MINISTÉRIO DA
CIÊNCIA, TECNOLOGIA
E INOVAÇÕES

PÁTRIA AMADA
BRASIL
GOVERNO FEDERAL

sid.inpe.br/mtc-m21c/2021/02.19.18.01-TDI

ASSESSING AND IMPROVING LAND USE AND COVER SAMPLES USING SATELLITE IMAGE TIME SERIES

Lorena Alves dos Santos

Doctorate Thesis of the Graduate Course in Applied Computing, guided by Drs. Karine Reis Ferreira Gomes, and Giberto Camara, approved in February 03, 2021.

URL of the original document:
<http://urlib.net/8JMKD3MGP3W34R/447JHQ8>

INPE
São José dos Campos
2021

PUBLISHED BY:

Instituto Nacional de Pesquisas Espaciais - INPE
Coordenação de Ensino, Pesquisa e Extensão (COEPE)
Divisão de Biblioteca (DIBIB)
CEP 12.227-010
São José dos Campos - SP - Brasil
Tel.:(012) 3208-6923/7348
E-mail: pubtc@inpe.br

MINISTÉRIO DA
CIÊNCIA, TECNOLOGIA
E INOVAÇÕES

PÁTRIA AMADA
BRASIL
GOVERNO FEDERAL

sid.inpe.br/mtc-m21c/2021/02.19.18.01-TDI

ASSESSING AND IMPROVING LAND USE AND COVER SAMPLES USING SATELLITE IMAGE TIME SERIES

Lorena Alves dos Santos

Doctorate Thesis of the Graduate Course in Applied Computing, guided by Drs. Karine Reis Ferreira Gomes, and Giberto Camara, approved in February 03, 2021.

URL of the original document:
<http://urlib.net/8JMKD3MGP3W34R/447JHQ8>

INPE
São José dos Campos
2021

Cataloging in Publication Data

Santos, Lorena Alves.

Sa59a Assessing and improving land use and cover samples using satellite image time series / Lorena Alves dos Santos. – São José dos Campos : INPE, 2021.

xx + 97 p. ; (sid.inpe.br/mtc-m21c/2021/02.19.18.01-TDI)

Thesis (Doctorate in Applied Computing) – Instituto Nacional de Pesquisas Espaciais, São José dos Campos, 2021.

Guiding : Drs. Karine Reis Ferreira Gomes, and Giberto Camara.

1. Satellite image time series. 2. Spatiotemporal patterns. 3. Self-organizing maps. 4. Land use and cover changes. 5. Class noise. I.Title.

CDU 528.8:332.3

MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E INOVAÇÕES

INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS

Serviço de Pós-Graduação - SEPGR

DEFESA FINAL DE TESE DE LORENA ALVES DOS SANTOS

BANCA Nº 008/2021

No dia 03 de fevereiro de 2021, as 09h, por videoconferência, o(a) aluno(a) mencionado(a) acima defendeu seu trabalho final (apresentação oral seguida de arguição) perante uma Banca Examinadora, cujos membros estão listados abaixo. O(A) aluno(a) foi APROVADO(A) pela Banca Examinadora, por unanimidade, em cumprimento ao requisito exigido para obtenção do Título de Doutora em Computação Aplicada. O trabalho precisa da incorporação das correções sugeridas pela Banca Examinadora e revisão final pelo(s) orientador(es).

Título: "Assessing and improving land use and cover samples using satellite image time series"

Eu, Rafael Duarte Coelho dos Santos, como Presidente da Banca Examinadora, assino esta ATA em nome de todos os membros.

Membros da Banca

Dr. Rafael Duarte Coelho dos Santos - Presidente - INPE

Dra. Karine Reis Ferreira Gomes - Orientadora - INPE

Dr. Gilberto Câmara - Orientador - INPE

Dr. Marcos Gonçalves Quiles - Membro da banca - UNIFESP

Dr. Laerte Guimarães Ferreira - Convidado - UFG

Dra. Ana Carolina Lorena - Convidada - ITA

Documento assinado eletronicamente por **Rafael Duarte Coelho dos Santos**, **Tecnologista**, em 09/02/2021, às 16:21 (horário oficial de Brasília), com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site http://sei.mctic.gov.br/verifica.html, informando o código verificador **6499601** e o código CRC **8219DE94**.

Referência: Processo nº 01340.000198/2021-76 SEI nº 6499601

My parents Maria de Fatima and Aeroastro

ACKNOWLEDGEMENTS

First, I would like to thank my family's support, mainly my parents. My sister Luiza and my friends Natanael and Italo are also considered part of my family. They have always been encouraged over all these years.

I thank my supervisors, Dr. Karine Ferreira and Dr. Gilberto Câmara, for the dedication, patience, and competence they have given me during these years. I also would like to thank Dr.Raul Zurita-Milla and Dr.Ellen-Wien from the University of Twente for the advisor during my internship in the Netherlands.

Many thanks to my friends and colleagues from INPE Rolf Simões, Rodrigo Begotti, Alber Sanchez, Felipe Souza and Mikhail. They had an essential contribution to the development of this thesis. In special, I would like to thank Michelle Picoli for all technical support, discussion about the experiments performed for this thesis, and her friendship.

I would like to thank my friends from São José dos Campos, Italo, Pry, Gabi, Iuri, Naiallem, Dieg, Fernanda, Mayara, Leandro, Rita, Chris, Marcelo, and the people from GFI's republic for the fun times to keep me relaxed and for the comforting words during my doctorate. I also would like the friends I met during my internship in the Netherlands, Nubia, Laura, Evelina, Lina, Yang, and Nestor, for the host, support, friendship, and funny moments during my internship.

I would like to thank my friends from PUC-Goiás, Leticia, Lino, Doug, and Gilvan, for encouraging me to follow my way in science since the graduation.

I thank the institution National Institute for Space Research (INPE), who provided infrastructure support during this thesis's development. I thank the support by the Amazon Fund through the financial collaboration of the Brazilian Development Bank (BNDES) and the Foundation for Science, Technology and Space Applications (FUNCATE), process 17.2.0536.1; Coordenação de Aperfeiçoamento de Pessoal de Nível Superior-Brasil (CAPES), finance code 001; Institutional Program for Internationalization (PrInt-CAPES).

ABSTRACT

Land use and cover changes (LUCC) have caused a major impact on tropical ecosystems, increasing global greenhouse gas emissions and reducing the planet's biodiversity. Remote sensing and digital image processing are powerful tools to measure and monitor LUCC effectively. Nowadays, with many Earth observation satellite images freely available, image time series analysis brings new opportunities and challenges for LUCC mapping over large areas. The use of remote sensing image time series analysis and machine learning to produce LUCC information has greatly increased. Machine learning using supervised techniques require a training step using land use and cover samples labeled a priori. For this reason, it is necessary high-quality samples to avoid negative effects in classification performance. Due to the recent availability of open Earth observation data, methods using satellite image time series still are a gap in the literature. This thesis contributes to Earth observation field proposing two methods to assess the samples' quality and reduce the noise in the land use and cover reference datasets. The main idea is to identify mislabeled samples, data with low discrimination when mixed with other classes, and explore the samples' spatiotemporal variability using satellite image time series. The first method is based on unsupervised neural networks, the self-organizing map (SOM), and Bayesian Inference. It provides measures to identify mislabeled samples and assess the reliability of the samples. In the second method, the hierarchical clustering is combined with SOM to generate subgroups to identify spatiotemporal patterns to explore the samples' intra-class variability. Both methods use satellite image time series. It allows the Earth observation scientists to understand the sample's behavior over time, contributing to noise reduction in land use and cover reference databases. These methods were applied in different case studies using samples in the Cerrado biome in Brazil.The results indicated that the methods are efficient to reduce class noise and to assess the spatio-temporal variation of satellite image time series training samples.

Keywords: Satellite image time series. Spatiotemporal patterns. Self-organizing maps. Land use and cover changes. class noise.

AVALIAÇÃO E MELHORIA DE AMOSTRAS DE USO E COBERTURA DA TERRA UTILIZANDO SÉRIES TEMPORAIS DE IMAGENS DE SATÉLITE

RESUMO

Mudanças no uso e cobertura da terra têm causado grande impacto nos ecossistemas tropicais, aumentando as emissões globais de gases de efeito estufa e reduzindo a biodiversidade do planeta. O sensoriamento remoto e o processamento digital de imagens são ferramentas poderosas para medir e monitorar mudanças no uso e cobertura da terra. Atualmente, com uma grande quantidade de imagens de satélite de observação da Terra disponíveis gratuitamente, a análise de séries temporais de imagens traz novas oportunidades e desafios para o mapeamento das mudanças de uso e cobertura em grandes áreas. O uso de análise de séries temporais de imagens de sensoriamento remoto para produzir informações de mudança da terra tem aumentado bastante. Diversas abordagens de aprendizado de máquina com o foco em técnicas supervisionadas têm sido aplicadas para gerar mapas classificados de uso e cobertura da terra. Uma vez que esses métodos exigem amostras de treinamento rotuladas, elas devem possuir alta qualidade para evitar efeitos negativos no desempenho da classificação. No cenário de Observação da Terra, os métodos que utilizam séries temporais ainda são uma lacuna na literatura devido à recente disponibilidade de dados abertos de observação da Terra, principalmente métodos para a qualidade das amostras de treinamento. Esta tese contribui para o área de observação da Terra propondo métodos para avaliar a qualidade das amostras com o intuito de reduzir o ruído nos conjuntos de dados de observação da terra. A ideia principal é identificar amostras rotuladas de forma errônea, dados que apresentam baixa discriminação quando misturados com outras classes e explorar a variabilidade espaço-temporal dentro de cada classes utilizando séries temporais de imagens de satélite. O primeiro método apresentado nesta tese é baseada em redes neurais não supervisionadas, o mapa de auto-organização combinado com inferência bayesiana. Esta abordagem fornece medidas para identificar amostras com rótulos incorretos e avaliar a confiabilidade das amostras. No segundo método, o agrupamento hierárquico é combinado com o mapa auto-organizável para gerar padrões espaço-temporais de subgrupos com o intuito de explorar e identificar a variabilidade intraclasse das amostras. Ambos os métodos utilizam séries temporais de imagens de satélite. Isto permite que os cientistas de observação da Terra entendam o comportamento da amostra ao longo do tempo, contribuindo para a redução de ruído nos conjuntos de dados de amostras de uso e cobertura da terra. Os métodos abordados nesta tese foram aplicados em diferentes estudos de caso utilizando amostras no bioma Cerrado no Brasil.

Keywords: Séries temporais de imagens de satélite. Padrões espaço-temporal. Mapas auto-organizáveis. Mudança de uso e cobertura da terra. ruído de classe.

LIST OF FIGURES

Page

LIST OF TABLES

Page

LIST OF ABBREVIATIONS

BMU	–	Best Matching Unit
CCDC	–	Continuous Change Detection and Classification
DTW	–	Dynamic Time Warping
EO	–	Earth Observation
EVI	–	Enhanced Vegetation Index
GNU	–	General Public License
IBGE	–	Instituto Brasileiro de Geografia e Estatística
INPE	–	Brazilian Institute for Space Research
JRC	–	Joint Research Centre
LAI	–	Leaf Area Index
LSTM	–	Long Short-Term Memory
LUCC	–	Land Use and Cover Changes
MIR	–	Mid Infrared
MODIS	–	Moderate Resolution Imaging Spectroradiometer
NDVI	–	Normalized Difference Vegetation Index
NIR	–	Near Infrared
ODC	–	Open Data Cube
RF	–	Random Forest
SITS	–	Satellite Image Time Series
SOM	–	Self-Organizing Maps
SVM	–	Support Vector Machine
TWDTW	–	Time-Weighted Dynamic Time Warping
WSAS	–	Web Sample Assessment Service
VI	–	Vegetation Index

CONTENTS

Page

[1]This chapter is based on the paper: Santos, L., Ferreira, K. R., Picoli, M. and Camara, G., 2019. *Self-organizing maps in earth observation data cubes analysis.* In: International Workshop on Self-Organizing Maps and Learning Vector Quantization, Clustering and Data Visualization (WSOM), Vol. 13, Barcelona, Spain, pp. 70–79

[2]This chapter is based on the paper: Santos, L., Ferreira, K. R., Picoli, M. and Camara, G. and Simoes R,. Quality control and class noise reduction of satellite image time series. (Under review by the ISPRS Journal of Photogrammetry and Remote Sensing)"

[3]This chapter is based on the paper: Santos, L.A.; Ferreira, K.; Picoli, M.; Camara, G.; Zurita-Milla, R.; Augustijn, E.-W. Identifying Spatiotemporal Patterns in Land Use and Cover Samples from Satellite Image Time Series. Remote Sens. 2021, 13, 974.

1 INTRODUCTION

Due to growing pressures for food and energy production promoted by increasing population, humans have rapidly modified Earth's environment. Recently, land use and cover changes (LUCC) have caused a major impact on tropical ecosystems, increasing global greenhouse gases emissions and reducing the planet's biodiversity (FOLEY et al., 2005; HANSEN; LOVELAND, 2012). Land cover is the (bio)physical material at the Earth's surface, e.g., tropical forest, snow, desert, savannas, and water, while land use is how the humans utilize the land, e.g., agriculture, cattle ranching, and wood extraction (COMBER et al., 2008). Land use is the set of activities performed for humans on the land cover (LAMBIN et al., 2001), such as urban and agricultural. The characterization and mapping of these changes are essential for planning and managing natural resources.

Remote sensing and digital image processing allow to identify and map land use and cover changes (CHEN et al., 2015; GOMEZ et al., 2016). According to Kennedy et al. (2014), traditional remote sensing literature views ecosystems as static entities, with occasional disruptions causing dramatic contrasts in two images, taken before and after the change. However, it is difficult to detect the mapping of abrupt transformation with only two images. Change in land ecosystems is a continuous process. Instead of using only two dates, the best situation is to obtain a sequence of images that show change events as they happen.

Good quality datasets with the best possible temporal and spatial resolution are crucial to developing remote sensing analysis methods that enable a continuous view of ecosystem dynamics (KENNEDY et al., 2014). Integrating the temporal component with spatial and spectral dimensions results in richer datasets that improve LUCC monitoring. Progress on this area depends on the availability of spatial high-resolution dense time series analysis capable of capturing subtle and long-term change patterns (HOSTERT et al., 2015). According to Pasquarella et al. (2016), time series derived from Earth Observation (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. Camara et al. (2016) use the term *time-first, space-later* to refer to approaches that exploit the benefits of remote sensing time series for change detection.

Besides that, vegetation indices derived from spectral information of EO satellite images are widely used to generate LUCC information. According to Huete et al. (2002), vegetation indices are spectral transformations of two or more bands designed

to enhance vegetation properties. They provide spatial and temporal comparisons of global vegetation conditions that can be used to monitor and capture abrupt and long-term vegetation trends across large areas. Two examples of the most used vegetation indices are NDVI (Normalized Difference Vegetation Index) and EVI (Enhanced Vegetation Index). Some space agencies provide these vegetation indices as products derived from their satellite images. An example is the product MOD13Q1 of MODIS (Moderate Resolution Imaging Spectroradiometer) sensor provided by NASA with a temporal resolution of 16 days and spatial resolution of 250 meters (HUETE et al., 2002; FENSHOLT et al., 2015).

Nowadays, with a large amount of EO satellite images freely, the use of image time series brings new opportunities and challenges for LUCC mapping over large areas (GOMEZ et al., 2016). To support the satellite image time series analysis, EO Data Cubes infrastructures have been created. They model Analysis-Ready Data (ARD) generated from remote sensing images as multidimensional cubes (space, time, and spectral properties) (NATIVI et al., 2017; LEWIS et al., 2017). Besides that, these infrastructures take advantage of big data technologies and methods to store, process, and analyze the big amount of Earth observation satellite images (GOMES et al., 2020). Recent initiatives have made great strides towards creating EO data cubes for specific countries such as Australian Data Cube (LEWIS et al., 2017), Africa Regional Data Cube, Swiss Data Cubes (GIULIANI et al., 2017), Catalan Data Cube, and Brazilian Data Cube (FERREIRA et al., 2020). EO data cubes' purpose is to organize and provide data for simple and intuitive use to broadly the use of open EO data.

The Brazilian National Institute for Space Research (INPE) is developing the Brazil Deata Cube (BDC) initiative to create ARD and EO data cubes for all Brazilian territory. Besides datasets, software products and web services have been developed by the BDC project to create, access, discover, analyse and process EO data cubes. Big data technologies, cloud computing environments, and machine learning techniques have also been explored extensively in the BDC project.

The BDC project has four main objectives: (1) create ARD image collections from medium resolution remote sensing images (10 to 64 m) for all the Brazilian territory; (2) model these ARD images as multidimensional data cubes with three or more dimensions that include space, time and spectral-derived properties, mainly to support image time series analysis; (3) use, propose and develop big data technologies to create, store, and process these data cubes; and (4) create land use and cover

information from these EO data cubes for Brazil, using satellite image time series analysis, machine learning methods, and image processing procedures.

The processing of large amounts of data requires advanced computational techniques, for this reason the use of machine learning to produce LUCC information has greatly increased in recent years (GOMEZ et al., 2016; AGUIAR et al., 2010; ARVOR et al., 2011; MAUS et al., 2016; SPERA et al., 2014; PICOLI et al., 2018; SIMOES et al., 2020). Aguiar et al. (2010) identify pasture land and its different levels of degradation in Mato Grosso do Sul state, Brazil, using MODIS NDVI time series and a J48 classifier with wavelet technique. Arvor et al. (2011) use MODIS EVI time series to quantify the evolution of the agricultural area from 2000 to 2006 in Mato Grosso state, Brazil. Maus et al. (2016) propose an algorithm called Time-Weighted Dynamic Time Warping (TWDTW), based on the classical Dynamic Time Warping (DTW) method, for land cover and land use classification and present a case study using MODIS time series in the Porto dos Gaúchos municipality in Mato Grosso state, Brazil. Spera et al. (2014) use MODIS EVI time series to examine patterns of cropland expansion in Mato Grosso from 2001 to 2011 using a decision-tree algorithm. Picoli et al. (2018) and Simoes et al. (2020) use MODIS time series to provide information about crop and pasture expansion over natural vegetation in the Mato Grosso state from 2001 to 2017 using Support Vector Machine (SVM).

Most of the approaches for LUCC mapping use supervised learning techniques. However, when these techniques are applied, the biggest challenge is the necessity of the large training data labeled a priori with good quality. In supervised machine learning methods, training samples are also called reference databases. They must properly represent the land use and cover classes that are supposed to be identified by the classifier. The quality of training samples is crucial in the supervised classification process because it leads to results with better accurate maps and decreases the computational complexity and training time (ZHU; WU, 2004a; GOMEZ et al., 2016; PELLETIER et al., 2017).

The task of getting well-labeled land use and cover samples is also a challenge. The dataset requires many samples to describe the variability mainly over larger areas (PELLETIER et al., 2017). Usually, the main ways to collect the samples are through the interpretation of high-resolution satellite images or land use and cover maps and field collection. However, these sources of data are plausible for errors. The publication of new maps can be long due to the long production, the field collection can be an expensive process, and sometimes, the experts can introduce errors when the classes

contain small differences. Several approaches have applied semi-supervised learning and active learning to support the training samples acquisition (DEMIR et al., 2010; TUIA et al., 2011; HUANG et al., 2015; LU et al., 2017). However, these methods are not suitable for large areas due to the large number of samples necessary to characterize each class (RADOUX et al., 2014), and they require good quality preexisting labeled training data (VIANA et al., 2019).

Although much research on big data and machine learning is increasing, there is a knowledge gap for EO scientists (ELMES et al., 2020) because of the recent availability of open EO data. For this reason, the methods for satellite image time series are still maturing, and the literature to deal with the problem to improve the quality of training data for large areas in multiples years is limited (PELLETIER et al., 2017). Different approaches have been proposed to improve the reference data quality (PENGRA et al., 2020). Such strategies include best practices in collecting training data (OLOFSSON et al., 2014; ELMES et al., 2020; PENGRA et al., 2020; HUANG et al., 2020) and refinement methods of samples using satellite image time series (VIANA et al., 2019; SIMOES et al., 2020; BELGIU et al., 2021). However, most studies are limited to small areas, and inter-annual extents (PENGRA et al., 2020). Besides that, the intra-class variability of land use and cover data is intrinsic to different regions and periods due to distinct factors, such as different agricultural practices and climatological variations (HOSTERT et al., 2015). For this reason, it is necessary to develop methods considering a big EO database and the high intra-class variability considering the spatiotemporal variations, mainly in large areas and multiple years.

1.1 Our proposal

As part of the BDC initiative, this thesis contributes to assess and improve the quality of land use and cover samples used to produce LUCC maps using machine learning and image time series.

This thesis addresses the research question: *How to improve the quality of big EO reference data using satellite image time series to produce more accurate land use and cover change maps?* The hypothesis is that methods based on time series clustering are useful when dealing with high-variability data or incomplete information. Satellite image time series have high intra-annual variability, and observations of the same land cover can differ yearly. For such datasets, non-supervised clustering is useful to assign similar observations to the same cluster.

The methods proposed in this thesis are based on Self-Organizing Maps (SOM) neural network (KOHONEN, 1982) combined with other methods such as Bayesian Inference and Hierarchical clustering. The SOM was chosen due to its two fundamental properties, (1) dimensionality reduction and (2) topological preservation. SOM has the capability of mapping from a high-dimensional input space to a low-dimensional map space (usually two-dimensional grids), generating clusters of similar patterns in the output space. Due to these properties, SOM is a suitable tool for exploratory analysis of remote sensing time series. In the SOM method, each sample is allocated in a neuron. The clusters are composed of neurons with similar characteristics. Similar patterns tend to stay close in the output space. Hence, SOM is a suitable tool for outliers identification in the training samples. Several approaches have applied SOM in the spatiotemporal analysis due to its properties (ASTEL et al., 2007; AUGUSTIJN; ZURITA-MILLA, 2013; CHEN et al., 2018; QI et al., 2019), it can help to deal with the variability of vegetation phenology better than other methods that do not have these properties. Due to climatic phenomena, the vegetation phenology suffers variations over time. Phenological patterns can vary spatially across a region and are strongly correlated with climate variations over time (SUEPA et al., 2016). For example, rainy years may have a pattern for pasture different from a non-rainy year. It is impossible to have a set of samples that capture all phenological variations over time. Therefore, it is necessary methods that can take into account these variations.

The main contributions of this thesis are: (1) the proposed methods to improve the quality of the samples. The first method uses SOM combined with Bayesian inference to provide measures indicating the reliability of the samples. In the second method, the SOM is combined with hierarchical clustering to assess spatiotemporal patterns to explore the sample's intra-class variability; (2) the implementation of these methods as part of an **R** package, *sits* (Satellite Image Time Series), available on GitHub[1]; and (3) good-quality land use and cover samples dataset created using the proposed methods. These samples were used to create LUCC maps for the state of Mato Grosso, Brazil, from 2001 to 2017, based on MODIS image time series (collection 6) (SIMOES et al., 2020) available at PANGAEA repository (CAMARA et al., 2019).

1.2 Document structure

The structure of this thesis is based on three papers:

[1] https://github.com/e-sensing/sits.

a) Santos, L., Ferreira, K. R., Picoli, M. and Camara, G., 2019. *Self-organizing maps in earth observation data cubes analysis*. In: International Workshop on Self-Organizing Maps and Learning Vector Quantization, Clustering and Data Visualization (WSOM), Vol. 13, Barcelona, Spain, pp. 70–79, doi:10.1007/978-3-030-19642-4_7.

b) Santos, L., Ferreira, K. R., Picoli, M. and Camara, G. and Simoes R,. *Quality control and class noise reduction of satellite image time series*. (The first version was submitted in July, 2020 and an updated version, based on the reviewers comments, was submitted on November, 2020 to the ISPRS Journal of Photogrammetry and Remote Sensing).

c) Santos, L.A.; Ferreira, K.; Picoli, M.; Camara, G.; Zurita-Milla, R.; Augustijn, E.-W. Identifying Spatiotemporal Patterns in Land Use and Cover Samples from Satellite Image Time Series. Remote Sens. 2021, 13, 974. doi:https://doi.org/10.3390/rs13050974

Chapter 2 presents an overview of LUCC mapping using remote sensing image time series and how EO data cubes have been applied to generate this information. This chapter also introduces an initial experiment about the utility of SOM to extract LUCC information from EO Data Cubes infrastructures using image time series analysis. In this context, SOM is used to assess land use and cover samples and evaluate which spectral bands and vegetation indices are best suitable for the separability of land use and cover classes. A case study is described in this work and shows the potential of SOM in this application.

Chapter 3 describes the method proposed to assess and improve satellite image time series training data quality. The method uses SOM to produce time series clusters 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 in reducing class noise and assessing the spatiotemporal variation of satellite image time series training samples. The code used in this chapter to generate the results presented in Section is provided under the GNU General Public License v3.0. It is available in (SANTOS, 2020). The 50,160 samples used in chapter 3 are also available in (SANTOS, 2020).

Chapter 4 presents a method to identify and explore spatiotemporal patterns in land use and cover samples from satellite image time series. The proposed method recognizes intra-class variability to different regions and periods, mainly in large areas and multiple years. The method is based on self-organizing maps to reduce the large data dimensionality and hierarchical clustering to evaluate the intra-class variability. We present a case study using pasture and agriculture samples over the Cerrado biome in Brazil. The results show that the proposed methods are suitable for identifying spatiotemporal patterns in land use and cover samples and, consequently, improving the training samples' quality.

Chapter 5 presents the final discussions and future works.

2 SELF-ORGANIZING MAPS IN EARTH OBSERVATION DATA CUBE ANALYSIS[2]

Earth Observation Data Cubes (EO Data Cubes) are emergent infrastructures that model analysis-ready data generated from remote sensing images as multidimensional cubes, especially for satellite image time series analysis (NATIVI et al., 2017). Such data cubes have three or more dimensions that include space, time, and properties. EO Data Cubes can be defined as a set of time series associated with spatially aligned pixels ready for analysis.

EO Data Cubes infrastructure is an innovative way to organize the big amount of Earth observation satellite images freely available nowadays and take advantage of big data technologies and methods to store, process, and analyze time series extracted from these images. Examples of computational platforms for EO Data Cubes are the Open Data Cube (ODC) (LEWIS et al., 2017), the Joint Research Centre (JRC) Earth Observation Data and Processing Platform (JEODPP) (SOILLE et al., 2018) and the System for Earth Observation Data Access, Processing and Analysis for Land Monitoring (SEPAL) (FOOD AND AGRICULTURE ORGANIZATION - FAO, 2020).

A typical application that benefits from EO Data Cubes infrastructures and satellite image time series analysis is LUCC monitoring. Characterizing and mapping changes in the land surface is essential for planning and managing natural resources. The growing pressures for food and energy production promoted by increasing population make humans modify the Earth's environment rapidly. LUCC can affect hydrological and biological process causing great impacts on tropical ecosystems (PASQUARELLA et al., 2016).

Recently, EO Data Cubes infrastructures and satellite image time series analysis have brought new opportunities and challenges for LUCC mapping over large areas. Time series derived from Earth observation 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). The use of remote sensing image time series analysis to produce LUCC information has increased greatly in recent years (GOMEZ et al., 2016).

[2]This chapter is based on the paper: Santos, L., Ferreira, K. R., Picoli, M. and Camara, G., 2019. *Self-organizing maps in earth observation data cubes analysis*. In: International Workshop on Self-Organizing Maps and Learning Vector Quantization, Clustering and Data Visualization (WSOM), Vol. 13, Barcelona, Spain, pp. 70–79

Most classification techniques to create LUCC maps from satellite image time series are based on supervised learning methods. Such methods require a training phase using land use and cover samples labeled *a priori*. These training samples must properly represent the land use and cover classes to be identified by the classifier. The quality of these samples is crucial in the classification process. Representative samples lead to good LUCC maps.

This chapter presents the utility of the Self-Organizing Maps (SOM) neural network method to extract LUCC maps from EO Data Cubes infrastructures. SOM is a clustering method suitable for time series datasets. It describes the use of SOM in the training phase to produce metrics that indicate the quality of the land use and cover samples and evaluate which spectral bands and vegetation indexes are best suitable for the separability of land use and cover classes. A case study is described in Section 2.2, and it shows the potential of SOM in this context.

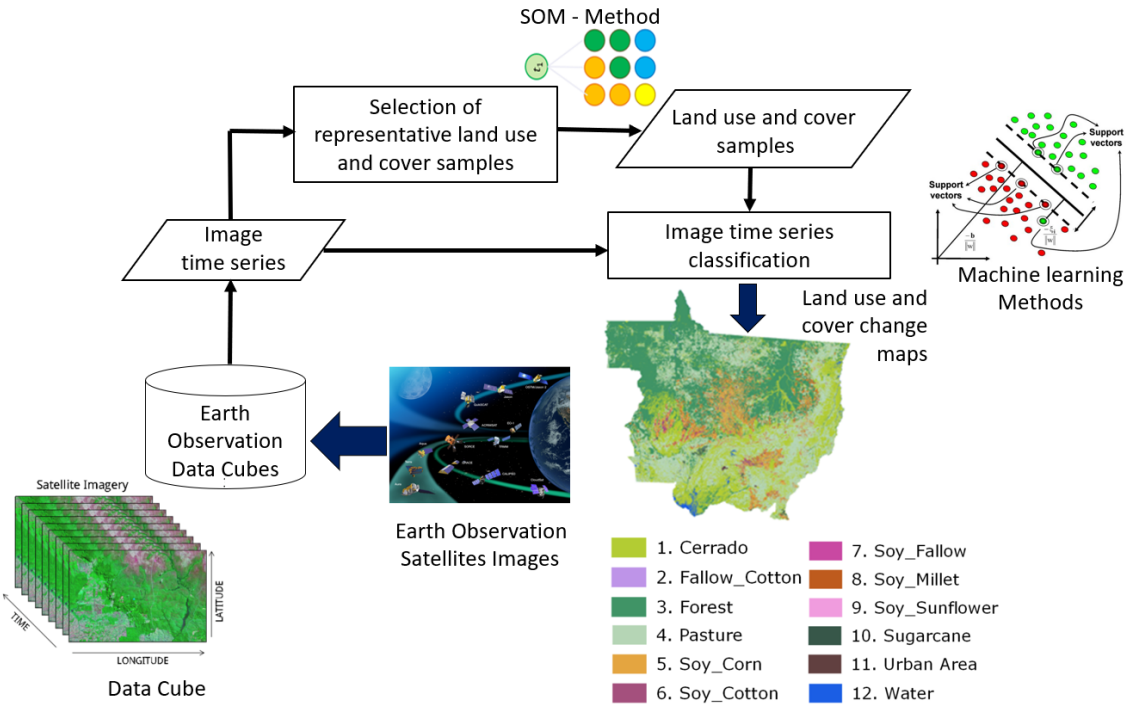
In the LUCC domain, SOM has not been widely explored for image time series analysis. The good review provided by Gomez et al. (2016) cites Bagan et al. (2005) as the main reference in this context. However, Bagan et al. (2005) proposed an approach to classify land cover from MODIS EVI time series using SOM. Besides that, Lawawirojwong (2013) proposed the use of supervised SOM for pure and mixed pixels, called soft supervised self-organizing map to improve the classification of MODIS-EVI time series. Both references, Bagan et al. (2005) and Lawawirojwong (2013), proposed the use of SOM to classify one agricultural year using only the EVI attribute. Unlike them, our proposal uses SOM to explore the separability time series using several attributes to improve the classification.

2.1 Land use and cover change information from Earth observation data cubes

This section describes the process, illustrated in Figure 2.1, to extract LUCC information from Earth Observation Data Cubes using image time series analysis and the utility of SOM method in this process.

To perform LUCC classification using Earth observation image time series, machine learning methods such as Support Vector Machine (SVM) and Random Forest (RF) have been used quite frequently (PICOLI et al., 2018). Most of these methods are based on supervised learning methods, which require a training phase using land use and cover samples labeled *a priori*.

Figure 2.1 - LUCC information from EO Data Cubes.



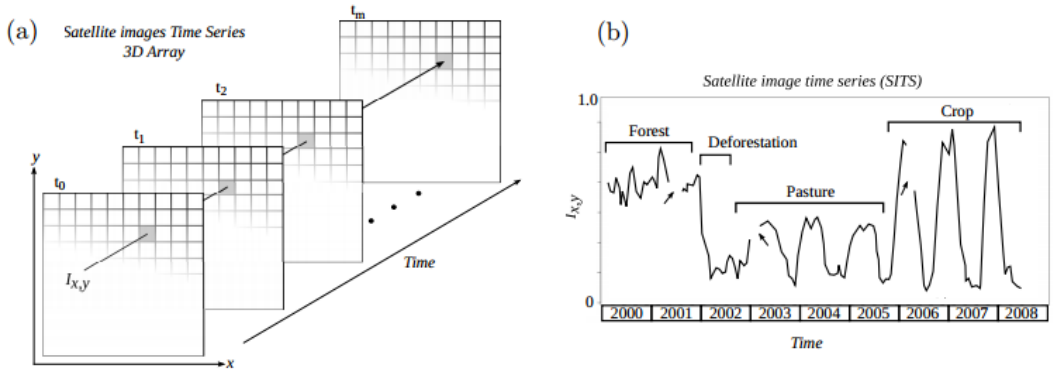
SOURCE: Author.

The selection of representative samples is crucial to obtain good classification accuracy. The exploratory analysis using time series clustering techniques, such as the SOM method, helps users improve the quality of land cover change samples.

2.1.1 Earth observation satellite image time series

Remote sensing satellites revisit the same place on Earth during their life cycle. The measure of the same place can be obtained at different times. These measures are mapped to three-dimensional array in space-time (MAUS et al., 2016), as shown in Figure 2.2(a). The time series is made from values obtained of each pixel location $I(x, y)$ over time, as presented in Figure 2.2(b). From these time series, LUCC can be extracted through vegetation phenology. Figure 2.2(b) shows an example of an area covered by forest from 2000 to 2001, then it was deforested, and during three years it was maintained as pasture. From 2006 to 2008, it was used for crop production.

Figure 2.2 - (a) A dimensional array of satellite images, (b) vegetation index time series at pixel location (x,y).



SOURCE: Maus et al. (2016).

2.1.2 Vegetation indices

Vegetation phenology is a biological event that indicates the stages of plants' growth and development during the life cycles. Remote sensing satellites are becoming essential for remotely capturing phenological variations on a large scale and extract phenological metrics from time series data of vegetation parameters. The most common parameters used are the vegetation indices (VI) (ZHANG et al., 2003).

During plant growth periods, different vegetation styles can be distinguished by time series vegetation indices (BOLES et al., 2004). Along with the VI, MODIS provides surface reflectance bands as RED, BLUE Near Infrared (NIR), and Mid Infrared (MIR). The VI are derived from these reflectance bands.

Limitations of the NDVI include sensitivity to atmospheric conditions, soil background, and saturation tendency in closed vegetation canopies with large leaf area index values (BOLES et al., 2004). The EVI signal has improved sensitivity in high biomass regions and improved vegetation monitoring. The blue band is used to remove residual atmosphere contamination caused by smoke, and sub-pixel thin cloud (UDELHOVEN et al., 2015). While NDVI is chlorophyll sensitive, EVI is more responsive to canopy structural variations, including leaf area index (LAI), canopy type, plant physiognomy, and canopy architecture (HUETE et al., 2002). The two vegetation indices complement each other in global vegetation studies.

2.1.3 Using SOM to improve the quality of land use and cover samples

In extracting LUCC information from EO Data Cubes, SOM is used to improve the training step of the land cover change classification. It is used to assess the quality of the land use and cover samples and evaluate which spectral bands and vegetation indexes are best suitable for the separability of land use and cover classes. This approach explores two main feature of SOM: (1) the topological preservation of neighborhood, which generates spatial clusters of similar patterns in the output space; and (2) the property of adaptation, where the winner neuron and its neighbors are updated to make the weight vectors more similar to the input.

Besides that, multiples attributes such as combined vegetation indices and multiples spectral bands can improve the patterns generated by SOM.

Instead of using only one vegetation index or spectral band, Wehrens and Buydens (2007) implemented an approach suitable for the use of several attributes simultaneously.

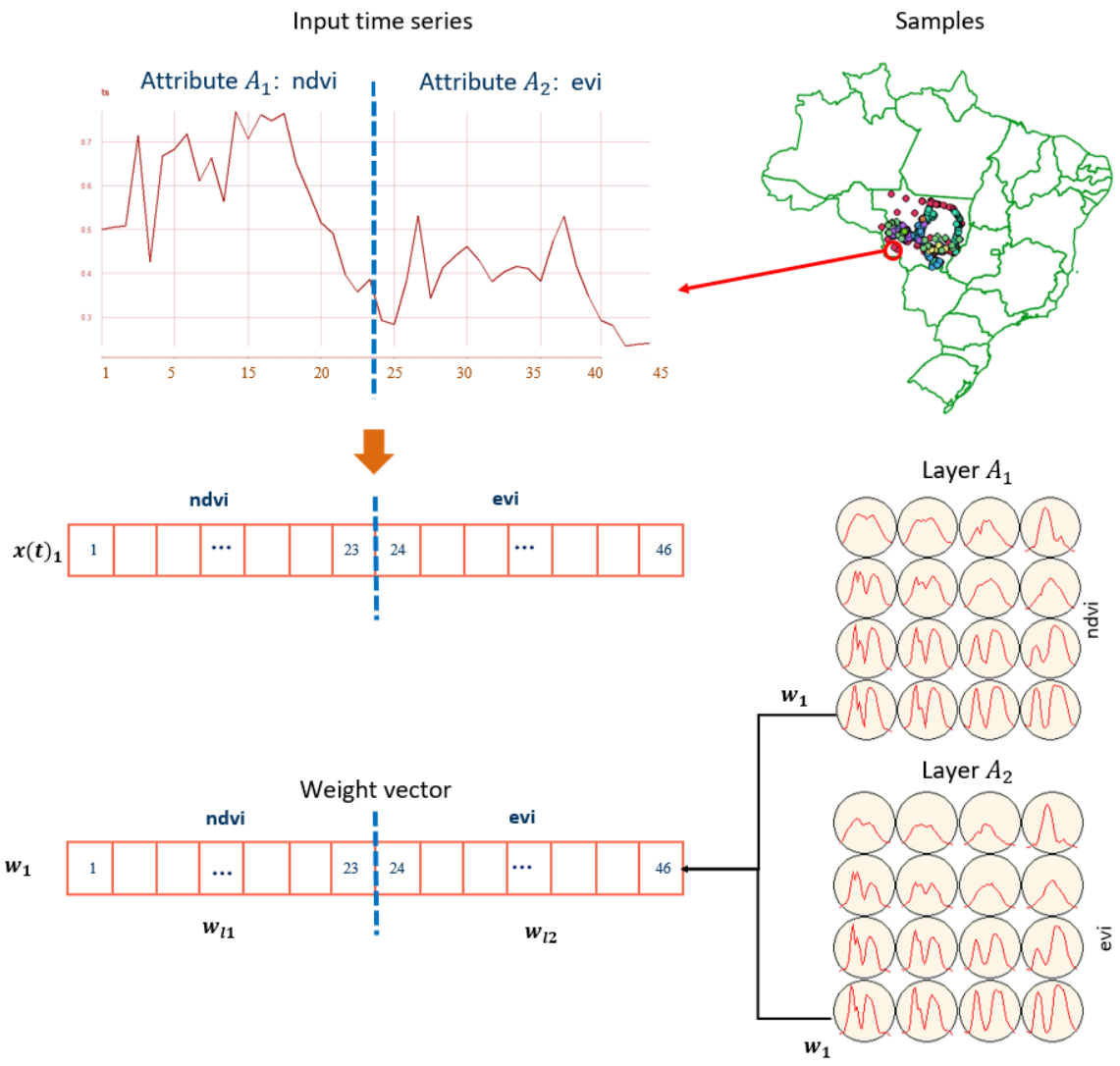
Considering a sample $x(t)_1$ with two attributes, x_{A1} and x_{A2}, for example, the vegetation indices NDVI and EVI, as shown in Figure 2.3. The input and weight vectors are concatenated according to the number of attributes. For example, considering 16-days of MODIS temporal resolution, each NDVI time series has 23 points representing a year, therefore using two attributes, a sample has an input vector with 46 points. Similarly, the weight vectors are initialized randomly with the same dimension of the input vector.

Internally, the approach implemented by Wehrens and Buydens (2007) creates an output layer consisting of a 2-D grid of neurons separately for each attribute. To identify the most similar neuron for an input vector, the Best Matching Unit (BMU), the distances between the input vector and all weight vectors, must be computed. In this case, it is computed separately for each layer. Considering the example of Figure 2.3, for the input $x(t)_1$ 16 distances are found for the attribute NDVI, and 16 distance for the EVI. To define a unique distance between the input $x(t)_1$ and the weight vector w_1, the BMU's found for each layer must be summed. The equation 2.1 shows how to calculate the distance for multiple attributes.

$$D_l = \sum_{l=1}^{n_l} D_l(i,j). \tag{2.1}$$

where l is the layer and n_l is the number of layers.

Figure 2.3 - Structure of SOM with two attributes.



SOURCE: Author.

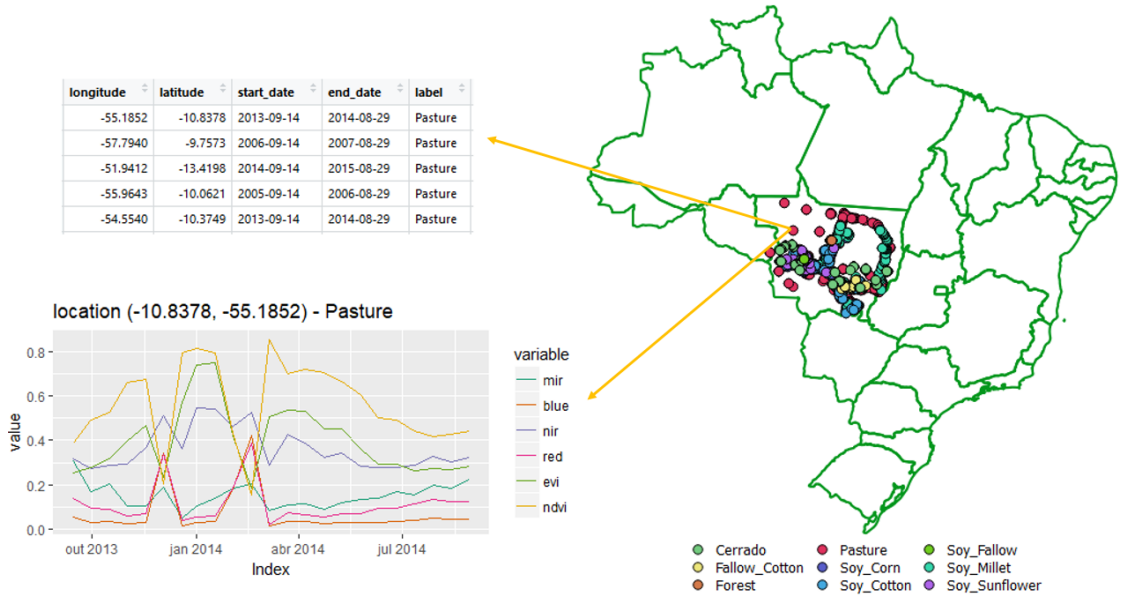
After the SOM training step, to evaluate the separability of samples it is necessary to label the neurons to create clusters. A cluster can be one neuron or a set of neurons that belongs to the same class. In this step, each neuron is labeled using the majority vote technique. Each neuron receives the label of the majority of the samples associated with it. In some cases, no samples are associated with a neuron. Then, this empty neuron receives the label 'Noclass'. To verify the quality of the

clusters generated by SOM, the confusion matrix can be accessed. From the confusion matrix, the percentage of the mixture within a cluster is calculated.

2.2 Case study

To show the potential of the SOM method in selecting good quality land use and cover samples from satellite image time series, this section describes a case study using VI time series of the product MOD13Q1 of the MODIS sensor from 2001 to 2016. The study area is the Mato Grosso State in Brazil, as shown in Figure 2.4. Each sample has a spatial location (latitude and longitude), start and end date that corresponds to agricultural year (from August to September), the class label that corresponds to the sample, and a set of time series with multiple attributes. In this case study, we used EVI, NDVI, NIR, MIR, BLUE, and RED. The ground samples include natural vegetation and agricultural classes for the Mato Grosso state of Brazil. The dataset includes 2215 ground samples divided into nine land use and cover classes: (1) Cerrado, (2) Pasture, (3) Forest, (4) Soy-Corn, (5) Soy-Cotton, (6) Soy-Fallow, (7) Soy-Millet, (8) Fallow-Cotton and (9) Soy-Sunflower. The ground samples were collected by Picoli et al. (2018).

Figure 2.4 - Samples Dataset.

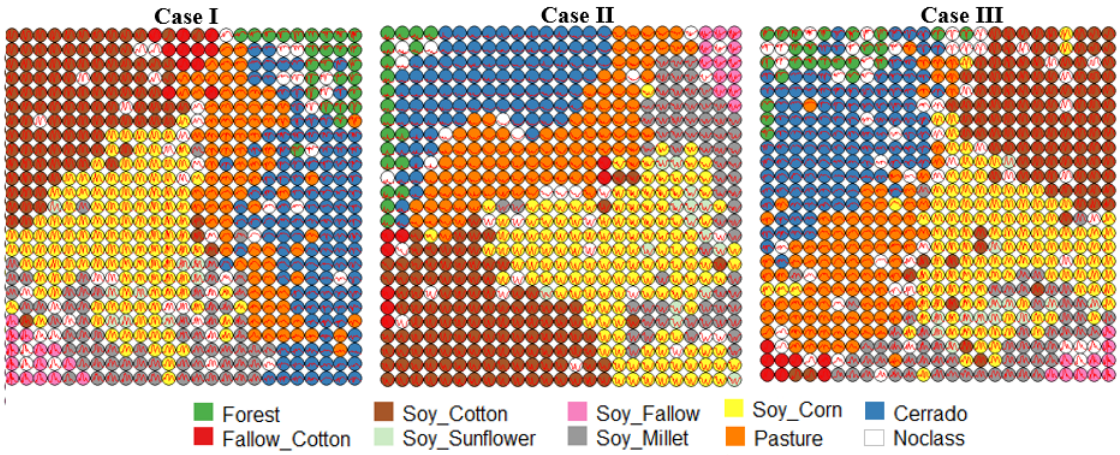


SOURCE: Author.

To evaluate the separability of these classes using SOM, clusters combining spectral bands and vegetation indices were generated in three cases: (1) Case I: NVDI and EVI; (2) Case II: NDVI, EVI, NIR, and MIR; (3) Case III: NDVI, EVI, NIR, MIR, RED, and BLUE. The SOM parameters that we used were: grid size = 25×25, learning rate = 1, and number of iterations = 100.

Figure 2.5 shows the maps created for each case. As we have the label of samples, the confusion matrix for each case was generated. Although the large variability within the land use and cover classes and the similarity of the phenological patterns among the classes, SOM separated these land use and cover classes with good accuracy. For Case I the accuracy was 88%, the Case II was 93%, and the Case III was 90%. Besides that, we can note that most of the neurons that belong to a neighborhood are the same category, but the time series samples contain small variations.

Figure 2.5 - Grids generated for each case.



SOURCE: Author.

From the confusion matrix, we can evaluate the quality of each land use and cover cluster generated by SOM. For each case, Table 1 shows the percentage of samples assigned to the appropriate cluster, that is, the class associated with the sample is the same class associated with the cluster. For example, the Cerrado cluster has 97.3% of samples labeled as Cerrado in Case II, 84% in Case I, and 93.3% in Case III.

In general, we can notice that in Case II, where the attributes MIR and NIR were considered, the quality of clusters was improved. The separability had a significant increase in the Cerrado and Fallow-Cotton clusters. There was a loss of separability quality for Forest and Soy-Sunflower clusters, but it is not so significant. In the same way, in Case III, the attributes BLUE and RED improved some clusters' separability compared with Case I but not so significantly.

Table 2.1 - Quality of clusters.

Cluster	**Case I**	**Case II**	**Case III**
Cerrado	84%	97.3%	93.3%
Fallow-Cotton	72.2%	85.7%	78.9%
Forest	100%	99.3%	89.9%
Pasture	92.7%	97.3%	93.7%
Soy-Corn	82.0%	84.0%	85.4%
Soy-Cotton	94.6%	95.5%	93.5%
Soy-Fallow	97.8%	100%	98.9%
Soy-Millet	85.5%	90.3%	88.2%
Soy-Sunflower	77.1%	76.9%	72.9%

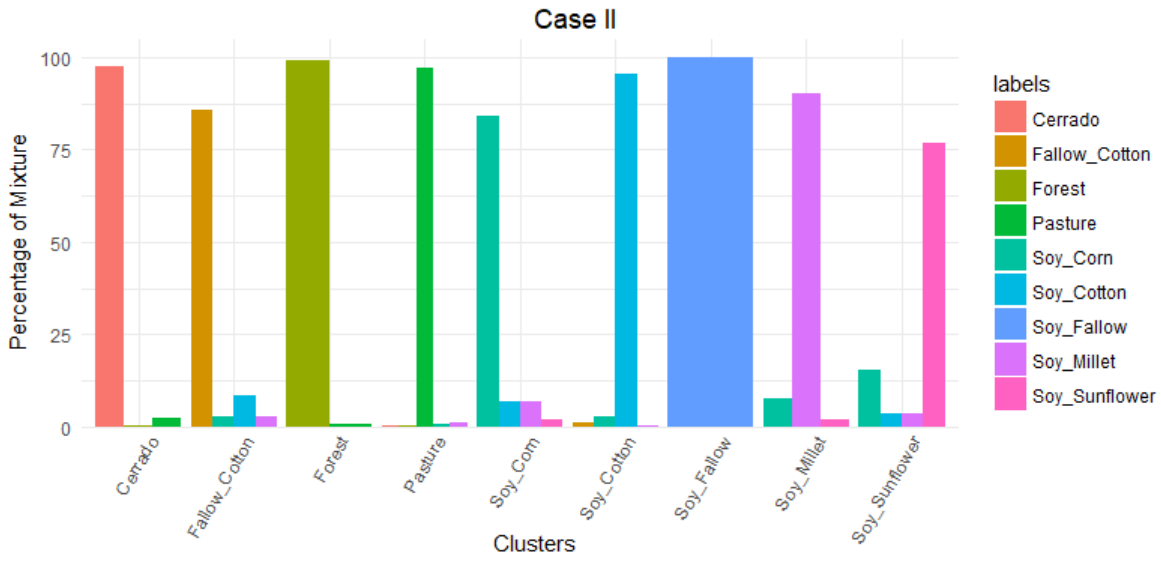
For Case II, each cluster's confusion is shown in Figure 2.6. The clusters Fallow-Cotton, Soy-Corn, and Soy-Sunflower, are the most confusing, that are crop classes. Crop classes have similar phenological patterns. This confusion can be noted in the maps of Figure 2.5 where there are neurons labeled as Fallow-Cotton and Soy-Sunflower within the neighborhood of Soy-Corn. Some Cerrado and Pasture samples have similar spectral curves, but the attributes MIR and NIR reduced the confusion between these samples, as shown in Figure 2.6.

2.3 Final Remarks

This chapter presents the SOM method's utility to improve LUCC classification from satellite image time series using EO Data Cubes infrastructures. The proposed approach uses SOM to evaluate which spectral bands and vegetation indexes are best suitable for the separability of land use and cover classes and improve the quality of the land use and cover samples.

It was presented a case study that evaluate the combination of six attributes, EVI, NDVI, NIR, MIR, RED, and BLUE, using MODIS time series of land use and cover

Figure 2.6 - Confusion among the classes.



SOURCE: Author.

samples in the Mato Grosso State in Brazil. The results show the potential of SOM to identify the separability of land use and cover types.

Even though the general accuracy of separability using only NDVI and EVI attributes was 88%, the classification in big scale areas can generate many errors. Including the attributes MIR and NIR, we noticed a great improvement in the accuracy of separability that was 93%. Considering that a neuron is a cluster, when the MIR and NIR attributes were added, the percentage of samples assigned to the appropriate clusters increased. A sample assigned to the appropriate cluster means that the class associated with the sample is the same class associated with the cluster.

In the third case, when we used all attributes, including BLUE and RED, the accuracy of separability was worse than in the second case, resulting in 90%. This analysis is important because it is possible to conclude that adding more attributes does not mean increasing the accuracy. Besides that, the computational cost is proportional to the number of attributes used in the LUCC classification. It is crucial to identify the minimal number of attributes that lead to the best results.

Finally, we have implemented our approach in **R** using the SOM method available in the Kohonen R package (WEHRENS; BUYDENS, 2007). This package has the online

and batch approaches of SOM. However, we use the online method for this work. The Kohonen package was integrated with the Satellite Image Time Series (sits) package. The package sits package was developed in the e-sensing project developed by the Brazilian Institute for Space Research (INPE) to provide tools for working with analyses, clustering, and classification of satellite image time series. Its source code is available at github[3].

[3]https://github.com/e-sensing/sits.

3 QUALITY CONTROL AND CLASS NOISE REDUCTION OF SATELLITE IMAGE TIME SERIES[4]

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 remote sensing image processing methods play a crucial role in the identification, mapping, assessment, and monitoring of LUCC.

In this chapter, 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 increasingly used in 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; 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 (COMPARISON...,). Therefore, using machine learning for SITS analysis requires large and good quality training sets.

[4]This chapter is based on the paper: Santos, L., Ferreira, K. R., Picoli, M. and Camara, G. and Simoes R,. Quality control and class noise reduction of satellite image time series. (Under review by the ISPRS Journal of Photogrammetry and Remote Sensing)"

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 datasets associated with satellite image time series. Thus, our work 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 SITS analysis methods 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 chapter, we propose a new method for class noise reduction in the SITS reference database.

The proposed method creates a self-organizing map 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 land use and cover classification using a large SITS dataset.

3.1 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 its effects in supervised learning is widely discussed in the literature of Neurocomputing, Artificial Intelligence and Machine Learning (ZHU; WU, 2004b; FRÉNAY; 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; WU, 2004b; GARCIA et al., 2012), based on Bayesian classifier (SUN et al., 2007) or based on clustering methods (REBBAPRAGADA; BRODLEY, 2007). Filters are applied before the learning process. Differently from filters, noise-tolerant variants

of classifiers are proposed to be more tolerant and robust to noise, dealing with label noise during learning or considering label noise in an embedded way (KHARDON; WACHMAN, 2007; NATARAJAN et al., 2013; FRÉNAY; VERLEYSEN, 2013; PATRINI et al., 2017). For example, one variant of the 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 with near ones based on decision boundaries. Although these methods are robust, they are not still free to be affected by noise (FRÉNAY; VERLEYSEN, 2013).

In Remote Sensing, many works have highlighted the importance of good quality samples to train machine learning methods to produce land use and cover maps with great accuracy from SITS analysis (OLOFSSON et al., 2014; GOMEZ et al., 2016; BELGIU; DRAGUT, 2016; ELMES et al., 2020). However, few papers 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; MCDERMID, 2009; ATZBERGER; 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 drops 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 chapter addresses the class label noise problem in large sets of SITS training samples and presents a solution for it. We propose a novel method for class label noise reduction in large SITS data and present how it improves land use quality 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 identify and remove label noise in large SITS training datasets.

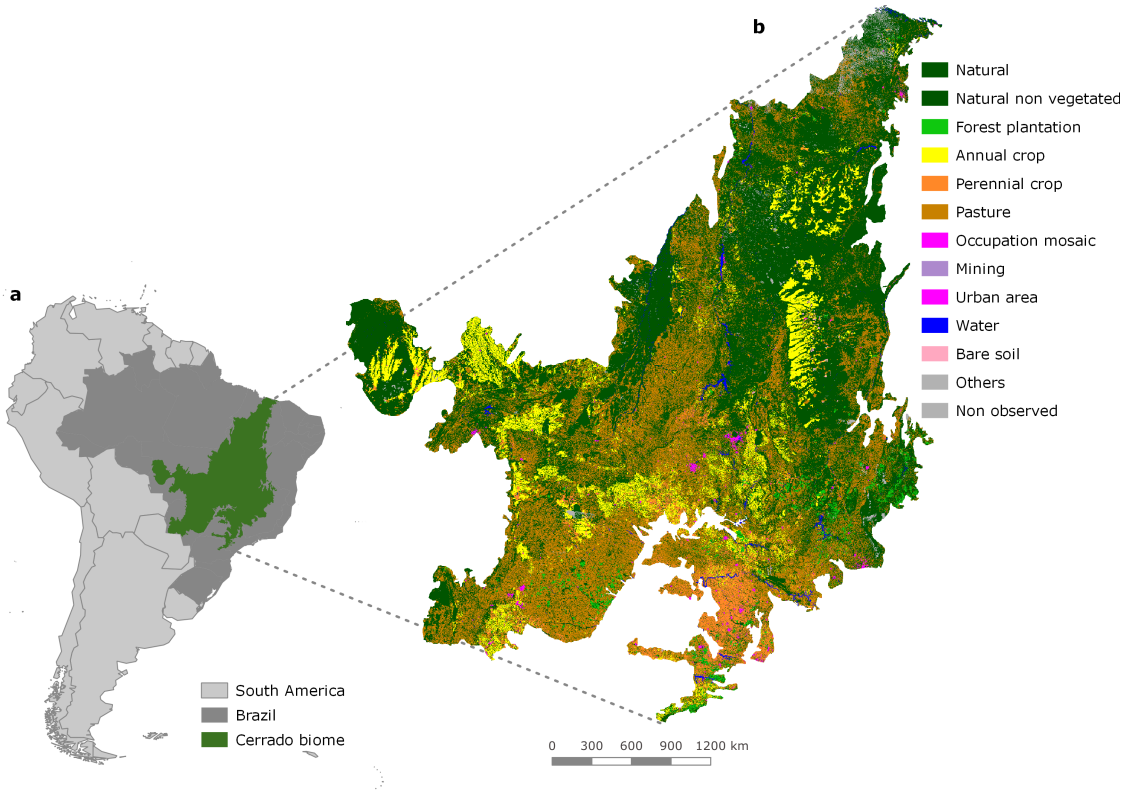
3.2 Material and methods

3.2.1 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 ($\sim$ 22% of Brazil) (see Figure 3.1) (BRASIL. MINISTÉRIO DO MEIO AMBIENTE, 2019). The Cerrado is a global biodiversity hotspot due to 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 (SOTERRONI et al., 2019). Only 8.21% of the Cerrado is legally protected by conservation units (BRASIL. MINISTÉRIO DO MEIO AMBIENTE, 2019), and it is estimated that 88 Mha (46%) of its natural vegetation cover has been lost (STRASSBURG et al., 2017).

Figure 3.1 - **a.** Cerrado location relative to Brazil and South America. **b.** Land use and cover map of the Cerrado. Source: TerraClass (INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS - INPE, 2013).



SOURCE: Adapted from INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS - INPE (2013).

3.2.2 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

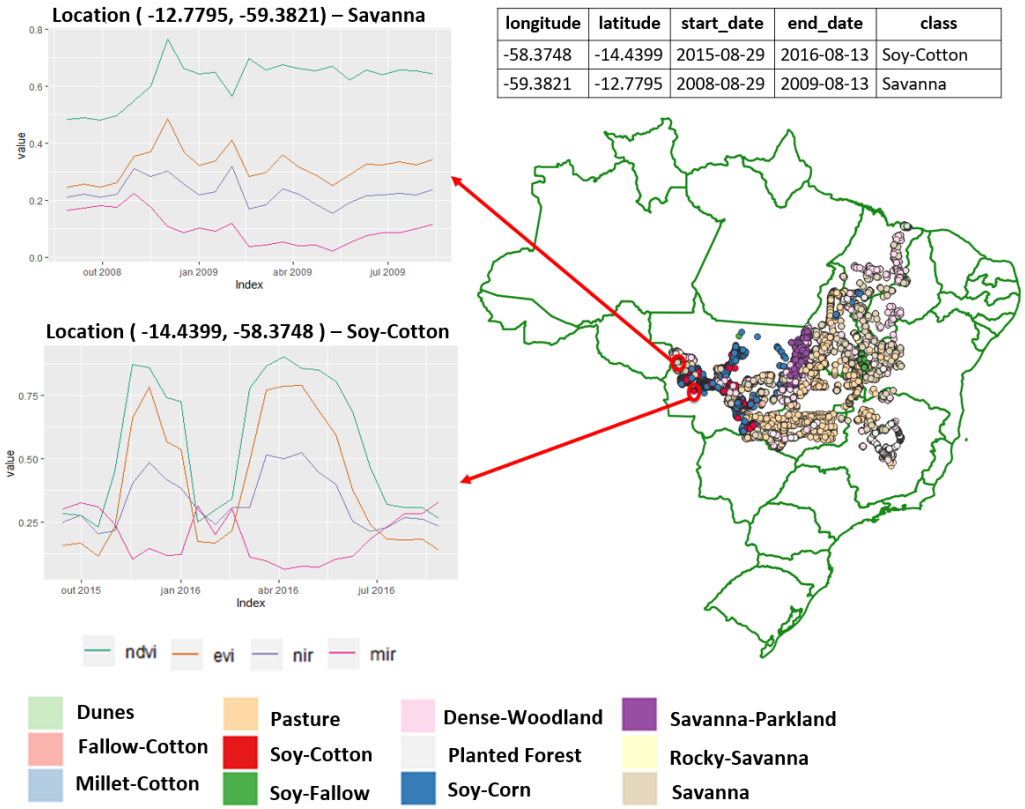
(INPE) team and partners. This set ranges from 2000 to 2017 and includes 50,160 land use and cover samples divided into 12 classes: (1) Dunes, (2) Fallow-Cotton, (3) Millet-Cotton, (4) Soy-Corn, (5) Soy-Cotton, (6) Soy-Fallow, (7) Pasture, (8) Rocky Savanna (in Portuguese *cerrado rupestre*), (9) Savanna, (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; WALTER, 2008) who provide a taxonomy of classes for the biome. The samples number for each class is presented in Table 3.1.

Table 3.1 - 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%

As shown in Figure 3.2, 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). Figure 3.2 illustrates the four satellite image time series, one for each attribute (NDVI, EVI, NIR, and MIR), associated with samples of Savanna and Soy-Cotton classes.

Figure 3.2 - Reference data set.



longitude	latitude	start_date	end_date	class
-58.3748	-14.4399	2015-08-29	2016-08-13	Soy-Cotton
-59.3821	-12.7795	2008-08-29	2009-08-13	Savanna

SOURCE: Author.

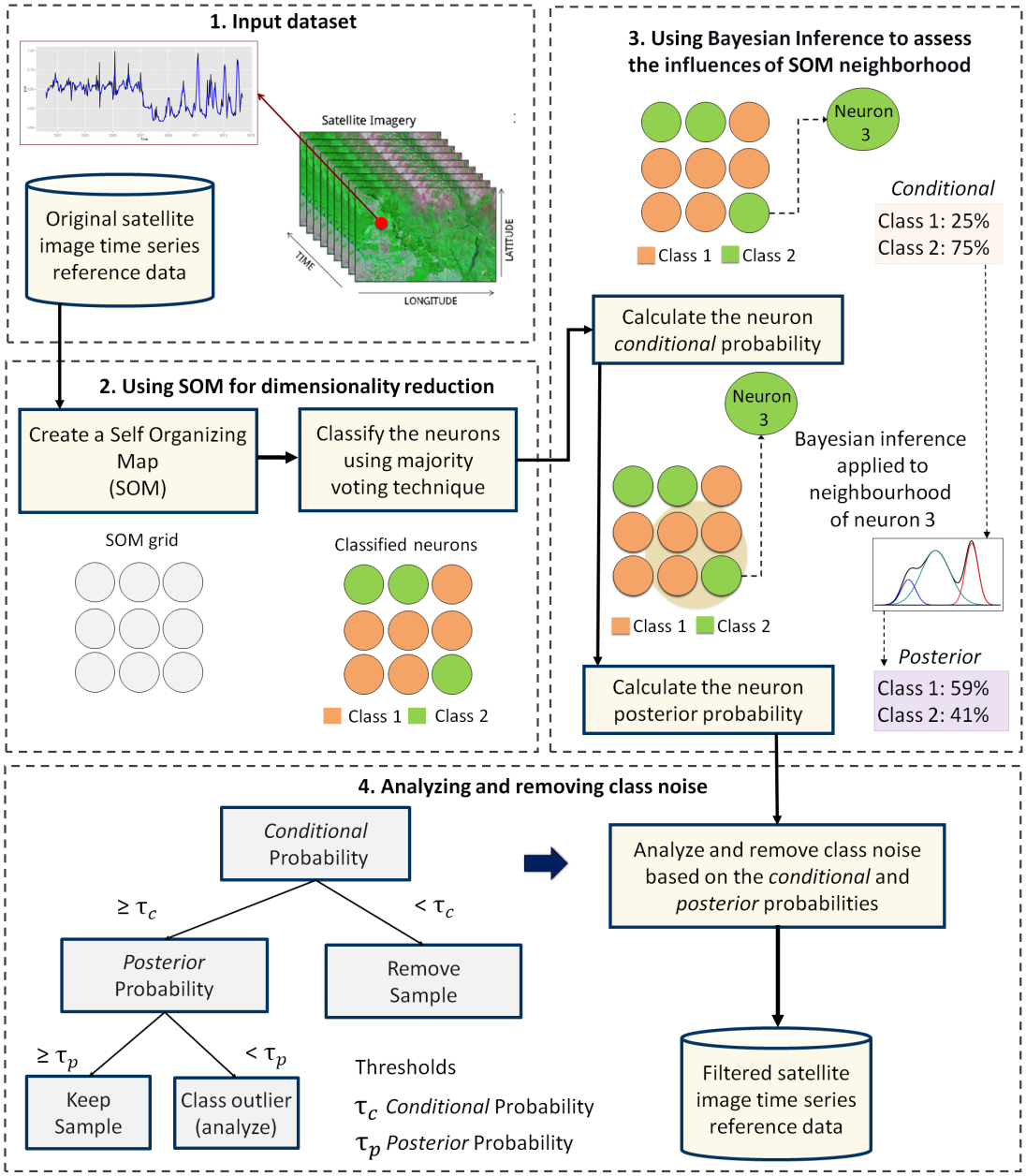
Multi-dimensional time series help in distinguishing the different land classes. The dry season occurs from May to September and the rainy season from October to April in the Cerrado biome. The variability of EVI values during the rainy season helps to distinguish between natural vegetation cover types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

3.2.3 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 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. Although 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 Figure 3.3. The method uses self-organizing maps (SOM) (KOHONEN, 1990) to perform dimensionality reduction while preserving the topology of original datasets. Since SOM preserves neighborhoods' topological structure 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 SOM map neighborhoods to improve the evaluation of each time series sample's quality.

Figure 3.3 - A method for class noise reduction in satellite image time series reference data.



SOURCE: Author.

3.2.4 Using SOM for dimensionality reduction

SOM is an unsupervised neural network that maps a high-dimensional input dataset to a low-dimensional one, usually a 2D grid. As Figure 3.4 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 to preserve their similarity (KOHONEN, 2013). This mapping and adjustment procedure 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 nearby ones. The resulting SOM grid combines dimensionality reduction with topological preservation.

To project a multidimensional set of time series onto a SOM map, each neuron j is associated a random vector of weights $w_j = [w_{j1}, \ldots, w_{jn}]$, with the same length of each time series sample $x(t)_i = [x_{t_1}, \ldots, x_{t_n}]$. 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 work (FERREIRA et al., 2019). We found out that the Euclidean metric provides reliable and robust results. Therefore, our method uses Euclidean distances to find the BMU, d_b, as shown in Equation (3.1) and Equation (3.2).

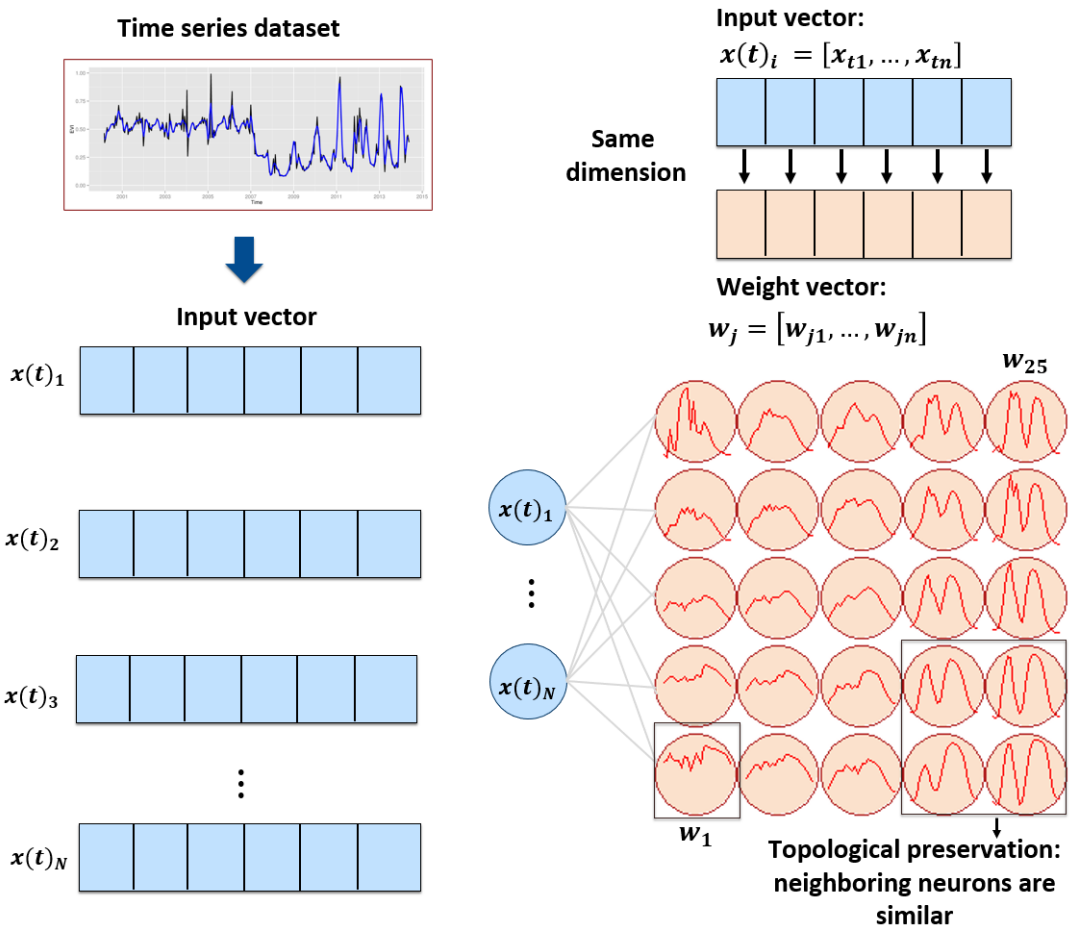
$$D_j = \sum_{i=1}^{n} \sqrt{(x(t)_i - w_j)^2}. \tag{3.1}$$

$$d_b = min\left\{D_1, \ldots, D_j\right\}. \tag{3.2}$$

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 Equation (3.3).

$$w_j = w_j + \alpha \times h_{b,j}[x(t)_i - w_j], \tag{3.3}$$

Figure 3.4 - Self-Organizing Maps structure.



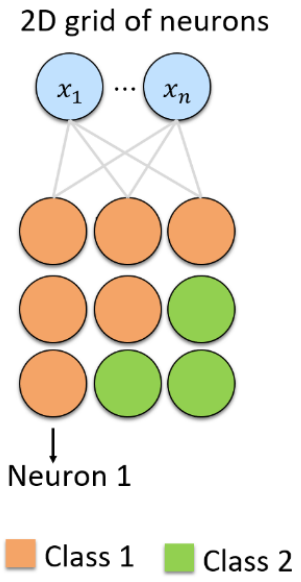
SOURCE: Author.

The parameter α is the learning rate and $h_{b,j}$ 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 $h_{b,j}$ determines which neurons must be updated and the intensity of the readjustment of each one (NATITA et al., 2016).

Our approach uses a unique vector composed of all attributes, including spectral bands and vegetation indices, to represent the input samples. Consequently, the weight vectors are initialized with the same dimension as 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 used to find the BMU.

At the end of the SOM training phase, each time series is associated with a neuron j in a 2D grid. Since each time series class is known, we assign a class to a neuron using majority voting. As an example, Figure 3.5 presents a grid with a set of neurons; the samples associated with 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 Class 2 with 20% probability. These probabilities will be used in the next phase of the method.

Figure 3.5 - Assignment of classes to neurons.



Samples grouped on neuron 1

Id sample	Sample class	Id Neuron
1	Class 1	1
2	Class 1	1
3	Class 1	1
4	Class 1	1
5	Class 2	1

Information about neuron 1

Id neuron	Class 1	Class 2
1	80%	20%

SOURCE: Author.

3.2.5 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 with a discrete probability distribution. 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 noisy samples. Furthermore, we consider that the neuron class probability is not the best measure

of class noise. It represents the prior probability $P(ClassNeuron/ClassSample)$. In fact, what we need is the inverse probability $P(ClassSample/ClassNeuron)$. 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:

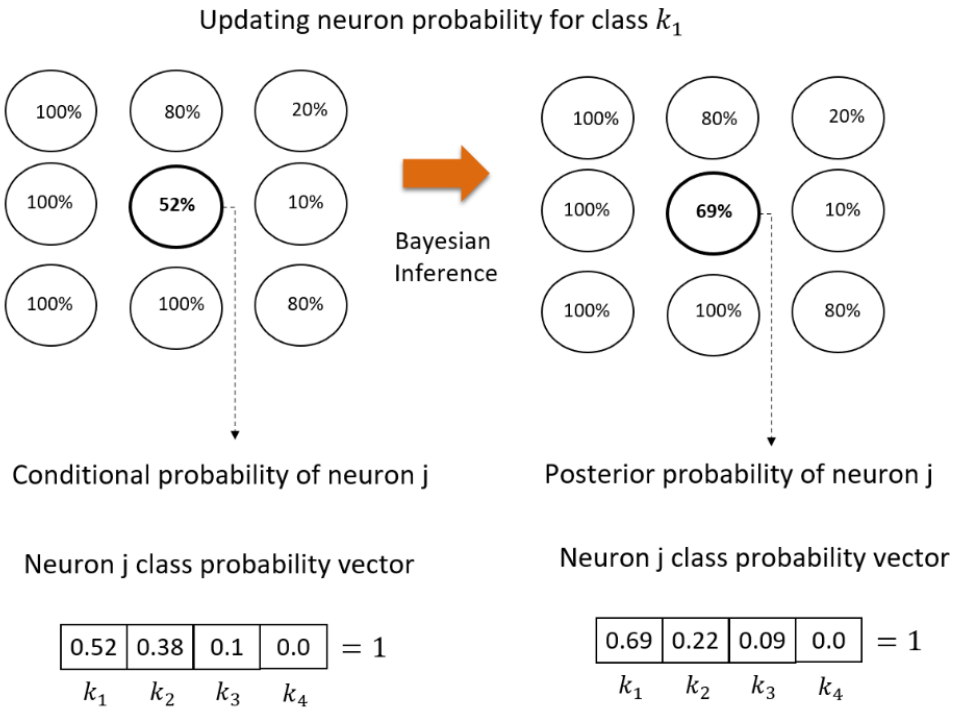
$$f(\theta_{j,k}|y_{j,k}) \propto f(y_{j,k}|\theta_{j,k})f(\theta_{j,k}). \tag{3.4}$$

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 a strategy of "borrowing strength from the neighbors" is commonly used in Bayesian inference (ASSUNCAO et al., 2005).

The approach is illustrated in Figure 3.6, where the neuron j has a prior probability of 52% of belonging to class k_1. Since most of its neighbors have a high probability of belonging to the class k_1, the posterior probability of the neuron j belonging to a class k_1 increases due to the neighborhood effects.

Figure 3.6 - Update neuron j for class k_1.



SOURCE: Author.

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

$$\theta_{j,k} \sim N(m_{j,k}, s^2_{j,k}). \tag{3.5}$$

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

$$m_{j,k} = \frac{\sum_{(i)\in V_j} y_{i,k}}{\#(V_{j,t})}, \tag{3.6}$$

$$s^2_{j,k} = \frac{\sum_{(i)\in V_j} [y_{i,k} - m_{j,k}]^2}{\#(V_j) - 1}. \tag{3.7}$$

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

$$y_{j,k}|\theta_{j,k} \sim N(\theta_{j,k}, \sigma_j^2) \tag{3.8}$$

where σ_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:

$$\mathrm{E}[\theta_{j,k}|y_{j,k}] = \frac{m_{j,k} \times \sigma_j^2 + y_{j,k} \times s_{j,k}^2}{\sigma_j^2 + s_{j,k}^2} \tag{3.9}$$

Rewriting this equation we have:

$$\mathrm{E}[\theta_{j,k}|y_{j,k}] = \left[\frac{s_{j,k}^2}{\sigma_j^2 + s_{j,k}^2}\right] \times y_{j,k} + \left[\frac{\sigma_j^2}{\sigma_j^2 + s_{j,k}^2}\right] \times m_{j,k} \tag{3.10}$$

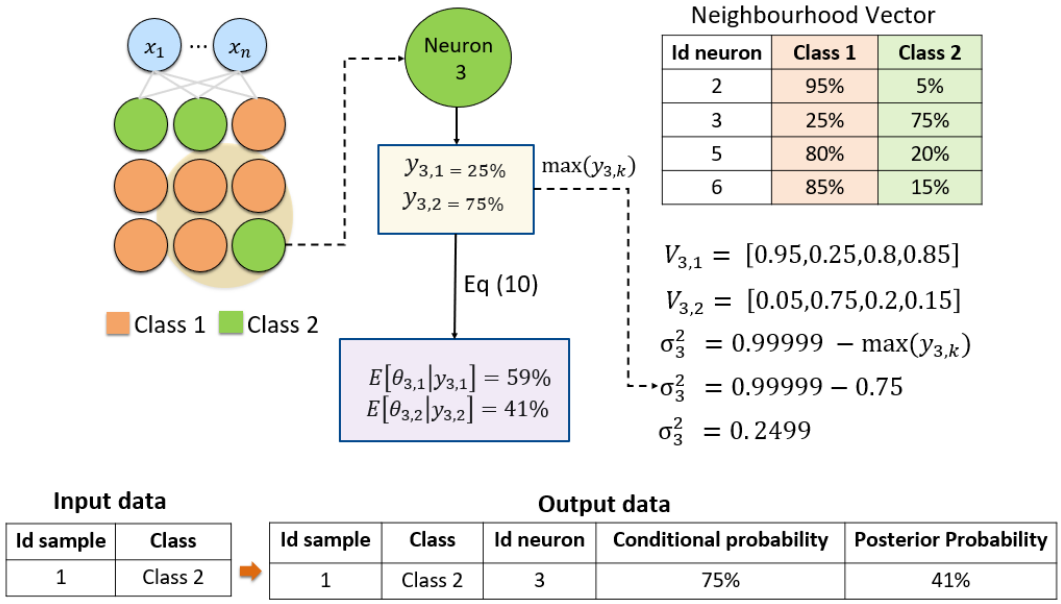
When the neighborhood variance $s_{j,k}^2$ for class k is high, Equation (3.10) gives more weight to the prior probability of $y_{j,k}$. Otherwise, if the neighborhood variance $s_{j,k}^2$ is small, the posterior estimate is controlled by the neighborhood mean $m_{j,k}$. This reflects the intuition that samples in areas of low variance are similar, while they differ in regions of high variance.

The value of the hyper-parameter σ_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 σ_j^2 should decrease. To maintain the σ_j^2 adjusted according to the class variance of neuron j, we define σ_j^2 as:

$$\sigma_j^2 =\mid 0.999999 - \max(y_{j,k}) \mid \tag{3.11}$$

Figure 3.7 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 belong to Class 1. When the Bayesian inference is applied, the probability of this neuron belongs to a Class 2 decreases due to the neighbors' strength. Therefore, sample 1, labeled as Class 2, inherits the neuron's probability belonging to Class 2.

Figure 3.7 - Applying Bayesian Inference in neuron 3.



Id neuron	Class 1	Class 2
2	95%	5%
3	25%	75%
5	80%	20%
6	85%	15%

Input data

Id sample	Class
1	Class 2

Output data

Id sample	Class	Id neuron	Conditional probability	Posterior Probability
1	Class 2	3	75%	41%

SOURCE: Author.

3.2.6 Removing and analyzing 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 neighbourhood. 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 mislabelled 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 includes thresholds τ_c for the prior probability and τ_p for the posterior probability:

a) If the prior probability is $< \tau_c$ then, the sample is removed from dataset;

b) If the prior probability is $\geq \tau_c$ and the posterior probability is $\geq \tau_p$, then, the sample is kept in the dataset;

c) If the prior probability is $\geq \tau_c$ and the posterior probability is $< \tau_p$, the samples will be flagged for further inspection.

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

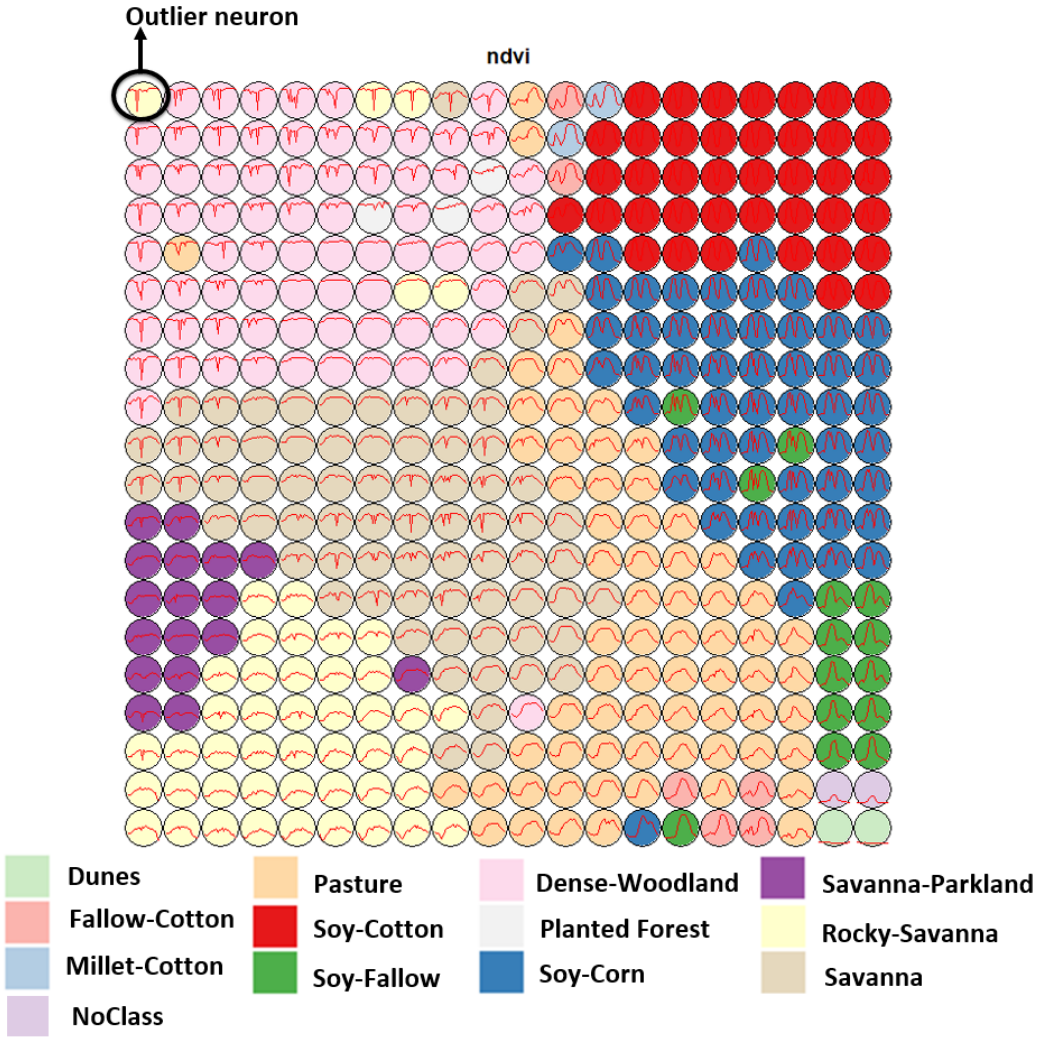
3.3.1 Detecting noisy and outlier samples

In what follows, we show how to apply our method to analyze and improve sample quality. Figure 3.8 shows the SOM grid map generated for 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; 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 twisted maps, and big values to a non ordered map (TAN; GEORGE, 2004). We suggest using a decreasing learning rate, starting at 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 data set 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. We hypothesize that mislabeled samples and outlier neurons are indicators of class noise. Thus, by examining them and identifying incorrect samples, we can improve the training set's quality.

Figure 3.8 - SOM grid.



SOURCE: Author.

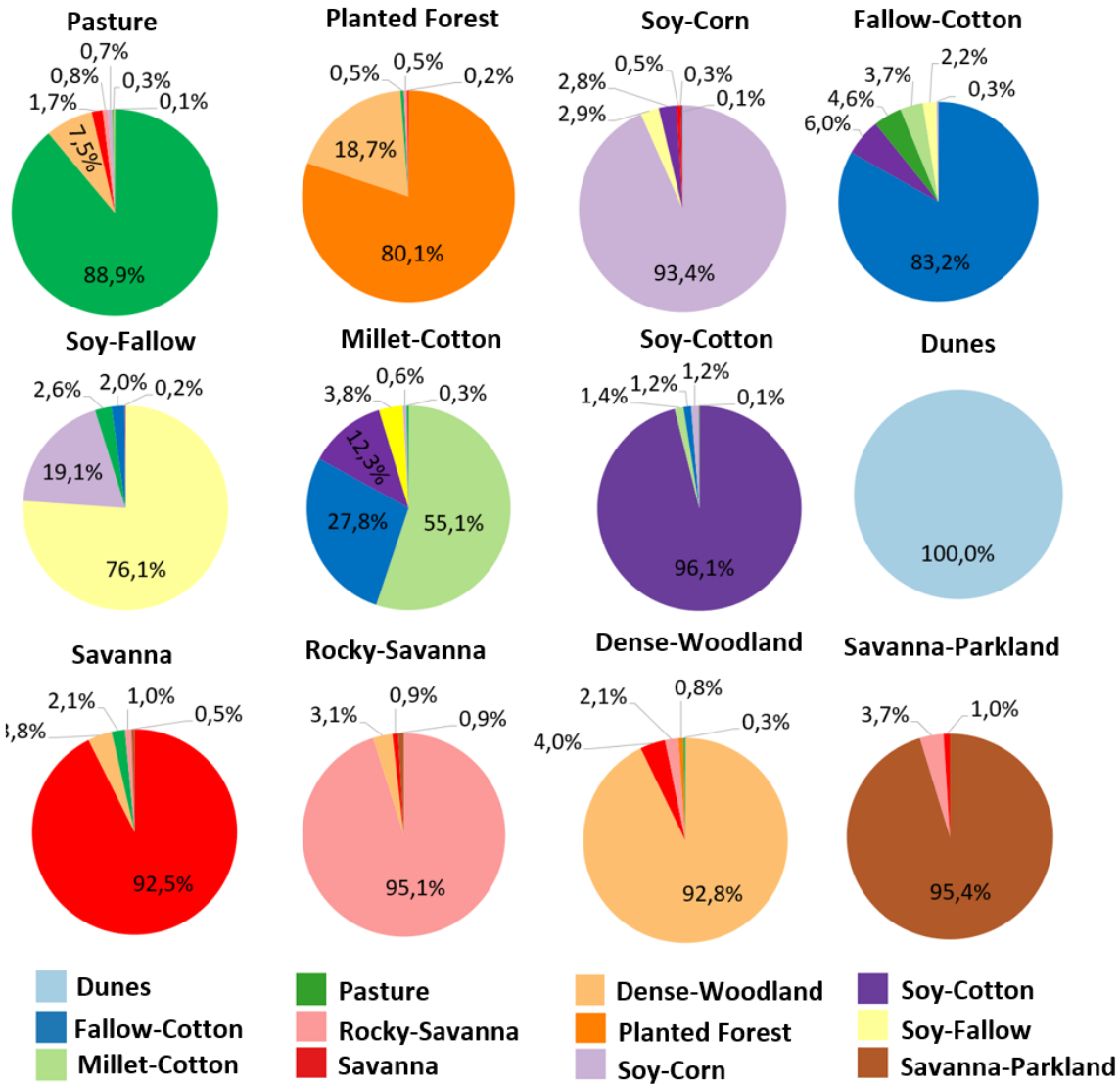
Based on the rules presented in 3.2.6 to identify good samples and class noise, we set the threshold as 60% to both prior and posterior probabilities to decide if each sample is to be kept or removed from the training set. Table 3.2 shows the percentage of samples by class that have been kept, removed, or flagged for analysis.

Table 3.2 - 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%

As shown in Table 3.2, 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 a 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 Figure 3.9. 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.

Figure 3.9 - Confusion between the classes.



SOURCE: Author.

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; 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 occur 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; 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; 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, Figure 3.10a and 3.10b. The values of NDVI, EVI, and NIR for Dense Woodlands samples 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 (Figures 3.10c and 3.10d). 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 (Figure 3.8), where the Dense Woodland neurons are far from Rocky Savanna and Savanna Parkland ones. Therefore, in general, these classes show different time series signals.

Figure 3.10 - Time series of ground samples for natural vegetation classes in the Cerrado Biome.

(a) Savanna (b) Dense-Woodland

(c) Savanna Parkland (d) Rocky Savanna

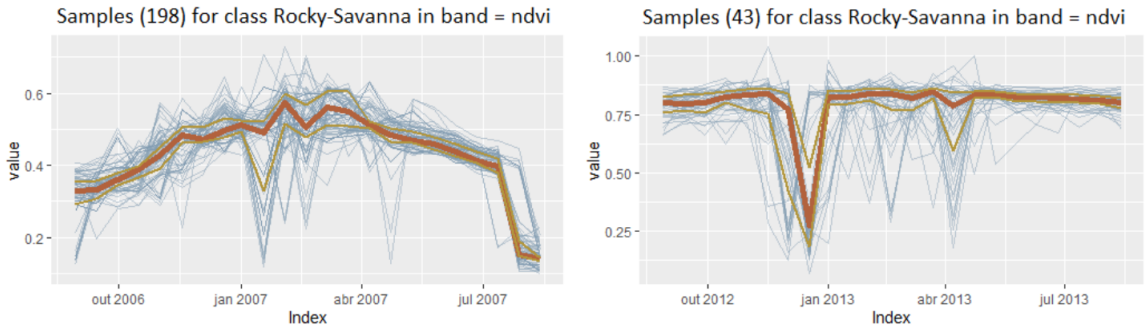
SOURCE: Author.

3.3.2 Identifying mislabelled samples

We now consider how our method helps to identify wrongly labelled samples. Figure 3.11 shows the NDVI signature of two different clusters of Rocky-Savanna class samples, identified in the SOM map. In our assessment, the time series samples presented in Figure 3.11a are consistent with the Rocky Savanna class's expected response. The posterior probability of these samples belonging to the Rocky Savanna class is 100%. By contrast, the samples presented in Figure 3.11b 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 neurons whose label is Dense Woodland. These samples have likely been mislabelled.

Figure 3.11 - NDVI time series samples labeled as Rocky-Savanna.



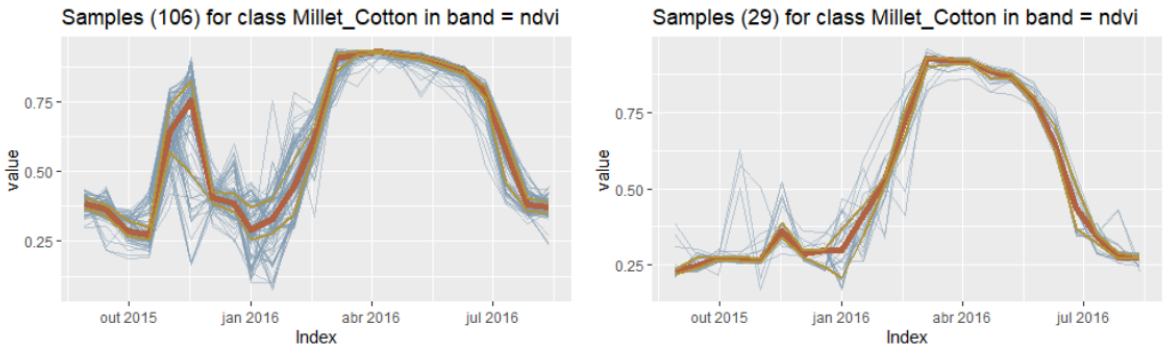
(a) Samples correctly labeled.

(b) Samples wrongly labeled.

SOURCE: Author.

We now consider the sources of confusion between natural vegetation and crops and between crop samples. In general, as seen in Figure 3.9, 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 3.2). Analyzing the SOM clusters, we found many mislabelled samples. Figure 3.12 shows two sets of NDVI values of Millet-Cotton samples. Clearly, samples shown in Figure 3.12a are correct, while those in Figure 3.12b 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.

Figure 3.12 - NDVI time series samples labeled as Millet-Cotton.



(a) Samples correctly labeled.

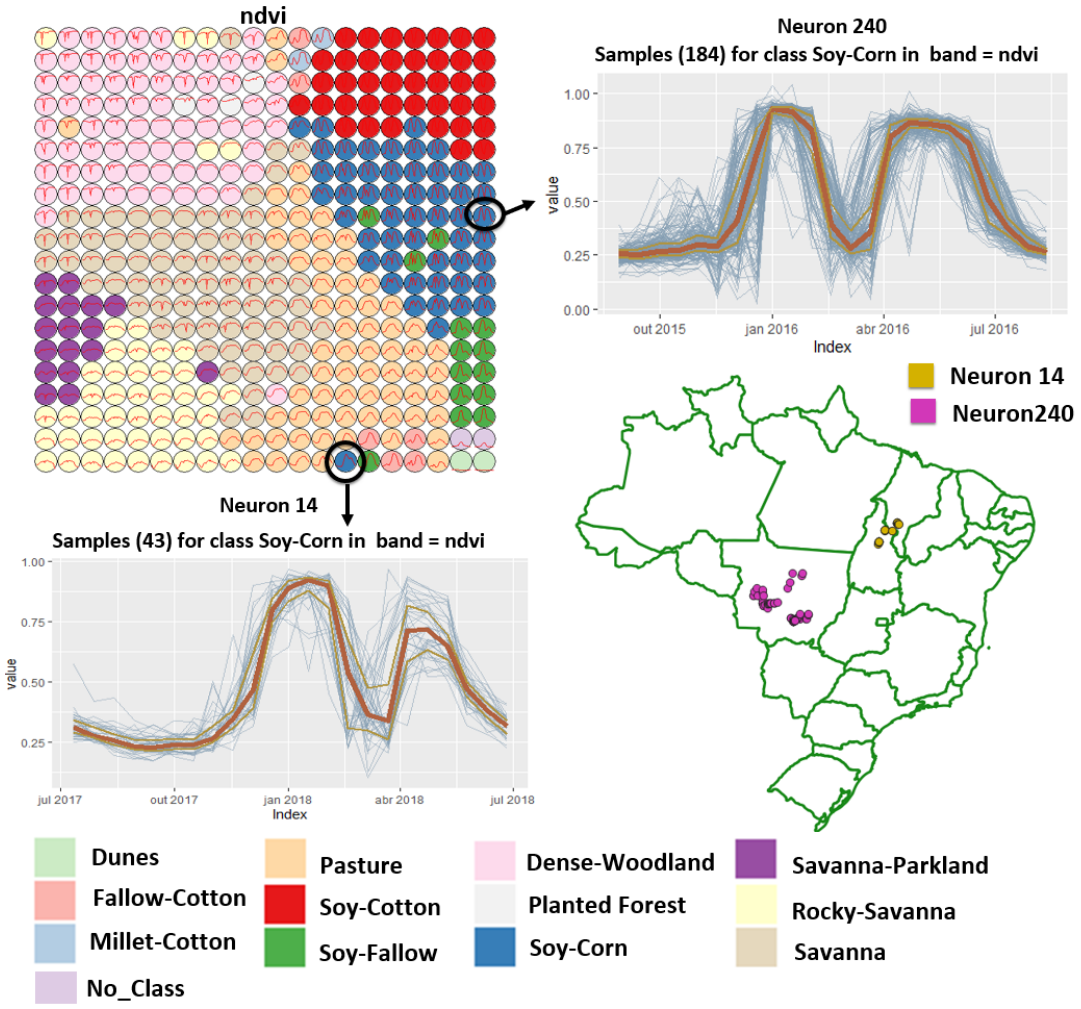
(b) Mislabeled samples of Millet-Cotton

SOURCE: Author.

3.3.3 Outlier analysis

This subsection considers the case of outliers on the SOM maps. Those outliers do not necessarily result from labeling errors but are more likely to arise from the ground data variability. Figure 3.13 shows that most Soy-Corn class patterns are neighbors in the SOM map. However, there is an outlier neuron (Neuron 14) of the same class. Comparing one of the Soy-Corn neighborhood neurons (Neuron 240) with Neuron 14, we find out Neuron 240 has prior and posterior 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 at 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 come from the Brazilian states of Tocantins and Maranhão (about 7S), while the cluster of samples close to Neuron 240 come from the state of Mato Grosso (about 12S). 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. Figure 3.13 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 analyzing outliers in the training set.

Figure 3.13 - Different patterns in the Soy-Corn class because of the agricultural calendar in different regions.



SOURCE: Author.

Table 3.3 presents the percentage of samples by class that was 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 have 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 3.2. 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.

Table 3.3 - 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%

3.3.4 Validation

We did a 5-fold cross-validation test to assess our proposed methods, comparing the original training set (50,160 samples) with the filtered set (44,040 samples). We used a Random Forest algorithm due to its robustness, and proven results on handling big data (BELGIU; 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 3.4 presents producer's and user's accuracy for both datasets. The 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.

Table 3.4 - Producer's and user's accuracy for original and filtered datasets.

	Producer's Accuracy		User's Accuracy	
Classes	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 %

3.4 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 in the accuracy of the resulting maps, the literature on detecting and removing class noise in SITS training sets is limited. To contribute to solving this challenge, this chapter 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 SOM map; these neurons are then labeled with their majority class. Using the SOM property of topological preservation property, the algorithm uses Bayesian inference to evaluate the neuron neighborhoods' classes. As a result, the method identifies both mislabeled samples and outliers that are flagged to further investigation.

The proposed refinement process of SITS training data improves the accuracy of the classification results. In the case study described in this chapter, the mislabeled samples and part of the outliers identified by the proposed method were removed from the training set. Two classifications were then 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 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 samples over large areas.

Despite the usefulness of our proposed method, organizing a good quality training data set 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 an in-depth local understanding of ecosystem behavior. Despite recent progress in machine learning, local knowledge continues to be irreplaceable when using remote sensing data for land use and cover classification.

Although we present a case study in land use and cover classification in this chapter, 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. Figure 3.3 describes, in general, the method steps for class noise reduction in time series reference data.

4 IDENTIFYING SPATIOTEMPORAL PATTERNS IN LAND USE AND COVER SAMPLES OF SATELLITE IMAGE TIME SERIES USING CLUSTERING METHODS[5]

The large amount of remote sensing images freely available nowadays with improved temporal and spatial resolution brings new opportunities for land use and cover mapping over large areas (GOMEZ et al., 2016). Many authors propose a paradigm shift, where change detection is replaced with continuous monitoring (WOODCOCK et al., 2020). To achieve this goal, researchers need access to satellite image time series to detect complex underlying processes (PASQUARELLA et al., 2016).

The use of machine learning methods has been the preferred approach for satellite image time series analysis to map land use and cover changes (LUCC) (GOMEZ et al., 2016). These methods are supervised approaches that require a training phase using samples labeled *a priori*. Comparative analysis of different machine learning methods shows that the quality of training samples has a large impact on classification accuracy (MILLARD; RICHARDSON, 2015; GOMEZ et al., 2016; PELLETIER et al., 2017; MAXWELL et al., 2018). These results motivate our work, which aims to answer the following question: How can training samples be assessed to improve the accuracy of LUCC maps produced by machine learning classification methods that use them?

Different approaches have been proposed to improve the quality of training samples (PENGRA et al., 2020). Such strategies include best practices in collecting training data (OLOFSSON et al., 2014; ELMES et al., 2020; PENGRA et al., 2020; HUANG et al., 2020) and refinement methods of samples using satellite image time series (VIANA et al., 2019; SIMOES et al., 2020; BELGIU et al., 2021). Other approaches such as semi-supervised learning and active learning have been applied to support training sample acquisition (DEMIR et al., 2010; TUIA et al., 2011; HUANG et al., 2015; LU et al., 2017; SOLANO-CORREA et al., 2019). However, most studies are limited to small areas and inter-annual extents (PENGRA et al., 2020), not being suitable for large areas due to the need for a large number of samples to characterize each class (RADOUX et al., 2014).

In large areas, the variability of land use and cover classes is high and intrinsic to different regions and periods due to heterogeneous biodiversity as well as distinct climatic conditions and management practices (HOSTERT et al., 2015; COMBER;

[5]This chapter is based on the paper: Santos, L.A.; Ferreira, K.; Picoli, M.; Camara, G.; Zurita-Milla, R.; Augustijn, E.-W. Identifying Spatiotemporal Patterns in Land Use and Cover Samples from Satellite Image Time Series. Remote Sens. 2021, 13, 974.

WULDER, 2019; ALENCAR et al., 2020; MERONI et al., 2021). Therefore, it is essential to explore ways to obtain samples that properly represent high intra-class variability considering spatiotemporal variations in large areas and multiple years.

In many situations, experts use generic labels for training samples (e.g., "forest", "cropland", and "grassland"). In practice, the actual spatiotemporal variability of the time series data does not match such generic labels. For this reason, it is useful to distinguish subclasses of high-level labels that correspond to regions of separability on the attribute space. Based on this, this paper presents a method to identify subclasses in training samples of satellite image time series. The method distinguishes different types of land use and cover classes over large areas in a more detailed granularity than user-provided labels. Using phenological and spectral information provided by satellite images time series, the method refines the generic labels and improves the accuracy of resulting LUCC maps.

The proposed method is based on time series clustering. In general, time series clustering has been applied for exploratory analysis (LIAO, 2005; AGHABOZORGI et al., 2015; PAPARRIZOS; GRAVANO, 2015), characterization of spatiotemporal patterns (BIRANT; KUT, 2007; ANDRIENKO et al., 2010; AUGUSTIJN; ZURITA-MILLA, 2013; LIU et al., 2018; QI et al., 2019), and to soften the lack of discriminated data of land use and cover types (XIONG et al., 2017; SOLANO-CORREA et al., 2019; WANG et al., 2019). It is a promising approach to exploit spectral and phenological information and refine training samples to ensure an acceptable level of quality (VIANA et al., 2019; SOLANO-CORREA et al., 2019; BELGIU et al., 2021; ALENCAR et al., 2020). Spectral and phenological information can be considered to discriminate different types of land use and cover classes during the sample collecting and labeling, contributing to improving the quality of training data sets. However, time series clustering results can be difficult to interpret and visualize, especially when the training data have a high dimension (HALLAC et al., 2017).

Our algorithm uses self-organizing maps (SOMs) (KOHONEN, 1990) combined with a hierarchical algorithm (KAUFMAN, 1990; EVERITT et al., 2011). SOMs have been applied in spatiotemporal data analysis (ZURITA-MILLA et al., 2012; AUGUSTIJN; ZURITA-MILLA, 2013; CHEN et al., 2018; QI et al., 2019; GUO et al., 2006; ASTEL et al., 2007; ANDRIENKO et al., 2010; LIU; WEISBERG, 2011) mainly due to two properties. SOMs map high-dimensional data into a low bi-dimensional grid, representing the input data in low dimension. They also preserve neighborhood information; similar patterns in attribute space tend to stay close in the 2D SOM space. After mapping

the samples from the high-dimensional attribute space to the 2D SOM space, we apply a hierarchical algorithm to the SOM clusters. The resulting subclusters refine the original training data into subclasses. These subclasses have lower intra-class variability and higher inter-class variability than the original SOM clusters. Given the SOM proprieties, they provide refinement of the original training samples.

As a proof of concept, we present a case study using Moderate Resolution Imaging Spectroradiometer (MODIS) image time series of 7 years (2010–2017) associated with samples of cropland and pasture classes over the Cerrado biome in Brazil. The results show that the proposed method is suitable for identifying spatiotemporal patterns in land use and cover samples that can be used to infer subclasses mainly for crop-types.

4.1 Material and methods

4.1.1 Data

The case study presented in this paper uses samples of the Cerrado biome in Brazil. Cerrado is a large and dynamic landscape with an area of approximately 204 million hectares (Mha), covering almost 22% of the central area in Brazil. It is one of the largest and most diverse tropical savannas in the world (DICKIE et al., 2016; SOTERRONI et al., 2019). However, half of the biome has been changed and deforested due to advanced agricultural production and livestock (KLINK; MACHADO, 2005).

The dataset is a merge of samples collected by remote sensing specialists through visual interpretation of high-resolution images, samples collected by the specialist in field observations, and farmer interviews provided by the Brazilian National Institute for Space Research (INPE) team partners.

Figure 4.1 illustrates the dataset used in our case study. The dataset includes 15,794 samples spread over the Cerrado biome from 2010 to 2017 and is divided into two classes: (1) cropland and (2) pasture. Most of the cropland samples are in the same locations associated with different years.

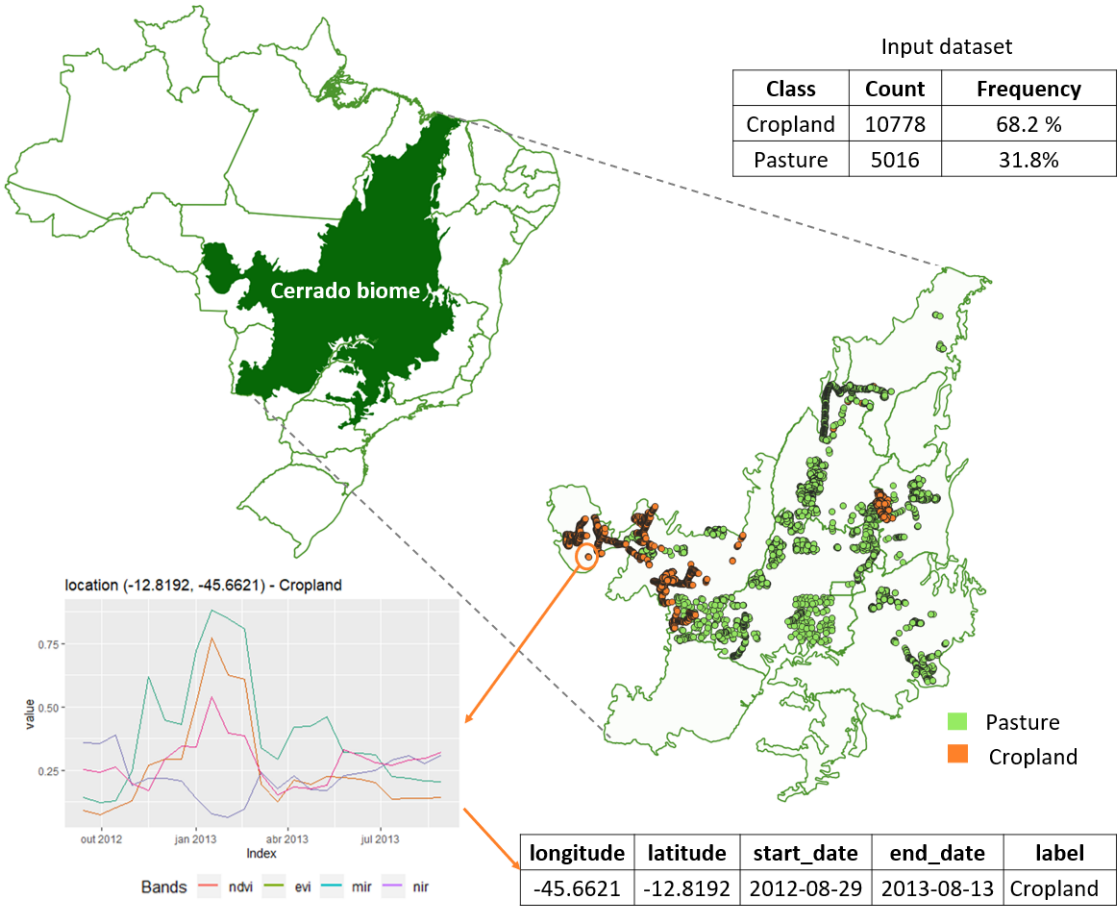
Each sample has a spatial location, a period containing the start date and end date according to an agricultural calendar (from August to September), a label describing the sample class, and the satellite image time series for each attribute (bands and vegetation indexes) associated with it. The time series were extracted from the product MOD13Q1 (Collection 6) of the MODIS sensor. The images were collected on an interval of 16 days with 250 m spatial resolution. For this dataset, we

used two vegetation indices and two bands available in MOD13Q1: the Normalized Difference Vegetation Index (NDVI) and the Enhanced Vegetation Index (EVI), and near-infrared (NIR) and and mid-infrared (MIR).

The MOD13Q1 product is created by considering the best available pixel from all acquisitions of MODIS images in a 16-day period. This approach selects the best pixel available in 16 days, avoiding cloud, shadow, or low-quality pixels. We obtained the time series from the MOD13Q1 product without performing further processing.

The Cerrado biome has distinct soil and weather over its area. Thus, it has great and heterogeneous types of land use and cover classes. However, it is not trivial to separate these types using 250 m and 16-day MOD13Q1 data. For this reason, we decided to use only the cropland and pasture classes to show how the proposed method works.

Figure 4.1 - Land use samples dataset of Cerrado biome.



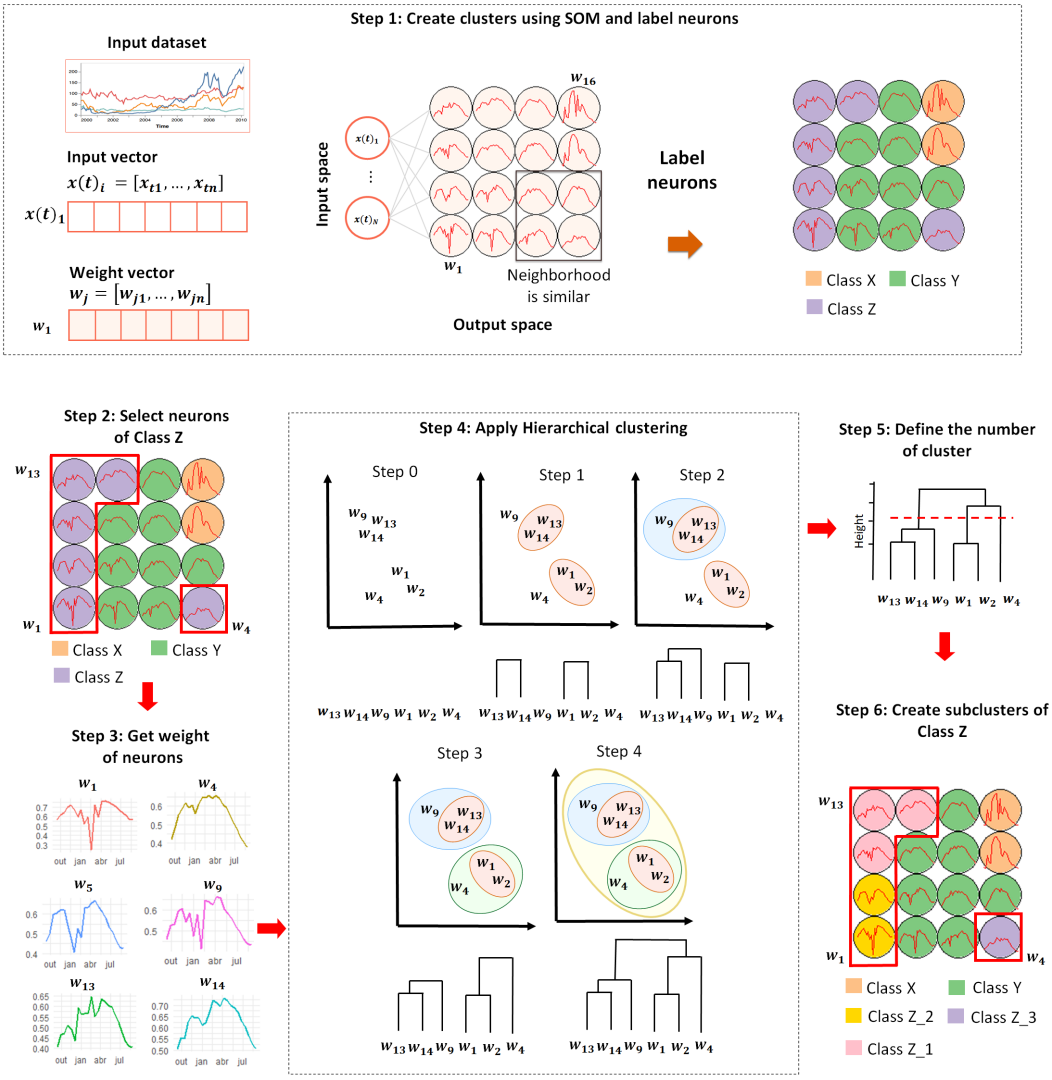
SOURCE: Author.

4.1.2 Overview

Clustering methods are useful to extract spatiotemporal information from satellite image time series samples (LIAO, 2005). There are methods of clustering that use the approach of spatial clustering objects and spatial and temporal objects. Although our time series have a geo-location, we used only the time series points as a similarity measure to identify the clusters (ANSARI et al., 2019; WU et al., 2020). Then, the geo-localization and period of each time series were used to explore the spatiotemporal patterns.

Figure 4.2 illustrates the proposed method to identify and analyze spatiotemporal patterns. Given a dataset of satellite image time series samples, we applied SOMs to group similar time series. SOMs map the input data (high dimensionality) onto a bi-dimensional grid (low dimensionality), keeping the topology of the neighborhood; that is, a similar time series tends to be close in 2D space (KOHONEN, 1990). The SOM grid contains units called neurons, where each neuron has a weight vector associated with it. At the end of the SOM process, each time series from the input dataset was assigned to a neuron. In this way, from the output of SOM, the clusters can be recognized and created. Once the class label of each sample was known, each neuron was labeled using the majority technique. To explore each cluster, the neurons with the same category were selected. The hierarchical method (JOHNSON, 1967) was then applied to the weight vectors of these neurons to extract subclusters. Due to the high intra-class variability, the hierarchical method, as secondary clustering, was applied to the neurons' weight vectors with the same category. The use of hierarchical clustering allows visualization through dendrogram, which can be cut in a specific height to define the number of subclusters. Our method suggests the number of the cluster through an inter-cluster validation index. However, specialists can apply different numbers of groups and explore them according to their necessities. The subclusters were grouped through the time series similarity. However, since all samples have the spatial location and the period containing the start date and end date, these subclusters can indicate patterns in space or time.

Figure 4.2 - The method for exploratory analysis using time series is based on clustering methods. In step 1, the clusters are created using SOM. In step 2, the neurons labeled as the same category are selected. In step 3, the weight vectors are extracted from selected neurons. In step 4, the hierarchical clustering is applied to weight vectors. In step 5, the number of sub-clusters for each category is defined. In step 6, the subclusters are created.



SOURCE: Author.

SOM is an unsupervised learning method based on competitive learning that reduces a high-dimensional feature input space onto a lower-dimensional feature output space. A large dataset can be mapped and represented by a set of neurons through the

weight vectors. An important characteristic of SOM is preserving the neighborhood topology; thus, similar input data are mapped to the same neuron or a nearby one.

Each neuron j of the output space has a weight vector $w_j = [w_{j1}, \ldots, w_{jn}]$ containing the same dimension n of the input data $x(t) = [x(t)_1, \ldots, x(t)_n]$. The weight vectors can be initialized randomly or according to some heuristic. The algorithm has two main steps. First, the distances D_j between a sample and each neuron of the SOM grid is computed. The neuron containing the smallest value of the distance d_b is selected as the best matching unit (BMU) for this sample. The equations to compute the distance and BMU are given by

$$D_j = \sum_{i=1}^{n} \sqrt{(x(t)_i - w_{ji})^2}. \tag{4.1}$$

$$d_b = min\left\{D_1, \ldots, D_j\right\}. \tag{4.2}$$

The second step is the adjustment of the weight vector of the BMU and its neighbors so that the neurons have similar characteristics. The equation of adjustment is given by

$$w_{ji} = w_{ji} + \alpha \times h_{b,j}[x(t)_i - w_{ji}]. \tag{4.3}$$

where the parameters α is the learning rate and $h_{b,j}$ is the neighborhood function. The learning must be set as $0 < \alpha < 1$, and it controls the change level of the BMU and its neighbors during the training step. The neighborhood function limits the size of the BMU's neighborhood that must be updated. The diversity of neighborhood functions can be seen in (KOHONEN, 1990; NATITA et al., 2016).

These steps are performed T times for neurons to organize themselves to have a similar neighborhood between them. Then, each sample from the dataset is assigned to a neuron.

The use of a secondary clustering directly on the SOM grid is not well suited in our case. It can generate confusion between the classes because of the high variability of patterns in a specific class; therefore, the neurons were labeled previously. For this reason, we labeled the neurons before splitting the clusters through hierarchical clustering. According to Kohonen (2013), when an input dataset has a specific number of classes, we can assume that each neuron belongs to one of these classes. The

neurons were categorized according to the majority class label that occurs in each one. In cases of the tiebreaker, the neuron received the label of the majority neighborhood.

Kohonen et al. (2000) argue that when the neurons are labeled according to the majority, the quality of the organization of the final map can be measured through the external criteria of clustering validation. In our approach, we consider a cluster as a set of neurons labeled with the same category. We use purity as the external criteria of cluster validation. Purity assesses the clusters according to the number of the most representative class assigned to them.

4.1.3 Hierarchical Clustering

Once all the neurons were labeled, we applied the hierarchical clustering on the weight vectors of neurons with the same category. This aids in extracting useful information, analyzing the spatiotemporal dynamics, and avoiding confusion between different classes' samples.

Hierarchical clustering is a method where the data are partitioned successively, building a hierarchy of clusters (EVERITT et al., 2011). This type of representation facilitates the visualization in each step where the level similarity occurs. There are two types of hierarchical clustering, agglomerative and divisive. In the divisive algorithm, the entire dataset starts in one cluster, and then it is split into two more similar clusters. In the agglomerative method, as illustrated in Figure 4.2 (step 4), each weight vector $(w_2, w_3, w_4, w_6, w_7, w_8)$ starts in its own cluster, a similarity matrix is computed, and the two most similar groups are identified. At each step, the clusters are merged, and the hierarchy is built based on linkage criteria.

The linkage criteria present the distance measures between the clusters. There are several linkage criteria proposed in the literature. The most common methods are single, complete, and average linkage (KAUFMAN, 1990). This paper uses the Euclidean distance and the average linkage to build the hierarchical clustering. The average linkage computes the mean distance from each element of a cluster to all the elements of the other cluster (KAUFMAN, 1990; EVERITT et al., 2011).

The hierarchical algorithms build a binary tree called a dendrogram. This structure represents the order of how the clusters were merged. A dendrogram divides the data into an internally homogeneous group. Through the hierarchy of the tree, it is possible to visualize the variability of the data. The dendrogram aims to explore and define the suitable number of clusters according to the analysis level. Figure 4.2 (step

5) illustrates a dendrogram and how the cluster can be defined. The dendrogram was cut at the height where three clusters were created. In our method, the internal cluster validity, the C-Index (HUBERT; LEVIN, 1976) criteria, were applied to define the number of cluster for each class from the dendrogram. It is defined by:

$$CIndex = \frac{Sd - Sd_{min}}{Sd_{max} - Sd_{min}} \tag{4.4}$$

where Sd is the summed distance between the neurons within the same cluster, Sd_{min} is the sum of the smallest distance between the pair of objects within the same cluster, and Sd_{max} is the sum of the largest distance. Finally, the new subclusters of class Z can be created, as shown in Figure 4.2 (step 6). The weights w_3, w_4 and w_5 were merged in cluster 1, w_6 and w_7 in cluster 2, and w_8 in cluster 3.

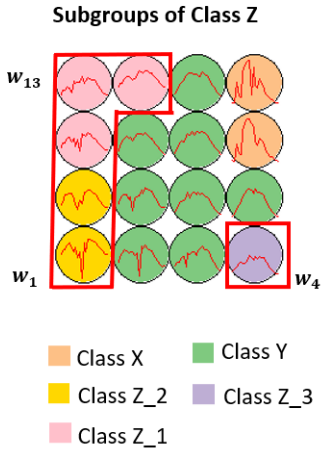
4.1.4 Clustering Output

After the clustering process, an interpretation and comparison of the clusters are necessary to identify the data's features. Generally, the amount of data used in remote sensing analysis is large; therefore, interpreting and analyzing the data in an aggregated way can facilitate the analysis. In general, our method performs the computational processing that allows the summarization of the input data to facilitate searching for knowledge, information, and pattern discovery. Besides the spatial location and time available in each dataset sample, the clustering result provides information that is useful in further analysis. After obtaining the identifiers for each subcluster and the information of the neurons associated with it, a specialist can identify, in an easy way, patterns that need to be interpreted. Moreover, to facilitate the analyses, interactive visual tools can process the output information generated by the clustering methods and jointly with the input data that contain the spatial location and period, providing spatial and temporal information.

Figure 4.3 illustrates an example of output after the clustering methods. For each sample (presented in Section 4.1.1), the result of clustering methods provides identifiers of neurons and subclusters and their labels generated by SOM and hierarchical clustering, respectively. Three samples of class Z were assigned to different neurons. However, these neurons were labeled as class Z because of the majority class. Although these samples belong to the same category's neurons, we can notice in neurons 1, 4, and 13 distinct patterns through the weight vector. When the hierarchical clustering was applied to the weight vectors of class Z, these samples were assigned to three different subclusters. Furthermore, through the SOM grid neighborhood, we notice

that neuron 4 is an outlier within the neurons of the subcluster of class Z. However, it is necessary to understand whether this neuron is an error or represents a different pattern, e.g., in space or time, within the class Z.

Figure 4.3 - Clustering output. The red lines in the SOM grid represent the subgroups that were generated from the neurons labeled as Class Z. For each sample of the dataset, an id and label of the neuron, and an id and label of subgroups are assigned to each one.



Input data

Id sample	longitude	latitude	start_date	end_date	label
1	-58.3748	-14.4399	2015-08-29	2016-08-13	Class Z
2	-59.3821	-12.7795	2008-08-29	2009-08-13	Class Z
3	-48.1653	-10.8569	2010-08-29	2011-08-13	Class Z

Clustering Methods

Output data

Id sample	Id neuron	Neuron label	Id subgroup	Subgroup Label
1	1	Class Z	2	Class Z_2
2	4	Class Z	2	Class Z_3
3	13	Class Z	1	Class Z_1

SOURCE: Author.

To evaluate the relative performance of the original and refined training data sets, we used 5-fold cross-validation (ARLOT et al., 2010). We split the training dataset into training and test sample sets to avoid overfitting and biased data (BENGIO; GRANDVALET, 2004; ARLOT et al., 2010). Using the 5-fold approach, we estimated the classification accuracy (overall, producer, and user accuracy). We chose the random forest classifier due to its robustness on land use and cover mapping (BELGIU; DRAGUT, 2016; PELLETIER et al., 2017) to evaluate the training datasets generated from the cluster analysis.

4.2 Results

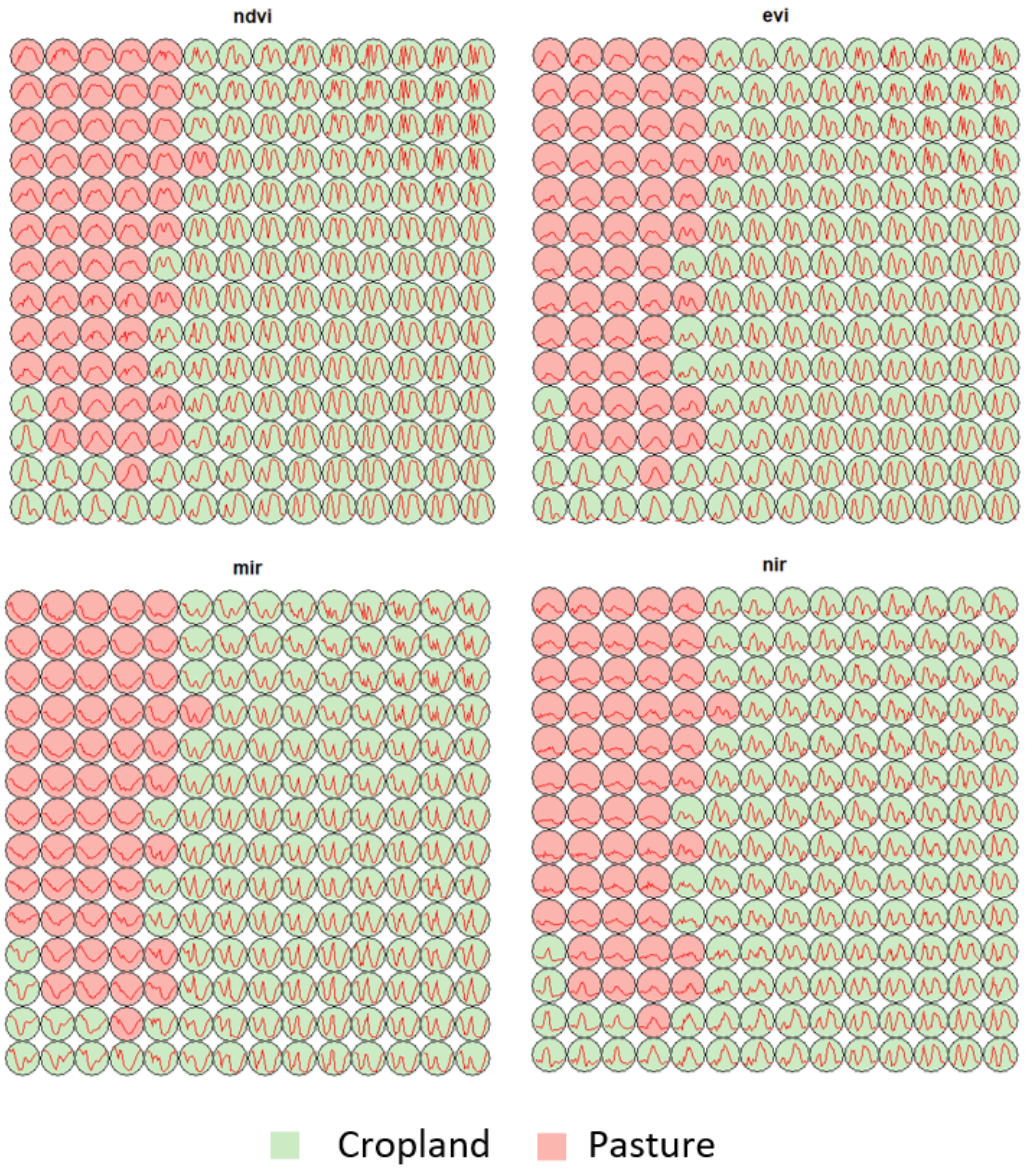
This section presents a case study to show how the method presented in this paper works.

4.2.1 Creating Clusters Using SOM

Figure 4.4 presents the SOM results for the vegetation indices and spectral bands used in the training step. The parameters used in SOM to generate the maps were Euclidean distance as similarity metric, grid size $= 14 \times 14$, learning rate initial $= 0.5$ and final $= 0.1$, number of iteration $= 100$ and the neurons were labeled according to majority technique voting. The vegetation indices and spectral bands were chosen based on the study conducted by Santos et al. (2019). While the input parameters of SOM were defined based on empirical experiments, the start points, such as the size grid and learning rate, were suggested by Vesanto and Alhoniemi (2000).

Figure 4.4 illustrates the primary clustering generated by SOM. The 15,794 samples, shown in Section 4.1.1, are represented by a set of 196 neurons, where 139 neurons represent the cluster of cropland, and the cluster of pasture is represented by 57 neurons. Only by analyzing the grid generated by SOM do we notice the high variability of the patterns, mainly in cropland neurons. Moreover, the single and double cropping groups can be identified in the cropland neurons, and some neurons of pasture are considerably similar to neurons of cropland with single cropping types. The similarity between these neurons can be investigated by an expert in more detail considering the spatial and temporal information provided by the samples. Some hypotheses can also be created, indicating whether these samples are separable or not due to noise, i.e., clouds, type of sensor used (spatial resolution, temporal resolution), or mislabeled samples.

Figure 4.4 - SOM grid. Each line inside the neurons is a weight vector generated by SOM to represent a set of sample in low dimensional space.

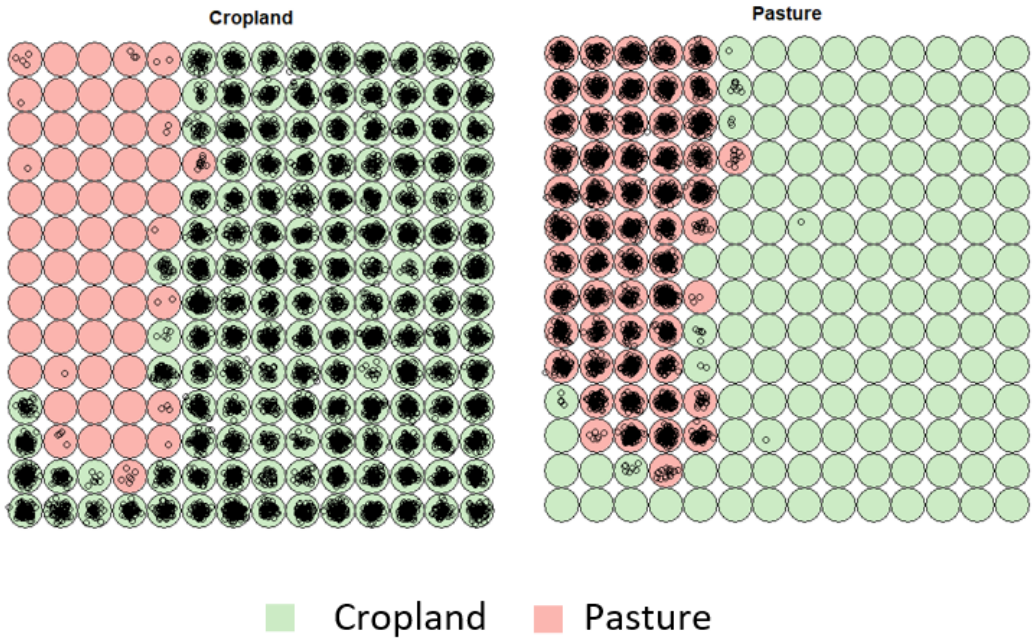


SOURCE: Author.

Figure 4.5 shows how the samples are spread over the SOM grid, identifying whether they are nearest or farthest of their neighborhood. Although the neurons are labeled as a specific class, they are not 100% pure. Samples of different types were associated with neurons for which the majority belong to other classes. Overall, the purity for

cropland and pasture is, respectively, 99.6% and 99.4%. Most of the confusion occurs near the class boundary. We can identify where these samples were mapped through the SOM grid. For instance, the cropland neurons with single cropping patterns are neighbors of the pasture neurons. In addition, the grids show precisely where the cropland samples were mapped in neurons labeled as pasture and vice-versa.

Figure 4.5 - Mapping samples in SOM grid. Each dot represents a sample.

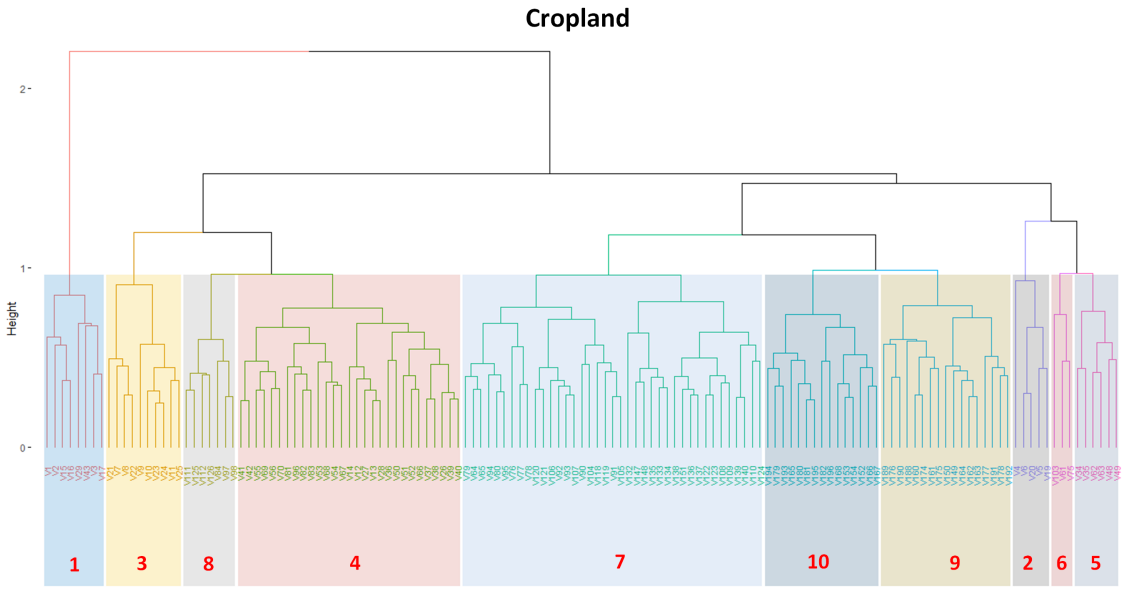


SOURCE: Author.

4.2.2 Revealing the Patterns of Cropland

The hierarchical clustering was applied to 139 weight vectors of cropland. Then, the C-Index suggested 10 clusters for cropland, as illustrated in Figure 4.6.

Figure 4.6 - Dendrogram partitioned into ten groups for Cropland.



SOURCE: Author.

Figures 4.7 and 4.8 give an overview of the cropland clusters' characteristics. Figure 4.7a illustrates the cropland's neurons partitioned into ten groups in the SOM grid. Note that the neighborhood within each subcluster tends to be similar. Each neuron contains a set of samples, and they are represented by a weight vector. Figure 4.7b illustrates the neurons weight vectors for each subcluster. Figure 4.8 shows the geographical location of samples mapped in each cluster. The samples are spread over the Cerrado biome in the states of Mato Grosso (MT), Bahia (BA), Tocantins (TO), and Maranhão (MA).

From the patterns generated by the subclusters and geographical location provided by the samples, an initial analysis that experts conducted to infer subclasses for Cropland suggests the following:

a) Cropland 1 represents samples of soy–fallow. These samples are mapped only in the state of Bahia. This region is known due to the mostly single cropping regimes (SANCHES et al., 2018). Some samples are spread over Goiás and Tocantins states; however, they are originally labeled as pasture.

b) Cropland 2 represents samples of fallow–cotton. This type of crop is mapped in the states of Mato Grosso and Bahia. The patterns of this group are well defined.

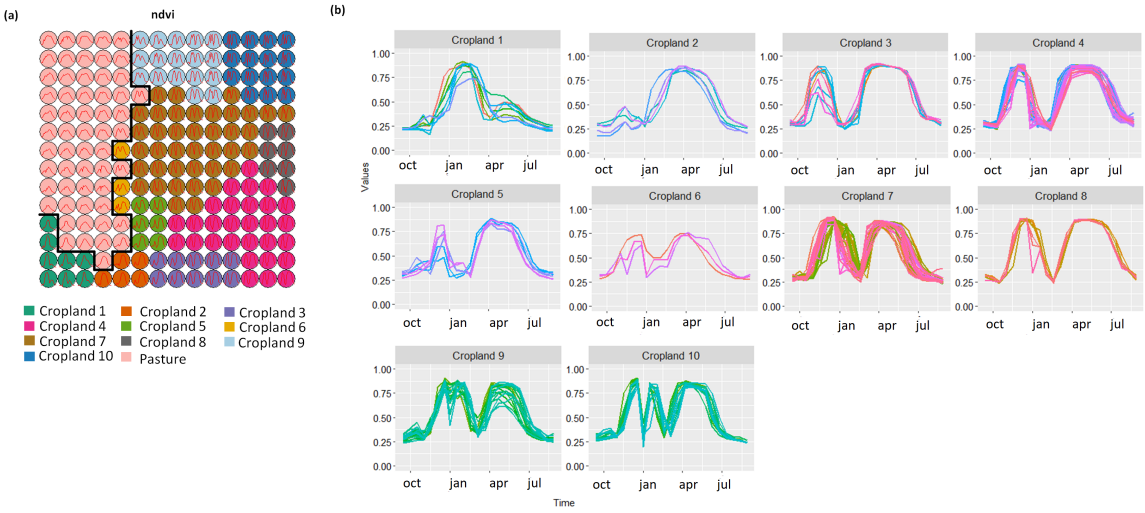
c) Croplands 3,4, and 8 represent samples of soy–cotton. In this study, this type of crop is mapped only in the state of Mato Grosso. Through the temporal patterns (Figure 4.7b), we can notice small variations, particularly during the first cycle (soy crop). This difference may be due to the soybean variety. The soybean varieties planted in Brazil can be of early, medium, and late maturity. The average cycle ranges from 99 to 128 days (ZITO et al., 2018).

d) Cropland 5 represents samples of millet–cotton. In this study, this type of crop is mapped only in the state of Mato Grosso.

e) Croplands 7, 9, and 10 represent samples of soy–corn. This type of crop is spread over the Cerrado biome. The variability of this class can be noticed through the temporal patterns extracted by SOM. It occurs due to the climatological and soil variations leading to differences in each area's agricultural calendar.

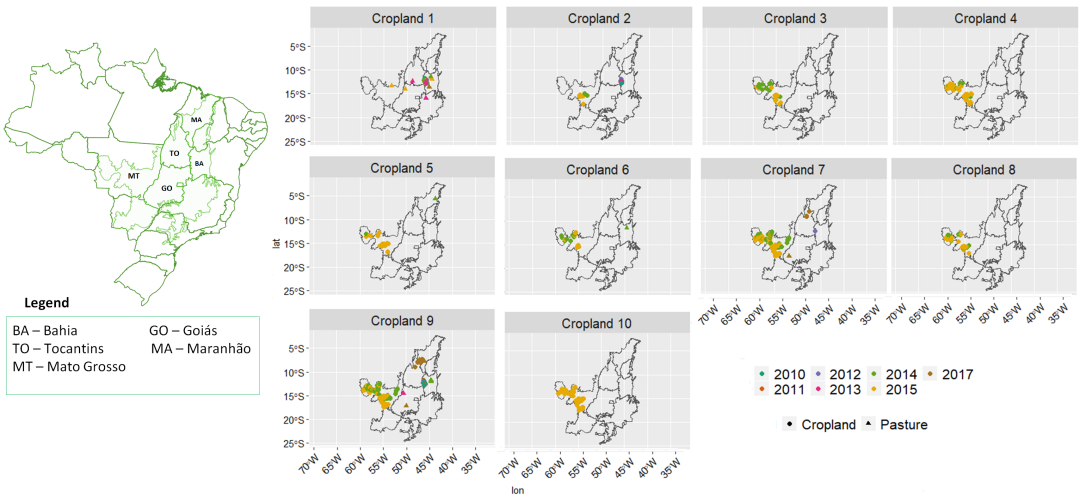
f) Cropland 6 is not well defined. Notice in the SOM grid (Figure 4.7a) that most of the neighbors of Cropland 6 belong to other groups. The temporal signatures can be confounded between soy and millet during the first cycle due to noise, likely caused by clouds. It is necessary to look at these samples in more detail. In contrast, in the second cycle, we can notice patterns of cotton and corn. Additionally, this cluster contains samples initially labeled as pasture.

Figure 4.7 - Clusters of cropland. (a) SOM grid with subclusters of Cropland. (b) Weight vectors of each subcluster. Each line represent a neuron.



SOURCE: Author.

Figure 4.8 - Clusters of cropland. Spatial location, by cluster, where the samples are.

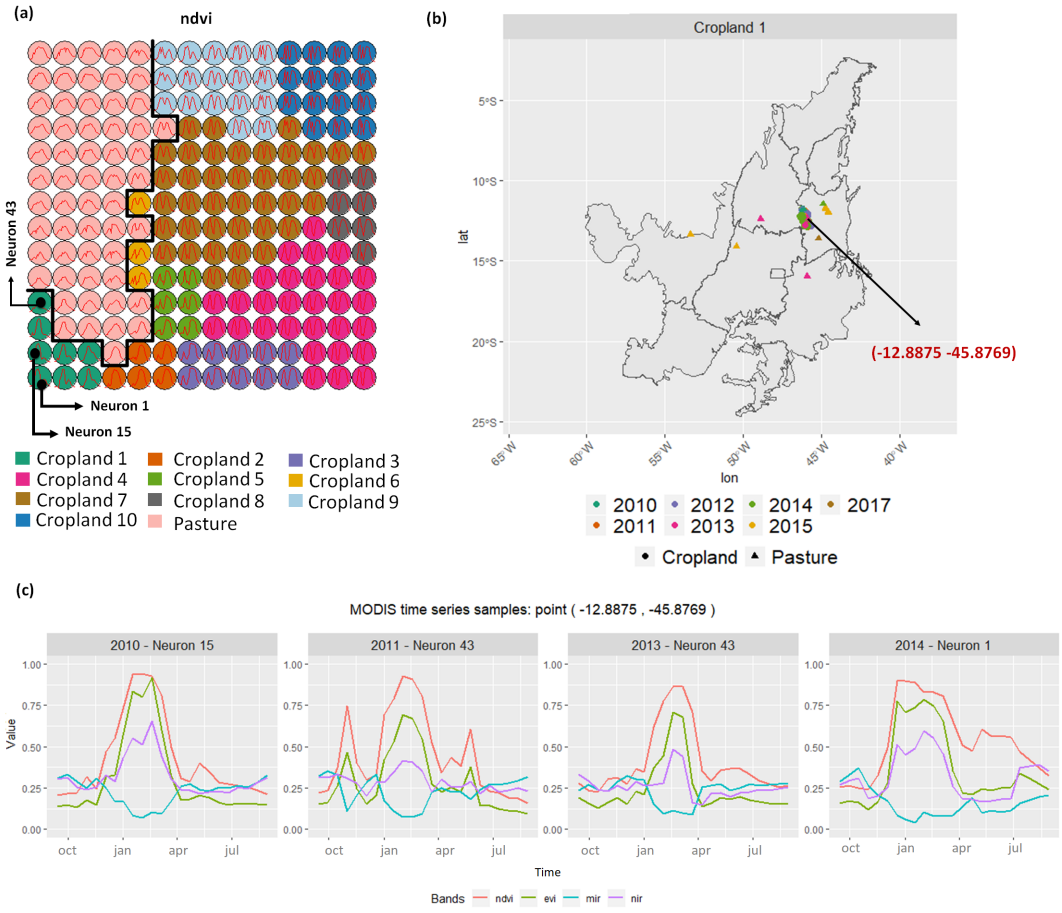


SOURCE: Author.

Some examples are presented throughout this section, detailing how knowledge extraction can be obtained from the output clustering methods.

Figure 4.9 presents the temporal variability between the patterns of a single cropping soy–fallow. All the crop samples from Cropland 1 are mapped only in the west of Bahia state, a region with tradition in planting the single cropping system (SANCHES et al., 2018) (Figure 4.9b). Figure 4.9 shows the temporal dynamics of a point $(-12.8875, -45.8769)$ named single cropping over four harvests from 2010 to 2014. According to the agricultural calendars of the National Supply Company (CONAB) and the Brazilian Institute of Geography and Statistics (IBGE), soybean planting in this region varies from October to January. For rained systems, soybean planting is linked to the rainy season (ABRAHAO; COSTA, 2018). Therefore, the planting season varies from year to year, and this can be seen in Figure 4.9c. It can be observed that the period in which the crop was planted in the soil also varied, and this may be an indication of the variety of soybean that was planted in this region (with early, medium, or late maturation).

Figure 4.9 - The cluster of Soy-Fallow: Subgroups of Cropland 1. (a) SOM grid with Soy-Fallow subgroups.(b) Spatial location. (c) MODIS time series of point (-12.8875, -45.8769).

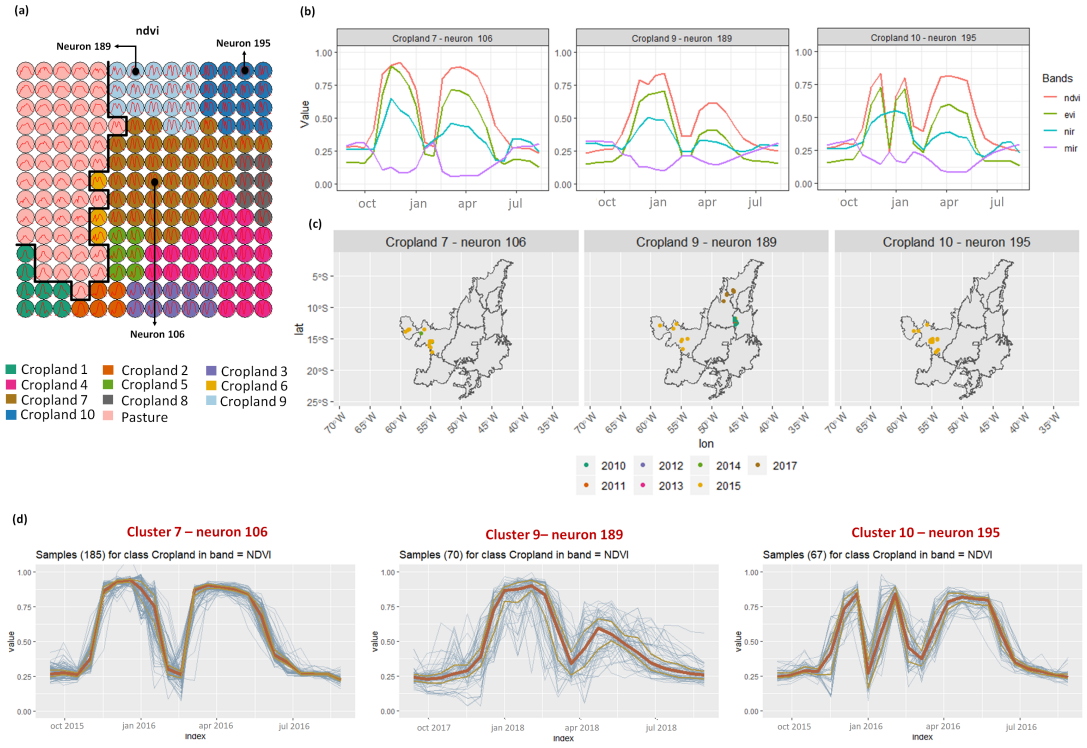


SOURCE: Author.

In Figure 4.10, we highlighted different neurons (Figure 4.10b) that belong to these clusters with distinct spatial location, agricultural calendar, and patterns affected by clouds (Figure 4.10c) to illustrate the variability between the patterns of soy–corn. In neuron 189, some samples belong to the states of Tocantins and Maranhão (7° S). The first cycle of these samples starts later than in Mato Grosso's state (neurons 106 and 195). This occurs because these regions have a different agricultural calendar. Despite the pattern of neuron 189 indicating a later beginning of cycle compared to samples in Mato Grosso, it also presents samples located in Mato Grosso. However, it likely happened because of the noise caused by clouds in Mato Grosso's sample, causing a pattern similar to that observed in Maranhão.

Figure 4.10b,d present the weight vectors and the NDVI time series samples associated with each neuron. In addition to the differences at the beginning of the first cycle, the biggest difference is in the cycle's harvest period. The samples assigned to neuron 189 were harvested late. This indicates that the soybean planted there has worse variety than those planted in samples assigned neurons 106 and 195. In the second cycle, we observe the opposite behavior. Most of the corn samples attributed to neuron 189 are shorter than in neurons 106 and 195. This may have occurred because most samples of neuron 12 are located in western Bahia, Tocantins, and Maranhão states, where corn crops are shorter (early maturing variety) than in the state of Mato Grosso. Moreover, the corn crop in these regions was affected due to the negative weather conditions in 2017.

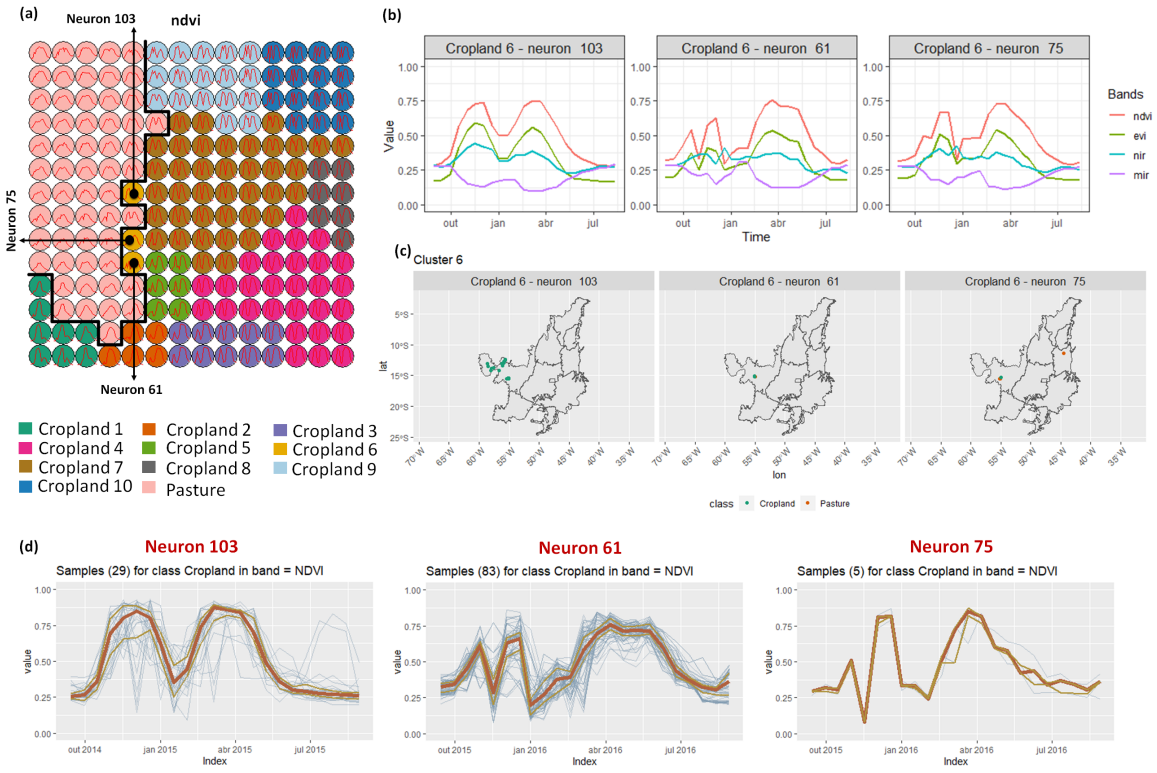
Figure 4.10 - The cluster of Soy-Corn: Subgroups of Cropland 7, 9, and 10. (a) SOM grid with Soy-Corn subgroups.(b) Weight vectors of each neuron. (c) Spatial location where the samples allocated by the neurons 106, 189 and 195 are respectively. (d) NDVI time series and the number of samples by assigned to these neurons of Soy-Corn.



SOURCE: Author.

Figure 4.11 illustrates in more detail the cluster of Cropland 6. This cluster has three neurons; however, the weight vectors (see Figure 4.11b) and the time series assigned to each neuron (see Figure 4.11d) indicate that neuron 103 contains samples of soy–corn. In contrast, neurons 61 and 75 contain samples of soy–cotton. Neurons 61 and 75 are distant from the clusters that represent soy–cotton (Cropland 3,4, and 8) and near the neurons of pasture and soy–corn and millet–cotton (Figure 4.11a). This occurs because of the peaks caused by clouds that impact the soybean patterns, causing a decrease in the pattern's average values. This can lead to confusion with millet. Moreover, this confusion can occur mainly when pasture is planted (notice in Figure 4.11c, samples of pasture mapped in neurons 61 and 75), since its planting period is between September and March, often coinciding with millet's planting period. Furthermore, millet is often used as pasture in the crop–livestock integration system (ALONSO et al., 2017).

Figure 4.11 - The cluster of Cropland 6 (a) SOM grid. (b) Weight vectors of neurons that belong to Cropland 6.(c) Spatial location. (d) NDVI time series and the number of samples by assigned to these neurons of Cropland 6.

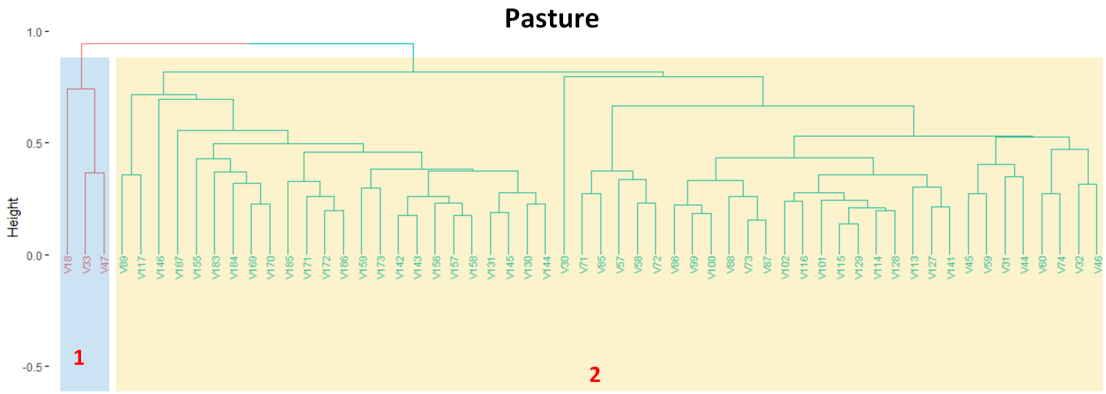


SOURCE: Author.

4.2.3 Revealing the patterns of Pasture

In the SOM grid, there are 57 neurons classified as pasture. However, the samples of pasture do not have a high variability like those of cropland. We separated the dendrogram (Figure 4.12) into two clusters, as suggested by the C-index. Although the C-Index suggests only two clusters, small variations can be found in cluster 2. However, an expert can explore these variations if necessary.

Figure 4.12 - Dendrogram for Pasture partitioned in two clusters.

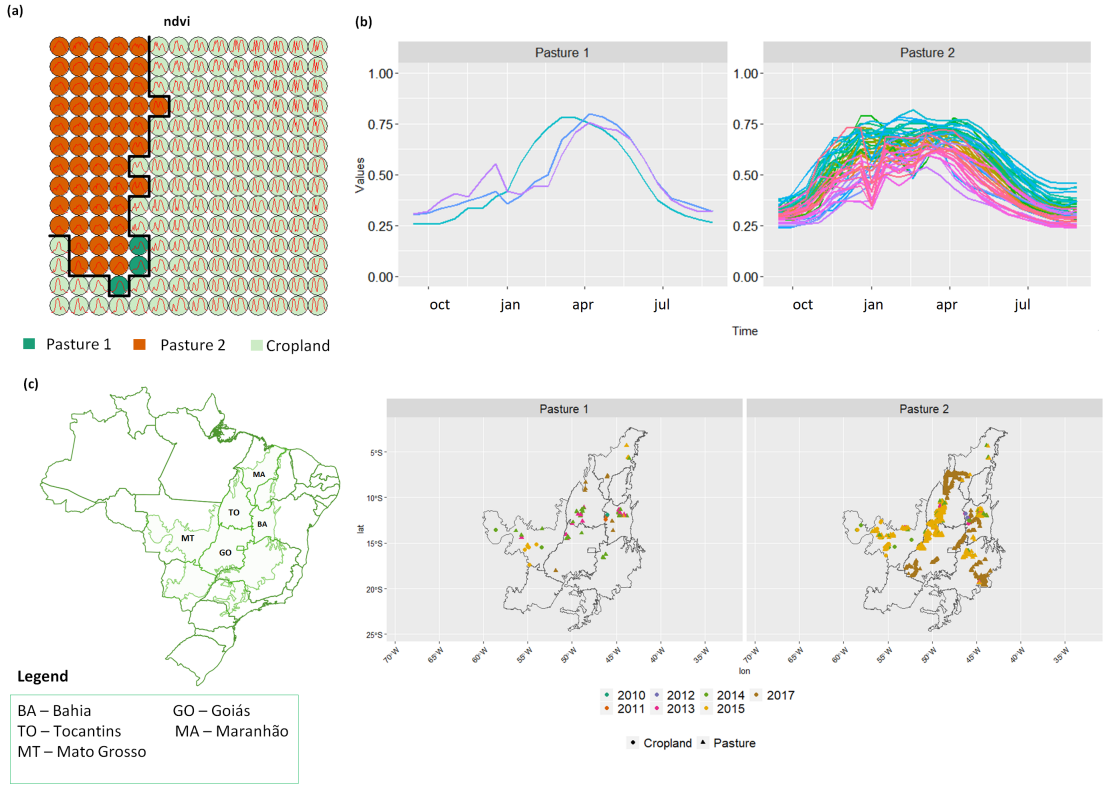


SOURCE: Author.

Figure 4.13 illustrates an overview of the pasture cluster. In Pasture 1, the weight vectors (Figure 4.13b) present patterns that can be confounded with single cropping. This type of confusion may occur due to October to April belonging to the rainy season (EMPRESA BRASILEIRA DE PESQUISA AGROPECUÁRIA - EMBRAPA, 2020), and, as such, the spectral profiles in these months are low due to the noise of clouds and rain. Furthermore, it can be observed in the SOM grid in Figure 4.13a that the neighbors of the neurons labeled as Pasture 1 are classified as cropland.

The group of Pasture 2, on average, presents similar spectral patterns. Furthermore, this cluster has proportionally less confusion with cropland. In addition, there are some patterns with the lowest spectral values, e.g., NDVI values below 0.5. NDVI is an index related to biophysical variables that control vegetation productivity, such as net primary productivity and the leaf area index (JENSEN, 2009). Therefore, there is a probability of these samples representing areas with lower productivity.

Figure 4.13 - Cluster of pasture. (a) SOM grid with subclusters of pasture. (b) Weight vectors of each subcluster. Each line represent a neuron. (c) Spatial subclusters

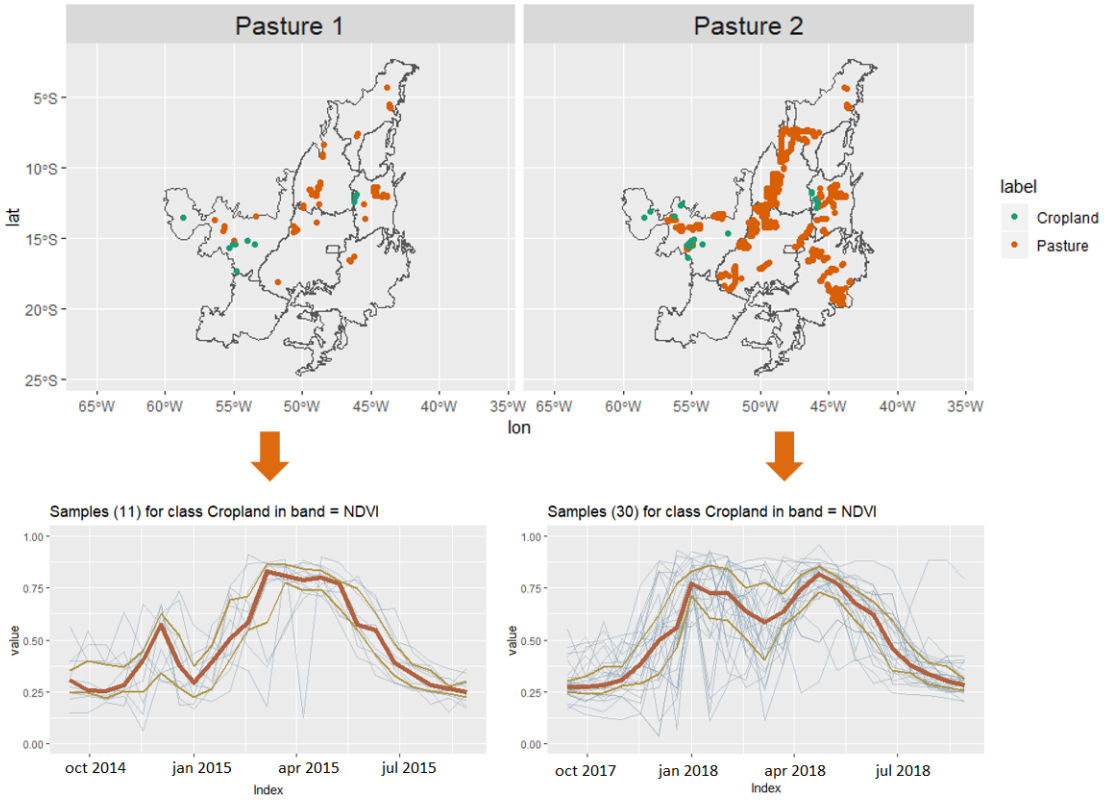


SOURCE: Author.

Figure 4.14 illustrates the cropland samples mapped in Pasture clusters. These samples belong to the states of Mato Grosso and Bahia. The NDVI time series of the cropland samples assigned to Pasture 1 present a high incidence of clouds from September to February. Due to these peaks, it is not easy to distinguish which specific type of crop these samples belong to or if they really are pasture. It would be necessary to check each region's agricultural calendar and their respective years or high-resolution images in detail. The NDVI time series of cropland's samples mapped in Pasture 2 are noisy, and their temporal signatures are similar to those of pasture. These samples may have been affected by noise, or they may be mislabeled. In addition, in the work of Picoli et al. (2018), the authors presented a pattern of 370 samples of pasture in the state of Mato Grosso, also using time series of the product MOD13Q1, where it is also possible to observe much noise caused by clouds, similar

to the patterns presented in the samples labeled as cropland that were assigned in Pasture 2.

Figure 4.14 - Samples originally labeled as cropland that were assigned to clusters of Pasture.



SOURCE: Author.

4.2.4 Assessing the performance of the training samples

After the cluster analysis, we evaluate the performance of two datasets generated by the clusters. First, using the entire dataset which the samples mapped in Pasture neurons keep the label as Cropland. Second, we remove from the dataset samples of Cropland that were mapped in Pasture's clusters. Besides that, we partitioned the Cropland 6 into Soy-Corn and Soy-Cotton (see figure 4.11). Table 4.1 presents the number of samples associated with each cluster and the classes that were assigned to each one.

Table 4.1 - Number of training samples by cluster

Cluster	Count	Frequency	Class
1. Cropland_Pasture	41	0.26%	Cropland
2. Cropland_1	563	3.5%	Soy-Fallow
3. Cropland_2	348	2.2 %	Fallow-Cotton
4. Cropland_3_4_6_8	3866	24.5 %	Soy-Cotton
5. Cropland_5	429	2.8 %	Millet-Cotton
6. Cropland_6_7_9_10	5331	34.5%	Soy-Corn
7. Pasture_1	90	0.58%	Pasture_1
8. Pasture_2	4926	31.2%	Pasture_2

Tables 4.2 and 4.3 present the confusion matrix for the samples relabeled according to the analysis provided by the clusters. For the entire dataset the overall accuracy was 97% (Table 4.2), and for the filtered dataset the overall was 98% (Table 4.3). Although the producer's accuracy to Cropland, presented in Table 4.2, is low, few samples were classified as Pasture. Both confusion matrices indicate the majority confusion among Soy-Corn, Soy-Cotton, and Millet-Cotton, as highlighted in Cropland 6 analysis. Soy-Cotton and Soy-Fallow mix each other; it can happen because the first cycle is the same, and small variations during the second cycle can affect the separability between them. The producer's accuracy to Millet-Cotton is also low, and it probably happens because the noise caused by clouds during the first cycle of soybean becoming the separability between soybean and Millet difficult. The confusions between Fallow-Cotton and Soy-Cotton can occur because of some land variation during the Fallow cycle, and the same occurs to confusions between Soy-Fallow and Soy-Corn. Although we removed Cropland samples mapped in Pasture neurons (Table 4.3), the confusion between Pasture 2 and the double-cropping types still occurs due to clouds' noise affecting the separability between them.

Table 4.2 - Confusion Matrix - The Cropland samples mapped in Pasture were kept in the dataset

	1	2	3	4	5	6	7	8	UA
1. Cropland	5	0	0	0	0	0	0	0	100%
2. Soy-Corn	23	5372	33	63	1	7	6	1	97%
3. Soy-Cotton	0	50	3912	27	11	0	1	1	97%
4. Millet-Cotton	0	1	4	339	0	0	1	0	98%
5. Fallow-Cotton	6	1	2	0	334	0	0	0	97%
6. Soy-Fallow	2	13	0	0	1	552	0	0	97%
7. Pasture_2	5	6	3	0	1	4	4978	61	98%
8. Pasture_1	0	0	0	0	0	0	0	27	100%
PA	12%	98%	98%	79%	95%	98%	99%	30%	**98%**

Table 4.3 - Confusion Matrix - The Cropland samples mapped in Pasture were removed from the dataset

	1	2	3	4	5	6	7	UA
1. Soy-Corn	5365	31	68	1	7	6	1	97%
2. Soy-Cotton	58	3913	28	9	0	1	0	97%
3.Millet-Cotton	1	3	333	0	0	1	0	98%
4. Fallow-Cotton	1	3	0	336	0	0	0	98%
5. Soy-Fallow	13	0	0	1	551	0	0	97%
6. Pasture_2	5	4	0	1	5	4918	61	98%
7. Pasture_1	0	0	0	0	0	0	28	100%
PA	98%	98%	77%	96%	97%	99%	31%	**97%**

In addition, we evaluate the original dataset without the relabeled samples, as shown in Table 4.4. As expected, the training data performance is much better with non-discriminated data. Despite the confusion between crop and Pasture, the overall accuracy was 99%. When the labels are specified, the confusion is spread among the same types. However, we can see where the most significant confusion happens, such as Pasture 2 and double-cropping, and Pasture 1 and single-cropping.

Table 4.4 - Confusion Matrix - Original Dataset

	Pasture	Cropland	UA
Pasture	4980	17	99%
Cropland	36	10761	99%
PA	99%	99%	**99%**

4.3 Discussions

In this study, we show how the proposed method can be applied. The aim is to assess the clusters provided by SOM combined with hierarchical clustering to analyze the possibility of generating subclasses. The method suggests ten subclusters for Cropland and two subclusters for Pasture. We relabeled the samples that were initially labeled as Cropland into five subclasses according to the spatial, temporal, and phenological information . Besides the subclasses, it also possible to identify mislabeled or noisy samples in the training dataset contributing to dataset enhancement. For instance, in Cropland 1 there are some samples of Pasture. In this case, an analysis by experts could be done to verify why these samples are not well separated. It can be done using phenological metrics, checking high resolution images or maps of agriculture and pasture. This analysis is essential to verify if the samples are not separable due to the sensors' spatial and temporal resolution, noise such as clouds, or mislabeled.

The output table provided by the clustering methods, illustrated in Figure 4.3, allows quantifying the samples' information, such as the number of Pasture samples mapped in neurons labeled as Cropland 1. As shown in Table 4.4, the original dataset presents confusion between Pasture and Cropland. However, when the subclasses are inferred through the subclusters, the reason for this type of confusion becomes clearer. In our dataset, there is a confusion between samples with single-cropping and Pasture. It occurs due to the similarity between their temporal patterns. Notice it in more details in Cropland 2 (Figure 4.7b) and Pasture 1 (Figure 4.13b). Since the method indicates exactly which samples are being confounded, and their information such as spatial location (longitude and latitude) and time, an expert could check through high resolution or LUCC maps the real label. Thus, future errors can be avoided in the classification.

In general, the samples' patterns presented in the results section help significantly distinguish the types of Agriculture and Pasture. Nevertheless, it is important to note

that small variations within the subclusters can still be observed in Cropland 1. It is possible to note the different temporal patterns in each neuron of Cropland 1 and consequently the samples assigned to them, indicating variation in soybean maturity. This type of variation could be captured through the hierarchical clustering using the MODIS sensor if the number of subclusters provided by the dendrogram was higher than 10. Once we have 139 neurons to represent Cropland samples, the number of subcluster provided by hierarchical clustering could be 139 if the dendrogram in presented in Figure 4.6 was cut when the height = 0. In our case study, the C-index was applied to define the number of subcluster for Cropland, but it could be defined manually according to the final user's real goal.

In contrast to Soy-Fallow, Soy-Corn and Soy-Cotton were distributed among different groups. The number of Soy-Corn and Soy-Fallow samples cause an unbalance in the dataset. For this reason, the differences between the Soy-Corn and Soy-Fallow neurons are quite small. The most significant differences are attributed to more distant neurons in the SOM grid. In this way, considering the cut in the dendrogram, the significant intra-class variations, such as spatial location, agriculture calendar, are pointed in different subclusters. In addition, phenological metrics could be used as thresholds to distinguish the clusters and capture crop variations.

Although the neurons present intra-class variations, since identifying soybean variability until the crop separability, it is important to highlight that the sensor's choice may present different results than what was shown in this paper due to the spatial and temporal resolution. In addition, the choice of the similarity measure used in the clustering methods can also omit or highlight these variations (FERREIRA et al., 2019).

4.4 Conclusion

There is a lack of high-quality training samples in the remote sensing field, mainly when it is driven for change monitoring in large areas and multiples years due to the high intra-class variability of land use and cover types. A way to improve sample acquisition task is through temporal patterns. The use of satellite image time series provides phenological and spectral information that can be considered during the collecting samples or labeling process.

The focus of this paper was to present the method and show how it can be applied. The proposed method combined SOM with hierarchical clustering to identify spatiotemporal patterns through exploratory analysis in training samples dataset. The

method works well to infer subclasses using MODIS images of 250 meters and 16 days, mainly for Cropland classes presented in this dataset. From the subclusters, we were able to identify mislabeled samples and refine the generic labels to infer subclasses. The method is not free of generating uncertain labels; however, the phenological, spectro-temporal, and spatial information provided by the satellite image time series samples identified through the subclusters patterns assists the experts during the labeling process.

We explored Cropland and Pasture samples over the Cerrado biome in Brazil using MODIS because of their high variability in different regions and years. An initiative called Brazil Data Cube (FERREIRA et al., 2020) is producing Analysis-Ready Data (ARD) and multidimensional data cubes from medium-resolution satellite images for all Brazilian territory. Using these data cubes, we intend to apply the proposed method to 10-20 meters and five days Sentinel 2 image time series in order to separate all distinct types of land use and cover classes in Cerrado.

5 FINAL REMARKS AND CONCLUSION

LUCC information is essential for understanding relationships between natural phenomena and human activities in order to improve resource management and decision-making (FOLEY et al., 2005; UDELHOVEN et al., 2015). Recently, time series from big data sets of EO satellite images and machine learning methods have been widely used to effectively map LUCC (ELMES et al., 2020).

This thesis contributes to the EO field, addressing the challenge of obtaining good quality land use and cover samples to train machine learning methods and producing accurate LUCC maps from big data sets of image time series. The approach proposed is based on SOM, and it has been applied in spatiotemporal datasets with significant results because its properties of dimensionality reduction and topology preservation generating spatial clusters in its attribute space. It motivated the use of SOM as the main method in this thesis.

This document presented two methods. The first method, described in Chapter 3, is based on SOM combined with Bayesian Inference to provide measures that assess each sample's level of reliability. The second method, described in Chapter 4, presented SOM combined with hierarchical clustering to identify and explore spatiotemporal patterns in training samples dataset to evaluate the intra-class variability in large areas and multiples years. Although not present in Chapter 4, the method to identify spatiotemporal patterns embraces the measures provided by SOM and Bayesian. Both proposes can be used together by an EO scientist to help filter high-quality land use and cover samples.

Two additional works that complement this thesis's content were carried out. The first study explores different distance measures for time series clustering (FERREIRA et al., 2019) (see annex A). The second study analyses the Growing Self-Organizing Maps (GSOM) algorithm for clustering satellite image time series as an alternative to SOM (ADEU et al., 2020). In the GSOM method, users do not have to define a grid size a priori. The GSOM presented satisfactory results for clustering time series samples presented in the Mato Grosso state. However, the neighborhood topology must be evaluated carefully when GSOM is implemented because the neighborhood is an important characteristic of this thesis's methods.

From the case studies presented in this thesis, satisfactory results were obtained by proposed methods to improve the quality of the land use and cover samples. In the first method, mislabeled samples and part of the outliers were identified and removed

from the training dataset through probabilities provided by the method. In the second method, sub-classes were inferred from the clusters because of temporal patterns and different spatial locations. Mislabeled samples were also identified, contributing to the improvement of the sample dataset. Classifications were performed to validate both proposes using the new datasets provided by the methods. The results demonstrated a positive impact on the overall classification accuracy.

5.1 Future work

Based on the case studies results of this thesis, some points for future work are described:

a) Exploit techniques to avoid and handle unbalanced classes.

b) Assess the sample's quality using phenological metrics instead of the raw time series.

c) Investigate the internal cluster validation indices and the use of auxiliary datasets (such as climatological datasets) to provide a start point for the EO scientists to choose the best number of clusters to explore the intra-class variability according to their necessity.

d) Implement a Web Sample Assessment Service (WSAS) that consumes the method in the SITS **R** package to evaluate the quality of the training samples stored in BDC infrastructure using different data cubes.

e) Develop a web interface combining both methods to facilitate the samples' exploratory analysis.

REFERENCES

ABRAHAO, G. M.; COSTA, M. H. Evolution of rain and photoperiod limitations on the soybean growing season in brazil: the rise (and possible fall) of double-cropping systems. **Agricultural and Forest Meteorology**, v. 256-257, p. 32 – 45, 2018. ISSN 0168-1923. 66

ADAM, E.; MUTANGA, O.; RUGEGE, D. Multispectral and hyperspectral remote sensing for identification and mapping of wetland vegetation: a review. **Wetlands Ecology and Management**, v. 18, n. 3, p. 281–296, 2010. 26

ADEU, R. d. S. da S.; FERREIRA, K. R.; ANDRADE, P. R.; SANTOS, L. Assessing satellite image time series clustering using growing SOM. In: INTERNATIONAL CONFERENCE ON COMPUTATIONAL SCIENCE AND ITS APPLICATIONS. **Proceedings...** Cagliari, Italy: Springer, 2020. p. 270–282. 79

AGHABOZORGI, S.; SHIRKHORSHIDI, A. S.; WAH, T. Y. Time-series clustering: A decade review. **Information Systems**, v. 53, p. 16–38, 2015. 50

AGUIAR, D. A.; ADAMI, M.; SILVA, W. F.; RUDORFF, B.; MELLO, M. P.; SILVA, J. d. S. V. Modis time series to assess pasture land. In: INTERNATIONAL GEOSCIENCE AND REMOTE SENSING SYMPOSIUM. **Proceedings...** Honolulu, HI, USA, 2010. p. 2123–2126. 3

ALENCAR, A.; SHIMBO, J. Z.; LENTI, F.; MARQUES, C. B.; ZIMBRES, B.; ROSA, M.; ARRUDA, V.; CASTRO, I.; RIBEIRO, J. P. F. M.; VARELA, V.; ALENCAR, I.; PIONTEKOWSKI, V.; RIBEIRO, V.; BUSTAMANTE, M. M. C.; SANO, E. E.; BARROSO, M. Mapping three decades of changes in the Brazilian savanna native vegetation using Landsat data processed in the Google Earth Engine platform. **Remote Sensing**, v. 12, n. 6, 2020. ISSN 2072-4292. Available from: <https://www.mdpi.com/2072-4292/12/6/924>. 49, 50

ALONSO, M. P.; MORAES, E.; PEREIRA, D. H.; PINA, D. S.; MOMBACH, M. A.; HOFFMANN, A.; GIMENEZ, B. de M.; SANSON, R. M. M. Pearl millet grain for beef cattle in crop-livestock integration system: intake and digestibility. **Semana-Ciencias Agrarias**, v. 38, p. 1471–1482, 2017. 69

ANDRIENKO, G.; ANDRIENKO, N.; BREMM, S.; SCHRECK, T.; LANDESBERGER, T. V.; BAK, P.; KEIM, D. Space-in-time and time-in-space self-organizing maps for exploring spatiotemporal patterns. **Computer Graphics Forum**, v. 29, n. 3, p. 913–922, 2010. 50

ANSARI, M. Y.; AHMAD, A.; KHAN, S. S.; BHUSHAN, G.; MAINUDDIN. Spatiotemporal clustering: a review. **Artificial Intelligence Review**, v. 53, p. 2381–2423, 2019. 53

ARLOT, S.; CELISSE, A. et al. A survey of cross-validation procedures for model selection. **Statistics Surveys**, v. 4, p. 40–79, 2010. 59

ARVOR, D.; JONATHAN, M.; MEIRELLES, M. S. P.; DUBREUIL, V.; DURIEUX, L. Classification of MODIS EVI time series for crop mapping in the state of Mato Grosso, Brazil. **International Journal of Remote Sensing**, v. 32, n. 22, p. 7847–7871, nov. 2011. ISSN 0143-1161, 1366-5901. 3

ASSUNCAO, R. M.; SCHMERTMANN, C. P.; POTTER, J. E.; CAVENAGHI, S. M. Empirical bayes estimation of demographic schedules for small areas. **Demography**, v. 42, n. 3, p. 537–558, aug. 2005. ISSN 1533-7790. 32

ASTEL, A.; TSAKOVSKI, S.; BARBIERI, P.; SIMEONOV, V. Comparison of self-organizing maps classification approach with cluster and principal components analysis for large environmental data sets. **Water Research**, v. 41, n. 19, p. 4566 – 4578, 2007. ISSN 0043-1354. Available from: <http://www.sciencedirect.com/science/article/pii/S0043135407004137>. 5, 50

ATKINSON, P. M.; JEGANATHAN, C.; DASH, J.; ATZBERGER, C. Inter-comparison of four models for smoothing satellite sensor time-series data to estimate vegetation phenology. **Remote Sensing of Environment**, v. 123, p. 400–417, 2012. 23

ATZBERGER, C.; EILERS, P. H. Evaluating the effectiveness of smoothing algorithms in the absence of ground reference measurements. **International Journal of Remote Sensing**, v. 32, n. 13, p. 3689–3709, 2011. 23

AUGUSTIJN, E.-W.; ZURITA-MILLA, R. Self-organizing maps as an approach to exploring spatiotemporal diffusion patterns. **International Journal of Health Geographics**, v. 12, n. 1, p. 60, dec. 2013. ISSN 1476-072X. 5, 50

BAGAN, H.; WANG, Q.; WATANABE, M.; YANG, Y.; MA, J. Land cover classification from MODIS EVI times-series data using SOM neural network. **International Journal of Remote Sensing**, v. 26, n. 22, p. 4999–5012, 2005. 10

BELGIU, M.; BIJKER, W.; CSILLIK, O.; STEIN, A. Phenology-based sample generation for supervised crop type classification. **International Journal of**

Applied Earth Observation and Geoinformation, v. 95, p. 102264, 2021. ISSN 0303-2434. Available from: <http://www.sciencedirect.com/science/article/pii/S0303243420309077>. 4, 49, 50

BELGIU, M.; DRAGUT, L. Random Forest in remote sensing: a review of applications and future directions. **ISPRS Journal of Photogrammetry and Remote Sensing**, v. 114, p. 24–31, 2016. 21, 23, 45, 59

BENGIO, Y.; GRANDVALET, Y. No unbiased estimator of the variance of k-fold cross-validation. **Journal of Machine Learning Research**, v. 5, p. 1089–1105, dec. 2004. ISSN 1532-4435. 59

BIRANT, D.; KUT, A. St-dbscan: an algorithm for clustering spatial–temporal data. **Data Knowledge Engineering**, v. 60, n. 1, p. 208 – 221, 2007. ISSN 0169-023X. Intelligent Data Mining. Available from: <http://www.sciencedirect.com/science/article/pii/S0169023X06000218>. 50

BOLES, S. H.; XIAO, X.; LIU, J.; ZHANG, Q.; MUNKHTUYA, S.; CHEN, S.; OJIMA, D. Land cover characterization of temperate east asia using multi-temporal {VEGETATION} sensor data. **Remote Sensing of Environment**, v. 90, n. 4, p. 477 – 489, 2004. 12

BRASIL. MINISTÉRIO DO MEIO AMBIENTE. **Brazilian Biomes**. 2019. Available from: <https://www.mma.gov.br/>. 23, 24

CAMARA, G.; ASSIS, L. F.; RIBEIRO, G.; FERREIRA, K. R.; LLAPA, E.; VINHAS, L. Big earth observation data analytics: Matching requirements to system architectures. In: INTERNATIONAL WORKSHOP ON ANALYTICS FOR BIG GEOSPATIAL DATA, 5. **Proceedings...** Burlingname, CA, USA: ACM, 2016. p. 1–6. 1

CAMARA, G.; PICOLI, M.; MACIEL, A.; SIMOES, R.; SANTOS, L.; ANDRADE, P. R.; FERREIRA, K.; BEGOTTI, R.; SANCHES, I.; YCARVALHO, A. X.; COUTINHO, A.; ESQUERDO, J.; ANTUNES, J.; DAMIENARVOR. data set, **Land cover change maps for Mato Grosso State in Brazil: 2001-2017 (version 3)**. PANGAEA, 2019. Available from: <https://doi.org/10.1594/PANGAEA.899706>. 5

CHEN, I.-T.; CHANG, L.-C.; CHANG, F.-J. Exploring the spatio-temporal interrelation between groundwater and surface water by using the self-organizing

maps. **Journal of Hydrology**, v. 556, p. 131 – 142, 2018. ISSN 0022-1694. Available from: <http://www.sciencedirect.com/science/article/pii/S0022169417306807>. 5, 50

CHEN, J.; CHEN, J.; LIAO, A.; CAO, X.; CHEN, L.; CHEN, X.; HE, C.; HAN, G.; PENG, S.; LU, M.; ZHANG, W.; TONG, X.; MILLS, J. Global land cover mapping at 30m resolution: A pok-based operational approach. **ISPRS Journal of Photogrammetry and Remote Sensing**, v. 103, p. 7–27, 2015. ISSN 0924-2716. Global Land Cover Mapping and Monitoring. Available from: <https://www.sciencedirect.com/science/article/pii/S0924271614002275>. 1

COMBER, A.; WULDER, M. Considering spatiotemporal processes in big data analysis: Insights from remote sensing of land cover and land use. **Transactions in GIS**, v. 23, n. 5, p. 879–891, 2019. ISSN 1467-9671. 49, 50

COMBER, A. J.; WADSWORTH, R. A.; FISHER, P. F. Using semantics to clarify the conceptual confusion between land cover and land use: the example of forest. **Journal of Land Use Science**, v. 3, n. 2–3, p. 185–198, 2008. 1

COMPARISON of random Forest, k-Nearest Neighbor, and Support Vector Machine Classifiers for Land Cover Classification Using Sentinel. 21

DEMIR, B.; PERSELLO, C.; BRUZZONE, L. Batch-mode active-learning methods for the interactive classification of remote sensing images. **IEEE Transactions on Geoscience and Remote Sensing**, IEEE, v. 49, n. 3, p. 1014–1031, 2010. 4, 49

DICKIE, A.; MAGNO, I.; GIAMPIETRO, J.; DOLGINOW, A. **Challenges and Opportunities for Conservation, Agricultural Production, and Social Inclusion in the Cerrado Biome**. [S.l.], 2016. 51

ELMES, A.; ALEMOHAMMAD, H.; AVERY, R.; CAYLOR, K.; EASTMAN, J. R.; FISHGOLD, L.; FRIEDL, M. A.; JAIN, M.; KOHLI, D.; BAYAS, J. C. L.; LUNGA, D.; MCCARTY, J. L.; PONTIUS, R. G.; REINMANN, A. B.; ROGAN, J.; SONG, L.; STOYNOVA, H.; YE, S.; YI, Z.-F.; ESTES, L. Accounting for training data error in machine learning applied to earth observations. **Remote Sensing**, v. 12, n. 6, 2020. ISSN 2072-4292. Available from: <https://www.mdpi.com/2072-4292/12/6/1034>. 4, 23, 49, 79

EMPRESA BRASILEIRA DE PESQUISA AGROPECUÁRIA - EMBRAPA. **O Cerrado**. 2020. Available from:

<http://www.cpac.embrapa.br/unidade/ocerrado/>. Access in: 14 September 2020. 70

EVERITT, B. S.; LANDAU, S.; LEESE, M.; STAHL, D. **Cluster analysis**. 5. ed. [S.l.]: Wiley, 2011. (Wiley Series in Probability and Statistics). ISBN 0470749911,9780470749913,9780470977804,9780470977811,9780470978443. 50, 57

FENSHOLT, R.; HORION, S.; TAGESSON, T.; EHAMMER, A.; GROGAN, K.; TIAN, F.; HUBER, S.; VERBESSELT, J.; PRINCE, S. D.; TUCKER, C. J.; RASMUSSEN, K. Assessment of vegetation trends in drylands from time series of earth observation data. In: _____. **Remote sensing time series: revealing land surface dynamics**. [S.l.]: Springer International Publishing, 2015. p. 159–182. 2

FERREIRA, K.; SANTOS, L.; PICOLI, M. Evaluating distance measures for image time series clustering in land use and cover monitoring. In: MACHINE LEARNING FOR EARTH OBSERVATION WORKSHOP. **Proceedings...** Würzburg, Germany, 2019. 29, 36, 76, 79, 97

FERREIRA, K. R.; QUEIROZ, G. R.; VINHAS, L.; MARUJO, R. F. B.; SIMOES, R. E. O.; PICOLI, M. C. A.; CAMARA, G.; CARTAXO, R.; GOMES, V. C. F.; SANTOS, L. A.; SANCHEZ, A. H.; ARCANJO, J. S.; FRONZA, J. G.; NORONHA, C. A.; COSTA, R. W.; ZAGLIA, M. C.; ZIOTI, F.; KORTING, T. S.; SOARES, A. R.; CHAVES, M. E. D.; FONSECA, L. M. G. Earth observation data cubes for brazil: requirements, methodology and products. **Remote Sensing**, v. 12, n. 24, 2020. ISSN 2072-4292. Available from: <https://www.mdpi.com/2072-4292/12/24/4033>. 2, 77

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. Global consequences of land use. **Science**, v. 309, n. 5734, p. 570–574, 2005. 1, 21, 79

FOOD AND AGRICULTURE ORGANIZATION - FAO. **SEPAL**. GitHub, 2020. Available from: <https://github.com/openforis/sepal>. 9

FRÉNAY, B.; VERLEYSEN, M. Classification in the presence of label noise: a survey. **IEEE Transactions on Neural Networks and Learning Systems**, v. 25, n. 5, p. 845–869, 2013. 22, 23

GARCIA, L. P.; CARVALHO, A. C. de; LORENA, A. C. Effect of label noise in the complexity of classification problems. **Neurocomputing**, v. 160, p. 108–119, 2015. 22

GARCIA, L. P. F.; C.LORENA, A.; CARVALHO, A. C. P. L. F. A study on class noise detection and elimination. In: BRAZILIAN SYMPOSIUM ON NEURAL NETWORKS. **Proceedings...** Curitiba, Brazil, 2012. p. 13–18. 22

GIULIANI, G.; CHATENOUX, B.; BONO, A. D.; RODILA, D.; RICHARD, J.-P.; ALLENBACH, K.; DAO, H.; PEDUZZI, P. Building an Earth observations data cube: lessons learned from the Swiss Data Cube (SDC) on generating Analysis Ready Data (ARD). **Big Earth Data**, v. 1, n. 1-2, p. 100–117, dec. 2017. ISSN 2096-4471, 2574-5417. 2

GOMES, V. C. F.; QUEIROZ, G. R.; FERREIRA, K. R. An overview of platforms for Big Earth Observation data management and analysis. **Remote Sensing**, v. 12, n. 8, p. 1253, jan. 2020. 2

GOMEZ, C.; WHITE, J. C.; WULDER, M. A. Optical remotely sensed time series data for land cover classification: a review. **ISPRS Journal of Photogrammetry and Remote Sensing**, v. 116, p. 55–72, 2016. ISSN 0924-2716. 1, 2, 3, 9, 10, 21, 23, 27, 49

GRIFFITHS, P.; NENDEL, C.; HOSTERT, P. Intra-annual reflectance composites from Sentinel-2 and Landsat for national-scale crop and land cover mapping. **Remote Sensing of Environment**, v. 220, p. 135–151, jan. 2019. ISSN 00344257. 21

GUO, D.; CHEN, J.; MACEACHREN, A. M.; LIAO, K. A visualization system for space-time and multivariate patterns (vis-stamp). **IEEE Transactions on Visualization and Computer Graphics**, v. 12, n. 6, p. 1461–1474, 2006. 50

HALLAC, D.; VARE, S.; BOYD, S.; LESKOVEC, J. Toeplitz inverse covariance-based clustering of multivariate time series data. In: INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 23. **Proceedings...** Halifax, NS, Canada: ACM, 2017. p. 215–223. 50

HANSEN, M. C.; LOVELAND, T. R. A review of large area monitoring of land cover change using Landsat data. **Remote Sensing of Environment**, v. 122, p. 66–74, 2012. ISSN 0034-4257. 1

HIRD, J. N.; MCDERMID, G. J. Noise reduction of NDVI time series: an empirical comparison of selected techniques. **Remote Sensing of Environment**, v. 113, n. 1, p. 248–258, 2009. 23

HOSTERT, P.; GRIFFITHS, P.; van-der-Linden, S.; PFLUGMACHER, D. Time Series Analyses in a new era of optical satellite data. In: KUENZER, C.; DECH, S.; WAGNER, W. (Ed.). **Remote Sensing Time Series: revealing Land Surface Dynamics**. [S.l.]: Springer International Publishing, 2015. p. 25–41. ISBN 978-3-319-15967-6. 1, 4, 49, 50

HUANG, H.; WANG, J.; LIU, C.; LIANG, L.; LI, C.; GONG, P. The migration of training samples towards dynamic global land cover mapping. **ISPRS Journal of Photogrammetry and Remote Sensing**, v. 161, p. 27 – 36, 2020. ISSN 0924-2716. Available from: <http://www.sciencedirect.com/science/article/pii/S0924271620300101>. 4, 49

HUANG, X.; WENG, C.; LU, Q.; FENG, T.; ZHANG, L. Automatic labelling and selection of training samples for high-resolution remote sensing image classification over urban areas. **Remote Sensing**, v. 7, n. 12, p. 16024–16044, 2015. 4, 49

HUBERT, L. J.; LEVIN, J. R. A general statistical framework for assessing categorical clustering in free recall. **Psychological Bulletin**, v. 83, n. 6, p. 1072, 1976. 58

HUETE, A.; DIDAN, K.; MIURA, T.; RODRIGUEZ, E.; GAO, X.; FERREIRA, L. Overview of the radiometric and biophysical performance of the modis vegetation indices. **Remote sensing of Environment**, v. 83, n. 1, p. 195–213, 2002. 1, 2, 12

INGLADA, J.; VINCENT, A.; ARIAS, M.; TARDY, B.; MORIN, D.; RODES, I. Operational high resolution land cover map production at the country scale using satellite image time series. **Remote Sensing**, v. 9, n. 1, p. 95, jan. 2017. 21

INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS - INPE. **TerraClass Cerrado Project: Use and Vegetation Cover Map of the Cerrado**. 2013. Available from: <http://www.dpi.inpe.br/tccerrado/>. Access in: Accessed on 28 November 2019. xiii, 24

JENSEN, J. R. **Remote sensing of the environment: an earth resource perspective**. 2. ed. [S.l.]: Pearson, 2009. 70

JIANG, Z.; HUETE, A. R.; DIDAN, K.; MIURA, T. Development of a two-band enhanced vegetation index without a blue band. **Remote Sensing of Environment**, v. 112, n. 10, p. 3833–3845, oct. 2008. ISSN 0034-4257. 23

JOHNSON, S. C. Hierarchical clustering schemes. **Psychometrika**, v. 32, n. 3, p. 241–254, 1967. 54

KAUFMAN, P. J. R. L. **Finding groups in data: an introduction to cluster analysis**. 9. ed. [S.l.]: Wiley-Interscience, 1990. ISBN 9780471878766,0471878766. 50, 57

KENNEDY, R. E.; ANDREFOUET, S.; COHEN, W. B.; GOMEZ, C.; GRIFFITHS, P.; HAIS, M.; HEALEY, S. P.; HELMER, E. H.; HOSTERT, P.; LYONS, M. B.; MEIGS, G. W.; PFLUGMACHER, D.; PHINN, S. R.; POWELL, S. L.; SCARTH, P.; SEN, S.; SCHROEDER, T. A.; SCHNEIDER, A.; SONNENSCHEIN, R.; VOGELMANN, J. E.; WULDER, M. A.; ZHU, Z. Bringing an ecological view of change to Landsat-based remote sensing. **Frontiers in Ecology and the Environment**, v. 12, n. 6, p. 339–346, 2014. ISSN 1540-9309. 1

KHARDON, R.; WACHMAN, G. Noise tolerant variants of the perceptron algorithm. **Journal of Machine Learning Research**, v. 8, n. 8, p. 227–248, 2007. Available from: <http://jmlr.org/papers/v8/khardon07a.html>. 23

KLINK, C.; MACHADO, R. Conservation of the Brazilian cerrado. **Conservation Biology**, v. 19, n. 3, p. 707–713, 2005. 51

KOHONEN, T. Self-organized formation of topologically correct feature maps. **Biological Cybernetics**, v. 43, n. 1, p. 59–69, 1982. 5

_____. The self-organizing map. **Proceedings of the IEEE**, v. 78, n. 9, p. 1464–1480, sep. 1990. ISSN 1558-2256. 27, 50, 54, 56

_____. Essentials of the self-organizing map. **Neural Networks**, v. 37, p. 52–65, jan. 2013. ISSN 0893-6080. 29, 56

KOHONEN, T.; KASKI, S.; LAGUS, K.; SALOJARVI, J.; HONKELA, J.; PAATERO, V.; SAARELA, A. Self organization of a massive document collection. **IEEE Transactions on Neural Networks**, v. 11, n. 3, p. 574–585, 2000. 57

LAMBIN, E. F.; TURNER, B. L.; GEIST, H. J.; AGBOLA, S. B.; ANGELSEN, A.; BRUCE, J. W.; COOMES, O. T.; DIRZO, R.; FISCHER, G.; FOLKE, C.; GEORGE, P. S.; HOMEWOOD, K.; IMBERNON, J.; LEEMANS, R.; LI, X. B.;

MORAN, E. F.; MORTIMORE, M.; RAMAKRISHNAN, P. S.; RICHARDS, J. F.; SKANES, H.; STEFFEN, W.; STONE, G. D.; SVEDIN, U.; VELDKAMP, T. A.; VOGEL, C.; XU, J. C. The causes of land-use and land-cover change: moving beyond the myths. **Global Environmental Change-Human and Policy Dimensions**, v. 11, n. 4, p. 261–269, 2001. 1

LAWAWIROJWONG, S. Soft supervised self-organizing mapping (3som) for improving land cover classification with modis time-series. In: . [S.l.: s.n.], 2013. 10

LEWIS, A.; OLIVER, S.; LYMBURNER, L.; EVANS, B.; WYBORN, L.; MUELLER, N.; RAEVKSI, G.; HOOKE, J.; WOODCOCK, R.; SIXSMITH, J.; WU, W.; TAN, P.; LI, F.; KILLOUGH, B.; MINCHIN, S.; ROBERTS, D.; AYERS, D.; BALA, B.; DWYER, J.; DEKKER, A.; DHU, T.; HICKS, A.; IP, A.; PURSS, M.; RICHARDS, C.; SAGAR, S.; TRENHAM, C.; WANG, P.; WANG, L.-W. The Australian Geoscience Data Cube — foundations and lessons learned. **Remote Sensing of Environment**, v. 202, p. 276–292, dec. 2017. ISSN 00344257. 2, 9

LIAO, T. W. Clustering of time series data: a survey. **Pattern Recognition**, v. 38, n. 11, p. 1857–1874, 2005. 50, 53

LIESENBERG, V.; PONZONI, F. J.; GALVÃO, L. S. Análise da dinâmica sazonal e separabilidade espectral de algumas fitofisionomias do cerrado com índices de vegetação dos sensores modis/terra e aqua. **Revista Árvore**, v. 31, n. 2, p. 295–305, 2007. 26, 40

LIU, H.; ZHAN, Q.; YANG, C.; WANG, J. Characterizing the spatio-temporal pattern of land surface temperature through time series clustering: based on the latent pattern and morphology. **Remote Sensing**, v. 10, n. 4, p. 654, 2018. 50

LIU, Y.; WEISBERG, R. H. A review of self-organizing map applications in meteorology and oceanography. **Self-Organizing Maps: Applications and Novel Algorithm Design**, p. 253–272, 2011. 50

LU, Q.; MA, Y.; XIA, G.-S. Active learning for training sample selection in remote sensing image classification using spatial information. **Remote Sensing Letters**, v. 8, n. 12, p. 1210–1219, 2017. 4, 49

MAUS, V.; CAMARA, G.; CARTAXO, R.; SANCHEZ, A.; RAMOS, F. M.; QUEIROZ, G. R. 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**, v. 9, n. 8, p. 3729–3739, 2016. 3, 11, 12, 21

MAXWELL, A. E.; WARNER, T. A.; FANG, F. Implementation of machine-learning classification in remote sensing: an applied review. **International Journal of Remote Sensing**, v. 39, n. 9, p. 2784–2817, 2018. 21, 49

MELLOR, A.; BOUKIR, S.; HAYWOOD, A.; JONES, S. 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**, v. 105, p. 155 – 168, 2015. ISSN 0924-2716. Available from: <http://www.sciencedirect.com/science/article/pii/S0924271615000945>. 23

MERONI, M.; D'ANDRIMONT, R.; VRIELING, A.; FASBENDER, D.; LEMOINE, G.; REMBOLD, F.; SEGUINI, L.; VERHEGGHEN, A. Comparing land surface phenology of major european crops as derived from SAR and multispectral data of Sentinel-1 and -2. **Remote Sensing of Environment**, v. 253, p. 112232, 2021. ISSN 0034-4257. Available from: <https://www.sciencedirect.com/science/article/pii/S0034425720306052>. 49, 50

MILLARD, K.; RICHARDSON, M. On the importance of training data sample selection in random forest image classification: a case study in peatland ecosystem mapping. **Remote Sensing**, v. 7, n. 7, p. 8489–8515, 2015. 49

MOUNTRAKIS, G.; IM, J.; OGOLE, C. Support vector machines in remote sensing: a review. **ISPRS Journal of Photogrammetry and Remote Sensing**, v. 66, n. 3, p. 247–259, 2011. 21

NATARAJAN, N.; DHILLON, I. S.; RAVIKUMAR, P. K.; TEWARI, A. Learning with noisy labels. In: INTERNATIONAL CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS. **Proceedings...** [S.l.], 2013. p. 1196–1204. 23

NATITA, W.; WIBOONSAK, W.; DUSADEE, S. 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**, v. 6, n. 1, 2016. 30, 56

NATIVI, S.; MAZZETTI, P.; CRAGLIA, M. A view-based model of data-cube to support big earth data systems interoperability. **Big Earth Data**, v. 1, n. 1-2, p. 75–99, 2017. 2, 9

OLOFSSON, P.; FOODY, G. M.; HEROLD, M.; STEHMAN, S. V.; WOODCOCK, C. E.; WULDER, M. A. Good practices for estimating area and assessing accuracy of land change. **Remote Sensing of Environment**, v. 148, p. 42–57, 2014. 4, 23, 49

PAPARRIZOS, J.; GRAVANO, L. k-shape: efficient and accurate clustering of time series. In: INTERNATIONAL CONFERENCE ON MANAGEMENT OF DATA. **Proceedings...** New York, NY, USA: ACM, 2015. p. 1855–1870. 50

PARENTE, L.; MESQUITA, V.; MIZIARA, F.; BAUMANN, L.; FERREIRA, L. 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 Sensing of Environment**, v. 232, p. 111301, oct. 2019. ISSN 0034-4257. 21

PASQUARELLA, V. J.; HOLDEN, C. E.; KAUFMAN, L.; WOODCOCK, C. E. 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**, v. 2, n. 3, p. 152–170, 2016. ISSN 2056-3485. 1, 9, 21, 49

PATRINI, G.; ROZZA, A.; MENON, A. K.; NOCK, R.; QU, L. Making deep neural networks robust to label noise: a loss correction approach. In: IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION. **Proceedings...** Honolulu, HI, USA, 2017. p. 1944–1952. 23

PELLETIER, C.; VALERO, S.; INGLADA, J.; CHAMPION, N.; SICRE, C. M.; DEDIEU, G. Effect of training class label noise on classification performances for land cover mapping with satellite image time series. **Remote Sensing**, v. 9, n. 2, p. 173, feb. 2017. 3, 4, 22, 23, 27, 49, 59

PENGRA, B. W.; STEHMAN, S. V.; HORTON, J. A.; DOCKTER, D. J.; SCHROEDER, T. A.; YANG, Z.; COHEN, W. B.; HEALEY, S. P.; LOVELAND, T. R. Quality control and assessment of interpreter consistency of annual land cover reference data in an operational national monitoring program. **Remote Sensing of Environment**, v. 238, p. 111261, 2020. 4, 49

PETITJEAN, F.; INGLADA, J.; GANCARSKI, P. Satellite image time series analysis under time warping. **IEEE Transactions on Geoscience and Remote Sensing**, v. 50, n. 8, p. 3081–3095, 2012. ISSN 0196-2892. 21

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. Big earth observation time series analysis for monitoring Brazilian agriculture. **ISPRS Journal of Photogrammetry and Remote Sensing**, v. 145, p. 328–339, 2018. 3, 10, 15, 21, 71

QI, J.; LIU, H.; LIU, X.; ZHANG, Y. Spatiotemporal evolution analysis of time-series land use change using self-organizing map to examine the zoning and scale effects. **Computers, Environment and Urban Systems**, v. 76, p. 11 – 23, 2019. ISSN 0198-9715. Available from: <http://www.sciencedirect.com/science/article/pii/S0198971518301923>. 5, 50

RADOUX, J.; LAMARCHE, C.; BOGAERT, E. V.; BONTEMPS, S.; BROCKMANN, C.; DEFOURNY, P. Automated training sample extraction for global land cover mapping. **Remote Sensing**, v. 6, n. 5, p. 3965–3987, 2014. 4, 49

REBBAPRAGADA, U.; BRODLEY, C. E. Class noise mitigation through instance weighting. In: EUROPEAN CONFERENCE ON MACHINE LEARNING. **Proceedings...** Berlin, Heidelberg, Germany: Springer, 2007. p. 708–715. 22

RIBEIRO, J. F.; WALTER, B. M. T. As principais fitofisionomias do Bioma Cerrado. In: SANO, S. M.; ALMEIDA, S. P.; RIBEIRO, J. F. (Ed.). **Cerrado: Ecologia e Flora**. [S.l.]: Embrapa, 2008. p. 152–212. ISBN 9788573833973. 25, 39, 40

RIBEIRO, L.; TABARELLI, M. A structural gradient in cerrado vegetation of Brazil: changes in woody plant density, species richness, life history and plant composition. **Journal of Tropical Ecology**, v. 18, n. 5, p. 775–794, sep. 2002. ISSN 0266-4674, 1469-7831. 40

SANCHES, I. D.; FEITOSA, R. Q.; ACHANCCARAY, P.; MONTIBELLER, B.; LUIZ, A. J. B.; SOARES, M. D.; PRUDENTE, V. H. R.; VIEIRA, D. C.; MAURANO, L. E. P. Lem benchmark database for tropical agricultural remote sensing application. **ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences**, v. 42, p. 387–392, 2018. Available from: <https://www.int-arch-photogramm-remote-sens-spatial-inf-sci.net/XLII-1/387/2018/>. 63, 66

SANTOS, L. **Source code for: quality control and class noise reduction**. Zenodo, jul. 2020. Available from: <https://doi.org/10.5281/zenodo.3941278>. 6

SANTOS, L.; FERREIRA, K.; PICOLI, M.; CAMARA, G. Self-organizing maps in earth observation data cubes analysis. In: VELLIDO, A.; GIBERT, K.; ANGULO, C.; MARTIN, J. (Ed.). **Advances in Self-Organizing Maps, Learning Vector Quantization, Clustering and Data Visualization**. [S.l.]: Springer International Publishing, 2019. p. 70–79. 60

SIMOES, R.; PICOLI, M. C. A.; CAMARA, G.; MACIEL, A.; SANTOS, L.; ANDRADE, P. R.; SÁNCHEZ, A.; FERREIRA, K.; CARVALHO, A. Land use and cover maps for Mato Grosso State in Brazil from 2001 to 2017. **Scientific Data**, v. 7, n. 1, p. 34, dec. 2020. ISSN 2052-4463. 3, 4, 5, 21, 49

SOILLE, P.; BURGER, A.; MARCHI, D. D.; KEMPENEERS, P.; RODRIGUEZ, D.; SYRRIS, V.; VASILEV, V. A versatile data-intensive computing platform for information retrieval from big geospatial data. **Future Generation Computer Systems**, v. 81, p. 30–40, apr. 2018. ISSN 0167-739X. 9

SOLANO-CORREA, Y. T.; BOVOLO, F.; BRUZZONE, L. A semi-supervised crop-type classification based on Sentinel-2 NDVI satellite image time series and phenological parameters. In: IEEE INTERNATIONAL GEOSCIENCE AND REMOTE SENSING SYMPOSIUM. **Proceedings...** Yokohama, Japan, 2019. p. 457–460. 49, 50

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. Expanding the soy moratorium to Brazil's Cerrado. **Science Advances**, v. 5, n. 7, p. eaav7336, jul. 2019. ISSN 2375-2548. 24, 51

SPERA, S. A.; COHN, A. S.; VANWEY, L. K.; MUSTARD, J. F.; RUDORFF, B. F.; RISSO, J.; ADAMI, M. Recent cropping frequency, expansion, and abandonment in Mato Grosso, Brazil had selective land characteristics. **Environmental Research Letters**, v. 9, n. 6, p. 064010, 2014. 3

STRASSBURG, B. B. N.; BROOKS, T.; FELTRAN-BARBIERI, R.; IRIBARREM, A.; CROUZEILLES, R.; LOYOLA, R.; LATAWIEC, A. E.; FILHO, F. J. B. O.; SCARAMUZZA, C. A. M.; SCARANO, F. R.; SOARES-FILHO, B.; BALMFORD,

A. Moment of truth for the Cerrado hotspot. **Nature Ecology & Evolution**, v. 1, n. 4, p. 1–3, 2017. 24

SUEPA, T.; QI, J.; LAWAWIROJWONG, S.; MESSINA, J. P. Understanding spatio-temporal variation of vegetation phenology and rainfall seasonality in the monsoon southeast Asia. **Environmental Research**, v. 147, p. 621–629, 2016. 5, 27

SUN, J.; F.ZHAO; WANG, C.; CHEN, S. Identifying and correcting mislabeled training instances. In: FUTURE GENERATION COMMUNICATION AND NETWORKING. **Proceedings...** [S.l.]: IEEE, 2007. v. 1, p. 244–250. 22

TAN, H. S.; GEORGE, S. E. Investigating learning parameters in a standard 2d som model to select good maps and avoid poor ones. In: AUSTRALIAN JOINT CONFERENCE ON ADVANCES IN ARTIFICIAL INTELLIGENCE. **Proceedings...** Berlin, Heidelberg: Springer-Verlag, 2004. p. 425–437. ISBN 3540240594. Available from: <https://doi.org/10.1007/978-3-540-30549-1_38>. 36

TUIA, D.; VOLPI, M.; COPA, L.; KANEVSKI, M.; MUNOZ-MARI, J. A survey of active learning algorithms for supervised remote sensing image classification. **IEEE Journal of Selected Topics in Signal Processing**, v. 5, n. 3, p. 606–617, 2011. 4, 49

UDELHOVEN, T.; STELLMES, M.; RODER, A. Assessing rainfall-evi relationships in the okavango catchment employing modis time series data and distributed lag models. In: _____. **Remote sensing time series: revealing land surface dynamics**. [S.l.]: Springer, Cham, 2015. v. 22, p. 225–245. 12, 79

VESANTO, J.; ALHONIEMI, E. Clustering of the self-organizing map. **IEEE Transactions on Neural Networks**, v. 11, n. 3, p. 586–600, 2000. 36, 60

VIANA, C. M.; GIRãO, I.; ROCHA, J. Long-term satellite image time-series for land use/land cover change detection using refined open source data in a rural region. **Remote Sensing**, v. 11, n. 9, 2019. ISSN 2072-4292. Available from: <https://www.mdpi.com/2072-4292/11/9/1104>. 4, 49, 50

WANG, S.; AZZARI, G.; LOBELL, D. B. Crop type mapping without field-level labels: random forest transfer and unsupervised clustering techniques. **Remote Sensing of Environment**, v. 222, p. 303–317, 2019. ISSN 0034-4257. Available from: <https://www.sciencedirect.com/science/article/pii/S0034425718305790>. 50

WEHRENS, R.; BUYDENS, L. Self-and super-organizing maps in R: the kohonen package. **Journal of Statistical Software**, v. 21, n. 5, 2007. 13, 18

WOODCOCK, C. E.; LOVELAND, T. R.; HEROLD, M.; BAUER, M. E. Transitioning from change detection to monitoring with remote sensing: a paradigm shift. **Remote Sensing of Environment**, v. 238, p. 111558, mar. 2020. ISSN 00344257. 21, 49

WU, X.; CHENG, C.; ZURITA-MILLA, R.; SONG, C. An overview of clustering methods for geo-referenced time series: from one-way clustering to co- and tri-clustering. **International Journal of Geographical Information Science**, v. 34, n. 9, p. 1822–1848, 2020. Available from: <https://doi.org/10.1080/13658816.2020.1726922>. 53

XIONG, J.; THENKABAIL, P. S.; GUMMA, M. K.; TELUGUNTLA, P.; POEHNELT, J.; CONGALTON, R. G.; YADAV, K.; THAU, D. Automated cropland mapping of continental africa using google earth engine cloud computing. **ISPRS Journal of Photogrammetry and Remote Sensing**, v. 126, p. 225–244, 2017. ISSN 0924-2716. Available from: <https://www.sciencedirect.com/science/article/pii/S0924271616301575>. 50

ZHANG, X.; FRIEDL, M. A.; SCHAAF, C. B.; STRAHLER, A. H.; HODGES, J. C. F.; GAO, F.; REED, B. C.; HUETE, A. Monitoring vegetation phenology using MODIS. **Remote Sensing of Environment**, v. 84, n. 3, p. 471–475, 2003. 12, 21

ZHU, X.; WU, X. Class noise vs. attribute noise: a quantitative study. **Artificial Intelligence Review**, v. 22, n. 3, p. 177–210, 2004. 3

_____. Class noise vs. attribute noise: A quantitative study. **Artificial Intelligence Review**, v. 22, n. 3, p. 177–210, 2004. 22

ZITO, R. K.; FILHO, O. L. M.; PEREIRA, M. J. Z.; MEYER, M. C.; HIROSE, E.; NICOLI, C. M. L.; COSTA, S. V. d.; MEDEIROS NETO, C. D.; NUNES JUNIOR, J.; VIEIRA, N. E.; SEII, A. H.; MULLICH, J. R.; PIMENTA, C. B.; SANCHEZ, I.; MOREIRA, A. J. A.; NUNES, M. R.; DESSIMONE, M. G. L.; JUNIOR, O. P. d. M.; NEIVA, L. C. d. S.; BARROS, A. C. d.; FILHO, R. S. **Cultivares de soja: macrorregiões 3, 4 e 5 Goiás e Região Central do Brasil**. Londrina: Embrapa Soja, 2018. Available from: <https://www.embrapa.br/en/busca-de-publicacoes/-/publicacao/1067791/cultivares-de-soja-macrorregioes-3-4-e-5-goias-e-regiao-central-do-brasil>. Access in: 15 July 2020. 64

ZURITA-MILLA, R.; GIJSEL, J. V.; HAMM, N. A.; AUGUSTIJN, P.; VRIELING, A. Exploring spatiotemporal phenological patterns and trajectories using self-organizing maps. **IEEE Transactions on Geoscience and Remote sensing**, v. 51, n. 4, p. 1914–1921, 2012. 50

ANNEX A - EVALUATING DISTANCE MEASURES FOR IMAGE TIME SERIES CLUSTERING IN LAND USE AND COVER MONITORING

This annex presents a paper published in the European Conference on Machine Learning and Principles and Pratice of Knowledge Discovery in Databases (ECML PKDD) in Würzburg, Germany on September 2019 (FERREIRA et al., 2019):

Evaluating distance measures for image times series clustering in land use and cover monitoring

Karine Reis Ferreira Lorena Santos and Michelle C. A. Picoli

Brazilian National Institute for Space Research (INPE)
São José dos Campos, São Paulo, Brazil

Evaluating distance measures for image time series clustering in land use and cover monitoring

Karine Reis Ferreira[1], Lorena Santos[1], and Michelle C. A. Picoli[1]

Brazilian National Institute for Space Research (INPE)
São José dos Campos, São Paulo, Brazil
karine.ferreira@inpe.br, lorena.santos@inpe.br, michelle.picoli@inpe.br

Abstract. Time series derived from Earth observation satellite images have been widely used for land use and cover classification and change detection. Clustering is a common technique performed to discovery intrinsic patterns on time series data sets, by grouping similar time series together based on a certain similarity measure. This short paper describes an ongoing work on evaluating distance measures for remote sensing image time series clustering using Self-Organizing Maps (SOM), specifically to land use and cover monitoring. We present an experiment to evaluate three similarity measures, Dynamic Time Warping (DTW), Euclidean (ED) and Manhattan (MD). In this experiment, we show that ED and ED are more accurate than DTW for remote sensing image time series clustering in land use and cover application.

Keywords: remote sensing image time series, time series clustering, similarity measures, land use and cover monitoring

1 Introduction

Nowadays, the big amount of Earth observation satellite images freely available has motivated the use of time series analysis for land use and cover classification and change detection [3]. Time series derived from remote sensing images have been widely used for detecting agricultural intensification [8], forest disturbance [4], ecological dynamics [7], and phenological change detection [10].

Clustering is a common technique performed to discovery intrinsic patterns on time series data sets [1]. Time series clustering is a unsupervised method that groups similar time series together into homogeneous collections based on a certain similarity measure. According to Ding et al. [2], the similarity measure is a key aspect for achieving effectiveness in time series analysis. Time series represent sequences of values ordered over time. Thus, the distance between time series needs to be carefully defined in order to reflect the fundamental similarity of these sequences. Ding et al. [2] evaluated 9 similarity measures and their variants, testing their effectiveness on 38 time series data sets from different application domains, and concluded that on small data sets, *elastic* measures, e.g. Dynamic Time Warping (DTW), can be significantly more accurate than L_n-*norm*, e.g. Euclidean and Manhattan distances.

This paper presents an ongoing work on evaluating similarity measures for remote sensing image time series clustering using the Self-Organizing Maps (SOM) neural network [6]. In a previous work, we describe the use of SOM method with Euclidean distance to assess land use and cover samples and to evaluate which time series of spectral bands and vegetation indexes are best suitable for the separability of land use and cover classes [9]. However, more studies are necessary to evaluate which distance measure has the best accuracy for clustering such time series using SOM. Thus, in this work, we analyse the SOM method with three distinct distance measures, the Manhattan distance (MD), the Euclidean Distance (ED) and the *elastic* measure DTW. Differently from Ding et al. [2], our experiment shows that ED and MD distances are more accurate than DTW for remote sensing image time series clustering in land use and cover application.

2 Similarity measures for time series

Distance metrics aid to identify how the data is similar or dissimilar with each other. Given two time series $x = [x_1, \ldots, x_i, \ldots, x_n]$ and $y = [y_1, \ldots, y_i, \ldots, y_n]$, the Euclidean distance (ED) between these two time series is:

$$ED = \sqrt{\sum_{i=1}^{N} (x_i - y_i)^2} \tag{1}$$

The Manhattan distance (MD) between these two time series is:

$$MD = \sum_{i=1}^{N} |x_i - y_i| \tag{2}$$

The *elastic* DTW measure aligns similar sequences in time series that match even if they are out of phase in the time axis [5]. The first step of DTW is to compute a cost matrix Ψ, $n \times n$, given by the squared distance between each point of the two time series:

$$\Psi_{i,j} = (x_i - y_j)^2. \tag{3}$$

From Ψ,the best matching between two time series can be found, and the an optimal path that minimizes the cost warping is obtained. The warping path is a contiguous set of matrix elements that defines a mapping between the time series:

$$d_{i,j} = \Psi_{i,j} + \min \begin{cases} d_{i-1,j} \\ d_{i-1,j-1} \\ d_{i,j-1} \end{cases} \tag{4}$$

Figure 1 shows two time series X and Y. The distance measures between these two time series are: DTW = 0, ED = 4.242 and MD = 6. We can note that DTW measure considers that, even though the time series are out of phase in

time axis, X and Y are matching and the distance between them is zero (DTW = 0). On the other hand, ED and MD distances reveal a significant difference between these two time series.

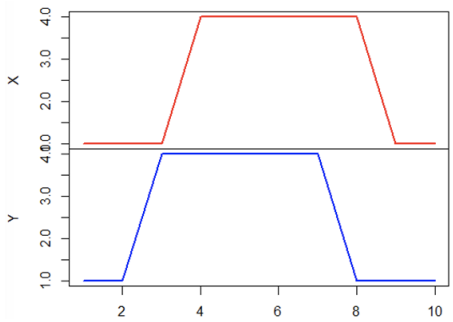


Fig. 1. Time series: X = {1,1,1,4,4,4,4,4,1,1} and Y = {1,1,4,4,4,4,4,1,1,1}. The distance measures between X and Y are: DTW = 0; ED = 4.242 and MD = 6.

3 Experiment and results

In this work, we performed an experiment using 2115 ground samples located in the Mato Grosso state, Brazil, as shown in Figure 2. These samples are divided in nine land cover classes: (1) Forest, (2) Cerrado, (3) Pasture, (4) Soybean-fallow, (5) Fallow-cotton, (6) Soybean-cotton, (7) Soybean-corn, (8) Soybean-millet, and (9) Soybean-sunflower. For each sample, we extracted six time series associated to its location from the MODIS sensor images (MOD13Q1 product) of NASA, provided every 16 days at 250-meter spatial resolution. The six time series are the original spectral bands (1) BLUE, (2) RED, (3) Near-Infrared (NIR), and (4) Mid-Infrared (MIR), and the vegetation indexes (5) Normalized Difference Vegetation Index (NDVI) and (6) Enhanced Vegetation Index (EVI).

The main goal of this study is to evaluate which distance measure, ED, MD or DTW, has the best accuracy for clustering the time series of the nine land cover classes using the SOM method. SOM is an unsupervised neural network suitable for time series clustering [6] [1]. It allows mapping from a high-dimensional space to a low-dimensional space, preserving the data topology while reducing computational cost. It is composed by input and output layers, where the input layer is the sample data to be clustered and the output layer is a set of neurons.

To evaluate the separability of the clusters, we performed SOM for each distance metric combining different time series of spectral bands and vegetation indices. We tested three combinations: (1) Case I: NVDI and EVI; (2) Case II: NDVI, EVI, NIR and MIR; (3) Case III: NDVI, EVI, NIR, MIR, RED and BLUE.

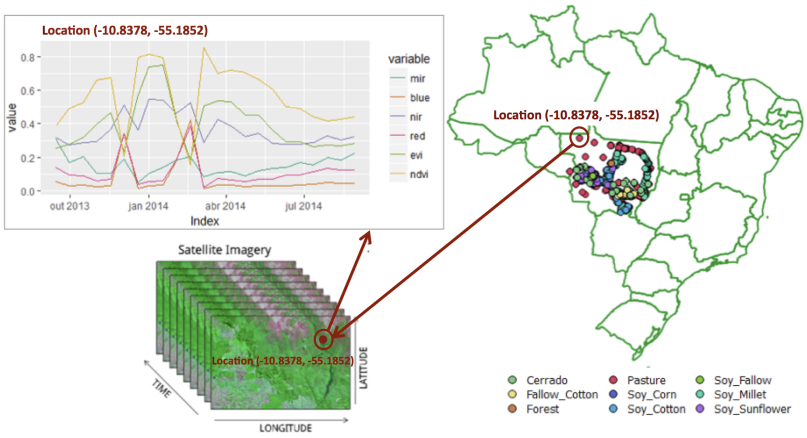


Fig. 2. Remote sensing image time series associated to land use and cover ground samples in Mato Grosso state, Brazil.

Figure 3 shows the spectral-temporal patterns and the amplitudes of the EVI index time series of the ground samples of four land use and cover ground samples: soybean-corn, soybean-cotton, soybean-sunflower and soybean-millet. It shows the spectral-temporal patterns of these time series during the Brazilian crop season that begins at September and spans to August of the next year. We can observe that these agricultural crops have a very similar spectral-temporal response. This similarly is due to the plant's own phenology and to the agricultural calendar of the state of Mato Grosso that relates the planting periods to the rainy season, and the harvest periods with the dry season.

Table 1 presents the cluster accuracy generated by SOM for each distance measure and for each set of time series (Cases I, II and II). To generate these clusters, we created a 2D SOM grid of neurons and initialized their weight vectors randomly. The SOM parameters that we used were: grid size $= 25 \times 25$, learning rate $= 1$, and number of iterations $= 100$. Then, for each time series, the algorithm finds the 2D grid neuron which has the smallest distance to the time series, based on its weight vector. After the match, the neuron's weight vector and those of its neighbors are then updated. After all time series are associated with neurons, each neuron is labelled using a majority vote, taking the most frequent class from the time series associated with it. A neuron labelled as class X is part of the cluster X. The accuracy of the cluster X is calculated based on the percentage of time series associated to the class X in neurons labelled as class X.

We can observe in Table 1 that the best general accuracy is 93% generated by the Euclidean and Manhattan distances both in Case II, using the time series NDVI, EVI, NIR and MIR. Because the spectro-temporal patters of the crops are very similar, as shown in Figure 3, DTW distance can not distinguish them well and so can not produce clusters with good accuracy.

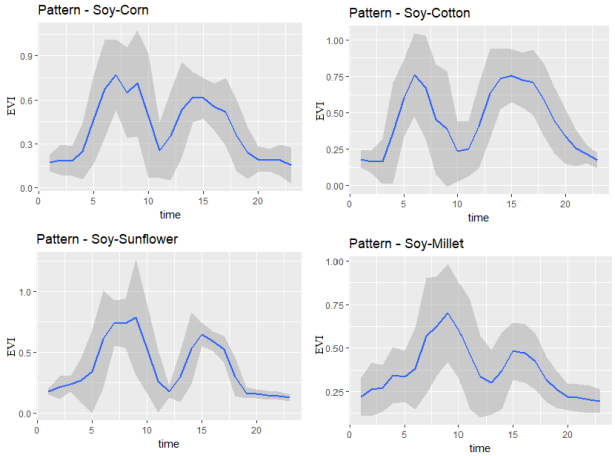


Fig. 3. Spectro-temporal patterns of the EVI index time series.

Table 1. Cluster accuracy for each distance measure and for each case (I, II and III)

	Euclidean			Manhattan			DTW		
	I	II	III	I	II	III	I	II	III
Cerrado	84	97.3	93.3	92.8	97.2	95.6	88	97.5	98.0
Fallow-Cotton	72.2	85.7	78.9	69	80.9	73.9	66	73.68	76.1
Forest	100	99.3	89.9	99.2	99.2	97.1	98	98.5	96.5
Pasture	92.7	97.3	93.7	94.9	95.9	96.9	92.1	98.9	98.9
Soy-Corn	82.0	84.0	85.4	84.6	84.9	86.5	70.1	74.3	80.2
Soy-Cotton	94.6	95.5	93.5	95.45	97.3	96.82	74.8	85.7	92.0
Soy-Fallow	97.8	100	98.9	100	97.7	100	73.6	88.8	98.9
Soy-Millet	85.5	90.3	88.2	87.5	92.8	88.5	72.2	85.1	100
Soy-Sunflower	77.1	76.9	72.9	73.21	86.0	75.9	-	50	63.2
Accuracy	**88.1**	**93**	**90**	**91.2**	**93**	**92.9**	**80.8**	**88.5**	**91.1**

In Table 1, we can observe that DTW in Case I can not distinguish the crop Soy-Sunflower from the others. That is, it is not able to create a group or cluster to represent the Soy-Sunflower crop. The confusion matrix of DTW in Case I is shown in Table 2 where we can observe the confusion between the classes Soy-Sunflower (9) and Soy-Corn (5). In this case, the majority of time series of the Soy-Sunflower class is in the cluster of the class Soy-Corn class.

To perform this experiment, we used the Kohonen R package [11] and extended it with the DTW distance. The experiment presented in this work shows that Euclidean and Manhattan distances are more accurate than DTW for remote sensing image time series clustering in land use and cover application.

Table 2. Confusion Matrix - Case I - DTW

	1	2	3	4	5	6	7	8	9
Cerrado	379	0	1	20	0	0	0	0	0
Fallow_Cotton	0	2	0	0	3	19	5	5	0
Forest	8	0	130	0	0	0	0	0	0
Pasture	36	1	1	330	0	0	0	2	0
Soy_Corn	1	0	0	3	289	73	2	30	0
Soy_Cotton	0	0	0	1	33	343	15	7	0
Soy_Fallow	0	0	0	0	0	0	81	7	0
Soy_Millet	2	0	0	4	51	16	6	156	0
Soy_Sunflower	0	0	0	0	36	7	1	9	0

References

1. Aghabozorgi, S., Shirkhorshidi, A.S., Wah, T.Y.: Time-series clustering–a decade review. Information Systems **53**, 16–38 (2015)
2. Ding, H., Trajcevski, G., Scheuermann, P., Wang, X., Keogh, E.: Querying and mining of time series data: experimental comparison of representations and distance measures. Proceedings of the VLDB Endowment **1**(2), 1542–1552 (2008)
3. Gomez, C., White, J.C., Wulder, M.A.: Optical remotely sensed time series data for land cover classification: A review. ISPRS Journal of Photogrammetry and Remote Sensing **116**, 55–72 (2016)
4. Kennedy, R.E., Yang, Z., Cohen, W.B.: Detecting trends in forest disturbance and recovery using yearly Landsat time series. Remote Sensing of Environment **114**(12), 2897–2910 (2010)
5. Keogh, E., Ratanamahatana, C.A.: Exact indexing of dynamic time warping. Knowledge and Information Systems **7**(3), 358–386 (2005)
6. Kohonen, T.: The self-organizing map. Proceedings of the IEEE **78**(9), 1464–1480 (1990)
7. Pasquarella, V.J., Holden, C.E., Kaufman, L., Woodcock, C.E.: 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**(3), 152–170 (2016)
8. Picoli, M., Camara, G., Sanches, I., Simoes, R., Carvalho, A., Maciel, A., Coutinho, A., Esquerdo, J., Antunes, J., Begotti, R., Arvor, D., Almeida, C.: Big earth observation time series analysis for monitoring brazilian agriculture. ISPRS Journal of Photogrammetry and Remote Sensing **145**, 328 – 339 (2018)
9. Santos, L.A., Ferreira, K.R., Picoli, M., Camara, G.: Self-organizing maps in earth observation data cubes analysis. International Workshop on Self-Organizing Maps pp. 70–79 (2019)
10. Verbesselt, J., Hyndman, R., Zeileis, A., Culvenor, D.: Phenological change detection while accounting for abrupt and gradual trends in satellite image time series. Remote Sensing of Environment **114**(12), 2970 – 2980 (2010)
11. Wehrens, R., Buydens, L.M., et al.: Self-and super-organizing maps in r: the kohonen package. Journal of Statistical Software **21**(5), 1–19 (2007)

PUBLICAÇÕES TÉCNICO-CIENTÍFICAS EDITADAS PELO INPE

Teses e Dissertações (TDI)

Teses e Dissertações apresentadas nos Cursos de Pós-Graduação do INPE.

Manuais Técnicos (MAN)

São publicações de caráter técnico que incluem normas, procedimentos, instruções e orientações.

Notas Técnico-Científicas (NTC)

Incluem resultados preliminares de pesquisa, descrição de equipamentos, descrição e ou documentação de programas de computador, descrição de sistemas e experimentos, apresentação de testes, dados, atlas, e documentação de projetos de engenharia.

Relatórios de Pesquisa (RPQ)

Reportam resultados ou progressos de pesquisas tanto de natureza técnica quanto científica, cujo nível seja compatível com o de uma publicação em periódico nacional ou internacional.

Propostas e Relatórios de Projetos (PRP)

São propostas de projetos técnico-científicos e relatórios de acompanhamento de projetos, atividades e convênios.

Publicações Didáticas (PUD)

Incluem apostilas, notas de aula e manuais didáticos.

Publicações Seriadas

São os seriados técnico-científicos: boletins, periódicos, anuários e anais de eventos (simpósios e congressos). Constam destas publicações o Internacional Standard Serial Number (ISSN), que é um código único e definitivo para identificação de títulos de seriados.

Programas de Computador (PDC)

São a seqüência de instruções ou códigos, expressos em uma linguagem de programação compilada ou interpretada, a ser executada por um computador para alcançar um determinado objetivo. Aceitam-se tanto programas fonte quanto os executáveis.

Pré-publicações (PRE)

Todos os artigos publicados em periódicos, anais e como capítulos de livros.