SITS: An R Package for Data Access, Visualisation, Filtering, Clustering, Event Detection and Classification of Satellite Image Time Series

Rolf Simoes, Gilberto Camara
INPE, Brazil

Victor Maus
IIASA

Alexandre Iwata
IPEA, Brazil

Abstract

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. Given this motivation, this package provides a set of tools for data access, filtering, clustering and classification of satellite image time series.

Keywords: satellite image time series, big Earth Observation data.

1. 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 for Geoinformatics research: *How to design and build technologies that allow the Earth observation community to analyse big data sets?*

Since remote sensing satellites revisit the same place repeatedly, we can calibrate their images so measures of the same place in different times are comparable. These observation can be organised, so that each measure from sensor is mapped into a three dimensional array in space-time. From a data analysis perspective, researchers then have access to satellite image time series (SITS). Using time series derived from big Earth Observation data sets is one of the leading research trends in Land Use Science and Remote Sensing.

A time-series of measurements of the same location in the surface of the Earth can be considered as a historical record. When the images arise for a dense record of frequent revisits, the

temporal resolution of the big data set is able to capture the most important land use changes. Such dense time series allow researchers to which changes have taken place in each location.

The benefits of remote sensing time series analysis arise when the temporal resolution of the big data set is sufficient to capture the most important changes. In this case, the temporal autocorrelation of the data can be stronger than the spatial autocorrelation. In other words, given data with adequate repeatability, a pixel will be more related to its temporal neighbours rather its spatial ones. In this case, *time-first, space-later* methods will give better results than the *space-first, time-later* approach.

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 (**?**). 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 (**?**), ecology dynamics (**?**), agricultural intensification (**?**) and its impacts on deforestation (**?**).

his work combines SITS with statistical learning. In a broad sense, statistical learning refers to a class of algorithms for classification and regression analysis (**?**). 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.

There has been much recent interest in using classifiers such as support vector machines (**?**) and random forests (**?**). Most times, researchers use a *space-first, time-later* approach, where the dimension of the decision space is limited to the number of spectral bands or their transformations. Sometimes, the decision space is extended with temporal attributes. To do this, researchers filter the raw data to get smoother time series (**??**). Then, using software such as TIMESAT (**?**), they derive a small set of phenological parameters from vegetation indexes, like beginning, peak, and length of growing season (**??**). These approaches do not use the power of advanced statistical learning techniques to work on high-dimensional spaces and with big training data sets (**?**). They have one thing in common: raw time series data is considered too noisy to be used directly. This leads to the question: *do noise removal and homogenization steps reduce the information present in the satellite image time series?*

An alternative approach, proposed in this package, is to use 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 **?**, **?** and **?**. Our conclusion is that part of the information in raw time series is lost after filtering or statistical approximation. By choosing a statistical classifier which is robust to noise, one should be able to get better results than current approaches. 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. In this work, we used the MODIS MOD13Q1 product with 23 samples per year per pixel, and 4 bands (NVDI, EVI, nir and mir). By taking a series of labelled time series, we feed the statistical inference model with a 92-dimensional attribute 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.

Acknowledgments

Affiliation: