# SITS - An R Package for Data Analysis and Machine Learning using Satellite Image Time Series

**Rolf Simoes** *National Institute for Space Reseach, INPE, Brazil*
**Gilberto Camara** *National Institute for Space Reseach, INPE, Brazil*
**Alexandre Carvalho** *Institute for Applied Economics Research (IPEA), Brazil*
**Victor Maus** *International Institute for Applied System Analysis (IIASA)*

Using time series derived from big Earth Observation data sets is one of the leading research trends in Land Use Science and Remote Sensing. One of the more promising uses of satellite time series is its application for classification of land use and land cover, since our growing demand for natural resources has caused major environmental impacts. The SITS package provides support on how to use statistical learning techniques with image time series. These methods include linear and quadratic discrimination analysis, support vector machines, random forests and neural networks.

## Introduction

Earth observation satellites provide a continuous and consistent set of information about the Earth's land and oceans. Most space agencies have adopted an open data policy, making unprecedented amounts of satellite data available for research and operational use. This data deluge has brought about a major challenge: *How to design and build technologies that allow the Earth observation community to analyse big data sets?*

The approach taken in the current package is to develop data analysis methods that work with satellite image time series (SITS). The time series are obtained by taking calibrated and comparable measures of the same location in Earth at different times. These measures can be obtained by a single sensor (e.g., MODIS) or by combining different sensores (e.g., LANDSAT-8 and SENTINEL-2). If obtained by frequent revisits, the temporal resolution of these data sets can capture the most important land use changes.

Time series of remote sensing data show that land cover changes do not always occur in a progressive and gradual way, but they may also show periods of rapid and abrupt change followed either by a quick recovery (Lambin et al. 2003). Analyses of multiyear time series of land surface attributes, their fine-scale spatial pattern, and their seasonal evolution leads to a broader view of land-cover change. Satellite image time series have already been applied to applications such as mapping for detecting forest disturbance (Kennedy et al. 2010), ecology dynamics (Pasquarella et al. 2016), agricultural intensification (Galford et al. 2008) and its impacts on deforestation (Arvor et al. 2012).

The SITS package provides support on how to use statistical learning techniques with image time series. In a broad sense, statistical learning refers to a class of algorithms for classification and regression analysis (Hastie et al. 2009). These methods include

linear and quadratic discrimination analysis, support vector machines, random forests and neural networks. In a typical classification problem, we have measures that capture class attributes. Based on these measures, referred as training data, one's task is to select a predictive model that allows inferring classes of a larger data set.

Current approaches to image time series analysis still use limited number of attributes. A common approach is deriving a small set of phenological parameters from vegetation indexes, like beginning, peak, and length of growing season (Brown et al. 2013) (Kastens et al. 2017) (Estel et al. 2015) (Pelletier et al. 2016). These phenological parameters are then fed in specialised classifiers such as TIMESAT (Jönsson & Eklundh 2004). These approaches do not use the power of advanced statistical learning techniques to work on high-dimensional spaces and with big training data sets (James et al. 2013).

The SITS package uses the full depth of satellite image time series to create larger dimensional spaces. We tested different methods of extracting attributes from time series data, including those reported by Pelletier et al. (2016) and Kastens et al. (2017). Our conclusion is that part of the information in raw time series is lost after filtering or statistical approximation. Thus, the method we developed has a deceptive simplicity: *use all the data available in the time series samples*. The idea is to have as many temporal attributes as possible, increasing the dimension of the classification space. Our experiments found out that modern statistical models such as support vector machines, and random forests perform better in high-dimensional spaces than in lower dimensional ones.

In what follows, we describe the main characteristics of the SITS package. The first part describes the basic data structures used in SITS and the tools used for visualisation and data exploration. Then we show how to do data acquision from external sources, with an emphasis on the WTSS ("web time series service"). The next sections describe filtering and clustering techniques. We then discuss machine learning techiques for SITS data and how to apply them to image time series. Finally, we present validation methods.

## Data Handling and Visualisation Basics in SITS

The basic data unit in the sits package is the SITS tibble, which is a way of organizing a set of time series data with associated spatial information. In R, a "tibble" differs from the traditional data frame, insofar as a tibble can contain lists embedded as column arguments. Tibbles are part of the "tidyverse", a collection of R package designed to work together in data manipulation. The "tidyverse" includes packages such as "ggplot2", "dplyr" and "purrr" (Wickham & Grolemund 2017). The "SITS" package makes extensive use of the "tidyverse".

For a better explanation of how the "SITS tibble" works, we will read a data set containing 2,115 labelled samples of land cover in Mato Grosso state of Brazil. This state has 903,357 km$^2$ of extension, being the third largest state of Brazil. It includes three of Brazil's biomes: Amazonia, Cerrado and Pantanal. It is the most important agricultural frontier of Brazil and is Brazil's largest producer of soybeans, corn and cotton.

The samples contain time series extracted from the MODIS MOD13Q1 product from NASA from 2001 to 2016, provided every 16 days at 250-meter spatial resolution in the Sinusoidal projection. Based on ground surveys and high resolution imagery, we selected 2,115 samples of nine classes: forest, cerrado, pasture, soybean-fallow, fallow-cotton, soybean-cotton, soybean-corn, (8) soybean-millet, soybean-sunflower. Crop and pasture ground data was collected by researchers Alexandre Coutinho, Julio Esquerdo and Joao Antunes from the Brazilian Agricultural Research Agency (EMBRAPA) through farmer interviews in October 2009 and in October 2013. Samples for cerrado and forest classes were provided by Rodrigo Bergotti from INPE. Ground samples for soybean-fallow class were provided by Damien Arvor(Arvor et al. 2012).

```
# retrieve a set of samples from an RDS file
samples.tb <- readRDS(system.file("extdata/time_series/embrapa_mt.rds",
                                  package = "sits"))
samples.tb
```

```
## # A tibble: 2,115 x 7
##    longitude latitude start_date   end_date   label    coverage
##        <dbl>    <dbl>     <date>     <date>   <chr>       <chr>
##  1  -55.1852 -10.8378 2013-09-14 2014-08-29 Pasture mod13q1_512
##  2  -57.7940  -9.7573 2006-09-14 2007-08-29 Pasture mod13q1_512
##  3  -51.9412 -13.4198 2014-09-14 2015-08-29 Pasture mod13q1_512
##  4  -55.9643 -10.0621 2005-09-14 2006-08-29 Pasture mod13q1_512
##  5  -54.5540 -10.3749 2013-09-14 2014-08-29 Pasture mod13q1_512
##  6  -52.4572 -10.9512 2013-09-14 2014-08-29 Pasture mod13q1_512
##  7  -52.1443 -13.9981 2013-09-14 2014-08-29 Pasture mod13q1_512
##  8  -57.6907 -13.3382 2015-09-14 2016-08-28 Pasture mod13q1_512
##  9  -54.7034 -16.4265 2015-09-14 2016-08-28 Pasture mod13q1_512
## 10  -53.6543 -15.7155 2014-09-14 2015-08-29 Pasture mod13q1_512
## # ... with 2,105 more rows, and 1 more variables: time_series <list>
```

The "SITS tibble" contains data and metadata. The first six columns contain the metadata: spatial and temporal location, label assigned to the sample, and coverage from where the data has been extracted. The spatial location is given in longitude and latitude coordinates for the "WGS84" ellipsoid. For example, the first sample has been labelled "Pasture", at location (-55.1852, -10.8387), and is considered valid for the period (2013-09-14, 2014-08-29). Informing the dates where the label is valid is crucial for correct classification. In this case, the researchers involved in labelling the samples chose to use the agricultural calendar in Brazil, where the spring crop is planted in the months of September and October, and the autmun crop is planted in the months of February and March. For other applications and other countries, the relevant dates will most likely be different from those used in the example.

The SITS tibble also contains the time series data for each spatiotemporal location. The timeseries data is also organized as a tibble, with a column with the dates and the other columns with the values for each spectral band.

```
#print the first time series
samples.tb[1,]$time_series
```

```
## [[1]]
## # A tibble: 23 x 7
##         Index   ndvi    evi    nir    mir   blue    red
##  *     <date>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>
##  1 2013-09-14 0.4245 0.2800 0.2879 0.2439 0.0605 0.1163
##  2 2013-09-30 0.4673 0.2642 0.2570 0.1672 0.0357 0.0933
##  3 2013-10-16 0.5039 0.2993 0.2659 0.2019 0.0405 0.0877
##  4 2013-11-01 0.6489 0.4071 0.2762 0.1026 0.0266 0.0588
##  5 2013-11-17 0.7956 0.6692 0.4552 0.0875 0.0271 0.0518
##  6 2013-12-03 0.7248 0.5400 0.3661 0.0799 0.0516 0.0584
##  7 2013-12-19 0.7971 0.5574 0.3490 0.0478 0.0171 0.0394
##  8 2014-01-01 0.8049 0.7228 0.5255 0.1030 0.0327 0.0568
##  9 2014-01-17 0.7364 0.6867 0.5165 0.1449 0.0629 0.0784
## 10 2014-02-02 0.7870 0.7342 0.5320 0.1323 0.0452 0.0634
## # ... with 13 more rows
```

The SITS package provides functions for data manipulation and displaying information of a SITS tibble. For example, the function is `sits_bands` that lists the available bands.

```
sits_bands (samples.tb)
```

```
## [1] "ndvi" "evi"  "nir"  "mir"  "blue" "red"
```

Another useful command is `sits_labels` that shows the labels of the sample set and their frequencies.

```
sits_labels (samples.tb)
```

```
## # A tibble: 9 x 3
##           label count       freq
##           <chr> <int>      <dbl>
## 1       Cerrado   400 0.18912530
## 2 Fallow_Cotton    34 0.01607565
## 3        Forest   138 0.06524823
## 4       Pasture   370 0.17494090
## 5      Soy_Corn   398 0.18817967
## 6    Soy_Cotton   399 0.18865248
## 7    Soy_Fallow    88 0.04160757
## 8    Soy_Millet   235 0.11111111
## 9 Soy_Sunflower    53 0.02505910
```

In many cases, it is useful to relabel the data set. For examples, there may be situations when one wants to use a smaller set of labels, since samples in one label on the original set may not be distinguishable for samples with other labels. We then should use the `sits_relabel` function. This function requires a conversion list, as shown in the example below.

```
# a list for relabelling the samples
new_labels <- list("Cerrado"       = "Savanna",
                   "Pasture"       = "Grasslands",
                   "Soy_Corn"      = "Double_Cropping",
                   "Soy_Cotton"    = "Double_Cropping",
                   "Soy_Sunflower" = "Double_Cropping",
                   "Soy_Fallow"    = "Single_Cropping",
                   "Soy_Millet"    = "Single_Cropping",
                   "Fallow_Cotton" = "Single_Cropping")
# apply the sits_relabel function
samples2.tb <- sits_relabel(samples.tb, new_labels)
# view the result
sits_labels(samples2.tb)
```

```
## # A tibble: 5 x 3
##             label count       freq
##             <chr> <int>      <dbl>
## 1 Double_Cropping   850 0.40189125
## 2          Forest   138 0.06524823
## 3      Grasslands   370 0.17494090
## 4         Savanna   400 0.18912530
## 5 Single_Cropping   357 0.16879433
```

Given that we have used the tibble data format for the metadata and and the embedded time series, one can use the functions of the `dplyr`, `tidyr` and `purrr` packages of the "tidyverse" (Wickham & Grolemund 2017) to process the data. For example, the following example uses the `sits_select` function to get a subset of the sample data set with two bands ("ndvi" and "evi") and then uses the `dplyr::filter` function to select the samples labelled either as "Cerrado" or "Pasture". We can then use the `sits_plot` function to display the time series. Given a small number of samples to display, the `sits_plot` function tries to group as many spatial locations together. In the following example, the first 15 samples of the "Cerrado" class all refer to the same spatial location in consecutive times. For this reason, these samples are plotted together.

```
# select the "ndvi" bands
samples_ndvi.tb <- sits_select(samples.tb, bands = c("ndvi"))
# select only the samples with the cerrado label
samples_cerrado.tb <- dplyr::filter (samples_ndvi.tb, label == "Cerrado")
# plot the first 15 samples (different dates for the same points)
sits_plot (samples_cerrado.tb[1:15,])
```

For a large number of samples, where the amount of individual plots would be substantial, the default visualisation combines all samples together in a single temporal interval (even if they are valid for different years). Therefore, all samples of the same band and the same label are aligned to a common interval. This plot is useful to show the spread of values for the time series of each band. The strong red line in the plot shows the median of the values, and the two orange lines are the first and third interquartile ranges. The `sits_plot` function has different ways of working. Please refer to the documentation for more details.

```
# plot all cerrado samples together (shows the distribution)
sits_plot (samples_cerrado.tb)
```

**Importing data into SITS**

The SITS package allows different methods of data input, including: (a) obtain data from a WTSS (Web Series Time Service); (b) read data stored in a time series in the ZOO format (Zeileis & Grothendieck 2005); (c) read a time series from a RasterBrick (Hijmans 2015). This section describes options (a) and (b). Option (c) will be described in the section were we describe raster processing. The WTSS service is a light-weight service, designed to retrieve time series for selected locations and periods (Vinhas et al. 2016). This service has been implemented by the research team of the National Institute for Space Research to allow remote access to time series data. To access the service, the user needs to provide a URL that points to the WTSS server location and use the function `sits_infoWTSS` that provides information on the coverages available on the server.

```
URL <- "http://www.dpi.inpe.br/tws/wtss"
wtss_inpe <- sits_infoWTSS(URL)
```

After finding out which coverages are available at the WTSS service, one may request specific information on each coverage by using the function `sits_coverageWTSS` which lists the contents of the data set, including source, bands, spatial extent and resolution, time range, and temporal resolution. This information is then stored in a tibble for later use.

```
# get information about a specific coverage
coverage.tb <- sits_coverageWTSS(URL,"mod13q1_512")
```

The user can then request one or more points using the `sits_getdata` function. This function provides a general means of access to image time series. In it simplest fashion, the user provides the latitude and longitude of the desired location, the URL of the WTSS services, the coverage name, the bands, and the start date and end date of the time series. If the start and end dates are not provided, all of the samples are retrived. The result is a SITS tibble that can be visualised using `sits_plot`.

```
# a point in the transition forest pasture in Northern MT
long <- -55.57320
lat <- -11.50566
# obtain a time series from the WTSS server for this point
series.tb <- sits_getdata(longitude = long, latitude = lat, URL = URL,
             coverage = "mod13q1_512", bands = c("ndvi", "evi"),
             start_date = "2001-01-01", end_date = "2016-12-31")
# plot the series
sits_plot (series.tb)
```

A useful case is when users have a set of labelled samples, that are to be used as a training data set. In this case, one usually has some field data or trusted observations which are labelled.

## References


Arvor, D., Meirelles, M., Dubreuil, V., Bégué, A. & Shimabukuro, Y. E. (2012), 'Analyzing the agricultural transition in Mato Grosso, Brazil, using satellite-derived indices', *Applied Geography* **32**(2), 702–713.

Brown, J. C., Kastens, J. H., Coutinho, A. C., Victoria, D. d. C. & Bishop, C. R. (2013), 'Classifying multiyear agricultural land use data from Mato Grosso using time-series MODIS vegetation index data', *Remote Sensing of Environment* **130**, 39–50.

Estel, S., Kuemmerle, T., Alcantara, C., Levers, C., Prishchepov, A. & Hostert, P. (2015), 'Mapping farmland abandonment and recultivation across Europe using MODIS NDVI time series', *Remote Sensing of Environment* **163**, 312–325.

Galford, G. L., Mustard, J. F., Melillo, J., Gendrin, A., Cerri, C. C. & Cerri, C. E. (2008), 'Wavelet analysis of MODIS time series to detect expansion and intensification of row-crop agriculture in Brazil', *Remote sensing of environment* **112**(2), 576–587.

Hastie, T., Tibshirani, R. & J., F. (2009), *The Elements of Statistical Learning. Data Mining, Inference, and Prediction*, Springer, New York.

Hijmans, R. J. (2015), *raster: Geographic Data Analysis and Modeling*.

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2013), *An Introduction to Statistical Learning: with Applications in R*, Springer, New York, EUA.

Jönsson, P. & Eklundh, L. (2004), 'TIMESAT—a program for analyzing time-series of satellite sensor data', *Computers & Geosciences* **30**(8), 833 – 845.

Kastens, J., Brown, J., Coutinho, A., Bishop, C. & Esquerdo, J. (2017), 'Soy moratorium impacts on soybean and deforestation dynamics in Mato Grosso, Brazil', *PLOS ONE* **12**(4), e0176168.

Kennedy, R. E., Yang, Z. & Cohen, W. B. (2010), 'Detecting trends in forest disturbance and recovery using yearly Landsat time series', *Remote Sensing of Environment* **114**(12), 2897–2910.

Lambin, E. F., Geist, H. J. & Lepers, E. (2003), 'Dynamics of land-use and land-cover change in tropical regions', *Annual review of environment and resources* **28**(1), 205–241.

Pasquarella, V. J., Holden, C. E., Kaufman, L. & Woodcock, C. E. (2016), 'From imagery to ecology: leveraging time series of all available LANDSAT observations to map and monitor ecosystem state and dynamics', *Remote Sensing in Ecology and Conservation* **2**(3), 152–170.

Pelletier, C., Valero, S., Inglada, J., Champion, N. & Dedieu, G. (2016), 'Assessing the robustness of random forests to map land cover with high resolution satellite image time series over large areas', *Remote Sensing of Environment* **187**, 156–168.

Vinhas, L., Ribeiro, G., Ferreira, K. R. & Camara, G. (2016), Web services for big Earth observation data, *in* 'Proceedings of the 17th Brazilian Symposium on GeoInformatics', INPE, Campos do Jordão, SP, Brazil, pp. 26–35.

Wickham, H. & Grolemund, G. (2017), *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*, O'Reilly Media, Inc.

Zeileis, A. & Grothendieck, G. (2005), 'zoo: S3 infrastructure for regular and irregular time series', *Journal of Statistical Software* **14**(6), 1–27.