 2018; Woodcock et al., 2020).

Since machine learning methods have emerged as the best way to classify remote sensing images for providing land information (Zhang et al., 2003; Mountrakis et al., 2011; Belgiu and Dragut, 2016), there is a natural interest in using machine learning methods for SITS analysis. Recent results show that it is feasible to apply machine learning methods to SITS analysis in large areas of 100 million ha or more (Picoli et al., 2018; Simoes et al., 2020; Parente et al., 2019; Griffiths et al., 2019). Experience with machine learning methods has established that the limiting factor in obtaining good results is the number and quality of training samples. Large and accurate datasets are better, no matter the algorithm used (Maxwell et al., 2018); increasing the training sample size results in better classification accuracy (Thanh Noi and Kappas, 2018). Therefore, using machine learning for SITS analysis requires large and good quality training sets.

There are two main sources of noise and errors in satellite image time series (Pelletier et al., 2017). One effect is *feature noise*, caused by clouds and inconsistencies in data calibration. The second effect is *class noise*, when the label assigned to the sample is wrongly attributed. Class noise effects are common on large datasets. In particular, interpreters tend to group samples with different properties in the same category. For this reason, one needs good methods for quality control of large training

<sup>\*</sup> Corresponding author.

E-mail address: [lorena.santos@inpe.br](mailto:lorena.santos@inpe.br) (L.A. Santos).

<https://doi.org/10.1016/j.isprsjprs.2021.04.014>

Received 27 July 2020; Received in revised form 1 April 2021; Accepted 19 April 2021

0924-2716/© 2021 The Authors. Published by Elsevier B.V. on behalf of International Society for Photogrammetry and Remote Sensing, Inc. (ISPRS). This is an

open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

datasets associated with satellite image time series. Our work thus addresses the question: *How to reduce class noise in large training sets of satellite image time series?*

The availability of big open EO data is recent and the methods for SITS analysis are still maturing. For this reason, there is limited research dealing with the problem of how to improve the quality of large sets of SITS training data (Pelletier et al., 2017). In this paper, we propose a new method for class noise reduction in the SITS reference database.

The proposed method creates a self-organizing map (SOM) to reduce image time series dimensionality. SOM presents a fundamental property of neighborhood topological preservation. Time series samples with similar patterns tend to be close in the SOM output space. Hence, the neighborhood can offer additional information for outlier identification and intra-class and inter-class variability. Based on the SOM neighborhood preservation feature, we use Bayesian inference to reinforce the intra-class similarities evaluation and enhance the samples assessment quality. We present how the proposed method improves the results of land use and cover classification using a large SITS dataset.

## 2. Related work

*Class label noise* refers to mislabeling or sample instances whose labels are different from the ground truth labels (Pelletier et al., 2017). The problem of class label noise and their effects in supervised learning are widely discussed in the literature of Neurocomputing, Artificial Intelligence and Machine Learning (Zhu and Wu, 2004; Frénay and Verleysen, 2013; Garcia et al., 2015).

In Machine Learning, techniques to identify and remove label noise include filter approaches based on geometric, statistical and structural measures extracted from datasets (Zhu and Wu, 2004; Garcia et al., 2012), based on Bayesian classifier (Sun et al., 2007) or based on clustering methods (Rebbapragada and Brodley, 2007). Filters are applied before learning process. Differently from filters, noise tolerant variants of classifiers are proposed in order to be more tolerant and robust to noise, dealing with label noise during learning or considering label noise in an embedded way (Kharon and Wachman, 2007; Natarajan et al., 2013; Frénay and Verleysen, 2013; Patrini et al., 2017). For example, one variant of Support Vector Machine (SVM) method has a parameter to be tuned during its training, called regularization or lambda, that is responsible for identifying misclassified samples and replacing them to near ones based on decision boundaries. For hyperspectral classification, the use of joint sparse representation approach (Peng et al., 2019) and Log-Euclidean kernel-based joint sparse representation (Yang et al., 2019) have been applied to replace the least square loss to reduce the effects of outliers and heterogeneous pixels. Although these methods are robust, they are not still free to be affected by noise (Frénay and Verleysen, 2013).

In Remote Sensing, many works have highlighted the importance of good quality samples to train machine learning methods in order to produce land use and cover maps with great accuracy from SITS analysis (Olofsson et al., 2014; Gomez et al., 2016; Belgiu and Dragut, 2016; Elmes et al., 2020). However, there are few papers that focus on the problem of class label noise in large sets of SITS training (Pelletier et al., 2017). Most of the literature deals with the removal of *feature noise* focusing on cloud removal and smoothing (Hird and McDermid, 2009; Atzberger and Eilers, 2011; Atkinson et al., 2012). For *class label noise*, most papers evaluate the impact of mislabeled training data for land cover mapping using classical classifiers as SVM and Random Forest and show that their performance drop down for higher noise levels (Jiang et al., 2008; Mellor et al., 2015; Pelletier et al., 2017). There is a lack of solutions to identify and remove class label noise in large sets of SITS training samples.

This paper addresses the class label noise problem in large sets of SITS training samples and presents a solution for that. We propose a novel method for class label noise reduction in large SITS data and present how it improves the quality of land use and cover samples. In

land use and cover applications, label noise is common and occurs during field works mainly due to the lack of consensus in land cover definitions and the subjectivity of human judgment (Pelletier et al., 2017). The proposed method is useful for land use and cover applications, helping users to identify and remove label noise in large SITS training datasets.

## 3. Material and methods

### 3.1. Study area

Our case study uses a dataset of classes in the Cerrado biome in Brazil, the second largest biome in South America with an area of more than 2 million km<sup>2</sup> (~ 22% of Brazil) (see Fig. 1) (Ministry of the Environment, 2019). The Cerrado is a global biodiversity hotspot due of the abundance of endemic species; it has undergone a significant habitat loss in recent decades (Strassburg et al., 2017). The advance of agricultural and livestock activities has caused intense land change (Soteroni et al., 2019). Only 8.21% of the Cerrado is legally protected by conservation units (Ministry of the Environment, 2019), and it is estimated that 88 Mha (46%) of its natural vegetation cover has been lost (Strassburg et al., 2017).

### 3.2. Training Samples

The training samples were collected by ground surveys and high-resolution image interpretation by experts from the Brazilian National Institute for Space Research (INPE) team and partners. This set ranges from 2000 to 2017 and includes 50,160 land use and cover samples divided into 12 classes: (1) Dunes, (2) Fallow-Cotton, (3) Millet-Cotton, (4) Soy-Corn, (5) Soy-Cotton, (6) Soy-Fallow, (7) Pasture, (8) Rocky Savanna (in Portuguese *cerrado rupestre*), (9) Savanna, (10) Dense Woodland (in Portuguese *cerradão*), (11) Savanna Parkland (in Portuguese *savana parque*) and (12) Planted Forest. The class labels of natural classes of the Cerrado follow the work of (Ribeiro and Walter, 2008) who provide a taxonomy of classes for the biome. The samples number for each class is presented in Table 1.

As shown in Fig. 2, each sample has a spatial location (latitude and longitude), an interval (start and end dates) that corresponds to an agricultural year, a LUCC class, and a satellite image time series for each band or attribute. The time series were extracted from the MODIS sensor (MOD13Q1 product, collection 6) of the NASA's Terra satellite, available on a 16-day time interval with a 250 meter spatial resolution. We used a multidimensional time series with four MODIS bands: Normalized Difference Vegetation Index (NDVI) and Enhanced Vegetation Index (EVI), and the original bands near-infrared (NIR) and mid-infrared (MIR). Fig. 2 illustrate the four satellite image time series, one for each attribute (NDVI, EVI, NIR, and MIR), associated with samples of Savanna and Soy-Cotton classes.

Multi-dimensional time series help in distinguishing the different land classes. In the Cerrado biome, the dry season occurs from May to September and the rainy season from October to April. The variability of EVI values during the rainy season helps to distinguish between natural vegetation cover types (Liesenberg et al., 2007). For crop classes, the NDVI and EVI values are high during the growth of vegetation and start to decrease during the harvest. Spectral bands NIR and MIR also contribute to class discrimination as they are related to the structure of the leaves and soil (Adam et al., 2010). Areas with forests and woodlands have high values of NIR because of their leaf structures; they also have low MIR due to absorption of water (Adam et al., 2010).

### 3.3. General Description

Many factors lead to *class noise* in SITS. One of the main problems is the inherent variability of class signatures in space and time. When training data is collected over a large geographic region, natural

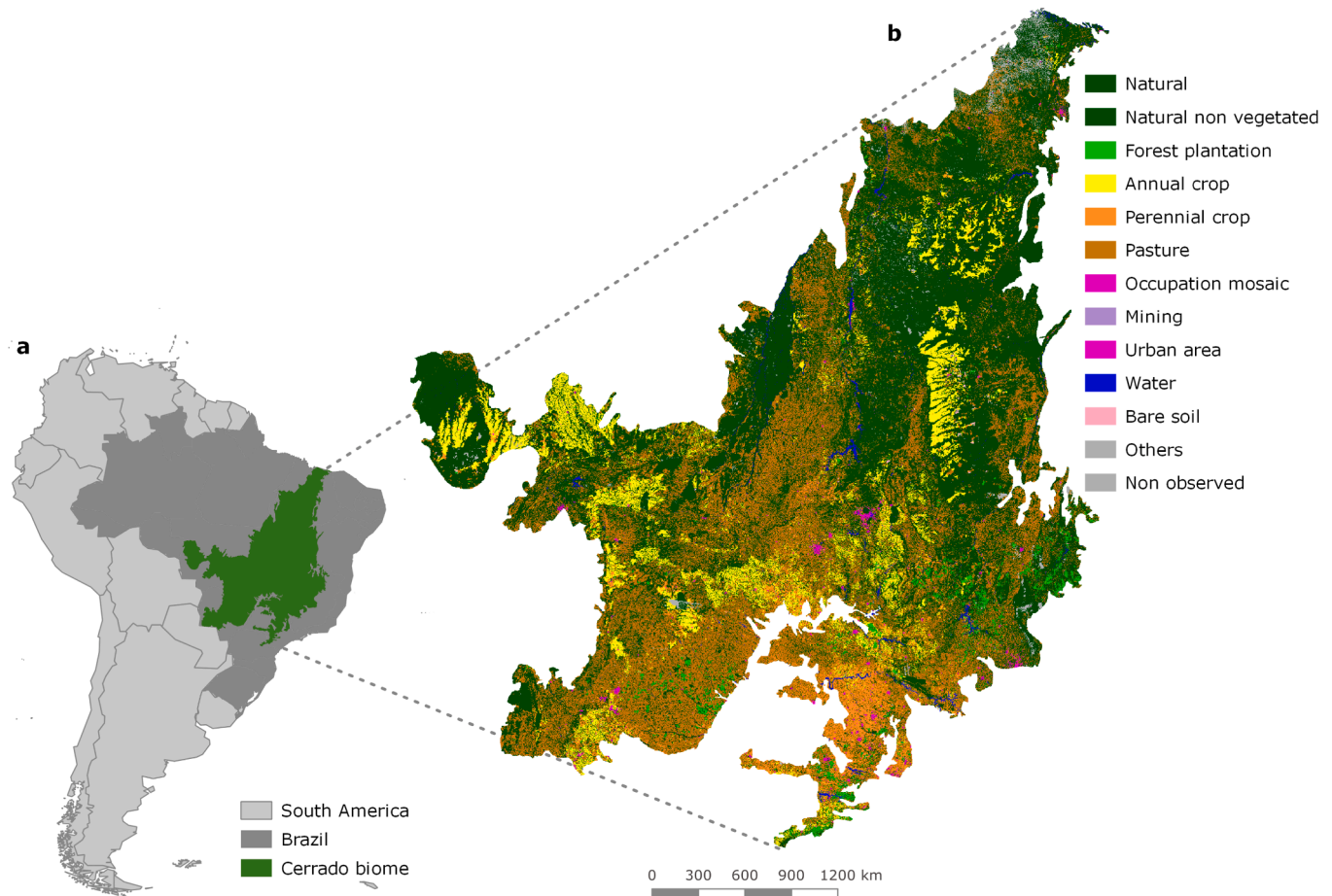


Fig. 1. a. Cerrado location relative to Brazil and South America. b. Land use and cover map of the Cerrado. Source: TerraClass (INPE, 2013).

Table 1  
Input dataset.

Class	Count	Frequency
Dunes	550	1.1%
Fallow-Cotton	630	1.26%
Millet-Cotton	316	0.63%
Soy-Corn	4971	9.9%
Soy-Cotton	4124	8.22%
Soy-Fallow	2098	4.1%
Pasture	7206	14.4%
Rocky Savanna	8005	16%
Savanna	9172	18.3%
Dense Woodland	9966	19.9%
Savanna Parkland	2699	5.3%
Planted Forest	423	0.84%

variability of vegetation phenology can result in different patterns being assigned to the same label. Phenological patterns can vary spatially across a region and are strongly correlated with climate variations (Suepa et al., 2016). A related issue is the limitation of crisp boundaries to describe the natural world. Class definition use idealized descriptions (e.g., “a savanna woodland has tree cover of 50% to 90% ranging from 8 to 15 meters in height”). However, in practice the boundaries between classes are fuzzy and sometimes overlap, making it hard to distinguish between them. Class noise can also result from labeling errors. Even trained analysts can make errors in class attributions. Despite the fact that machine learning techniques are robust to errors and inconsistencies in the training data (Gomez et al., 2016; Pelletier et al., 2017), quality control of training data can make a significant difference in the resulting maps.

The main steps of our proposed method for quality assessment of satellite image time series is shown in Fig. 3. The method uses self-organizing maps (SOM) (Kohonen, 1990) to perform dimensionality reduction while preserving the topology of original datasets. Since SOM preserves the topological structure of neighborhoods in multiple dimensions, the resulting 2D map can be used as a set of clusters. Training samples that belong to the same class will usually be neighbors in 2D space. The neighbors of each neuron of a SOM map provide additional information on intra-class and inter-class variability. We apply Bayesian inference to the neighborhoods of the SOM map to improve the evaluation of the quality of each time series sample.

### 3.4. Using SOM for dimensionality reduction

SOM is an unsupervised neural network that maps a high dimensional input dataset to low-dimensional one, usually a 2D grid. As Fig. 4 shows, the grid is composed by units called *neurons*. Each neuron has a weight vector, with the same dimension as the training samples. At the start, neurons are assigned a small random value and then trained by competitive learning. The algorithm computes the distances of each member of the training set to all neurons and finds the neuron closest to the input, called the best matching unit (BMU). The weights of the BMU and its neighbors are updated so as to preserve their similarity (Kohonen, 2013). This mapping and adjustment procedures is done in several iterations. At each step, the extent of the change in the neurons diminishes, until a convergence threshold is reached. The result is a 2D mapping of the training set, where similar elements of the input are mapped to the same neuron or to nearby ones. The resulting SOM grid combines dimensionality reduction with topological preservation.

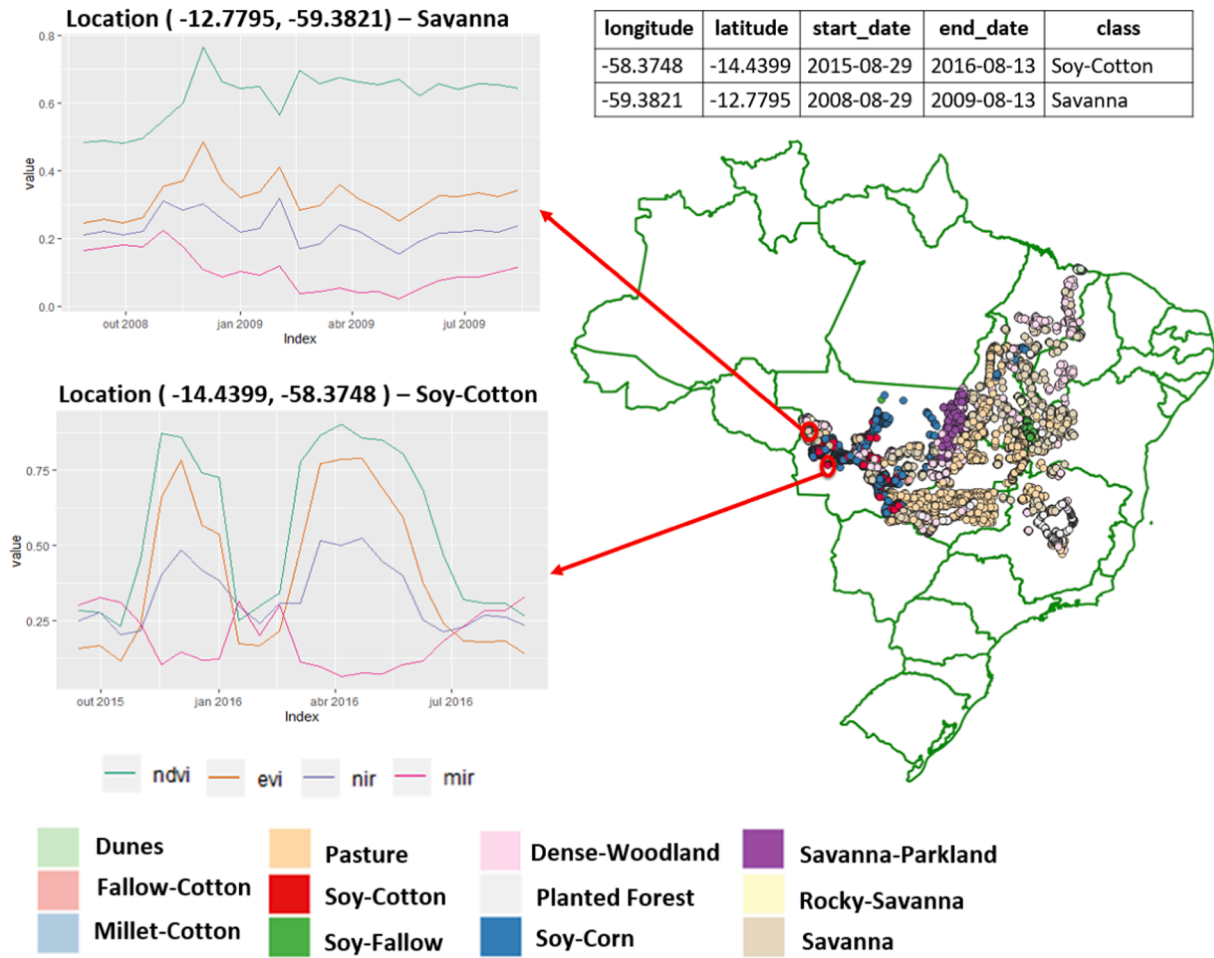


Fig. 2. Reference dataset.

To project a multidimensional set of time series onto a SOM map, each neuron  $j$  is associated a random vector of weights  $w_j = [w_{j1}, \dots, w_{jn}]$ , with the same length of each time series sample  $x(t)_i = [x_{t1}, \dots, x_{tn}]$ . Each time a sample is allocated to its best matching unit (BMU)  $b$ , which is the neuron with the smaller distance between the time series and its vector of weights. To compute the distance  $D_j$  between a time series  $x(t)_i$  and a neuron  $j$  we compared three metrics (Euclidean, Manhattan and Dynamic Time Warping) in a previous paper (Ferreira et al., 2019). We found out that Euclidean metric provides reliable and robust results. Therefore, our method uses Euclidean distances to find the BMU,  $d_b$ , as shown in Eq. (1) and Eq. (2).

$$D_j = \sqrt{\sum_{i=1}^n (x(t)_i - w_j)^2} \quad (1)$$

$$d_b = \min\{D_1, \dots, D_j\}. \quad (2)$$

The next step is to update the weights of the BMU and its neighbors. The weights are adjusted to approximate the input vector, as shown in Eq. (3).

$$w_j = w_j + \alpha \times h_{bj} [x(t)_i - w_j], \quad (3)$$

The parameter  $\alpha$  is the learning rate and  $h_{bj}$  is the neighborhood function. They are updated at each iteration of SOM. The learning rate controls how the weight vector changes. It must be set as  $0 < \alpha < 1$ . The neighborhood function  $h_{bj}$  determines which neurons must be updated and the intensity of the readjustment of each one (Natita et al., 2016).

SOM allows the use of multiple attributes as input data. Thus, our approach uses a unique vector composed of all attributes, including spectral bands and vegetation indexes, to represent the input samples. Consequently, the weight vectors are initialized with the same dimension of the input vectors. Although all attributes are put together in a unique vector, the distance between the input and weight vectors is computed separately for each attribute. Then, the distances of all attributes are summed to obtain a unique distance value that is used to find the BMU.

At the end of SOM training phase, each time series is associated with a neuron  $j$  in a 2D grid. Since the class of each time series is known, we assign a class to a neuron using majority voting. As an example, Fig. 5 presents a grid with a set of neurons, the samples associated to neuron 1 are presented. We then compute the class frequencies of the samples linked to the neuron. In this example, since four samples belong to Class 1, and one to Class 2, the neuron is assigned to Class 1 with 80% probability and to Class 2 with 20% probability. These probabilities will be used in the next phase of the method.

### 3.5. Using Bayesian Inference to assess the influences of the SOM neighborhood

After all time series are assigned to a neuron, the SOM map is used to assess the quality of each element of the training set. Each neuron will be associated to a discrete probability distribution, as shown in Fig. 2. More homogeneous neurons (those with a single class of high probability) are composed of good quality samples. Heterogeneous neurons (those with two or more classes with significant probability) are likely to contain



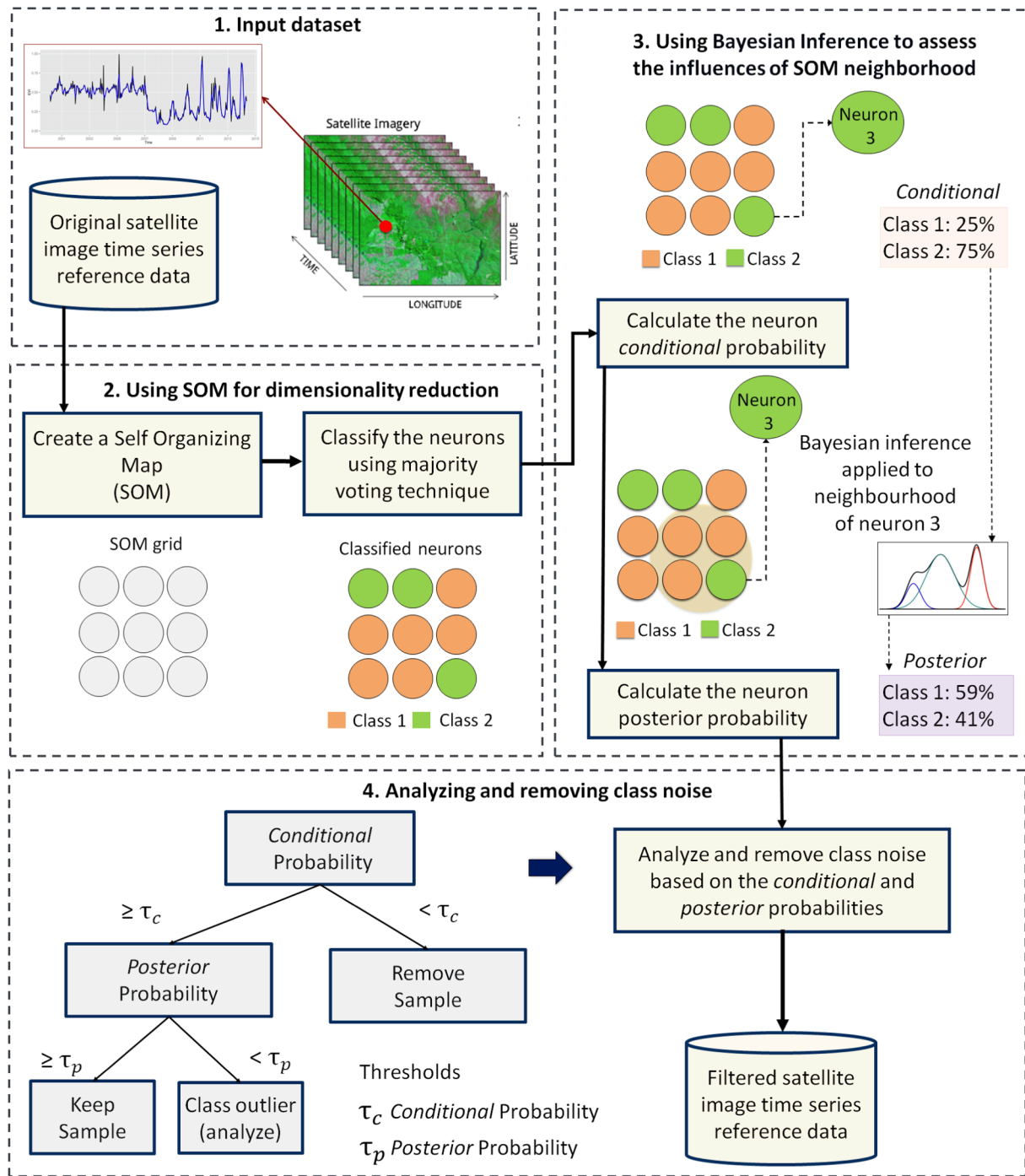


Fig. 3. A method for class noise reduction in satellite image time series reference data.

noisy samples. Furthermore, we consider that the neuron class probability is not the best measure of class noise. It represents the prior probability  $P(\text{ClassNeuron}/\text{ClassSample})$ . In fact, what we need is the inverse probability  $P(\text{ClassSample}/\text{ClassNeuron})$ . To obtain this inverse (also called posterior) probability, we use Bayesian inference.

Bayesian inference estimates the conditional probability  $f(\theta_{j,k}|y_{j,k})$  where  $\theta_{j,k}$  is the random variable associated to the occurrence of a class  $k$  in a neuron  $j$  and  $y_{j,k}$  is the value of probability of neuron  $j$  being of class  $k$ . Bayes' Rule is given by:

$$f(\theta_{j,k}|y_{j,k}) \propto f(y_{j,k}|\theta_{j,k})f(\theta_{j,k}), \quad (4)$$

where  $f(\theta_{j,k})$  is the prior probability distribution of  $\theta_{j,k}$ , that is, what we

know about the samples of class  $k$  that are part of neuron  $j$  before the SOM mapping. The conditional probability  $f(y_{j,k}|\theta_{j,k})$  represents the probability of a neuron  $j$  belonging to a class  $k$ , given the samples of class  $k$  that are associated to it. It is called the likelihood in Bayesian inference.

Since there is not enough information to compute the probabilities associated to the prior  $f(\theta_{j,k})$  and the likelihood  $f(y_{j,k}|\theta_{j,k})$ , we need to make some assumptions. First, we consider these probabilities to be modeled by Gaussian distributions. Second, we consider that the prior  $f(\theta_{j,k})$  can be estimated using the neighborhood of neuron  $j$ . This assumption is based on the SOM properties of topological consistency. Given how SOM works, we expect similar samples to be close together in SOM 2D space. Such strategy of "borrowing strength from the

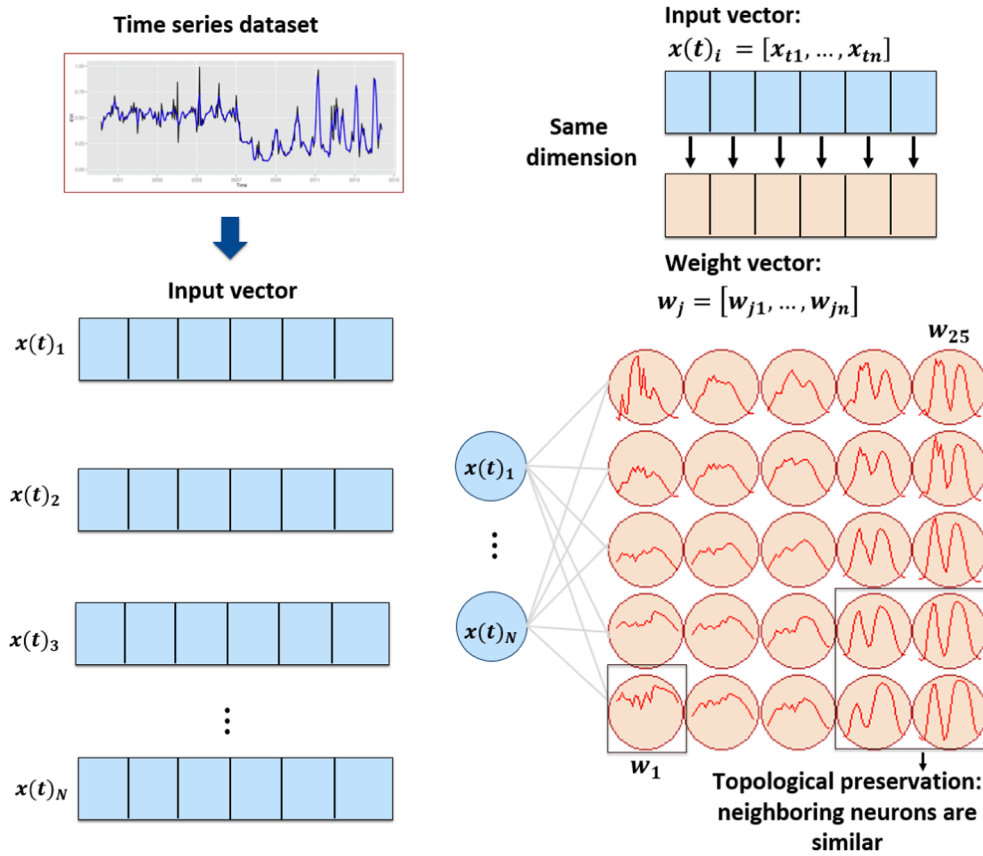


Fig. 4. Self-Organizing Maps structure.

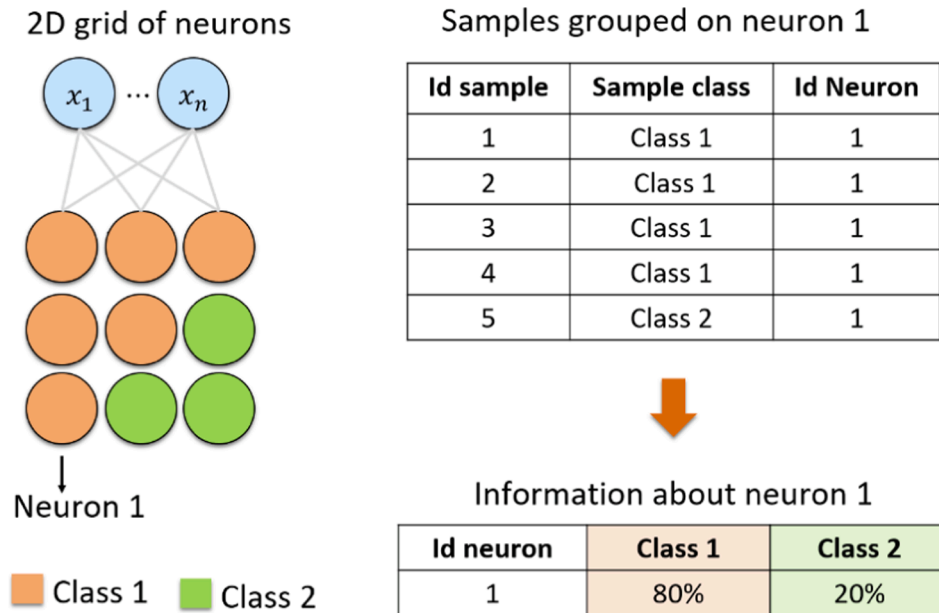


Fig. 5. Assignment of classes to neurons.

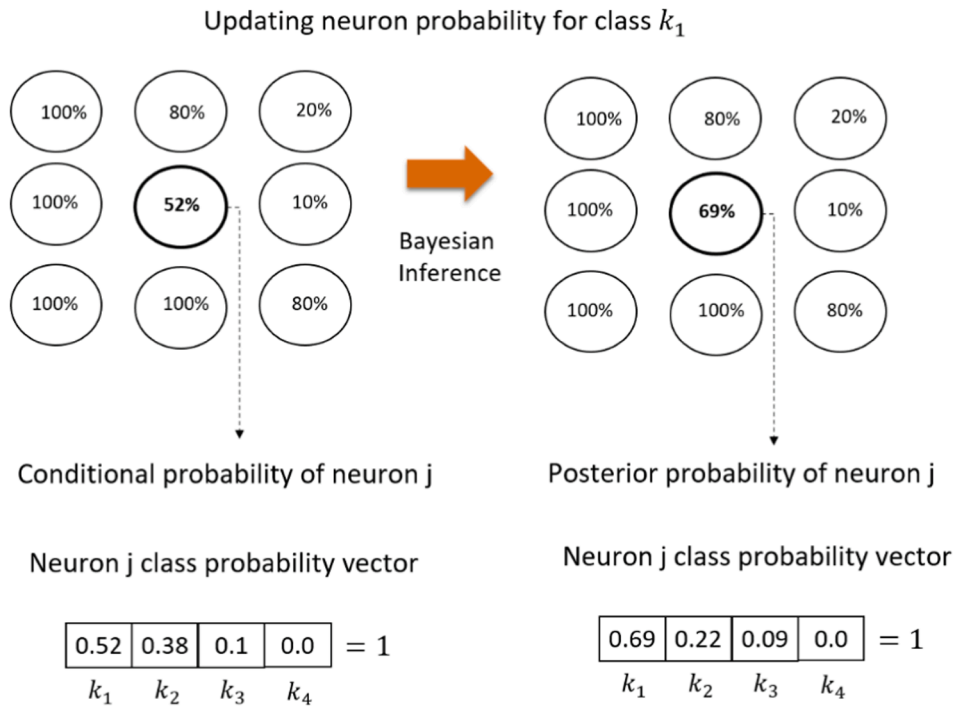
neighbors” is commonly used in Bayesian inference (Assuncao et al., 2005).

The approach is illustrated in Fig. 6, where the neuron  $j$  has a prior probability of 52% of belonging to class  $k_1$ . Since most of its neighbors have high probability of belonging to the class  $k_1$ , the posterior probability of the neuron  $j$  belonging to a class  $k_1$  increases due the neighborhood effects.

To estimate the prior distribution of  $\theta_{j,k}$ , we consider it to be expressed as a Gaussian distribution:

$$\theta_{j,k} \sim N(m_{j,k}, s_{j,k}^2). \quad (5)$$

where  $m_{j,k}$  is the mean of the probability of values for class  $k$  and  $s_{j,k}^2$  is the variance for class  $k$ . We estimate the means and variances

Fig. 6. Update neuron  $j$  for class  $k_1$ .

considering the neighborhood of neuron  $j$ . Let  $V_j$  be the neighborhood of neuron  $j$ , and  $\#(V_j)$  be the number of elements in  $V_j$ . We then have:

$$m_{j,k} = \frac{\sum_{(i) \in V_j} y_{i,k}}{\#(V_{j,t})}, \quad (6)$$

$$s_{j,k}^2 = \frac{\sum_{(i) \in V_j} [y_{i,k} - m_{j,k}]^2}{\#(V_j) - 1}. \quad (7)$$

For the likelihood  $f(y_{j,k}|\theta_{j,k})$ , we also consider a normal distribution given by:

$$y_{j,k}|\theta_{j,k} \sim N(\theta_{j,k}, \sigma_j^2) \quad (8)$$

where  $\sigma_j^2$  is an unknown hyper-parameter that controls the smoothness level. Given these estimates, according to Bayesian statistics the expected conditioned mean for  $\theta_{j,k}$  is given by:

$$E[\theta_{j,k}|y_{j,k}] = \frac{m_{j,k} \times \sigma_j^2 + y_{j,k} \times s_{j,k}^2}{\sigma_j^2 + s_{j,k}^2} \quad (9)$$

Rewriting the equation we have:

$$E[\theta_{j,k}|y_{j,k}] = \left[ \frac{s_{j,k}^2}{\sigma_j^2 + s_{j,k}^2} \right] \times y_{j,k} + \left[ \frac{\sigma_j^2}{\sigma_j^2 + s_{j,k}^2} \right] \times m_{j,k} \quad (10)$$

When the neighborhood variance  $s_{j,k}^2$  for class  $k$  is high, Eq. (10) gives more weight to the prior probability of  $y_{j,k}$ . Otherwise, if the neighborhood variance  $s_{j,k}^2$  is small, the posterior estimate is controlled by the neighborhood mean  $m_{j,k}$ . This reflects the intuition that samples in areas of low variance are similar, while they differ in regions of high variance.

The value of the hyper-parameter  $\sigma_j^2$  should be set so as to balance the neighborhood effects. A high value of  $y_{j,k}$  signals a strong confidence that all samples in neuron  $j$  belong to class  $k$ . In general, as the value of  $y_{j,k}$  increases, the smoothing  $\sigma_j^2$  should decrease. To maintain the  $\sigma_j^2$  adjusted according to the class variance of neuron  $j$ , we define  $\sigma_j^2$  as:

$$\sigma_j^2 = |0.999999 - \max(y_{j,k})| \quad (11)$$

Fig. 7 shows how the Bayesian inference is applied in our context. Given the prior probabilities of neuron 3 and its neighbors, initially neuron 3 belongs to a Class 2, however, all neurons of its neighborhood belongs to Class 1. When the Bayesian inference is applied, the probability of this neuron belongs to a Class 2 decreases due to the strength of the neighbors. Therefore, the sample 1, labeled as Class 2, inherits the probability of the neuron belonging to Class 2.

### 3.6. Analyzing and removing class noise

Our method uses the probabilities calculated in the previous step to evaluate the quality of the samples. Using these probabilities, we identify outlier neurons as those whose classes are distinct from their neighborhood. Identifying outlier neurons is a key part of our method. Our experiments show there are two possible causes for an outlier neuron: (a) its samples may be mislabeled or of bad quality; (b) due to the different patterns of land use and cover classes in space or time. Case (a) arises from class noise, and the associated samples should be discarded. By contrast, case (b) results from variability; thus, the associated samples should not be removed automatically from the dataset and need to be flagged for later analysis. To distinguish between these situations, we proposed the following rule, which include thresholds  $\tau_c$  for the prior probability and  $\tau_p$  for the posterior probability:

1. If the prior probability is  $< \tau_c$  then, the sample is removed from dataset;
2. If the prior probability is  $\geq \tau_c$  and the posterior probability is  $\geq \tau_p$ , then, the sample is kept in the dataset;
3. If the prior probability is  $\geq \tau_c$  and the posterior probability is  $< \tau_p$ , the samples will be flagged for further inspection.

## 4. Results and Discussions

As a proof of concept, this section presents a study that evaluates the quality of a time series sample set associated with land use and cover

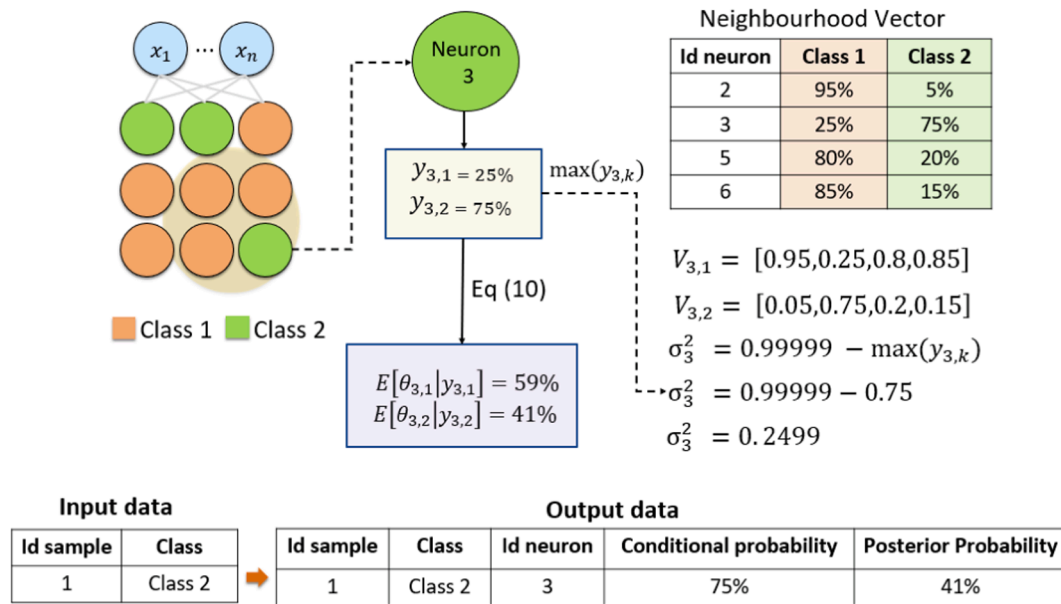


Fig. 7. Applying Bayesian Inference in neuron 3.

classes and shows how to identify class noise in this set, and thus improve the accuracy of the resulting classification.

#### 4.1. Detecting noisy and outlier samples

In what follows, we show how to apply our method to analyze and improve sample quality. Fig. 8 shows the SOM grid map generated for

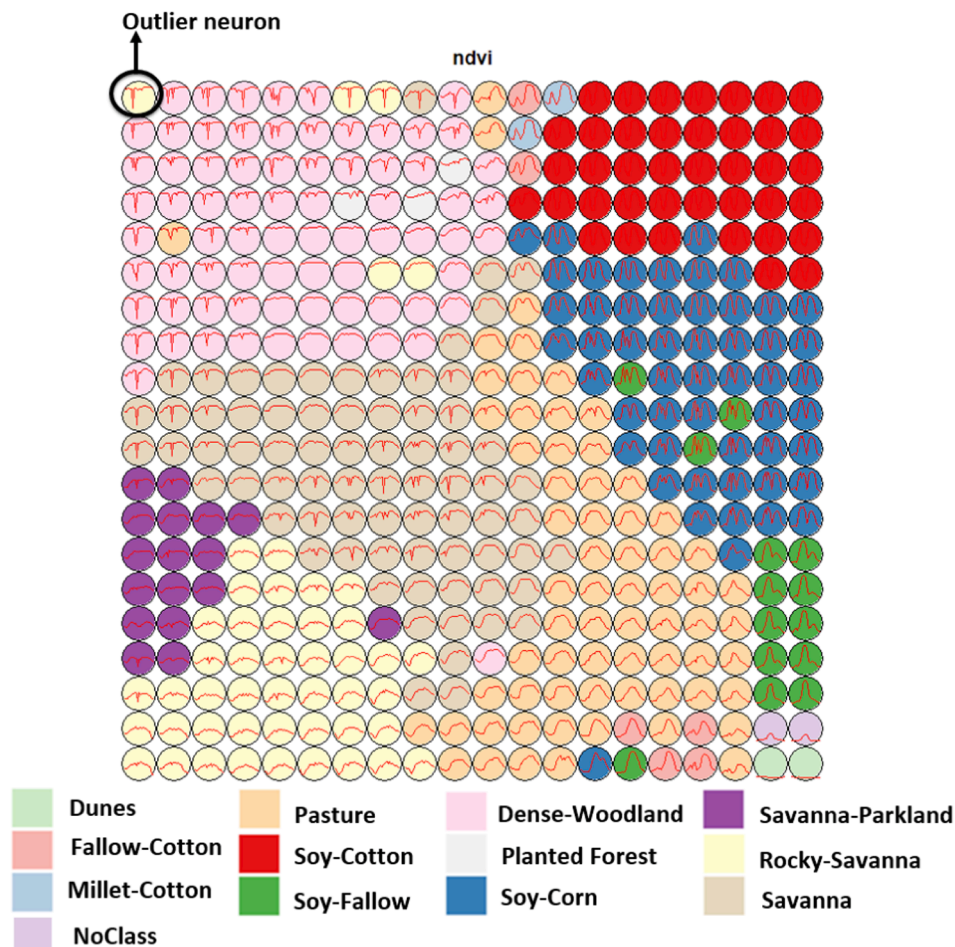


Fig. 8. SOM grid.



the training data. The parameters used in SOM are: grid size =  $20 \times 20$ , learning rate =  $(0.50, 0.01)$ , 100 iterations, and Euclidean distance for finding the BMU. Each sample is associated with its BMU neuron; after that, each neuron is labeled with its majority class.

To define the SOM grid size, (Vesanto and Alhoniemi, 2000) suggest about  $5 * \sqrt{N}$  neurons, where  $N$  is the number of observations or samples. However, based on empirical tests, we got a good result using around  $\frac{5 * \sqrt{N}}{2}$  neurons. Regarding learning rate, too small values can lead to a twisted maps and big values to a non ordered map (Tan and George, 2004). We suggest using a decreasing learning rate, starting with 0.5 to 0.1. The number of iterations is related to the convergence of SOM, that is, when additional iterations do not update the weight vectors. We tested our dataset and reached the convergence with 100 iterations. To select the distance measure, we evaluated three metrics and concluded that Euclidean and Manhattan are more accurate than Dynamic Time Warping for image time series clustering in land use and cover application (Ferreira et al., 2019).

Because of the variability among time series of the same land use and cover class, samples of the same class can be mapped into different neurons. However, due to the SOM properties, time series samples of the same class are expected to be similar and so neighbors in the output map. The map also contains potentially mislabeled and outlier samples. Mislabeled samples are those that have been mapped to neurons whose majority class is different from their own label. Outlier neurons are those whose majority class is different from that of their neighbors. Our hypothesis is that mislabeled samples and outlier neurons are indicators of class noise. Thus, by examining them and identifying incorrect samples, we can improve the quality of the training set.

Based on the rules presented in Section 2 to identify good samples and class noise, we set the threshold as 60% to both prior and posterior probabilities to decide if each sample is to be kept or removed from the training set. Table 2 shows the percentage of samples by class that have been kept, removed, or flagged for analysis.

As shown in Table 2, our method identifies noisy samples. Some should be removed, and others need to be analyzed to decide whether to keep or remove them. There are classes with large percentage of noisy samples, such as Millet-Cotton. Other classes have many samples flagged for further analysis, such as Planted Forest and Fallow-Cotton. Noisy samples can arise due to high intra-class variability or due to confusion between class signatures. Whatever the case, the SOM-based analysis helps to identify them. The SOM-based method also allows measuring confusion between classes, as shown in Fig. 9. As the figure shows, almost 20% of the Planted Forest samples are confused with those from the Dense Woodland class. Also, 19% of the Soy-Fallow samples have been mixed with samples from the Soy-Corn class. Such information helps experts to have a detailed view of class noise in their samples.

The confusion between classes in the Cerrado arises because its natural vegetation is a continuous mix of grasslands and trees. The boundaries between classes such as Savanna Parkland, Savanna, Rocky Savanna are fuzzy and there are many transitional regions. According to

(Ribeiro and Walter, 2008), areas of Savanna have trees whose height is between 3 and 8 meters and cover range from 20% to 50% of the area. Savanna Parkland areas have trees between 3 and 6 meters tall and the tree cover between 5% and 20% of the area. Rocky-Savannas occurs in regions with rock outcrops, where the tree cover ranges from 5% to 20% and the trees are between 2 and 4 meters tall. Dense Woodlands have a continuous canopy and the tree cover range from 50% to 90%, and trees between 7 and 15 meters of height (Ribeiro and Walter, 2008). In a complex biome such as the Cerrado, these labels are approximations of a continuous gradient of changes in the tree and grassland mix (Ribeiro and Tabarelli, 2002). Thus, some degree of confusion between the natural vegetation classes in the Cerrado biome is to be expected.

Outside transitions areas, one can distinguish these classes using vegetation indices, Figs. 10a and b. The values of NDVI, EVI and NIR for samples of Dense Woodlands are higher than for samples of Savanna. Samples of Savanna Parkland and Rocky Savanna classes have NDVI values lower than Savanna and Dense-Woodland. Although the NDVI values for Savanna Parkland and Rocky-Savanna can be similar, the EVI and NIR values for Savanna Parkland are more constant during the year than those of Rocky-Savanna (Figs. 10c and d). In our dataset, the samples of Savanna Parkland are located close to riverbanks. Therefore the vegetation does not have significant leaf loss during the dry season (Liesenberg et al., 2007). This explains the constant values of EVI and NIR during the year. The Rocky-Savanna and Savanna Parkland classes are more difficult to confuse with Dense-Woodland class because of the different NDVI values. This is confirmed in the SOM map (Fig. 8), where the Dense Woodland neurons are far from Rocky Savanna and Savanna Parkland ones. Therefore, in general, these classes show different time series signals.

#### 4.2. Identifying mislabeled samples

We now consider how our method helps to identify wrongly labeled samples. Fig. 11 shows the NDVI signature of two different clusters of samples of Rocky-Savanna class identified in the SOM map. In our assessment, the time series samples presented in Fig. 11a are consistent with the expected response of the Rocky Savanna class. The posterior probability of these samples belonging to the Rocky Savanna class is 100%. By contrast, the samples presented in Fig. 11b were removed from the dataset; their posterior probability of belonging to the Rocky Savanna class is 18.5%. These samples have been actually mapped to neuron whose label is Dense Woodland. These samples have likely been mislabeled.

We now consider the sources of confusion between natural vegetation and crops and between crop samples. In general, as seen in Fig. 9, natural classes and crop classes do not mix. There are exceptions, such as the confusion between Planted Forest and Dense Woodland samples. These classes have similar time series patterns due to the coarse spatial resolution of MODIS. As for confusion between crop samples, we identified many problems with Millet-Cotton and Fallow-Cotton samples (see Table 2). Analyzing the SOM clusters, we found many mislabeled samples. Fig. 12 shows two sets of NDVI values of Millet-Cotton samples. Clearly, samples shown in Fig. 12a are correct, while those in Fig. 12b are not. The latter set of samples had a posterior probability of 20% of belonging to the Millet-Cotton class. They were removed from the training set.

#### 4.3. Outlier Analysis

This subsection considers the case of outliers on the SOM maps. Those outliers do not necessarily result from errors in labeling, but are more likely to arise from variability of the ground data. Fig. 13 shows that most patterns of the Soy-Corn class are neighbors in the SOM map. However, there is an outlier neuron (Neuron 14) of the same class. Comparing one of the neurons of the Soy-Corn neighborhood (Neuron 240) with Neuron 14, we find out Neuron 240 has a prior and posterior

**Table 2**  
Result of class noise detection.

Samples by class	Keep	Remove	Flagged
Savanna Parkland	89.67%	4.63%	5.70%
Dense Woodland	86.70%	8.04%	5.24%
Savanna	88.25%	9.33%	2.40%
Rocky Savanna	86.55%	5.82%	7.62%
Dunes	100%	-	-
Fallow-Cotton	-	24.76%	75.24%
Millet-Cotton	-	67.40%	32.60%
Pasture	85.89%	11.89%	2.22%
Planted Forest	-	19.85%	80.15%
Soy-Corn	86.70%	9.31%	3.98%
Soy-Cotton	94.23%	4.58%	1.19%
Soy-Fallow	58.19%	29.93%	11.87%

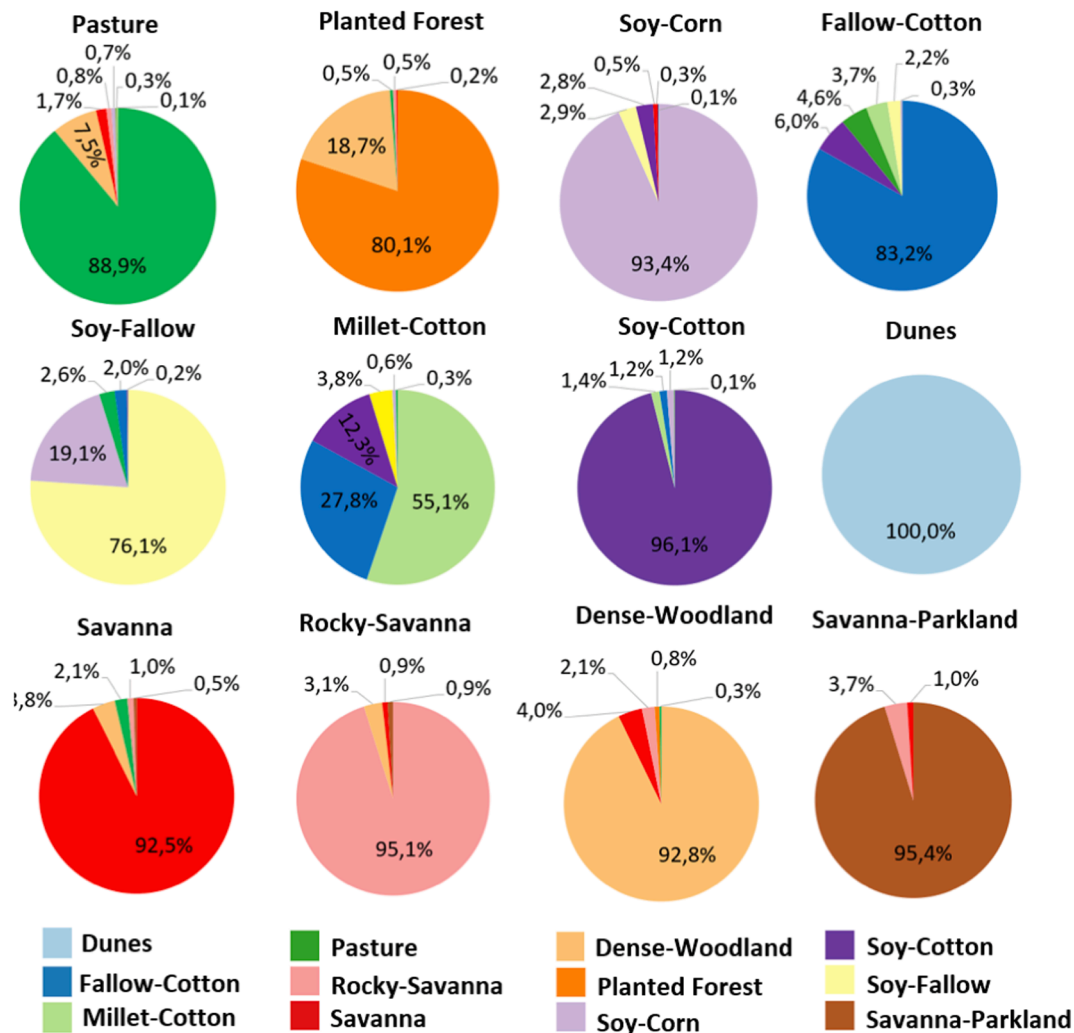


Fig. 9. Confusion between the classes.

probabilities of 96% and a and 91%. By contrast, Neuron 14 has prior and posterior probabilities of 81% and 39% and has been flagged for analysis. Looking in more detail in their spatial location, we discover that the samples mapped to Neuron 14 come from different areas that samples mapped to Neuron 240. Neuron 14 samples comes from the Brazilian states of Tocantins and Maranhão (about 7°S), while the cluster of samples close to Neuron 240 come from the state of Mato Grosso (about 12°S). The climatological variations lead the agricultural calendar to be different in these two areas. For this reason, the spectral response over time of the same class is different in the two areas. Fig. 13 displays the temporal signatures of the samples in both regions. These signatures show that the corn cycle is shorter in Tocantins and Maranhão than in Mato Grosso. This example illustrates the value of detecting and analysis outliers in the training set.

Table 3 presents the percentage of samples by class that were removed from the input dataset after the outlier analysis. All samples from the Savanna Parkland, Savanna, Planted Forest, Soy-Corn, and Soy-Cotton classes were kept in the dataset. Other classes analyzed had samples removed following the analysis. The Fallow-Cotton samples has the most noise. Given the thresholds set to evaluate sample quality, 24.8% of Fallow-Cotton samples were automatically removed from the dataset and 76.2% flagged for analysis, as shown in Table 2. After analysis, a further 56.2% of Fallow-Cotton samples were removed from the dataset, totaling 81.10% of samples removed. The input dataset contains 50,160 samples, whereas the filtered dataset includes 44,040. This means that 14% of the samples were removed due to class noise.

#### 4.4. Validation

To assess our proposed methods, we did a 5-fold cross-validation test, comparing the original training set (50,160 samples) with the filtered set (44,040 samples). We used a Random Forest (RF) algorithm due to its robustness and proven results on handling big data (Belgiu and Dragut, 2016). The number of trees used in our study was 2000, and the split rule for each node was the Gini index. The overall accuracy for the original dataset and the filtered dataset were respectively 94.3% and 98.4%. Table 4 presents producer's and user's accuracy for both datasets. Producer's accuracy improved for all classes; the largest increase occurred in noisy classes such as Fallow-Cotton and Millet-Cotton. The results corroborate our initial hypothesis that SOM-based clustering combined with Bayesian inference can improve the quality of large training samples of satellite image time series.

#### 5. Conclusion

Machine learning methods are now established as a useful technique for remote sensing image analysis. Despite the well-known fact that the quality of the training data is a key factor on the accuracy of the resulting maps, the literature on methods for detecting and removing class noise in SITS training sets is limited. To contribute to this limitation, this paper proposed a new technique. The proposed method uses the SOM neural network to group similar samples in a 2D map for dimensionality reduction. Each sample is mapped to a neuron on the 2D

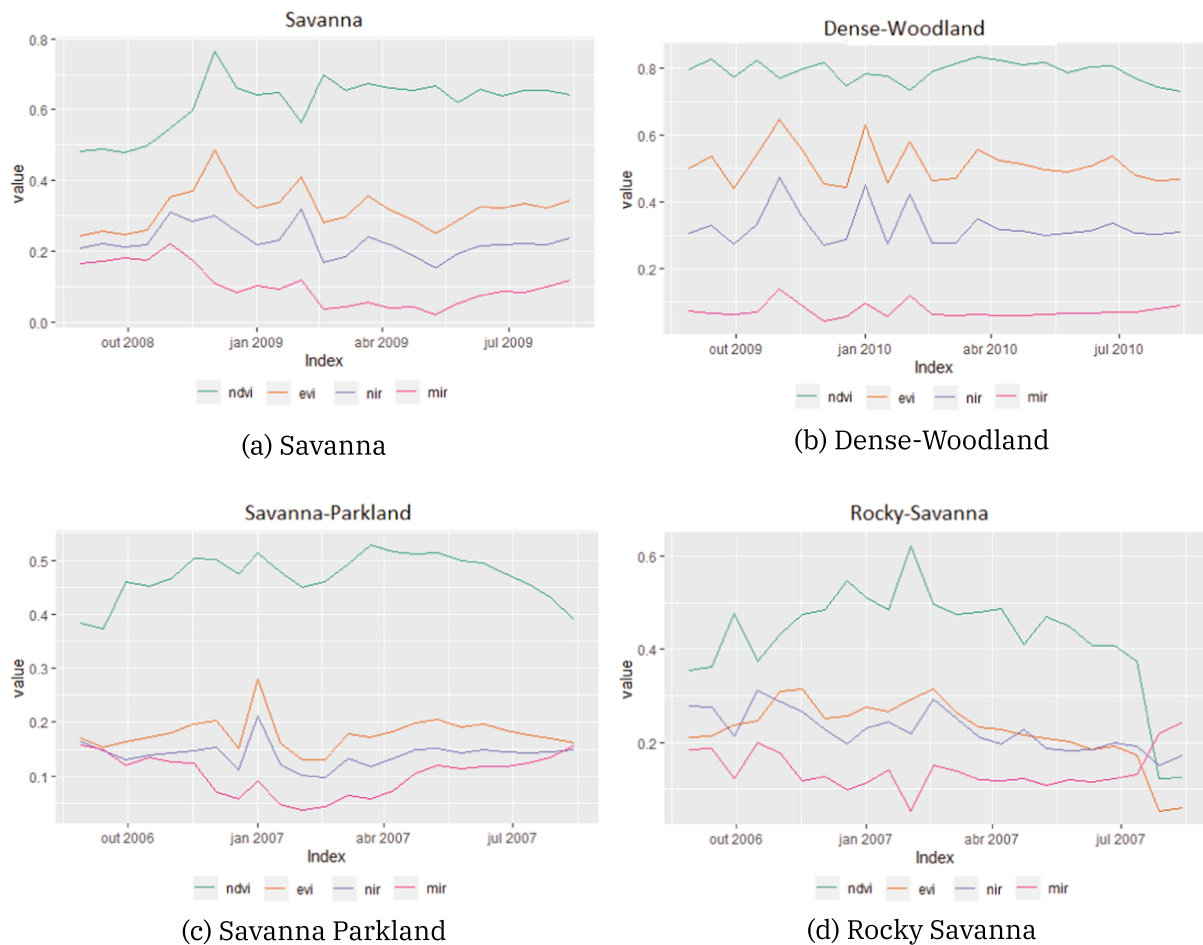


Fig. 10. Time series of ground samples for natural vegetation classes in the Cerrado Biome.

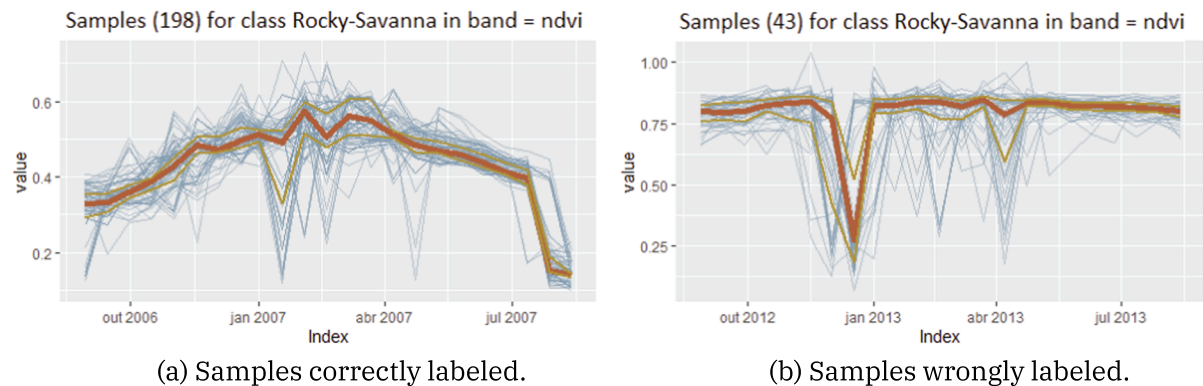


Fig. 11. NDVI time series samples labeled as Rocky-Savanna.

SOM map. In this way, a set of time series samples with similar patterns can be represented by a neuron's weight vector, as shown in Fig. 8. The neurons are then labeled with their majority class. Using the SOM property of topological preservation property, the algorithm uses Bayesian inference to evaluate the classes of the neuron neighborhoods through the probabilities provided by the labelling processing. As a result, the method identifies both mislabeled samples and outliers that are flagged to further investigation, as shown in Fig. 13.

The proposed refinement process of SITS training data improves the accuracy of the classification results. In the case study described in this paper, the mislabeled samples and part of the outliers identified by the proposed method were removed from the training set. Then, two

classifications were performed, one using the original SITS training set and the other using the filtered set. The results demonstrate the positive impact on the overall classification accuracy. Although the class noise removal adds an extra cost to the entire classification process, we believe that it is essential to improve the accuracy of classified maps using SITS analysis mainly for large areas.

One of the challenges of using machine learning techniques for analyzing large areas is the adequacy of sample data to the natural variations of classes in space and time. In our case study, each time series has a spatial location and a time period. Since our method associates different clusters of the same class in the SOM map to such space–time variations, it helps to deal with the problem of selecting good training



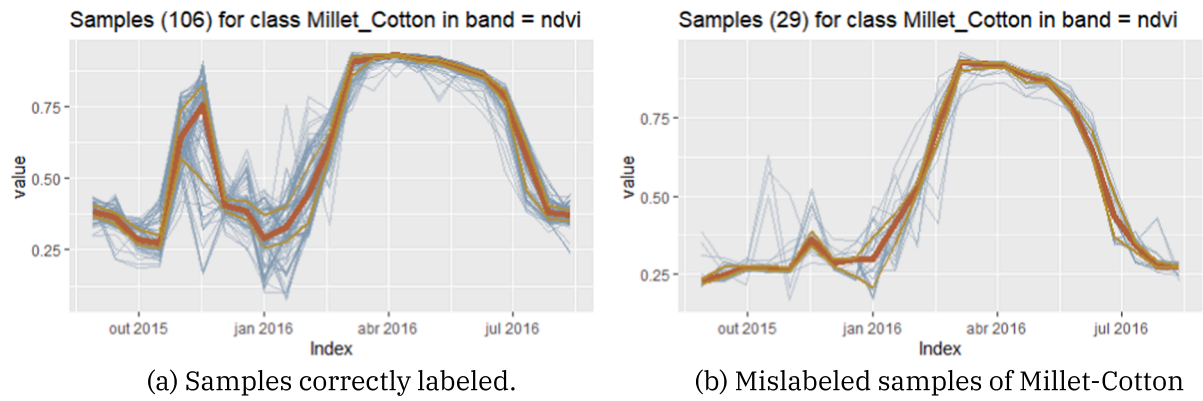


Fig. 12. NDVI time series samples labeled as Millet-Cotton.

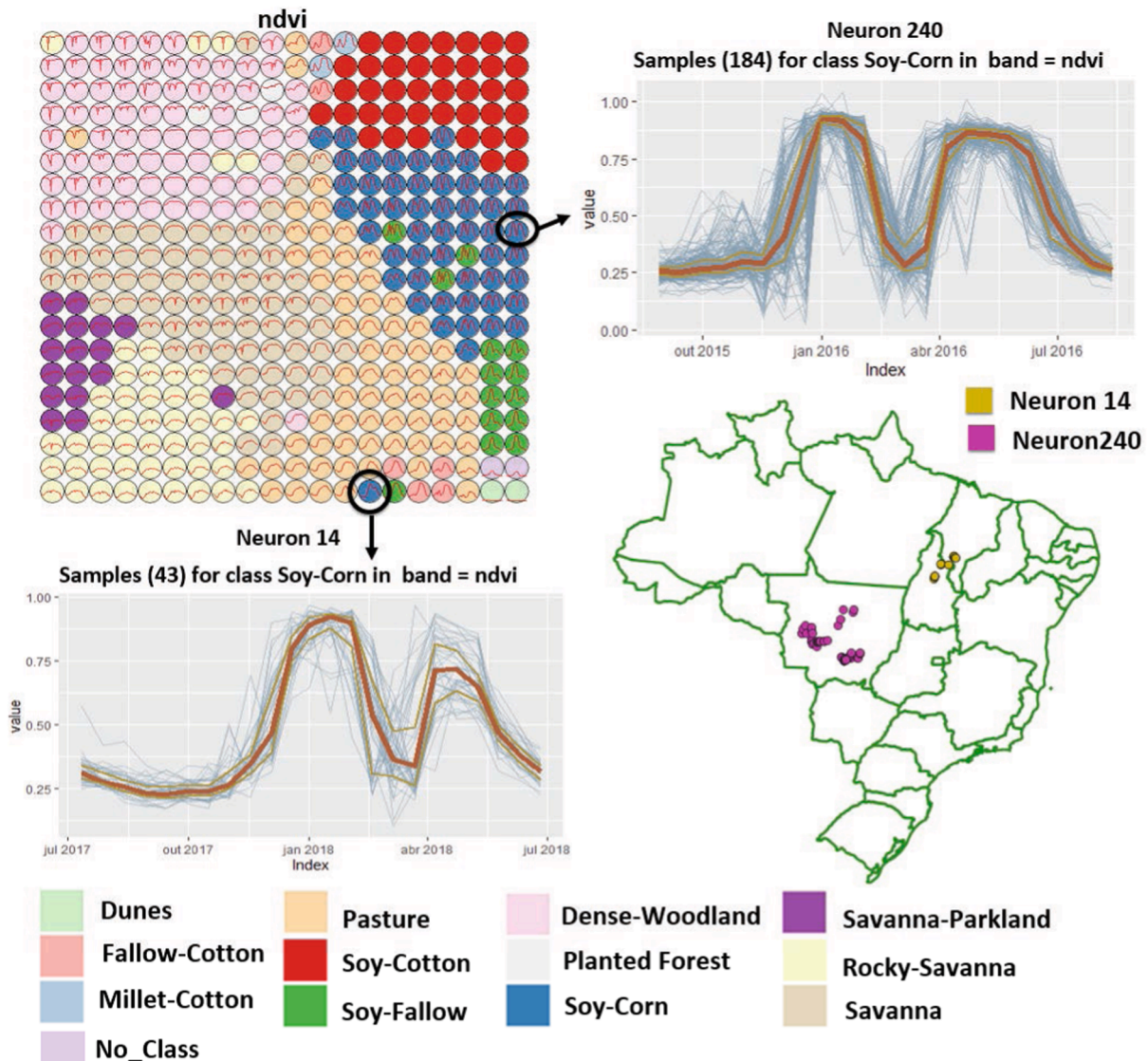


Fig. 13. Different patterns in the Soy-Fallow class because of the agricultural calendar in different regions.

samples over large areas.

Despite the usefulness of our proposed method, organizing a good quality training dataset remains one of the toughest problems in remote sensing data analysis. Natural land cover occurs in a continuum in spacetime. Transitions between ecosystems are rarely abrupt. Complex

biomes such as the Brazilian Cerrado contain subtle mixtures of trees and grasslands which defy crisp class definitions. As from pasture and croplands, agricultural practices vary from region to region and from year to year. For this reason, our method is an aid but not a substitute for in-depth local understanding of ecosystem behavior. Despite recent



**Table 3**

Overall samples removed before and after the analysis indicated by the conditional and posterior probabilities.

Samples by class	Before Analysis	After Analysis
Savanna Parkland	4.6 %	4.6 %
Dense-Woodland	8 %	9%
Savanna	9.3 %	9.3 %
Rocky-Savanna	5.8 %	11.5 %
Dunes	-	-
Fallow-Cotton	24.8 %	81%
Millet-Cotton	67.4 %	67.4 %
Pasture	11.9 %	13.6 %
Silviculture	19.9%	19.9%
Soy-Corn	9.3 %	9.4%
Soy-Cotton	4.6 %	4.6%
Soy-Fallow	29.9 %	41.8%

**Table 4**

Producer's and user's accuracy for original and filtered datasets.

Classes	Producer's Accuracy		User's Accuracy	
	Original	Filtered	Original	Filtered
Dense Woodland	97,4 %	98,9 %	87,1 %	96 %
Savanna Parkland	98,5 %	98,7 %	99,3 %	99,5 %
Rocky Savanna	90,9 %	96,3 %	99,2 %	99,7 %
Savanna	97,5 %	98,9 %	96,6 %	98,3 %
Dunes	100 %	100 %	100 %	100 %
Pasture	91,6 %	99,5 %	96,5 %	98,9 %
Planted Forest	75,4 %	76,9 %	98,7 %	99,2 %
Soy-Corn	97,5 %	98,8 %	91,1 %	98,9 %
Soy-Cotton	98,2 %	98,9 %	97,9 %	99,5 %
Fallow-Cotton	89,3 %	93,3 %	95,2 %	100 %
Millet-Cotton	85,1 %	97 %	96,7 %	100 %
Soy-Fallow	79,6 %	99,5 %	97,4 %	98,7 %
Total	91,7 %	96,3 %	96,3 %	99,0 %

progress in machine learning, local knowledge continues to be irreplaceable when using remote sensing data for land use and cover classification.

Although we present in this paper a case study in land use and cover classification, the proposed method is generic for class noise identification in any kind of time series reference database. Broadly, the method returns the probability of the time series  $t$  labeled as class  $k$  actually belonging to the class  $k$ , based on similarities among time series. Fig. 3 describes, in general, the method steps for class noise reduction in time series reference data.

## 6. Code and data availability

The proposed method was implemented in the R package *sits* (Satellite Image Time Series), available on GitHub at <https://github.com/e-sensing/sits>. Besides the sample quality assessment, the *sits* package provides methods for visualization, clustering and classification of satellite image time series. The code used in this manuscript to generate the results presented in Section 3 is provided under the GNU General Public License v3.0 and is available in Santos (2020). The 50,160 samples used in Section 3 are also available in Santos (2020).

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgements

This research was funded by the “Coordenação de Aperfeiçoamento de

Pessoal de Nível Superior” - Brasil (CAPES) - Finance Code 001 (LS), and by the Environmental Monitoring of Brazilian Biomes project (*Brazil Data Cube*), funded by the Amazon Fund through the financial collaboration of the Brazilian Development Bank (BNDES) and the Foundation for Science, Technology and Space Applications (FUNCTATE), Process 17.2.0536.1 (KF, LS, MP), and by the Climate Investment Funds (CIF) and the World Bank Group for supporting the Project “Development of Systems to Prevent Forest Fires and Monitor Vegetation Cover in the Brazilian Cerrado” (Project ID P143185) under the Brazil Forest Investment Program (FIP).

## References

- Adam, E., Mutanga, O., Rugege, D., 2010. Multispectral and hyperspectral remote sensing for identification and mapping of wetland vegetation: a review. *Wetlands Ecol. Manage.* 18, 281–296.
- Assuncao, R.M., Schmertmann, C.P., Potter, J.E., Cavenaghi, S.M., 2005. Empirical bayes estimation of demographic schedules for small areas. *Demography* 42, 537–558. <https://doi.org/10.1353/dem.2005.0022>.
- Atkinson, P.M., Jeganathan, C., Dash, J., Atzberger, C., 2012. Inter-comparison of four models for smoothing satellite sensor time-series data to estimate vegetation phenology. *Remote Sens. Environ.* 123, 400–417.
- Atzberger, C., Eilers, P.H., 2011. Evaluating the effectiveness of smoothing algorithms in the absence of ground reference measurements. *Int. J. Remote Sens.* 32, 3689–3709.
- Belgiu, M., Dragut, L., 2016. Random Forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing* 114, 24–31.
- Elmes, A., Alemohammad, H., Avery, R., Caylor, K., Eastman, J.R., Fishgold, L., Friedl, M.A., Jain, M., Kohli, D., Laso Bayas, J.C., et al., 2020. Accounting for training data error in machine learning applied to earth observations. *Remote Sensing* 12, 1034. <https://doi.org/10.3390/rs12061034>.
- Ferreira, K., Santos, L., Picoli, M., 2019. Evaluating Distance Measures for Image Time Series Clustering in Land Use and Cover Monitoring. in: *MACHINE Learning for Earth Observation Workshop*. Würzburg, Germany.
- Foley, J.A., DeFries, R., Asner, G.P., Barford, C., Bonan, G., Carpenter, S.R., Chapin, F.S., Coe, M.T., Daily, G.C., Gibbs, H.K., Helkowski, J.H., Holloway, T., Howard, E.A., Kucharik, C.J., Monfreda, C., Patz, J.A., Prentice, I.C., Ramankutty, N., Snyder, P.K., 2005. Global Consequences of Land Use. *Science* 309, 570–574.
- Frénay, B., Verleysen, M., 2013. Classification in the presence of label noise: a survey. *IEEE transactions on neural networks and learning systems* 25, 845–869.
- Garcia, L.P., de Carvalho, A.C., Lorena, A.C., 2015. Effect of label noise in the complexity of classification problems. *Neurocomputing* 160, 108–119.
- Garcia, L.P.F., Lorena, A.C., Carvalho, A.C.P.L.F., 2012. A study on class noise detection and elimination. In: *2012 Brazilian Symposium on Neural Networks*, pp. 13–18. <https://doi.org/10.1109/SBRN.2012.49>.
- Gomez, C., White, J.C., Wulder, M.A., 2016. Optical remotely sensed time series data for land cover classification: A review. *ISPRS Journal of Photogrammetry and Remote Sensing* 116, 55–72.
- Griffiths, P., Nendel, C., Hostert, P., 2019. Intra-annual reflectance composites from Sentinel-2 and Landsat for national-scale crop and land cover mapping. *Remote Sens. Environ.* 220, 135–151. <https://doi.org/10.1016/j.rse.2018.10.031>.
- Hird, J.N., McDermid, G.J., 2009. Noise reduction of NDVI time series: An empirical comparison of selected techniques. *Remote Sens. Environ.* 113, 248–258.
- Inglada, J., Vincent, A., Arias, M., Tardy, B., Morin, D., Rodes, I., 2017. Operational High Resolution Land Cover Map Production at the Country Scale Using Satellite Image Time Series. *Remote Sensing* 9, 95. <https://doi.org/10.3390/rs9010095>.
- INPE, 2013. TerraClass Cerrado Project: Use and Vegetation Cover Map of the Cerrado. URL <http://www.dpi.inpe.br/tccerrado/>. accessed on 28 November 2019.
- Jiang, Z., Huete, A.R., Didan, K., Miura, T., 2008. Development of a two-band enhanced vegetation index without a blue band. *Remote Sens. Environ.* 112, 3833–3845. <https://doi.org/10.1016/j.rse.2008.06.006>.
- Kharon, R., Wachman, G., 2007. Noise tolerant variants of the perceptron algorithm. *Journal of Machine Learning Research* 8, 227–248 <http://jmlr.org/papers/v8/kharon07a.html>.
- Kohonen, T., 1990. The self-organizing map. *Proc. IEEE* 78, 1464–1480. <https://doi.org/10.1109/5.58325> conference Name: Proceedings of the IEEE.
- Kohonen, T., 2013. Essentials of the self-organizing map. *Neural Networks* 37, 52–65. <https://doi.org/10.1016/j.neunet.2012.09.018>.
- Liesenberg, V., Ponzoni, F.J., Galvão, L.S., 2007. Análise da dinâmica sazonal e separabilidade espectral de algumas fitofisionomias do cerrado com índices de vegetação dos sensores modis/terra e aqua. *Revista Árvore* 31, 295–305.
- Maus, V., Camara, G., Cartaxo, R., Sanchez, A., Ramos, F.M., Queiroz, G.R., 2016. A Time-Weighted Dynamic Time Warping Method for Land-Use and Land-Cover Mapping. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 9, 3729–3739.
- Maxwell, A.E., Warner, T.A., Fang, F., 2018. Implementation of machine-learning classification in remote sensing: An applied review. *Int. J. Remote Sens.* 39, 2784–2817.
- Mellor, A., Boukir, S., Haywood, A., Jones, S., 2015. 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. <https://doi.org/10.1016/j.isprsjprs.2015.03.014>  
<http://www.sciencedirect.com/science/article/pii/S0924271615000945>.
- Ministry of the Environment, B., 2019. Brazilian biomes. <https://www.mma.gov.br/>.
- Mountrakis, G., Im, J., Ogole, C., 2011. Support vector machines in remote sensing: A review. *ISPRS Journal of Photogrammetry and Remote Sensing* 66, 247–259.
- Natarajan, N., Dhillon, I.S., Ravikumar, P.K., Tewari, A., 2013. Learning with noisy labels. *Advances in neural information processing systems* 1196–1204.
- Natita, W., Wiboonsak, W., Dusadee, S., 2016. Appropriate learning rate and neighborhood function of self-organizing map (SOM) for specific humidity pattern classification over southern thailand. *International Journal of Modeling and Optimization* 6.
- Olofsson, P., Foody, G.M., Herold, M., Stehman, S.V., Woodcock, C.E., Wulder, M.A., 2014. Good practices for estimating area and assessing accuracy of land change. *Remote Sens. Environ.* 148, 42–57.
- Parente, L., Mesquita, V., Miziara, F., Baumann, L., Ferreira, L., 2019. Assessing the pasturelands and livestock dynamics in Brazil, from 1985 to 2017: A novel approach based on high spatial resolution imagery and Google Earth Engine cloud computing. *Remote Sens. Environ.* 232, 111301. <https://doi.org/10.1016/j.rse.2019.111301>.
- Pasquarella, V.J., Holden, C.E., Kaufman, L., Woodcock, C.E., 2016. From imagery to ecology: Leveraging time series of all available LANDSAT observations to map and monitor ecosystem state and dynamics. *Remote Sensing in Ecology and Conservation* 2, 152–170. <https://doi.org/10.1002/rse2.24>.
- Patrini, G., Rozza, A., Krishna Menon, A., Nock, R., Qu, L., 2017. Making deep neural networks robust to label noise: A loss correction approach, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1944–1952.
- Pelletier, C., Valero, S., Inglada, J., Champion, N., Marais Sicre, C., Dedieu, G., 2017. Effect of Training Class Label Noise on Classification Performances for Land Cover Mapping with Satellite Image Time Series. *Remote Sensing* 9, 173. <https://doi.org/10.3390/rs9020173>.
- Peng, J., Sun, W., Du, Q., 2019. Self-paced joint sparse representation for the classification of hyperspectral images. *IEEE Trans. Geosci. Remote Sens.* 57, 1183–1194. <https://doi.org/10.1109/TGRS.2018.2865102>.
- Petitjean, F., Inglada, J., Gancarski, P., 2012. Satellite Image Time Series Analysis Under Time Warping. *IEEE Trans. Geosci. Remote Sens.* 50, 3081–3095. <https://doi.org/10.1109/TGRS.2011.2179050>.
- Picoli, M., Camara, G., Sanches, I., Simoes, R., Carvalho, A., Maciel, A., Coutinho, A., Esquerdo, J., Antunes, J., Begotti, R.A., Arvor, D., Almeida, C., 2018. Big earth observation time series analysis for monitoring Brazilian agriculture. *ISPRS journal of photogrammetry and remote sensing* 145, 328–339.
- Rebbapragada, U., Brodley, C.E., 2007. Class noise mitigation through instance weighting. In: *European Conference on Machine Learning*. Springer, pp. 708–715.
- Ribeiro, J.F., Walter, B.M.T., 2008. As principais fitofisionomias do Bioma Cerrado. In: Sano, S.M., Almeida, S.P., Ribeiro, J.F. (Eds.), *Cerrado: Ecologia e Flora*. Embrapa, pp. 152–212.
- Ribeiro, L., Tabarelli, M., 2002. A structural gradient in cerrado vegetation of Brazil: Changes in woody plant density, species richness, life history and plant composition. *J. Trop. Ecol.* 18, 775–794. <https://doi.org/10.1017/S026646740200250X>.
- Santos, L., 2020. Source code for: Quality control and class noise reduction. URL <https://doi.org/10.5281/zenodo.3941278>, doi:10.5281/zenodo.3941278.
- Simoes, R., Picoli, M.C.A., Camara, G., Maciel, A., Santos, L., Andrade, P.R., Sánchez, A., Ferreira, K., Carvalho, A., 2020. Land use and cover maps for Mato Grosso State in Brazil from 2001 to 2017. *Scientific Data* 7, 34. <https://doi.org/10.1038/s41597-020-0371-4>.
- Soterroni, A.C., Ramos, F.M., Mosnier, A., Fargione, J., Andrade, P.R., Baumgarten, L., Pirker, J., Obersteiner, M., Kraxner, F., Camara, G., Carvalho, A.X.Y., Polasky, S., 2019. Expanding the Soy Moratorium to Brazil's Cerrado. *Science. Advances* 5, eaav7336. <https://doi.org/10.1126/sciadv.aav7336>.
- Strassburg, B.B.N., Brooks, T., Feltran-Barbieri, R., Iribarrem, A., Crouzeilles, R., Loyola, R., Latawiec, A.E., Oliveira Filho, F.J.B., Scaramuzza, C.A.M., Scarano, F.R., Soares-Filho, B., Balmford, A., 2017. Moment of truth for the Cerrado hotspot. *Nature Ecology & Evolution* 1, 1–3.
- Sun, J., Zhao, F., Wang, C., Chen, S., 2007. Identifying and correcting mislabeled training instances, in: *Future Generation Communication and Networking (FGCN 2007)*, pp. 244–250. doi:10.1109/FGCN.2007.146.
- Suepa, Tanita, Qi, Jiaguo, Lawawirojwong, Siam, Messina, Joseph P., 2016. Understanding Spatio-Temporal Variation of Vegetation Phenology and Rainfall Seasonality in the Monsoon Southeast Asia. *Environmental Research* 147, 621–629.
- Tan, H.S., George, S.E., 2004. Investigating learning parameters in a standard 2d som model to select good maps and avoid poor ones. In: *Proceedings of the 17th Australian Joint Conference on Advances in Artificial Intelligence*. Springer-Verlag, Berlin, Heidelberg, pp. 425–437. [https://doi.org/10.1007/978-3-540-30549-1\\_38](https://doi.org/10.1007/978-3-540-30549-1_38).
- Thanh Noi, P., Kappas, M., 2018. Comparison of Random Forest, k-Nearest Neighbor, and Support Vector Machine Classifiers for Land Cover Classification Using Sentinel-2 Imagery. *Sensors* 18, 18. <https://doi.org/10.3390/s18010018>.
- Vesanto, J., Alhoniemi, E., 2000. Clustering of the self-organizing map. *IEEE Transactions on neural networks* 11, 586–600.
- Woodcock, C.E., Loveland, T.R., Herold, M., Bauer, M.E., 2020. Transitioning from change detection to monitoring with remote sensing: A paradigm shift. *Remote Sens. Environ.* 238, 111558. <https://doi.org/10.1016/j.rse.2019.111558>.
- Yang, W., Peng, J., Sun, W., Du, Q., 2019. Log-euclidean kernel-based joint sparse representation for hyperspectral image classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 12, 5023–5034. <https://doi.org/10.1109/JSTARS.2019.2952408>.
- Zhang, X., Friedl, M.A., Schaaf, C.B., Strahler, A.H., Hodges, J.C.F., Gao, F., Reed, B.C., Huete, A., 2003. Monitoring vegetation phenology using MODIS. *Remote Sens. Environ.* 84, 471–475.
- Zhu, X., Wu, X., 2004. Class noise vs. attribute noise: A quantitative study. *Artif. Intell. Rev.* 22, 177–210. <https://doi.org/10.1007/s10462-004-0751-8>.