Self-Organizing Maps in Earth Observation Data Cubes Analysis

Lorena Santos(✉), Karine Reis Ferreira, Michelle Picoli, and Gilberto Camara

National Institute for Space Research (INPE), São Jose dos Campos, Brazil
{lorena.santos,karine.ferreira,gilberto.camara}@inpe.br,
mipicoli@gmail.com

Abstract. Earth Observation (EO) Data Cubes infrastructures model analysis-ready data generated from remote sensing images as multidimensional cubes (space, time and properties), especially for satellite image time series analysis. These infrastructures take advantage of big data technologies and methods to store, process and analyze the big amount of Earth observation satellite images freely available nowadays. Recently, EO Data Cubes infrastructures and satellite image time series analysis have brought new opportunities and challenges for the Land Use and Cover Change (LUCC) monitoring over large areas. LUCC have caused a great impact on tropical ecosystems, increasing global greenhouse gases emissions and reducing the planet's biodiversity. This paper presents the utility of Self-Organizing Maps (SOM) neural network method in the process to extract LUCC information from EO Data Cubes infrastructures, using image time series analysis. Most classification techniques to create LUCC maps from satellite image time series are based on supervised learning methods. In this context, SOM is used as a method to assess land use and cover samples and to 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 this work and shows the potential of SOM in this application.

Keywords: Self-Organizing Maps · Earth Observation Data Cubes Analysis · Satellite image time series · Land Use and Cover Changes

1 Introduction

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 [1]. 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 to 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 to

A. Vellido et al. (Eds.): WSOM 2019, AISC 976, pp. 70–79, 2020.
https://doi.org/10.1007/978-3-030-19642-4_7

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) [2], the Joint Research Centre (JRC) Earth Observation Data and Processing Platform (JEODPP) [3] and the System for Earth Observation Data Access, Processing and Analysis for Land Monitoring (SEPAL) [4].

A typical application that benefits from EO Data Cubes infrastructures and satellite image time series analysis is LUCC monitoring. Characterizing and mapping changes in land surface is essential for planing and managing natural resources. The growing pressures for food and energy production promoted by increasing population make humans modify the Earth's environment in a rapid pace. LUCC can affect hydrological and biological process causing great impacts on tropical ecosystems [6].

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 bitemporal or other traditional change detection approaches [6]. The use of remote sensing image time series analysis to produce LUCC information has increased greatly in the recent years [7].

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 *apriori*. 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 paper presents the utility of Self-Organizing Maps (SOM) neural network method in the process to extract LUCC maps from EO Data Cubes infrastructures. SOM is a clustering method suitable for time series data sets. This work describes the use of SOM in the training phase to produce metrics that indicate the quality of the land use and cover samples and to 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 this work and shows the potential of SOM in this context.

In the LUCC domain, SOM has not being widely exploited for image time series analysis. The good review provided by [7] cites [8] as the main reference in this context. However, [8] proposed an approach to classify land cover from MODIS EVI time series using SOM. Besides that, [9] proposes 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, [8] and [9], propose the use of SOM to classify one agricultural year using only EVI attribute. Differently from them, our proposal use SOM to explore the separability time series using several attributes in order to improve the classification.

2 Land Use and Cover Change Information from Earth Observation Data Cubes

This section describes the process, illustrated in Fig. 1, to extract LUCC information from Earth Observation Data Cubes using image time series analysis and the utility of SOM method in this process.

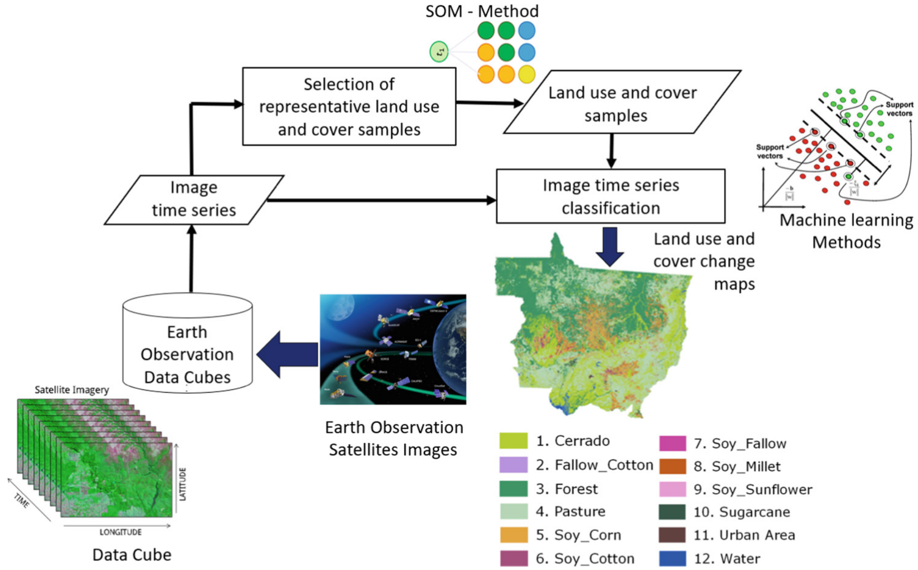


Fig. 1. LUCC information from EO Data Cubes

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 [12]. Most of these methods are based on supervised learning methods which require a training phase using land use and cover samples labeled *apriori*.

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

2.1 Earth Observation Satellite Image Time Series

Remote sensing satellite revisit the same place on Earth during their life cycle. The measure of same place can be obtained in different times. These measures are mapped to three-dimensional array in space-time [10], as shown in Fig. 2(a). The time series is made from values obtained of each pixel location $I(x, y)$ over time, as presented in Fig. 2(b). From these time series, LUCC can be extracted through vegetation phenology. Figure 2(b) shows an example of an area that was covered by forest during 2000 to 2001 the area, then it was deforested and during three years it was maintained as pasture. From 2006 to 2008 it was used for crop production.

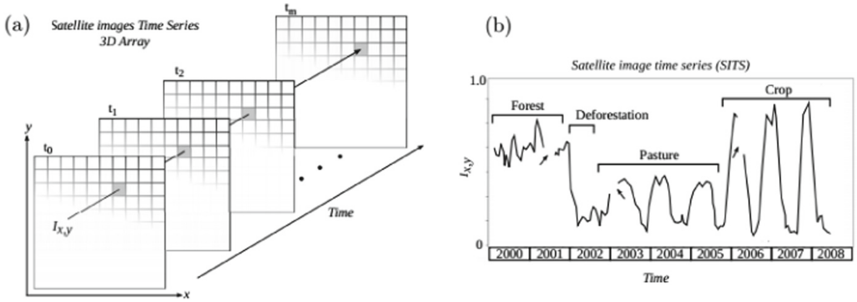


Fig. 2. (a) A dimensional array of satellite images, (b) vegetation index time series at pixel location (x, y). Source: [10]

2.2 Vegetation Indexes

Vegetation phenology is a biological event that indicate the stages of growth and development of plants during the life cycles. Remote sensing satellites are becoming essential for remotely capturing phenological variations in large scale and extract phenological metrics from time series data of vegetation parameters. The most commonly parameters used are the vegetation indexes.

Vegetation Indexes (VI) derived from spectral information of Earth observation satellite images are widely used to generate LUCC information. Vegetation indices are spectral transformations of two or more bands designed to enhance vegetation properties. Two examples of most used vegetation indices are NDVI (Normalized Difference Vegetation Index) and EVI (Enhanced Vegetation Index). Some spatial agencies provide these vegetation indexes as products derived from their satellite images. An example is the product MOD13Q1 of MODIS (Moderate Resolution Imaging Spectroradiometer) sensor provided by NASA with temporal resolution of 16 days and spatial resolution of 250 m [11].

During plant growth periods, different vegetation styles can be distinguished by time-series vegetation indexes [13]. 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 [13]. 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 [14]. 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 [11]. The two vegetation indices complement each other in global vegetation studies.

2.3 Using SOM to Improve the Quality of Land Use and Cover Samples

In the process to extract 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 to 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.

SOM can deal with the variability of vegetation phenology better than other methods that do not have these two features. Due to climatic phenomena, the vegetation phenology can suffer variations over time. Phenological patterns can vary spatially across a region and are strongly correlated with climate variations over time [5]. For example, rainy years may have a pattern for pasture different from a non-rainy year. Therefore, it is necessary methods that can take into account these small variations. The SOM method is able to learn with new patterns during the training process. Besides that, the use of multiples attributes such as combined vegetation indices and multiples spectral bands can improve that patterns generated by SOM.

Instead of use one attribute as input for SOM, the super-organized maps were implemented by [15] in order to use several attributes in a separate layer during the network training. Considering a sample x with two attributes, x_{a1} and x_{a2}, for example the vegetation indices NDVI and EVI. For each attribute a output layer consisting of a 2-D grid of neurons is created. The layers A_1 and A_2 are associated with attributes a_1 and a_2 respectively. The weight vectors, ω_{a1} and ω_{a2}, for each attribute are initialized with random values as shown in Fig. 3. To find the Best Matching Unit (BMU) the distances between input vectors and weight vectors are computed separately for each layer and then they are summed in order to define an overall distance of an sample to a neuron. The Eq. 1 shows how to calculate the distance for multiple layers.

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

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

After the training step of SOM, to evaluate the separability of samples is necessary to label the neurons in order 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

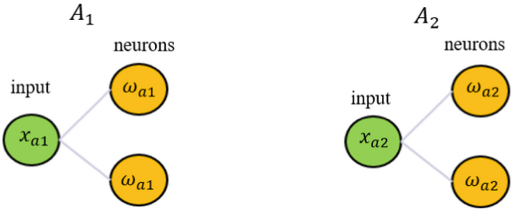


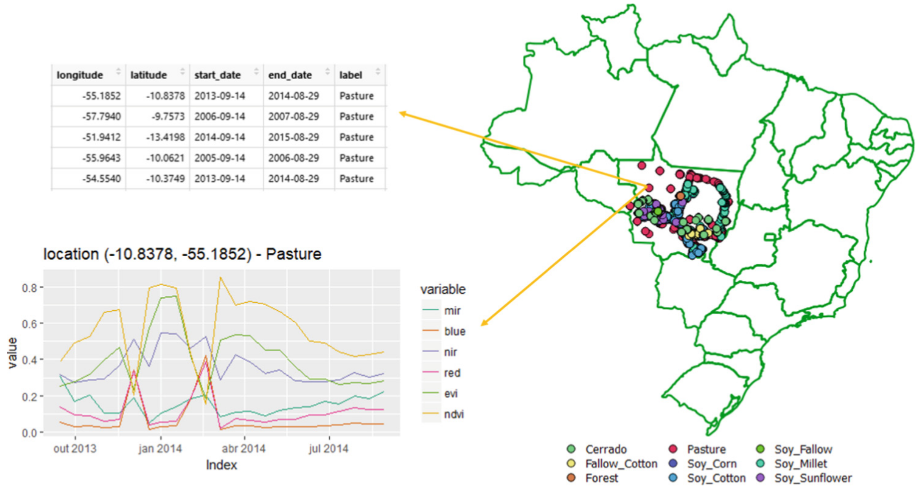
Fig. 3. Structure of SOM with two attributes

label of the majority of the samples associated to it. In some cases, no samples is associated to a neuron. Then, this empty neuron receives the label ‘Noclass’. To verify the quality of clusters generated by SOM, the confusion matrix can be accessed. From the confusion matrix, percentage of mixture within a cluster is calculated.

3 Case Study

To show the potential of SOM method in the selection of 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 area of study is the Mato Grosso State in Brazil, as shown in Fig. 4. Each sample has a spatial location (latitude and longitude), start and end date that corresponds to agricultural year (from August to September), the label of the class that corresponds to the sample, and the set of time series with multiple attributes. In this case study, we used the attributes EVI, NDVI, NIR, MIR, BLUE and RED. The ground samples include natural vegetation and agricultural classes for the Mato Grosso state of Brazil. The data set includes 2215 ground samples divided in 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 [12].

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 iteration = 100.



longitude	latitude	start_date	end_date	label
-55.1852	-10.8378	2013-09-14	2014-08-29	Pasture
-57.7940	-9.7573	2006-09-14	2007-08-29	Pasture
-51.9412	-13.4198	2014-09-14	2015-08-29	Pasture
-55.9643	-10.0621	2005-09-14	2006-08-29	Pasture
-54.5540	-10.3749	2013-09-14	2014-08-29	Pasture

Fig. 4. Samples dataset

Figure 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 phenological patterns similarity among the classes, SOM was able to separate these land use and cover classes with good accuracy. For the first case the accuracy was 88%, the second case was 93% and the third was 90%. Besides that, we can note that the most of neurons that belongs to a neighborhood are the same category, but the time series samples contains small variations.

Table 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

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 that were assigned to the right cluster, that is, the class associated to

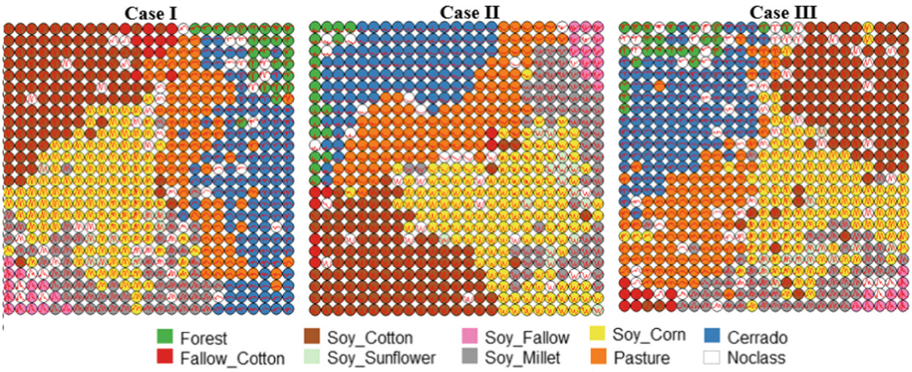


Fig. 5. Grids generated for each case

the sample is the same class associated to the cluster. For example, the Cerrado cluster has 97.3% of samples labelled as Cerrado in the Case II; 84% in the Case I and 93.3% in the Case III.

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

For the Case II, the confusion of each cluster is shown in Fig. 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 Fig. 5 where there are neurons labelled as Fallow-Cotton and Soy-Sunflower within the neighborhood of Soy-Corn. Some samples of Cerrado and Pasture have similar spectral curves but the attributes MIR and NIR reduced the confusion between these samples as shown in Fig. 6.

4 Final Remarks

This paper presents the utility of SOM method 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 to improve the quality of the land use and cover samples.

We present a case study that evaluates the combination of six attributes, EVI, NDVI, NIR, MIR, RED and BLUE, using MODIS time series of land use and cover 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.

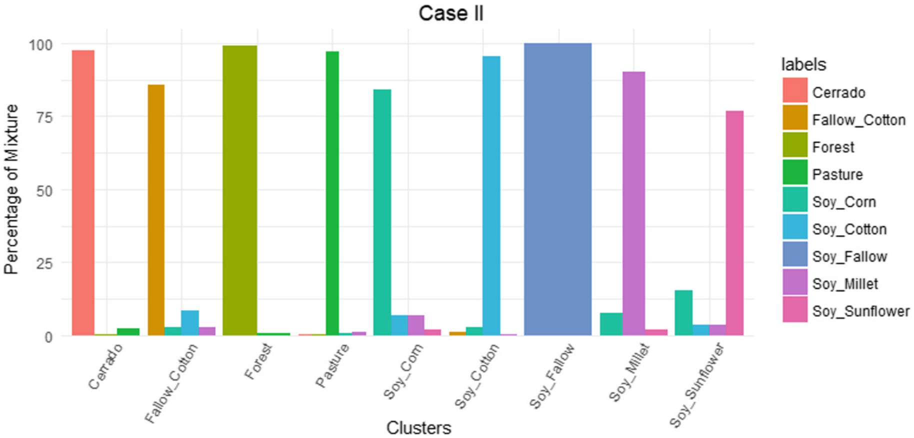


Fig. 6. Confusion among the classes

Despite the general accuracy of separability using only NDVI and EVI attributes was 88%, the classification in big scale areas can generate a great amount of 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 that were assigned to right clusters increased. A sample assigned to a right cluster means that the class associated to the sample is the same class associated to the cluster.

In the third case, when we used all attributes including BLUE and RED, the accuracy of separability was worse (90%) than the second case. This analysis is important because we can conclude that adding more attributes does not mean increasing the results. 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 leads to the best results.

Finally, we have implemented our approach in **R** using the SOM method available in the Kohonen R package [15]. This package has the online and batch approaches of SOM, however we use the online method for this work. The Kohonen package was integrated in the Satellite Image Time Series (sits) package [16]. The package sits package was developed in the e-sensing project developed by the Brazilian Institute for Space Research (INPE) in order to provide a set of tools for working with analyses, clustering and classification of satellite image time series. Its source code is available at https://github.com/e-sensing/sits.

References

1. Nativi S, Mazzetti P, Craglia M (2017) A view-based model of data-cube to support big Earth data systems interoperability. Big Earth Data 1:75–99
2. Lewis A, Oliver S, Lymburner L, Evans B, Wyborn L, Mueller N, Wu W (2017) The Australian geoscience data cube — foundations and lessons learned. Remote Sens Environ 202:276–292
3. Soille P, Burger A, De Marchi D, Kempeneers P, Rodriguez D, Syrris V, Vasilev V (2018) A versatile data-intensive computing platform for information retrieval from big geospatial data. Future Gener Comput Syst 81:30–40
4. FAO: Sepal repository (2018). https://github.com/openforis/sepal. Accessed 14 Dec 2018
5. Suepa T, Qi J, Lawawirojwong S, Messina P (2016) Understanding spatio-temporal variation of vegetation phenology and rainfall seasonality in the monsoon Southeast Asia. Environ Res 147:621–629
6. Pasquarella J, Holden E, Kaufman L, Woodcock E (2016) From imagery to ecology: leveraging time series of all available landsat observations to map and monitor ecosystem state and dynamics. Remote Sens Ecol Conserv 2(3):152–170
7. Gomez C, White C, Wulder A (2016) Optical remotely sensed time series data for land cover classification: a review. J Photogram Remote Sens 116:55–72
8. Bagan H, Wang Q, Watanabe M, Yang Y, Ma J (2005) Land cover classification from Modis EVI time-series data using SOM neural network. Int J Remote Sens 26:4999–5012
9. Siam L (2013) Soft supervised self-organizing mapping (3SOM) for improving land cover classification with MODIS time-series. PhD thesis, Michigan State University, Michigan
10. Maus V, Camara G, Cartaxo R, Sanchez A, Ramos M, Queiroz G (2016) A time-weighted dynamic time warping method for land-use and land-cover mapping. IEEE J Sel Top Appl Earth Observ Remote Sens 9(8):3729–3739
11. Huete A, Didan K, Miura T, Rodriguez E, Gao X, Fereira L (2002) Overview of the radiometric and biophysical performance of the MODIS vegetation indices. Remote Sens Environ 86:195–213
12. Picoli M, Camara G, Sanches I, Simoes R, Carvalho A, Maciel A, Coutinho A, Esquerdo J, Antunes J, Begotti R, Arvor D, Almeida C (2018) Big Earth observation time series analysis for monitoring Brazilian agriculture. ISPRS J Photogram Remote Sens 145:328–339
13. Boles H, Xiao X, Liu J, Zhang Q, Munktuya S, Chen S, Ojima D (2004) Land cover characterization of temperate East Asia using multi-temporal vegetation sensor data. Remote Sens Environ 90(4):477–489
14. Udelhoven T, Stellmes M, Rodes A (2015) Assessing rainfall-EVI relationships in the Okavango catchment employing MODIS time series data and distributed lag models. In: Revealing land surface dynamics. Remote sensing time series. Springer, Cham, pp 225–245
15. Wehrens R, Buydens L (2007) Self and super-organizing maps in R: the Kohonen package. J Stat Softw 21:1–19
16. Camara G, Simoes R, Andrade P, Maus V, Sanchez A, Assis L, Santos L, Ywata A, Maciel A, Vinhas L, Ferreira K, Queiroz G (2018) Sits e-sensing/sits: Version 1.12.5, December 2018. https://doi.org/10.5281/zenodo.1974065