FAPESP Research Program in e-science - Grant 2014/08398-6

Project duration: January 01, 2015 – 31 December 2018

e-Sensing:

Big Earth observation data analytics

for land use and land cover change information

Yearly report: 01 January 2017 – 31 December 2017

Host institution: INPE – Instituto Nacional de Pesquisas Espaciais

Principal investigator: Gilberto Câmara (INPE)

Research team: Ieda Sanches, Karine Ferreira, Leila Fonseca, Lubia Vinhas, Isabel Escada, Gilberto Queiroz, Luiz Maurano, João Vianei Soares, Julio d'Alge, Pedro Andrade, Ricardo Cartaxo, Thales Körting.

Additional investigators: Eduardo Llapa (INPE)

Project website: http://www.esensing.org

Table of Contents

1 Overview of the project objectives

This document provides the third year report of the "e-sensing" FAPESP project (grant 2014/08398-6), and describes the activities carried out during the period 01.01.2017 to 31.12.2016. We will use numbers (such as [10]) to refer to the list of papers published by us in 2016, available in the References section.

Currently, scientists ignore the time reference inherent to Earth observation data, producing land cover maps taking either a single or at most two time references. As a result, only a small part of the big data sets produced by remote sensing satellite are ever used. This leads to an important research question: *How can we use e-science methods and techniques to substantially improve the extraction of land use and land cover change information from big Earth Observation data sets in an open and reproducible way?*

In response to this challenge, *our project will conceive, build and deploy a new type of knowledge platform for organization, access, processing and analysis of big Earth observation data*. The key elements of this knowledge platform are:

1. A scientific database based on the SciDB innovative array database management system, capable of managing large remote sensing data sets.
2. An innovative set of spatiotemporal image analysis methods, mostly based in analysis of satellite image time series. These methods are all developed as open source software to promote reproducibility.

The innovative infrastructure developed in the project will be used for new types of information extraction from Earth observation data, focused on land cover and land use change of large data sets. Our knowledge platform will allow scientists to perform data analysis directly on big data servers. Scientists will be then able to develop completely new algorithms that can seamlessly span partitions in space, time, and spectral dimensions.

We aim to make two important contributions:

1. New database methods and techniques that use array databases to build a geographical information system that handles big spatial data.
2. New data analysis, data mining, and image processing methods to extract land change information from large Earth observation data sets.

2 Main results of year 3 (January – December 2017)

During 2017, our most relevant results were:

1. Development of machine learning and deep learning methods that allow high accuracy in land use and land cover classification using satellite image time series [5] [14].
2. Implementation of SITS, an R package for working with satellite image time series. It includes data retrieval, clustering, and provides machine learning methods for time series classification, including SVM, LDA, QDA, GLM, Lasso, Random Forests and Deep Learning [6] [24].
3. Proposal, development and validation of a spatiotemporal calculus for reasoning about land use change dynamics [2] [8] [15].
4. Development of a new land use and land cover map for the state of Mato Grosso, from 2000 to 2016, in cooperation with EMBRAPA [5][14].
5. Advances in the understanding and modelling of tropical forest degradation [1][3][17][48]
6. Development of methods for space-time segmentation of satellite images [20].
7. Production of data sets and ground studies which are useful for validating multi-temporal land use classification methods [5] [13] [35].
8. Evaluation of existing land cover classifications to serve as baseline for the results of the e-sensing project [19][30].
9. Evaluation of smoothing methods on Landsat-8 EVI time series for crop classification based on phenological parameters [28].
10. Initial development of a multitemporal approach for land use mapping using Bayesian Networks [38].

Overall, the project is progressing as expected. During the third year, the team has made significant progress on land user and land cover change classifications, using advanced data analysis methods.

3 Detailed description of the results in Year 3 (2017)

This section describes the results of the project in 2016. In the presentation, we follow the project organization in three work packages (WP), and associated milestones, as laid out in the proposal:

1. *WP 1 – Databases*: research and development associated with using array databases to store large Earth observation data sets and developing workflows and methods for efficient storage, access and processing of large data, reproducibly.

2. *WP 2 – Data analysis*: R&D on spatiotemporal techniques for extracting change information on large Earth observation data sets, relevant for forestry applications; include novel time series applications for remote sensing data, and combined time series and multi-temporal image processing.

3. *WP 3 – Use case development*: case studies of forestry and agriculture applications that use large Earth observation data sets. These use cases will validate the methods and data developed by the other work packages.

To help the review of this report, we first present the table of milestones presented in the project proposal. We will then consider each the proposed milestones, stating whether it has been fulfilled or delayed.

For each milestone, we preview the result more directly associated with it. The rest of the results of the project can be found in the References section. All of the papers published by members of the research teams that are associated to the project are available at the project's website: http://www.esensing.org.

TABLE 1

PLANNED MILESTONES: RESULTS AFTER MONTH 24

Green background	Target was met
Yellow background	Target was partially met
Red background	Target was delayed
Blue background	New task
White background	For later years

TASK	Month 12	Month 24	Month 36	Month 48
T1.1 Building big EO databases	M1.1.1. V1 of the database for use cases in Brazil	M1.1.2 V2 of database for use cases in Brazil	M1.1.3 V3 of database for use cases in Brazil	M 1.1.4 V4 of database for regional use cases
T1.2 Extend SciDB for geographical data handling	M1.2.1 Integration of TerraLib and SciDB	M1.2.2 Algorithms for SciDB server-side processing	M1.2.3 Web service for SciDB server-side processing	M1.2.4 Extension of SciDB as a spatial data manager
T2.1 Exploratory big data analysis		M2.1.2 Interactive environment for data exploration	M2.1.3 Interactive environment for collaborative analysis	M2.1.4 Large-scale sharing with other research teams
T2.2 Data analysis for big EO data	M2.2.1 R-Big-EO time series analysis software (V1)	M2.2.2 R-Big-EO time series analysis software (V2)	M2.2.3 R-Big-EO space-time analysis software (V1)	M2.2.4 Big-EO space-time analysis software (V2)
T3.1 Tropical forest change	M3.1.1 Identification and selection of areas	M3.1.2 Preliminary detection of clear cut and degradation	M3.1.3 Detection of clear cut and degradation	M3.1.4 Assessment of forest change alert methods
T3.2 Tropical agriculture mapping	M3.2.1 Identification and selection of areas	M3.2.2 Mapping of soybeans, maize and sugarcane	M3.2.3 Mapping agriculture in Cerrado and Amazonia	M3.2.4 Assessment of agricultural mapping methods

3.1 Progress report on WP 1 - Big Earth observation databases

3.1.1 Task 1.1 - Building and deployment of big Earth observation databases to support data analysis and use cases

This task builds databases to be used by the project. In the proposal, we set the following milestone for month 24:

Milestone M1.1.3 – V3 of database for regional use cases in Brazil

The target of this milestone was to build a large remote sensing database containing the data needed for the use cases in Brazil in year 4. This result has been achieved. This target was already met at the end of year 2, as stated in the previous progress report.

At the end of year 3, we have loaded in the servers the following data sets:

1. MODIS MOD09Q1 images at 250-meter resolution from 2000 to 2017 for the whole of South America, with 13,800 images associated to 3.11 x 1011 (317 billion) different satellite image time series.
2. Selected LANDSAT images at 30-meter resolution from 2000 to 2015 for selected areas of Amazonia, with 202 images associated to 200 million different satellite image time series.

3.1.2 Task 1.2 – Extend SciDB for geographical data handling

The purpose of this task was to use the array database SciDB for Earth observation applications. To do this, we need to develop methods that would allow us to process satellite image time series in SciDB.

In the Report for Year 2 (2016), we explained that the original goal was to use the TerraLib software library, developed by INPE, as a source of data types and algorithms for geographical data. Our plan was to include an interface for SciDB in TerraLib. However, we found out that this integration was not required to achieve our aims. This simplifies the design of the infra-structure and allows for easier reproducibility. Therefore, we decided to use web services as the primary interface for our big data analytical methods.

For this reason, we decided to set the following milestone for month 36:

Milestone M1.2.3 – Web services for SciDB server-side processing

The aim of this new milestone was to develop a series of methods for processing big EO data on array databases. The strategy chosen was to based on the following aims:

1. *Analytical scaling:* provide support for the full cycle of research, allowing algorithms developed at the desktop to run on big databases with minor changes.
2. *Software reuse*: allow researchers to adapt existing methods for big data with minimal reworking.
3. *Collaborative work*: enable results to be shared with the scientific community.
4. *Replication*: encourage research teams to build their own infrastructure.

What we have envisioned was to use the R suite of statistical tools as the environment to develop our analytical methods. R is the *lingua franca* of data analytics. Using R, researchers can scale up their methods, reuse previous work, and collaborate with their peers. Our aim is to be able to execute the same script in both client and server side.

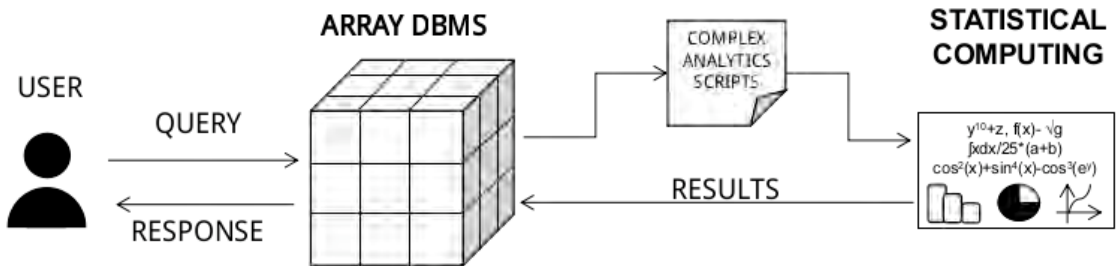


Figure 1 – Interface for server-side processing of analytical methods.

On Year 3 of the project, we developed and tested different methods for processing R scripts using SciDB. For efficient server-side processing, we optimized the organization of the array database SciDB for time series data analysis. We found out that we could obtain efficient performance of big Earth observation data using SciDB server side processing (see Table 1).

Table 1 – Processing times of large data sets using R-SciDB

Case Study	Area (km²)	Measures	Proc time (h)
Mato Grosso	900,000	135 GB	6
Cerrado	2,050,000	308 GB	13

We have been able to use the array database SciDB effectively for our processing. Table 1 shows the processing times required to classify 16 years of satellite image time series imagery for the state of Mato Grosso and the Cerrado biome in Brazil. This classification was done in 5 servers, each with 12 cores, 2.4GHz, 96 GB of RAM, 16 TB of data storage, and the SciDB back-end. The quality of these results will further be explained when we discuss the Work Package 3 (WP3) later in this report.

However, the full development of this task was not completed. Building a Web Service for SciDB turned out to be harder than anticipated. SciDB is proven to be efficient for handling large Earth observation data sets. It lacks full support for client-server access, since it does not have a user authentication facility. Additionally, the development of a Web service would require the conception of a protocol that would lead to differences in the code run in the server and that run in the client.

Therefore, in the end of Year 3, we decided that including the access to SciDB inside an R package would be better than developing a Web service. Therefore, this task has been merged with task 2.2 (see below). We consider that having a robust R package capable of doing both client-side and server-side processing would fit our requirements better.

3.2 Progress report on WP 2 - Data analysis for big Earth observation data

3.2.1 Task 2.1 – Web-based exploratory big data analysis

In our original proposal, we envisaged making an integration between SciDB, TerraLib and the R software. However, during the development of the project we found out that this interface would not be required to achieve our goals of building an efficient set of software for space-time analysis. Instead, we have developed a web-based interface exploratory big data analysis. The exploratory data analysis interface, developed in Year 2 of the project is shown in Figures 2. We consider that this task has been sufficiently achieved, and no new development were required in Year 3.

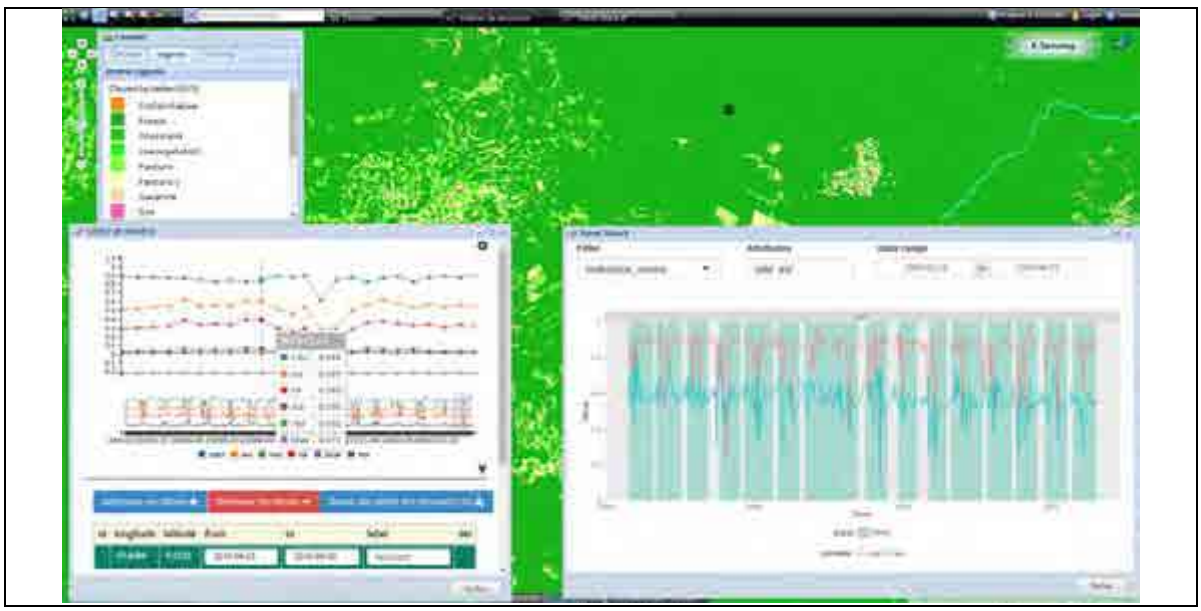


Figure 2 – Web-based exploratory big EO data analysis interface. In the left, there is a window allowing the user to select a training sample for later use in classification. In the right, the result of a classification of a point using the TWDTW algorithm is shown (source: e-sensing team).

The interfaces use the Web Time Series Service (WTSS, described in the Year 2 report) to obtain time series data. The user can then analyse the data using the methods available in the ***sits*** R package, developed by the project team (see section 3.2.2 below).

3.2.2 Task 2.2 - Space-time analysis of big Earth observation data for land change monitoring – Part 1: The SITS package

This task aims to develop new methods for space-time analysis of big Earth observation data. It is expected to produce new research results, since it will be the first time that EO scientists have full access to large data sets to validate their data analysis methods.

In this task, the project proposal set the following milestone for month 36:

Milestone M2.2.3: R-Big-EO space-time series analysis software

This milestone has been achieved. In the end of Year 2, we had developed a collection of tools for analysis of space-time. In Year 3, we decided to bring all these packages together in a single R package called ***sits***. A large part of our work on Year 3 has been spent on the development of this package.

The ***sits*** package ("Satellite Image Time Series") includes data retrieval from a WTSS (web time series service)[1], different visualization methods for image time series, smoothing methods for noisy time series, different clustering methods, including dendrograms and SOM. It matches noiseless patterns with noisy time series using the TWDTW[2] method for shape recognition and provides machine learning methods for time series classification.

To our knowledge, ***sits*** is the first R package that provides a unified support for many advanced data analysis methods for image time series. It is also the first to support advanced methods such as deep learning for time series classification.

---

[1] The WTSS (Web-Time Series Service) was one of the important results of Year 2 of the project and has been described in more detail in the Year 2 report. See also Vinhas et al., "Web Services for Big Earth Observation Data." In: GEOINFO 2016. Sao Jose dos Campos: INPE/SBC, 2016. v.1. p.166 – 177.

[2] The TWDTW (Time-weighted Dynamic Time Warping) is another important result of the e-sensing project that has been discussed in Year 2 report. See also Maus et al, "A Time-Weighted Dynamic Time Warping Method for Land-Use and Land-Cover Mapping". IEEE JSTARS, vol. 9(8): 3729-3739, 2016. DOI: 10.1109/JSTARS.2016.2517118.

The development of methods for clustering and machine learning were the most important results of the ***sits*** package in Year 3. Clustering is a way to improve training data to use in machine learning classification models. In this regard, cluster analysis can assist the identification of structural patterns and anomalous samples. The package support for the agglomerative hierarchical clustering (AHC) using the DTW (dynamic time warping) distance measure. As an example, Figure 5 shows the result of a dendrogram applied to 746 samples of land cover in Mato Grosso, divided in two classes ("Cerrado" and "Pasture"). From the dendogram, one can see that most of the samples form tight clusters, bu there are some outlines. These are some samples of "Pasture" that are more similar to those of "Cerrado" than to the class they were originally labelled. Based on the dendrogram, the ***sits*** package provides functions to find the optimal number of clusters and to remove outliers, thus generating homogenous clusters that can be used for classification.

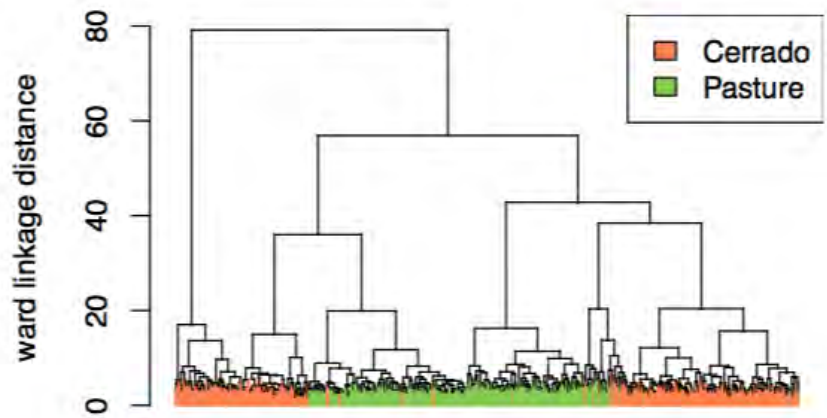


Figure 3 – Dendrogram produced by the sits R package

The machine learning methods in ***sits*** allow users to explore the full depth of satellite image time series data. The analysis techniques treat time series as a feature vector, made by taking all values of all pixels. The idea is to have as many temporal attributes as possible, increasing the dimension of the classification space. In this scenario, statistical learning models are the natural candidates to deal with high-dimensional data: learning to distinguish all land cover and land use classes from trusted samples exemplars (the training data) to infer classes of a larger data set.

The methods available for machine learning in sits include[3]:

- *Support vector machine* (***svm***): a classifier that uses linear and non-linear mapping of the input vectors into high-dimensional spaces, building hyperplanes that allow distinguishing between the data classes.
- *Random Forest (**rfor**):* ensemble learning method for classification, that works by building a multitude of decision trees at training time.
- *Linear Discriminant Analysis (lda)*: a method that finds a linear combination of features that characterizes or separates the desired classes.
- *Quadratic Discriminant Analysis* (***qda***): A methods that separate measurements of two or more classes of objects by a quadric surface.
- *Multinomial logistic regression* (***mlr***): method that generalizes logistic regression to multiclass problems. It does not assume statistical independence of the input random variables.
- *Deep learning* using *multilayer perceptrons (**dl-mlp**)*: A method that uses a cascade of multiple layers of neural networks with nonlinear processing units for feature extraction and transformation.

To compare the performance of these methods, we used a data set with 11.743 samples of 11 classes for the Brazilian Cerrado biome. The classes include natural vegetation (*"Tropical Forest"*, *"Cerrado strictu sensu"*, *"Cerrado campo"*, *"Cerrado rupestre"*) and agricultural classes *("Pasture"*, *"Soy-Fallow"*, *Soy-Corn"*, *"Soy-Millet"*, *"Soy-Cotton"*, *"Crop-Cotton"*, *"Millet-Cotton"*). We obtained these samples in cooperation with EMBRAPA and by the team project members, most notably Dr. Ieda Sanches and Dr. Rodrigo Bergotti.

Table 2 shows a performance comparison of the classifiers for the Cerrado data set. This assessment was done using cross-validation to estimate the expected prediction error. We ran 5 trials. In each trial, 80% of the samples were used to train the classifier and 20% was set aside for testing. A simple average of the five predictions gives us an estimation of the expected prediction error.

TABLE 2

[3] For a description of SVM and other statistical learning methods, please see T. Hastie et al., The *Elements of Statistical Learning. Data Mining, Inference, and Prediction*. Springer, NY, EUA, 2009. James et al, *An Introduction to Statistical Learning: with Applications in R*. Springer, New York, EUA, 2013. For a description of deep learning, please see Goodfellow et al., *Deep Learning*, Cambridge (MA), MIT Press, 2016 and Chollet and Allaire, *Deep Learning in R*, New York, Manning Pubs, 2018.

Performance of Classification Models for the Cerrado data set

Class	Accuracy	Kappa	Obs
svm	96%	0.94	Radial kernel, cost = 10
svm	95%	0.94	Radial kernel, cost = 100
rfor	94%	0.915	No. of trees = 2000
dl-mlp	93%	0.91	3 hidden layers (300, 200, 100 neurons), dropouts (0.4, 0.3, 0.2), "relu"activation, "adam"optimizer
mlr	89%	0.859	
qda	89%	0.856	
lda	88%	0.83	

The above results point out the high discriminatory power that is enabled by using all of the information in the time series. By contrast, many papers in the literature extract patterns or features from remote sensing data before applying machine learning methods. A typical way of working involves extracting features such as "start of growing season" and "peak of season" from remote sensing time series. These features are the inputs to machine learning classifiers[4]. The resulting accuracy is limited, as reported by previous work by the project team and in the scientific literature. The main reason for these unexceptional results is the loss of information involved, when part of the original data is discarded. Such choices are not required by the new generation of machine learning classifiers. These classifiers are robust and require large data sets.

The approach taken in the ***sits*** package is different: *use all the data available*.. The fact that different classifiers (*svm, random forest, deep learning*) are able to obtain high accuracy with the same data set shows that the quality and quantity of the sample sizes are the controlling factors in the classification performance. Even less sophisticated algorithms such as *quadratic discriminant analysis* can reach close to 90% accuracy in data of a large sample size. Therefore, we can conclude that having good samples is the key for obtaining good results in satellite image time series classification.

[4] See the review of the topic: Atzberger, C. "Advances in remote sensing of agriculture: Context description, existing operational monitoring systems and major information needs", *Remote Sensing* 5(2):949--981, 2013

One further point that might be of interest is the comparison of discriminative power between the three best classifiers (svm, rfor, dl-mlp). Table 2 below shows the detailed results for each class.

TABLE 3

Discriminative Power of Machine Learning Classifiers for the Cerrado Data Set

	Reliability			Accuracy		
Class	svm	rfor	dl-mlp	svm	rfor	dl-mlp
Forest	96%	98%	93%	99%	99%	99%
Cerrado strictu sensu	89%	94%	81%	77%	62%	63%
Cerrado campo	92%	85%	89%	97%	98%	94%
Cerrado rupestre	98%	95%	93%	97%	95%	97%
Fallow-Cotton	95%	94%	91%	99%	94%	93%
Millet-Cotton	99%	99%	91%	98%	90%	88%
Pasture	95%	94%	93%	95%	86%	94%
Soy-Corn	97%	94%	96%	97%	97%	95%
Soy-Cotton	97%	96%	94%	97%	95%	96%
Soy-Fallow	99%	92%	97%	99%	93%	99%
Soy-Millet	77%	45%	69%	82%	85%	57%

In the above table, the column *reliability* ("user's accuracy") shows the probability that a pixel labeled as a certain land-cover class in the map is really this class. The figures in column *accuracy* (also known as "producer's accuracy") refer to the probability that a certain land-cover of an area on the ground is classified as such. When looked in detail, we note that the ***svm*** classifier has a better discriminating power than ***rfor*** and ***dl-mlp***. However, this result cannot be generalized. Considering that deep learning methods have a large number of meta-parameters, the deep learning method used in the comparison is only one of many possible. Indeed, for each given data set, it is theoretically possible to fine tune a deep learning architecture that matches the performance of the svm. However, such fine tuning requires considerable time and resources, as well as a good understanding of the theory behind deep learning methods. Thus, we can state a rule-of-thumb for good satellite image time series classification is simple: *first, obtain a large sample size of very good quality; then, use all the data available. If possible, compare the performance of the advanced classifiers (svm, rfor, deep learning) and choose the one that best discriminates your data. If pressed by time or resources, use a support vector machine.*

3.2.3 Task 2.2 - Space-time analysis of big Earth observation data for land change monitoring – Part 2: Spatiotemporal calculus for reasoning about land use change dynamics

A second relevant result in WP 2 was the proposal and development of a spatiotemporal calculus for reasoning about land change dynamics [2][15]. When analyzing Earth observation data, scientists are particularly interested in *land use trajectories*, which are paths from one land use into another. Typical questions researchers would like to ask are: *Which forest areas were degraded from 2000 to 2017*? *When did new agricultural systems such as double-cropping were introduced in the regions? Which area changed for pasture to croplands in the past decade?* To allow researchers to reason about these and similar change, we propose a land use change calculus (LUC Calculus), composed of the following primitives:

1. The interval temporal predicates proposed by Allen[5].
2. The additional predicates FOLLOWS and PRECEDES for comparing time intervals.
3. The new set of predicates RECUR, CONVERT and EVOLVE for reasoning on composition of land use transitions.

Using these predicates, we can express complex queries about land use change trajectories. One example is distinguishing secondary vegetation from mature forest. Mature forests have high biomass and biodiversity, and have not been affected by recent human actions. Secondary vegetation areas are places where the original forest was cut and the area was later abandoned. After a few years, these areas will appear in remote sensing images as forests. However, their biodiversity and biomass is much smaller than that of a mature forest. Thus, it is important to identify areas of secondary vegetation, even though they appear to be mature forest.

As an example, we classified the different types of land use in the municipality of Itanhangá (MT), from 2001 to 2016 (see figure 4). We use the full history of the area considered as a set of land use change trajectories. For an area to be singled out as secondary vegetation, its initial state is classified as "Forest". Then the area is converted to pasture or cropland, and later abandoned so that the forest regrows. Table 4 shows the logical expression used to uncover areas of secondary vegetation, and Figure 5 presents the total area of secondary

[5] J. F. Allen. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11):832–843, 1983

vegetation since 2002. Results show that a significant portion of the deforested area was abandoned and has regrown as a forest. This result points out to the predatory nature of deforestation in Amazon. Farmers sometimes cut mature forest and cannot sustain a profitable economic activity in these areas. Using the expressive power of the LUC Calculus, these transitions can be highlighted and better understood.

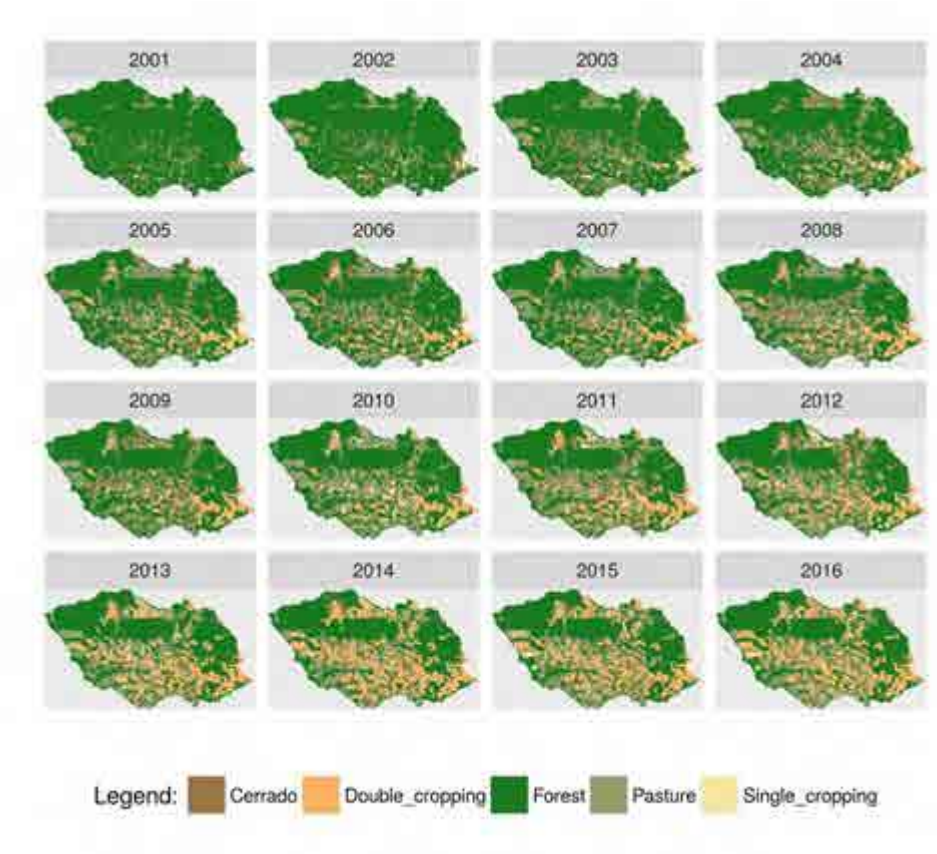


Figure 4 – Land use in Itanhangá, MT, Brazil, from 2001 to 2016

Table 4

LUC Calculus expression to find areas of secondary vegetation

Searching for all forest areas that have been replaced by pasture, *cerrado*, single cropping or double cropping and turned into forest areas again.
$\forall\ l \in L, t_1 = [2001, 2002], \forall\ t_i \in T, t_i \neq t_1,\ RECUR(l, \text{“}Forest\text{”}, t_1, t_i) \wedge$
$(\ EVOLVE(l, \text{“}Pasture\text{”}, t_1, \text{“}Forest\text{”}, t_i) \vee EVOLVE(l, \text{“}Cerrado\text{”}, t_1, \text{“}Forest\text{”}, t_i) \vee$
$EVOLVE(l, \text{“}Single\_cropping\text{”}, t_1, \text{“}Forest\text{”}, t_i) \vee$
$EVOLVE(l, \text{“}Double\_cropping\text{”}, t_1, \text{“}Forest\text{”}, t_i)\ )$

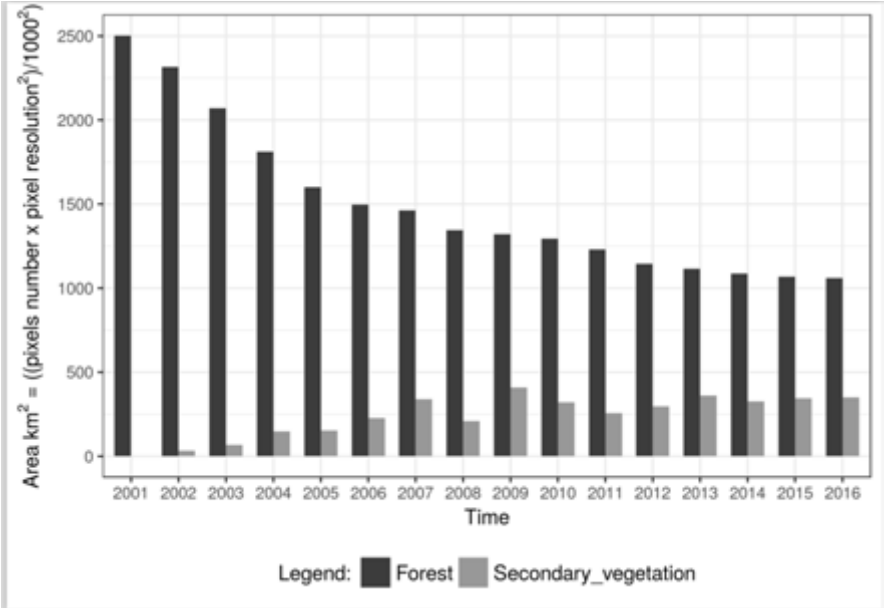


Figure 5 – Total area of forest and secondary vegetation in Itanhangá from 2001 to 2016.

3.2.4 Task 2.2 - Space-time analysis of big Earth observation data for land change monitoring – Part 3: Space-time segmentation

A third relevant result in WP 2 was the initial development of methods for space-time segmentation In [20], the authors propose a new segmentation method applied to time series of Earth Observation data. The method integrates regions in order to detect objects that are homogeneous in space and time. This approach aims to overcome the limitations of the snapshot model. Study cases were conducted using time series of MODIS and Landsat-8 OLI scenes by applying spatio-temporal segmentation using the Dynamic Time Warping measure as the homogeneity criterion.

The algorithm can be expressed by the following steps:

a) Select a sequence of images as input data.
b) Determine the number and location of the seeds at the image.
c) Compute DTW distance between the time series of the seeds and their neighbors. Similar neighbors are added to the region.
d) Continue examining all the neighbors until no similar neighbor is found. Label the obtained segmented as a complete region.

e) Observe the next unlabelled seed and repeat the process until all the seeds or pixels are labelled in a region.

The core of the method is to use DTW measure as the homogeneity criterion for growing regions in the study cases. These time series were used in DTW calculation between the seeds and its neighbouring pixels. The method was tested on MODIS NDVI and LANDSAT-8 and the results are encouraging [20]. In Year 4 of the project, we expect to further explore this topic.

3.3 Progress report on WP 3 - Use case development

3.3.1 Task 3.1 - Specification and validation of tropical forest change alert methods and data

Our use cases address two important problems: (a) how to use time series and spatiotemporal analysis to support the DEGRAD system of monitoring forest degradation; (b) how to use time series data to support the PRODES system for yearly assessments of tropical forest loss. Our aim is not to replace INPE's existing systems, but to explore how automated methods can complement and enhance them. In years 1 and 2, we did a lot of work on forest degradation, which is the most pressing and tougher scientific challenge. In this task, the project proposal has the following milestone for month 36:

Milestone M3.1.3: Detection of clear cut and degradation

This milestone has been accomplished. In this task, we have developed a methods for detecting forest degradation [3][17][47]. Forest degradation is defined as defined as *the long-term and gradual reduction of canopy cover due to forest fire and unsustainable logging.* Typically, a mature forest is degraded when part of its tree cover is removed. In the Brazilian Amazonia, degradation is caused by either unsustainable logging practices or by forest fires (intentional or uncontrolled). Understanding and identifying forest degradation is important, because it causes carbon emissions that have to be accounted for and leads to a loss of biodiversity caused by removal of important species. Also, degradation often (but not always) is associated to later actions that cause the full removal of forest cover (deforestation).

In Year 3, we carried out a detailed study on how to perform semi-automatic classification of forest degradation using LANDSAT-8 images. The method has the following steps:

a) Classify each image using a spectral mixture model, producing an index image that combines the soil and vegetation fractions.
b) Identify and map on the resulting image features associated to forest degradation. These features include the presence of areas for wood storage, corridors for transportation, and fire scars.
c) Perform a structural classification of degradation patterns using the GeoDMA[6] with 1 Km^2 cells. The structural classifier use landscape metrics and decision trees.

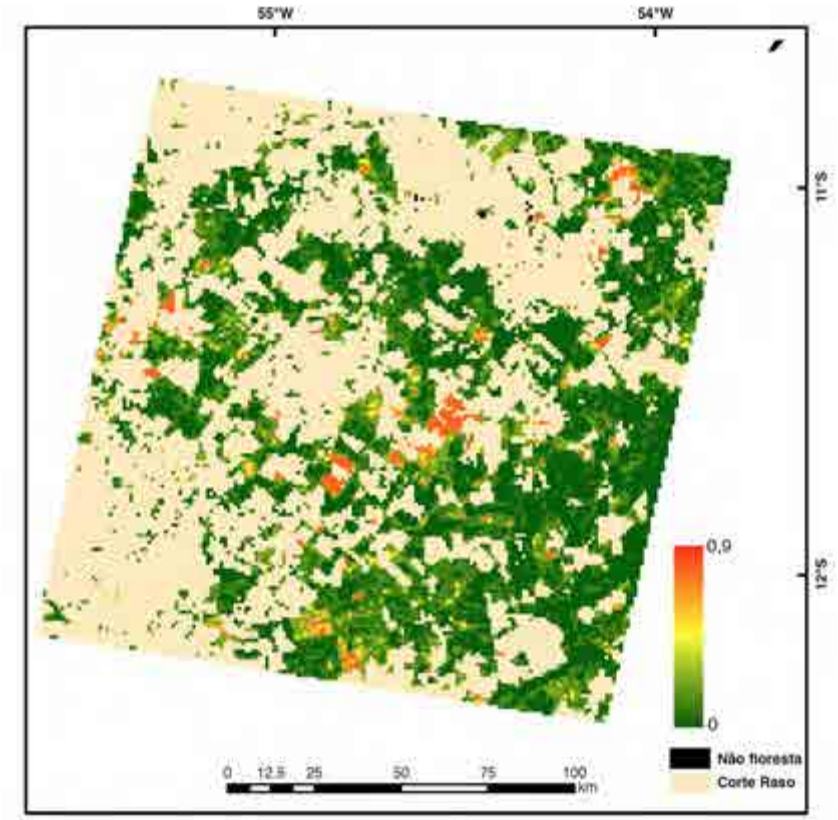


Figure 4 – Gradients of forest degradation intensity for year 2014 for Sinop, MT. The higher the index, the stronger the degradation. Source: [6].

[6] TS Körting, LMG Fonseca, G Câmara, GeoDMA—Geographic data mining analyst. Computers & Geosciences, 2013. This algorithm was developed by the project team on a previous FAPESP thematic project.

The classification accuracy was 96%, measured by comparing with ground samples. The method enabled the production of spatial gradients of forest degradation (see Figure 4). The results show that this approach, considering the intensity of the degradation, can be replicated in temporal studies of analysis of forest landscape conditions. Since each cell is fixed unit in time and space, it is possible to measure the variation of degradation in space and time.

3.3.2 Task 3.2 - Specification and Validation of Tropical Agriculture Monitoring Methods and Data

In this task, the project proposal has the following adjusted milestone for month 36:

Milestone M3.2.2: Mapping agriculture in Cerrado and Amazonia

This milestone has accomplished, led by the post-doc selected by project, Dr. Michelle Picoli, who has a PhD in Agricultural Engeneering from Campinas State University (UNICAMP) and has also worked in CTBE (Bioethanol Technology Centre). Dr. Picoli joined the project at the start of 2017.

The main work in this Task was the development of innovative methods for using satellite image time series to produce land use and land cover classification over large areas in Brazil from 2001 to 2016[5][14]. We used MODIS time series data to classify natural and human-transformed land areas in state of Mato Grosso, Brazil's agricultural frontier. Using the sits R package (see Section 3.2.3 above), we took the full depth of satellite image time series to create large dimensional spaces for statistical classification. Data consists of MODIS MOD13Q1 time series with 23 samples per year per pixel, and 4 bands (NVDI, EVI, nir and mir). By taking a series of labelled time series, we fe a support vector machine model with a 92-dimensional attribute space. Using a 5-fold cross validation, we obtained an overall accuracy of 94% for discriminating among 9 land cover classes: *forest, cerrado, pasture, soybean-fallow, fallow-cotton, soybean-cotton, soybean-corn, soybean-millet and soybean-sunflower*. Producer's and user's accuracies of all classes were close to or better than 90%.

The results point out to important trends in agricultural intensification in Mato Grosso. Double cropping systems are now the most common production system in the state, thus increasing the potential for land sparing. Pasture expansion and intensification has been less studied than crop expansion, although it has a stronger impact on deforestation and GHG emissions. Our

results points to a significant increase in stocking rate in Mato Grosso, and to the possible abandonment of pasture areas opened in the state's frontier. The detailed land cover maps contribute to the assessment of the interplay between production and protection in the Brazilian Amazonian and Cerrado biomes. Two of the resulting classification maps for Mato Grosso (in 2005 and 2016) are shown in Figure 5.

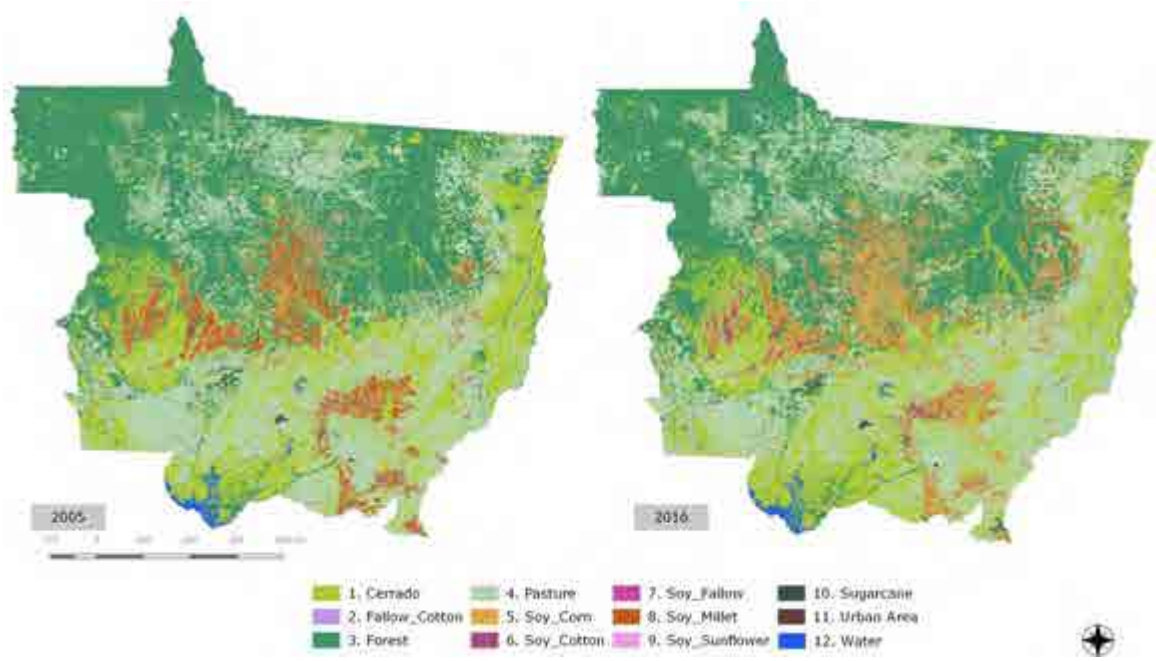


Figure 5 – Land use and land cover classification for Mato Grosso (2005 and 2016).

Among the many significant results we obtained by analyzing the resulting data set, we highlight the change in stocking rate. According to our classification, pasture area in Mato Grosso between 2005 and 2015 declined 4.6 million hectares, from 28.1 to 23.5 million hectares. According to IBGE, the number of cattle heads in the state has increased from 26.7 in 2005 to 29.3 million in 2015, a growth of 10%. In Figure6, we join our results with IBGE data in cattle herds to show that the stocking rate in Mato Grosso has grown steadily. The cattle heads grew by 10%, while pasture decreased by 16% between 2005 and 2015. This is a significant result, because it shows that the relative pressure of cattle raising for increasing deforestation is being reduced. In general, there is a trend towards pasture intensification coupled with abandonment of frontier areas, especially those at the most Northern part of the state.

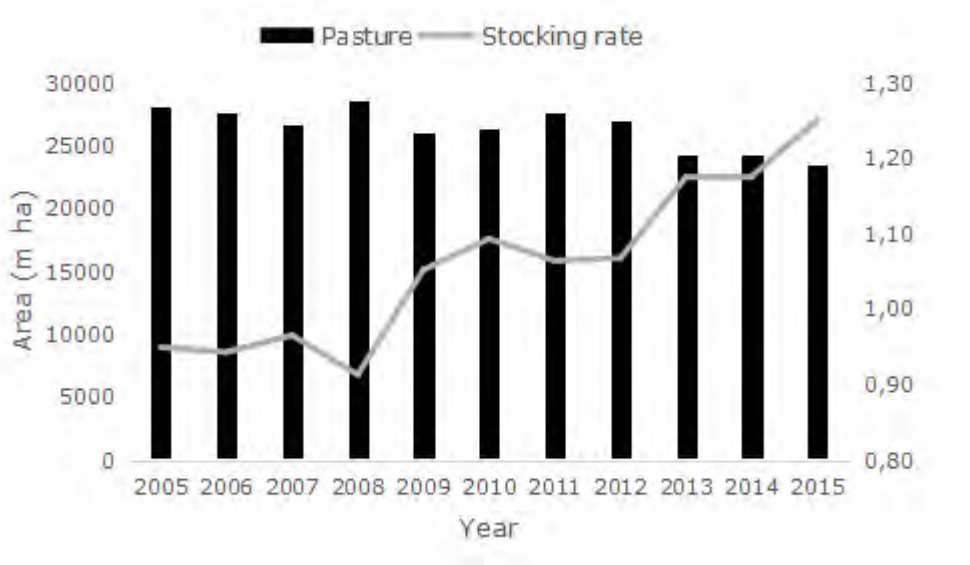


Figure 6 – Change in stocking rate (number of heads of cattle divided by pasture area) from 2005 to 2016 in Mato Grosso.

In Year 4 of the project, we intend to produce complete land use and land cover maps for the whole of Brazil, covering the main biomes.

4 Institutional support received in the period

The e-sensing project is hosted primarily by the Image Processing Division (DPI) of the National Institute of Space Research (INPE), with additional support provided by INPE's Remote Sensing Division (DSR). Both divisions report to the Earth Observation Directorate (OBT). During year 1 of the project, DPI gave substantial institutional support to the project, led by its head (Dr. Lubia Vinhas) as follows:

1. DPI/INPE hired, with additional funds from other projects, a full-time post-doc researcher (Dr. Eduardo Llapa) who is 100% dedicated to the project.
2. DPI/INPE also hired, with additional funds from other projects, a project support person (Ms. Denise Nascimento) who provides essential support for project management.
3. DPI/INPE also provides the IT infrastructure for hosting the data servers bought by the project with support from FAPESP, and for hosting the project's website (http://www.esensing.org).

The institutional support we are receiving from DPI/INPE is very good and fulfills the needs of the project.

5 Activities planned for project year 4 (January – December 2018)

5.1 Planned activities for WP 1 - Big Earth observation databases

5.1.1 Task 1.1 - Building and deployment of big Earth observation databases to support data analysis and use cases

Milestone M1.1.4 - Version 3 of the database for regional use cases (month 42)

This milestone will consist of loading in the database the data for the use cases to be done in year 4. In this case, we expect to include a significant number of LANDSAT images to complement the MODIS data that has been already inserted in the SciDB array manager. This data set will include images for Latin America, including Bolivia, Peru and Argentina.

5.1.2 Task 1.2 – Extend SciDB for geographical data handling

Milestone M1.2.4 – Server-side processing of time series classification learning methods using the sits package (month 48)

Given the problems we faced when developing a Web Service based on the SciDB array database, we are reviewing the original milestone (as explained in Section 3.1.2) and merging this goal with Task 2.2 (see below)

5.2 Planned activities for WP 2 - Data analysis for big Earth observation data

5.2.1 Task 2.1 - Exploratory big data analysis

As explained above, this task has been consider as having been completed at the end of Year 2. No further work on the task is considered to be required.

5.2.2 Task 2.2 – Space-time analysis of big Earth observation data for land change monitoring

*Milestone M2.2.4: Complete version of **sits** R package (month 36)*

As explained above, we decided to focus our software development effort in a single R package, called sits. Currently, as explained in Section 3.2.2, the package has a large set of clustering, filtering and machine learning methods. It has been successfully applied to the classification of agricultural areas (see Section 3.3.2). Currently, the server-side processing facilities that

use the SciDB data base are still not integrated in the package. In line with what we explained in Section 5.1.2, we will integrate client-side and server-side processing in sits. We also aim to include methods for detection of forest degradation (developed in Task 3.1) in the package, thus ensuring that all important conceptual advances in Forestry and Agriculture produced by the project can be reproduced.

5.3 Planned activities for WP 3 - Use case development

5.3.1 Task 3.1 - Specification and validation of tropical forest change alert methods and data

Milestone M3.1.4 Inclusion of detection of clear cut and degradation in the sits R package (month 48)

As explained above, the important efforts of mapping and identifying areas of forest degradation have not yet been integrated in the R sits package. In year 4, we intend to put together the software experts and forest experts of the project team. The result will be the availability of methods for detection of forest degradation in the sits package.

5.3.2 Task 3.2 - Specification and Validation of Tropical Agriculture Monitoring Methods and Data

Milestone M3.2.4 Full mapping of agriculture in selected Brazilian biomes (month 48)

To finish the project, in Year 4, we intend to produce a map of agricultural areas in selected Brazilian biomes. We intend to provide a full cropland and pastureland map of the Amazonia and Cerrado biomes, and a map of selected areas in other biomes.

6 Data Management Policy

We are following the policy we stated in the project proposal, as follows:

> *Our policy will be to deal with the databases and software created by this project as a resource to be shared with the Brazilian Earth Observation community. Thus, we will open the database after month 24 of the project to the community. We will encourage scientists to develop new data analysis methods and to use the methods and algorithms we will build to develop new applications. We will maintain the database accessible and updated for long-term use by the scientific community.*

During 2017, we undertook the following actions regarding implementing the proposed data management policy:

a) We established a partnership with the EMBRAPA Centre for Agricultural Informatics (CNPTIA), so that their experts are now using the sits package to produce their own applications on agricultural mapping. We held a training course on sits for EMBRAPA on September 2017.
b) We visited the research group on Remote Sensing at the University Federal de Goiás, led by prof. Laerte Ferreira. We established a partnership to allow them to use early releases of sits R package for their studies.

In our contacts with other Brazilian research groups, we found out that they have developed a good capacity of putting big data sets together as large sets of flat files. They told us they would rather use their existing data sets than having access to INPE's database. For this reason, we have included in the sits R package the capacity to perform classification on large sets of flat files. This feature is already part of the current open source version. The team at EMBRAPA has informed us that they are using it with success.

7 Final Remarks

The "e-Sensing" project has achieved important results in Year 3. The development of machine learning methods for satellite image time series analysis has been a major development in 2017. The other important result was the development of an R package that will encapsulate the most important project advances. These results will allow INPE and other research groups in Brazil to to produce new maps of land use change in Brazil. These maps will allow new insights into the trade-offs between environmental protection and agricultural production in Brazil.

As we remarked in the Year 1 Report, another important, although intangible result, was to have built an interdisciplinary approach to the problem of big Earth observation data handling. We held frequent seminars and workshops with the full project team, so that researchers could present their different viewpoints. It has been vitally important to have such discussions. As a result, all team members have deepened their understanding of the complex problem we will try to solve in the coming years.

8 PAPERS, SOFTWARE AND DATA PUBLISHED IN 2017

DOCTORAL DISSERTATIONS

1. MÁRCIO AZEREDO. Mineração e análise de trajetórias de mudança de cobertura da terra: explorando padrões comportamentais no contexto da degradação florestal. Doctoral dissertation in Applied Computing, INPE, 2017. Advisors: ISABEL ESCADA, MIGUEL MONTEIRO.
2. ADELINE MACIEL, Spatiotemporal Interval Logic for Reasoning About Land Use Change Dynamics. Doctoral dissertation in Earth System Science, INPE 2017. ADVISORS: LUBIA VINHAS, GILBERTO CÂMARA.

MASTER THESIS

3. VINICIUS CAPANEMA. Fatores de Degradação Florestal Atuantes em diferentes estágios da fronteira agropecuária na Amazônia: Estudo de caso na região de Sinop, MT. 2017. MSc in Remote Sensing. INPE 2017. Advisor: ISABEL ESCADA.
4. ALANA KASAHARA NEVES. Mineração e dados de sensoriamento remoto para detecção e classificação de áreas de pastagem na Amazônia Legal. 2017. MSc in Remote Sensing. INPE 2017. Advisor: THALES KÖRTING.

DATA SETS SUBMITTED TO PUBLIC REPOSITORIES

5. GILBERTO CÂMARA, MICHELLE PICOLI, ROLF SIMOES, ADELINE MACIEL, ALEXANDRE CARVALHO, ALEXANDRE COUTINHO, JULIO ESQUERDO, JOÃO ANTUNES, RODRIGO BEGOTTI, DAMIEN ARVOR (2017): Land cover change maps for Mato Grosso State in Brazil: 2001-2016, links to files. *PANGAEA*, https://doi.org/10.1594/PANGAEA.881291

SOFTWARE PACKAGES DEVELOPED

6. ROLF SIMÕES, GILBERTO CÂMARA, ALEXANDRE CARVALHO, VICTOR MAUS. SITS: Satellite Image Time Series Analysis. R package available at https://github.com/e-sensing/sits.
7. LUIZ ASSIS, GILBERTO RIBEIRO, VICTOR MAUS. wtss: An R Client for a Web Time-Series Service. Available at `https://CRAN.R-project.org/package=wtss`
8. ADELINE MACIEL, lucC: Land Use Change Calculus. R package available at https://github.com/ammaciel/lucC.

PAPERS PUBLISHED IN INTERNATIONAL JOURNALS

9. MARIANE REIS, LUCIANO DUTRA, ISABEL ESCADA. Examining Multi-Legend Change Detection in Amazon with Pixel and Region Based Methods. *Remote Sensing*, v. 9, p. 77, 2017. DOI:10.3390/rs9010077.

10. FABIEN WAGNER, BRUNO HERAULT, VIVIEN ROSSI, THOMAS HILKER, EDUARDO MAEDA, ALBER SANCHEZ, ALEXEI LYAPUSTIN, LÊNIO GALVÃO, YUJIEWANG, LUIZ ARAGÃO. Climate drivers of the Amazon forest greening. PLOS One, 12(7): e0180932. DOI: 10.1371/journal.pone.0180932.

PAPERS ACCEPTED IN INTERNATIONAL JOURNALS

11. DENISE MARTINI, LUIZ ARAGÃO, IEDA SANCHES, MARCELO VALADARES, CINTHIA SILVA, ELOI DALL-NORA. Land availability for sugarcane derived jet-biofuels in São Paulo—Brazil. *Land Use Policy*, vol.70, pp. 256-262 (January 2018).

12. VICTOR MAUS, GILBERTO CÂMARA, EDZER PEBESMA, MARIUS APPEL. dtwSat: Time-Weighted Dynamic Time Warping for Satellite Image Time Series Analysis in R. Accepted by the *Journal of Statistical Software*.

13. IEDA SANCHES, RAUL FEITOSA, PEDRO DIAZ, MARINALVA SOARES, ALFREDO LUIZ, BRUNO SCHULTZ, LUIS MAURANO. Campo Verde Database: Seeking to Improve Agricultural Remote Sensing of Tropical Areas. Accepted for publication in *IEEE Geoscience and Remote Sensing Letters*.

PAPERS SUBMITTED TO INTERNATIONAL JOURNALS

14. MICHELLE PICOLI, GILBERTO CAMARA, IEDA SANCHES, ROLF SIMÕES, ALEXANDRE CARVALHO, ADELINE MACIEL, ALEXANDRE COUTINHO, JULIO ESQUERDO, JOÃO ANTUNES, RODRIGO BEGOTTI, DAMIEN ARVOR, CLAUDIO ALMEIDA. Big Earth Observation Time Series Analysis for Monitoring Brazilian Agriculture. Submitted to *ISPRS Journal of Photogrammetry and Remote Sensing* (under review).

15. ADELINE MACIEL, GILBERTO CÂMARA, LÚBIA VINHAS, MICHELLE PICOLI, RODRIGO BEGOTTI, LUIZ ASSIS. Spatiotemporal interval logic for reasoning about land use change dynamics. Submitted to *Inter. Journal of Geographical Information Science* (2nd revision).

PAPERS PUBLISHED IN BRAZILIAN JOURNALS

16. RENNAN MARUJO, LEILA FONSECA, THALES KORTING, HUGO BENDINI, GILBERTO QUEIROZ, LÚBIA VINHAS, KARINE FERREIRA. Remote Sensing Image Processing

Functions in Lua Language. Accepted for publication *in Journal of Computational Interdisciplinary Sciences*, 2018.

PAPERS ACCEPTED IN BRAZILIAN JOURNALS

17. VINICIUS CAPANEMA, TAISE PINHEIRO, ISABEL ESCADA, SIDNEY SANT'ANNA. Mapeamento de Padrões de Intensidade de Degradação Florestal: Estudo de caso na Região de Sinop, Estado do Mato Grosso. *Revista Brasileira de Cartografia*. Submitted and accepted in 2017.

PEER-REVIEWED PAPERS IN SCIENTIFIC CONFERENCES

18. VITOR GOMES, GILBERTO QUEIROZ, KARINE FERREIRA, LUCIANE SATO, RAFAEL SANTOS, FABIANO MORELLI. Um ambiente para análise exploratória de grandes volumes de dados geoespaciais: explorando risco de fogo e focos de queimadas. Anais do 18o. Simpósio Brasileiro de Geoinformática", GEOINFO 2017. Salvador, 04-06 dez. 2017.

19. ALANA NEVES, THALES KORTING, LEILA FONSECA, GILBERTO QUEIROZ, LÚBIA VINHAS, KARINE FERREIRA, ISABEL ESCADA. TerraClass x MapBiomas: Comparative assessment of legend and mapping agreement analysis. Anais do 18o. Simpósio Brasileiro de Geoinformática", GEOINFO 2017. Salvador, 04-06 dez. 2017.

20. WANDERSON COSTA, LEILA FONSECA, THALES KORTING, MARGARETH SIMÕES, HUGO BENDINI, RICARDO SOUZA. Segmentation of optical remote sensing images for detecting homogeneous regions in space and time. Anais do 18o. Simpósio Brasileiro de Geoinformática", GEOINFO 2017. Salvador, 04-06 dez. 2017.

21. RENNAN MARUJO, LEILA FONSECA, THALES KORTING, HUGO BENDINI. Spectral normalization between Landsat-8/OLI, Landsat- 7/ETM+ and CBERS-4/MUX bands through linear regression and spectral unmixing. Anais do 18o. Simpósio Brasileiro de Geoinformática", GEOINFO 2017. Salvador, 04-06 dez. 2017.

22. MARIANE REIS, LUCIANO DUTRA, ISABEL ESCADA. Simultaneous multi-source and multi-temporal land cover classification using a Compound Maximum Likelihood classifier. Anais do 18o. Simpósio Brasileiro de Geoinformática", GEOINFO 2017. Salvador, 04-06 dez. 2017.

23. ALBER SANCHEZ, LÚBIA VINHAS, GILBERTO QUEIROZ, ROLF SIMÕES, VITOR GOMES, LUIZ ASSIS, EDUARDO LLAPA, GILBERTO CAMARA. Reproducible geospatial data science: Exploratory Data Analysis using collaborative analysis environments. Anais do 18o. Simpósio Brasileiro de Geoinformática", GEOINFO 2017. Salvador, 04-06 dez. 2017.

24. GILBERTO CAMARA, GILBERTO QUEIROZ, LÚBIA VINHAS, KARINE FERREIRA, RICARDO CARTAXO, ROLF SIMÕES, EDUARDO LLAPA, LUIZ ASSIS, ALBER SANCHEZ. "The e-Sensing architecture for big Earth observation data analysis". Proc. of the 2017 conference on Big Data from Space (BiDS 17). P. Soille and P.G. Marchetti (eds.). Toulouse, France, December 2017

25. ADELINE MACIEL, LUBIA VINHAS, GILBERTO CAMARA, VICTOR MAUS, LUIZ ASSIS. STILF - A spatiotemporal interval logic formalism for reasoning about events in remote sensing data. Anais do 18º. Simpósio Brasileiro de Sensoriamento Remoto, SBSR 2017. Santos, 28-31 maio 2017. ISBN= 978-85-17-00088-1.

26. ADELINE MACIEL, LÚBIA VINHAS. Time series classification using features extraction to identification of land use and land cover: A case study in the municipality of Itaqui, South Region of Brazil. Anais do 18º. Simpósio Brasileiro de Sensoriamento Remoto, SBSR 2017. Santos, 28-31 maio 2017. ISBN= 978-85-17-00088-1.

27. LUBIA VINHAS, MATHEUS ZAGLIA. Exploring the OpenSearch extension to disseminate Earth Observation Data. Anais do 18º. Simpósio Brasileiro de Sensoriamento Remoto, SBSR 2017. Santos, 28-31 maio 2017. ISBN= 978-85-17-00088-1.

28. HUGO BENDINI, LEILA FONSECA, THALES KORTING, IEDA SANCHES, RENNAN MARUJO. Evaluation of smoothing methods on Landsat-8 EVI time series for crop classification based on phenological parameters. In: Simpósio Brasileiro de Sensoriamento Remoto - SBSR, 2017, Santos. XVIII Simpósio Brasileiro de Sensoriamento Remoto - SBSR, 2017. p. 4267-4274.

29. PEDRO DIAZ, RAUL FEITOSA, FRANZ ROTTENSTEINER, IEDA SANCHES, CHRISTIAN HEIPKE. Spatio-temporal Conditional Random Fields for recognition of sub-tropical crop types from multi-temporal images. In: Simpósio Brasileiro de Sensoriamento Remoto - SBSR, 2017, Santos. XVIII Simpósio Brasileiro de Sensoriamento Remoto - SBSR, 2017. p. 2539-2546.

30. JULIO GUERRA, BRUNO SCHULTZ, IEDA SANCHES. Mapeamento automático da expansão da agricultura anual no MATOPIBA entre 2002 e 2015 utilizando a plataforma Google Earth Engine. In: Simpósio Brasileiro de Sensoriamento Remoto - SBSR, 2017, Santos. XVIII Simpósio Brasileiro de Sensoriamento Remoto - SBSR, 2017. p. 6850-6857.

31. ALFREDO LUIZ, IEDA SANCHES, MARCOS NEVES. Mudança no uso da terra pela agricultura brasileira de 1990 a 2014. In: Simpósio Brasileiro de Sensoriamento Remoto - SBSR, 2017, Santos. XVIII Simpósio Brasileiro de Sensoriamento Remoto - SBSR, 2017. p. 4002-4009.

32. RODOLFO MANJOLIN, CÉLIA GREGO, SANDRA NOGUEIRA, GUSTAVO BAYMA-SILVA, KLEBER TRABAQUIM, IEDA SANCHES. Variabilidade espacial da fertilidade, carbono e nitrogênio do solo em áreas de pastagem e cana-de-açúcar no estado de São Paulo. In: Simpósio Brasileiro de Sensoriamento Remoto - SBSR, 2017, Santos. XVIII Simpósio Brasileiro de Sensoriamento Remoto - SBSR, 2017. p. 7163-7170.

33. BRUNO MONTIBELLER, ALFREDO LUIZ, IEDA SANCHES, HILTON SILVEIRA. Análise da variabilidade espectro-temporal intraespecífica do milho. In: Simpósio Brasileiro de Sensoriamento Remoto - SBSR, 2017, Santos. XVIII Simpósio Brasileiro de Sensoriamento Remoto - SBSR, 2017. p. 2011-2018.

34. HILTON SILVEIRA, ISAQUE EBERHARDT, IEDA SANCHES, LENIO GALVAO. Análise da cobertura de nuvens no nordeste do Brasil e seus impactos no sensoriamento remoto agrícola operacional. In: Simpósio Brasileiro de Sensoriamento Remoto - SBSR, 2017, Santos. XVIII Simpósio Brasileiro de Sensoriamento Remoto - SBSR, 2017. p. 400-407.

35. KLEBER TRABAQUINI, GUSTAVO SILVA, IEDA SANCHES, SANDRA NOGUEIRA, DENILSON DORTZBACH. Avaliação espaço-temporal da cultura da cana-de-açúcar no oeste paulista. In: Simpósio Brasileiro de Sensoriamento Remoto - SBSR, 2017, Santos. XVIII Simpósio Brasileiro de Sensoriamento Remoto - SBSR, 2017. p. 4764-4771.

36. JOSE BERMUDEZ, RAUL FEITOSA, LAURA CUE, PEDRO DIAZ, IEDA SANCHES. A Comparative Analysis of Deep Learning Techniques for Sub-Tropical Crop Types Recognition from Multitemporal Optical/SAR Image Sequences. In: 2017 30th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), 2017, Niterói. 2017 30th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), 2017. p. 382.

37. JOSE BERMUDEZ, PEDRO DIAZ, IEDA SANCHES, LAURA CUE, RAUL FEITOSA. Evaluation of Recurrent Neural Networks for Crop Recognition from Multitemporal Remote Sensing Images. In: XXVII Congresso Brasileiro de Cartografia e XXVI Exposicarta, 2017, Rio de Janeiro. Anais do XXVII Congresso Brasileiro de Cartografia e XXVI Exposicarta, 2017. p. 800-804.

38. LAURA CUE, JOSE BERMUDEZ, PEDRO DIAZ, IEDA SANCHES, P. HAPP, RAUL FEITOsa. A Comparative Analysis of Deep Learning Techniques for Crop Type Recognition in Temperate and Tropical Regions From Multitemporal SAR Image Sequences. In: Congresso Brasileiro de Cartografia, 2017, Rio de Janeiro. Anais do XXVII Congresso Brasileiro de Cartografia e XXVI Exposicarta, 2017. p. 730-734.

39. ALEXSANDRO SILVA, LEILA FONSECA, THALES KORTING. A multitemporal approach for land use mapping using Bayesian Networks. In: Simpósio Brasileiro de Sensoriamento Remoto, 18 (SBSR), 2017, Santos. Anais do XVIII SBSR, 2017. v. 18.

40. AINDOMAR SILVA, CATHERINE ALMEIDA, NATÁLIA WIEDERKEHR, RENATA RIBEIRO, THALES KORTING. Application of the Linear Spectral Mixture Model in vegetation change detection based on the Green Vegetation Index. In: Simpósio Brasileiro de Sensoriamento Remoto, 18 (SBSR), 2017, Santos. Anais do XVIII SBSR, 2017. v. 18.

41. RENNAN MARUJO, LEILA FONSECA, THALES KORTING, RAFAEL SANTOS, HUGO BENDINI. CBERS-4/MUX automatic detection of clouds and cloud shadows using decision trees. In: Simpósio Brasileiro de Sensoriamento Remoto, 2017, Santos. Anais do XVIII SBSR, 2017. v. 18.

42. MIKHAELA PLETSCH, THALES KORTING, ISABEL ESCADA, SACHA SIANI. Data mining applied to temporal dynamics of deforestation pattern: a study case in Southern Amazon forest, Brazil. In: Simpósio Brasileiro de Sensoriamento Remoto, 2017, Santos. Anais do XVIII SBSR, 2017. v. 18.

43. CESARE GIROLAMO NETO, LEILA FONSECA, DALTON VALERIANO, ALANA NEVES, THALES KORTING. Desafios na classificação automática de fitofisionomias do Cerrado brasileiro com base em mapas de referência na escala 1:250.000. In: Simpósio Brasileiro de Sensoriamento Remoto, 2017, Santos. Anais do XVIII SBSR, 2017. v. 18.

44. THALES PENHA, ISABEL ESCADA, LEILA FONSECA, THALES KORTING. Detecção de mudanças e análise de padrões de expansão agrícola na Amazônia mato-grossense. In: Simpósio Brasileiro de Sensoriamento Remoto, 2017, Santos. Anais do XVIII SBSR, 2017. v. 18.

45. THALES PENHA, LEILA FONSECA, THALES KORTING. Inferência fuzzy na análise de vulnerabilidade de fragmentos florestais na Amazônia mato-grossense. In: Simpósio Brasileiro de Sensoriamento Remoto, 2017, Santos. Anais do XVIII SBSR, 2017. v. 18.

46. ALANA NEVES, THALES KORTING, CESARE GIROLAMO NETO, LEILA FONSECA. Mineração de dados de sensoriamento remoto para detecção e classificação de áreas de pastagem na Amazônia Legal. In: Simpósio Brasileiro de Sensoriamento Remoto, 18, 2017, Santos. Anais do XVIII SBSR, 2017. v. 18.

47. SOUZA, G. M., VINICIUS CAPANEMA, ISABEL ESCADA. Cicatrizes de queimadas e padrões de mudanças de uso e cobertura da terra no sudoeste do estado do Pará, Brasil. In: Simpósio brasileiro de Sensoriamento Remoto, 2017, Santos. Anais do SBSR. São José dos Campos: INPE, 2017. p. 5760-5767.

48. VINICIUS CAPANEMA, ISABEL ESCADA. Degradação Florestal na Amazônia: Geração de um Indicador Espacial de Suscetibilidade à Degradação Florestal. In: Simpósio brasileiro de Sensoriamento Remoto, 2017, Santos. Anais do XVIII SBSR. São José dos Campos: INPE, 2017. p. 6994-7001.

49. LUIZ MAURANO, ISABEL ESCADA, CAMILO RENNÓ. Desmatamento da Amazônia: O DETER pode ser utilizado como preditor das taxas anuais de desmatamento geradas pelo PRODES?. In: Simpósio brasileiro de Sensoriamento Remoto, 2017, Santos. Anais do XVIII SBSR. São José dos Campos: INPE, 2017. p. 4306-4313.

50. MARIANE REIS, SIDNEY SANT'ANNA, LUCIANO DUTRA, ISABEL ESCADA, ELIANA PANTALEÃO. The use of land cover change likelihood for improving land cover classification. In: IEEE INTERNATIONAL GEOSCIENCE AND REMOTE SENSING SYMPOSIUM (IGARSS), Fort Worth, Texas. 2017.

51. MARIANE REIS, ISABEL ESCADA, LUCIANO DUTRA, SIDNEY SANT'ANNA. Harmonização de legendas formalizadas em Land Cover Meta Language-LCML. In: Simpósio brasileiro de Sensoriamento Remoto, 2017, Santos. Anais do XVIII SBSR. São José dos Campos: INPE, 2017. p. 863-870.

52. C. SILVA JUNIOR, LUIZ ARAGAO, LIANA ANDERSON, ISABEL ESCADA. O papel da fragmentação e do efeito de borda na ocorrência e intensidade de incêndios florestais na Amazônia. In: Simpósio brasileiro de Sensoriamento Remoto, 2017, Santos. Anais do XVIII SBSR. São José dos Campos: INPE, 2017. p. 5952-5959.

ANNEXES

1. **Report of use of Technical Reserve and Complementary Benefits (in Portuguese)**

2. **Reports for Technical Training Scholarships:**

   2.1 Luiz Fernando Assis (2015/19540-0)
   2.2 Alber Sanchez Ipia (2016/16555-0)

3. **Reports for Postdoctoral Scholarships**

   3.1 Rodrigo Anzolin Begotti (2016/16968-2)
   3.2 Michelle Picoli

4. **Reports for Doctoral Scholarships**

   4.1 Rennan de Freitas Bezerra Marujo (2016/08719-2)

5. **Initial Pages of Papers published and submitted**

FORMULÁRIO OBRIGATÓRIO

Descrição Sucinta do uso da Reserva Técnica de Auxílio

NÚMERO DO PROCESSO: 14/08398-6

NOME DO OUTORGADO: GILBERTO CAMARA NETO

DESCRIÇÃO (Se houve gasto, descreva)

Os recursos da reserva técnica do projeto foram utilizados para cobrir as seguintes despesas

1. DIÁRIAS E DESPESAS DE TRANSPORTE

- Participação da Dra.Taíse Pinheiro em reuniões técnicas do projeto, no INPE, em São José dos Campos/SP.

2. MATERIAL PERMANENTE

- Foi adquirido um GPS para auxiliar os trabalhos de campo do projeto.

3. SERVIÇO DE TERCEIROS

- Pagamento de taxa de inscrição para participação em conferência: bolsista Alber Sanchez Ipia, GEOINFO 2017 (trabalho intitulado: "Reproducible geospatial data science: Exploratory Data Analysis using collaborative analysis environments".

☐ **Declaro que não houve utilização dos Recursos da Reserva Técnica**

Local, data e assinatura do Outorgado: São José dos Campos, 30.01.2018

FORMULÁRIO OBRIGATÓRIO

Descrição Sucinta do uso dos Benefícios Complementares

NÚMERO DO PROCESSO: 14/08398-6

NOME DO OUTORGADO: GILBERTO CÂMARA NETO

DESCRIÇÃO (Se houve gasto, descreva)

1. BENEFICIOS COMPLEMENTARES DA PESQUISADORA DRA.LEILA FONSECA

(a) Compra de material de consumo: suporte para monitor (autorização da FAPESP via SAGE).

(b) Passagem, diárias e taxa de inscrição para Participação do bolsista Cesare Neto em curso na Espanha (INIT/ AERFAI Summer School on Machine Learning) - autorização da FAPESP via SAGE.

(c) Passagem, diárias, seguro saúde e taxa de inscrição para Participação do bolsista Raian Maretto em curso no Canadá (Deep Learning and Reinforcement Learning Summer Schools - DLSS & RLSS) - autorização da FAPESP via SAGE.

(d) Pagamento de taxa de inscrição para participação do bolsista Alexsandro Silva na Conferência "Spatial Statistics 2017" em Oxford, UK - autorização da FAPESP via SAGE.

(e) Pagamento de diárias e taxa de inscrição para participação dos bolsistas Alexsandro Silva, Rennan Marujo e do pesquisador Thales Körting no XVIII SBSR 2017 (Simpósio Brasileiro em Sensoriamento Remoto) - autorização da FAPESP via SAGE.

(f) Pagamento de revisão de inglês para publicações de trabalhos em revistas internacionais dos bolsistas Rennan Marujo e Anderson Soares (autorização da FAPESP via SAGE).

(g) Pagamento de taxa de inscrição para participação dos bolsistas Alana Neves e Wanderson Costa no GEOINFO 2017 (XVIII Brazilian Symposium on Geoinformatics) - autorização da FAPESP via SAGE.

☐ **Declaro que não houve utilização dos Benefícios Complementares**

Local, data e assinatura do Outorgado: São José dos Campos, 30.01.2018

Integration between R, TerraLib and SciDB

Luiz Fernando Ferreira Gomes de Assis

Coordinator: Gilberto Câmara

Co-Advisor: Karine R. Ferreira

Project Number: 2016/16555-0

INPE - São José dos Campos

November, 2017

List of illustrations

List of tables

Contents

1 Introduction

This scientific report aims to describe the whole progress of the project entitled *Integration between SciDB, TerraLib and R*[1] developed by the awarded technical training fellow Luiz Fernando Ferreira Gomes de Assis[2]. This project started on the 1st of December 2016 and ended on the 31th of August 2017. It is part of the e-Science Program [3], a thematic grant funded by São Paulo Research Foundation. The research addresses how the scientific community can use e-Science methods and techniques to improve the extraction and analysis of land use and land cover change information from big Earth Observation (EO) data sets in an open and reproducible way. This project is called *e-Sensing: Big Earth Observation Data Analytics for Land Use and Land Cover Change Information* and is coordinated by Prof. Dr. Gilberto Câmara Neto[4]. The e-Sensing team members consists of MSc. and PhD students, as well as Postdocs, and Researchers. Its excellence and its published papers are related to Geoinformatics and GIScience.

The main goal to develop this project is in the enhancement of the performance to integrate an statistical environment with a multidimensional array database in order to provide imaging routines for land use classification. The most used open source statistical environment is **R** since it is easily extensible through a substantial set of statistical functions and packages. Although the **R** environment provides a wide variety of graphical and statistical tools, it still has main memory and performance limitations when executing routines with large volumes of remote sensing images compared to other languages.

In this sense, the integration between **R** and multidimensional array databases such as SciDB can offer processing and analysis in the data server environment minimizing the data transfer between client and server by means of an interface with the **R** environment. For all the aforementioned reasons, we developed an interface between the **R** data analysis language, the SciDB multidimensional array database, and a library able to handle data stored in a PostGIS database such as TerraLib library. This architecture based on open-source tools was evaluated to assess how it meets the needs of Earth Observation (EO) scientists taking into consideration related works and a set of criteria to build researcher-friendly architectures for big EO data analysis.

The remainder of this scientific report is organized as follows. Section 2 contains an interoperable approach for integrating SciDB-**R**. Section 3 describes existing global LULC classification maps with a summary of the best practices for analysis. Section 4 comprises an important application case study for evaluating the integration. Section 5 contains a description of the complementary activities to develop and guide this project to future perspectives. Section 6 consists of a conclusion about a description and an evaluation of the institutional support. Fi-

[1] http://bv.fapesp.br/pt/bolsas/168252/integracao-entre-scidb-terralib-e-r/
[2] http://www.bv.fapesp.br/pt/pesquisador/669236/luiz-fernando-ferreira-gomes-de-assis/
[3] http://www.fapesp.br/8436
[4] http://www.bv.fapesp.br/pt/pesquisador/997/gilberto-camara-neto/

nally, the papers written for better documenting this project are included in the appendices.

2 Toward an interoperable SciDB-**R** integration

Integrating multidimensional array databases such as SciDB with **R** algorithms offers not only more flexibility programming of complex analysis but also overcome the memory limitation of those statistical environment to deal with large data sets. This combination mitigates the burden on scientists, interested on developing remote sensing time series analysis, by providing a friendly environment, big data management, and statistical computing tools. We considered in our architecture the classification of satellite imagery organized in a three dimensional array in space-time [4]. This architecture facilitates the analysis of big Earth Observation (EO) data, complex algorithms reusability, collaborative work within the scientific community and results validation. These features are achieved by combining a multidimensional array database [5], and a statistical analytics environment [6]. As shown in Figure 1, this architecture decision was taken based on one of the most promising research trends in big EO data analysis, the extraction of information from remote sensing time-series.

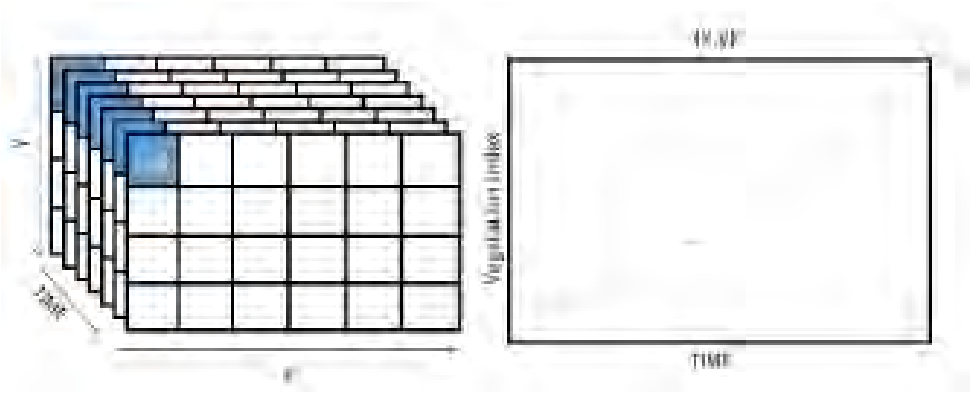


Figure 1: Remote Sensing Time Series Approach (1)

2.1 Web Time Series Processing Service (WTSPS) for Multidimensional Array Databases

The issue of handing an increasing number of geospatial data require new developments of geospatial technologies. An on-demand computing platform has emergerd to deal with remote sensing images based on a processing chain model. That approach coupled with interoperability provided by web services play an important role in big EO data processing. However, even considering the standards proposed by the Open Geospatial Consortium (OGC) for visualising, disseminating and processing geospatial data, we are still stuck on robust, inadequate and low latency performance approaches such as Web Map Service (WMS), Web Coverage Service

(WCS), Sensor Observation Service (SOS) and Web Processing Service (WPS) standards. For these reasons, we consider a new family of web services for scientists working with large sets of remote sensing time series. In this project, we designed a novel approach called Web Time Series Processing Service (WTSPS), which is better suited for processing large sets of EO time series using multidimensional array databases. Here, we illustrate how WTSPS works by presenting an approach overview, its use cases and a class diagram.

At first, we developed a intuitive interface in **R** so that domain specialists can use in an easily and fast manner. We plan to emphasize only WTSPS working in this document since WTSS was already presented to the scientific community. WTSPS aims to bridge the interoperability gap between scientists and big data processing throughout a set of standard operations for processing remote sensing time series organized in multidimensional array databases. That includes: 1. list algorithms, 2. describe algorithms, 3. get status process, 4. execute an algorithm and 5. manage permissions . Of course, more discussions are necessary but we do think these kickoff ideas are important to new insights. The initial idea is depicted in Figure 2.

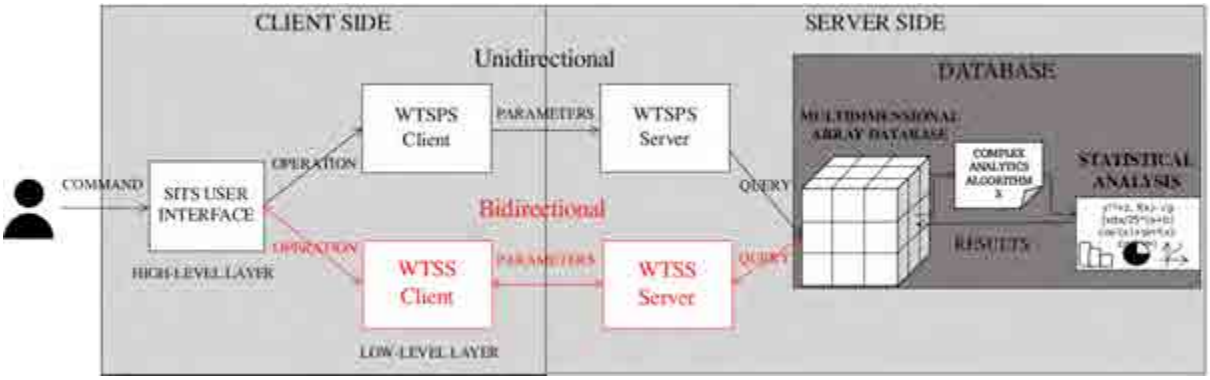


Figure 2: WTSPS Overview

After we specify some functionalities were necessary to develop a large processing service on the server side in the first year of this fellowship, a use case diagram of these requirements were created and presented in Figure 3. There, we can see that user can list the algorithms in which their scripts are stored on the server ready to be run for any data stored in a multidimensional array databases. Then, users are able get a description of their parameters, that is, their required input (e.g., format and type) and expected output (e.g., which attributes). Finally, users only can run based on specific roles, because it would be necessary a permission administration unit.

A class diagram of what we have discussed until here is depicted in Figure 4.

Figure 3: Use Case Diagram

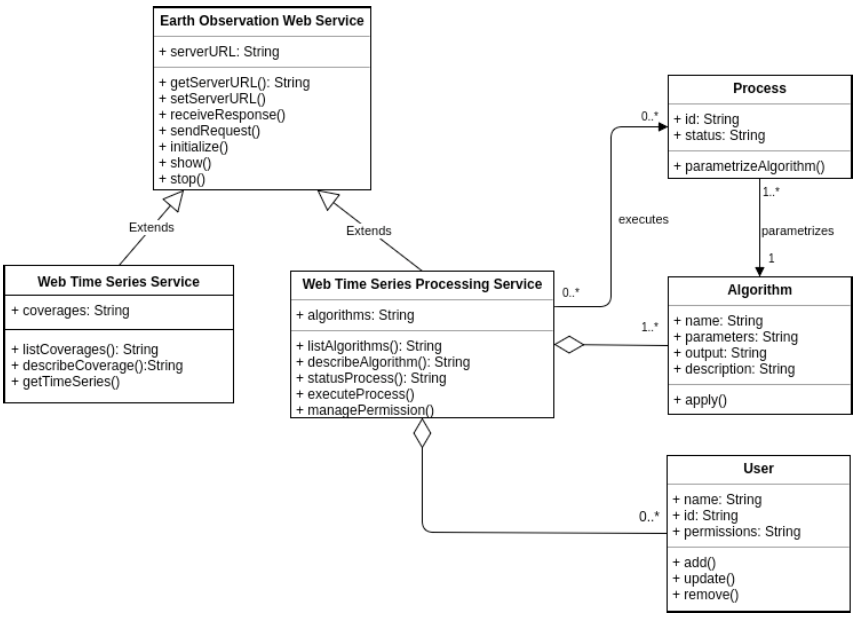


Figure 4: Class Diagram

3 Understanding the best practices for LULC change classification analysis

3.1 State of the art of large LULC change classification maps

With the appearance of technology and decreasing satellite data cost, many regional and global land cover products have been created in the last decade. With the availability of remote sensing products and improvements in variability, accessibility, and cost, land cover products have become essential inputs for interdisciplinary studies, for instance, to monitoring deforestation and change detection analysis [7]. In this section, we describe briefly two initiatives which provide land cover for all regions worldwide: GlobCover and GLC-SHARE. While these initiatives have specific legends, to provide as much detail as possible, in our approach we try to consider a particular legend in order to provide an adequate classification to the study regions. But before, we introduce some concepts about two MODIS products which we will use in our experiments.

3.1.1 Moderate-Resolution Imaging Spectroradiometer (MODIS)

The Moderate-Resolution Imaging Spectroradiometer (MODIS) data is a scientific instrument on board of the Terra and Aqua platforms that provides atmosphere, land, cryosphere and ocean features every 2 days (see Figure 5). This sensor collect not only raw data but also offers several products generated for specific applications. There are 460 non-fill tiles, tiles are separated by 10 degrees at the Earth's equator. The tile coordinate system starts at (0,0) (horizontal tile number, vertical tile number) in the upper left corner and proceeds to the right (horizontal) and downward (vertical). The tile in the bottom right corner is (35,17). They consist of 4,800 rows and 4,800 columns of 250 meter pixels [8].

We considered in this project the MOD13Q1 product that provides vegetation index values for each pixel each 16 days. The most important ones are the Normalized Difference Vegetation Index (NDVI) and the Enhanced Vegetation Index (EVI), which referred to the continuity index of the existing National Oceanic and Atmospheric Administration-Advanced Very High Resolution Radiometer (NOAA-AVHRR) derived NDVI, and the improved sensitivity over high biomass regions respectively. The MODIS contains reflectance bands 1 (Red), 2 (NIR), 3 (Blue), and 7 (MIR), as well as four observation layers, Table 1.

Another product is the MODIS Land Cover Type (MCD12Q1) product, that provides a collection of land cover types that support global change science by mapping global land cover using spectral and temporal information derived from MODIS. The MCD12Q1 product contains five classifications schemes, which describe land cover properties derived from observations spanning a year's of Terra and Aqua MODIS data. The first scheme identifies 17 classes defined by the International

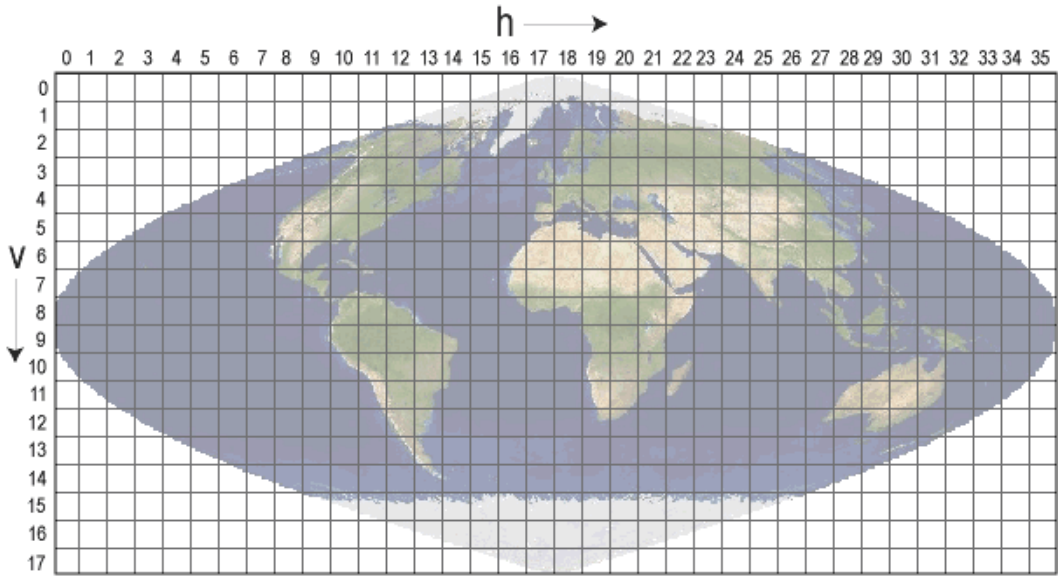


Figure 5: Sinusoidal Tile Grid

Table 1: Science Data Set Layer Characteristics

Description	Units	Data Type	Fill Value	Valid Range	Scaling Factor
250m 16 days NDVI	NDVI	16-bit signed integer	-3000	-2000 to 10000	0.0001
250m 16 days EVI	EVI	16-bit signed integer	-3000	-2000 to 10000	0.0001
VI quality indicators	Bit Field	16-bit unsigned	65535	0 to 65534	N/A
Surface Reflectance Band 1 (RED)	Reflectance	16-bit signed integer	-1000	0 to 10000 10000	0.0001
Surface Reflectance Band 2 (NIR)	Reflectance	16-bit signed integer	-1000	0 to 10000	0.0001
Surface Reflectance Band 3 (BLUE)	Reflectance	16-bit signed integer	-1000	0 to 10000	0.0001
Surface Reflectance Band 7 (MIR)	Reflectance	16-bit signed integer	-1000	0 to 10000	0.0001
View zenith angle of VI pixel	Degree	16-bit signed	-10000	0 to 18000	0.01
Sun zenith angle of VI pixel	Degree	16-bit signed	-10000	0 to 18000	0.01
Relative azimuth angle of VI pixel	Degree	16-bit signed	-4000	0 to -18000 to 18000	0.01
Day of year of VI pixel	Julian day of year	16-bit signed	-1	1 to 366	N/A
Quality reliability of VI pixel	Rank	8-bit signed integer	255	0 to 3	N/A

Geosphere Biosphere Programme (IGBP), which includes 11 natural vegetation classes, 3 develop and mosaicked land classes, and three non-vegetated land classes [8].

The land cover classes Land Cover Type 1 (IGBP) global vegetation classification scheme are: 0. *Water*, 1. *Evergreen Needleleaf forest*, 2. *Evergreen Broadleaf forest*, 3. *Deciduous Needleleaf forest*, 4. *Deciduous Broadleaf forest*, 5. *Mixed forest*, 6. *Closed shrublands*, 7. *Open shrublands*, 8. *Woody savannas*, 9. *Savannas*, 10. *Grasslands*, 11. *Permanent wetlands*, 12. *Croplands*, 13. *Urban and built-up*, 14. *Cropland/Natural vegetation mosaic*, 15. *Snow and ice*, 16. *Barren or sparsely vegetated* and 254. Unclassified [9].

3.1.2 GlobCover

The GlobCover, an European Space Agency (ESA) initiative started in 2005 in partnership with EEA, FAO, GOFC-GOLD, IGBP, JRC and UNEP is one of the projects we can study to samples and classes selection. This GlobCover developed a service for the generation for global composites and land cover maps based on observations from Envisat satellite MERIS (Medium Resolution Imaging Spectrometer Instrument) Fine Resolution surface reflectance mosaics, resolution 300m. Two land cover maps were available by ESA covering the periods: December 2004 to June 2006 and January until December 2009 [2].

The GlobCover 2009 land cover map is delivered as one global land cover map covering the entire Earth. Its legend, which counts 22 land cover classes, has been designed to be consistent at the global scale and therefore, it is determined by the level of information that is available and that makes sense at this scale [10]. Figure 6 presents the global GlobCover 2009 land cover map.

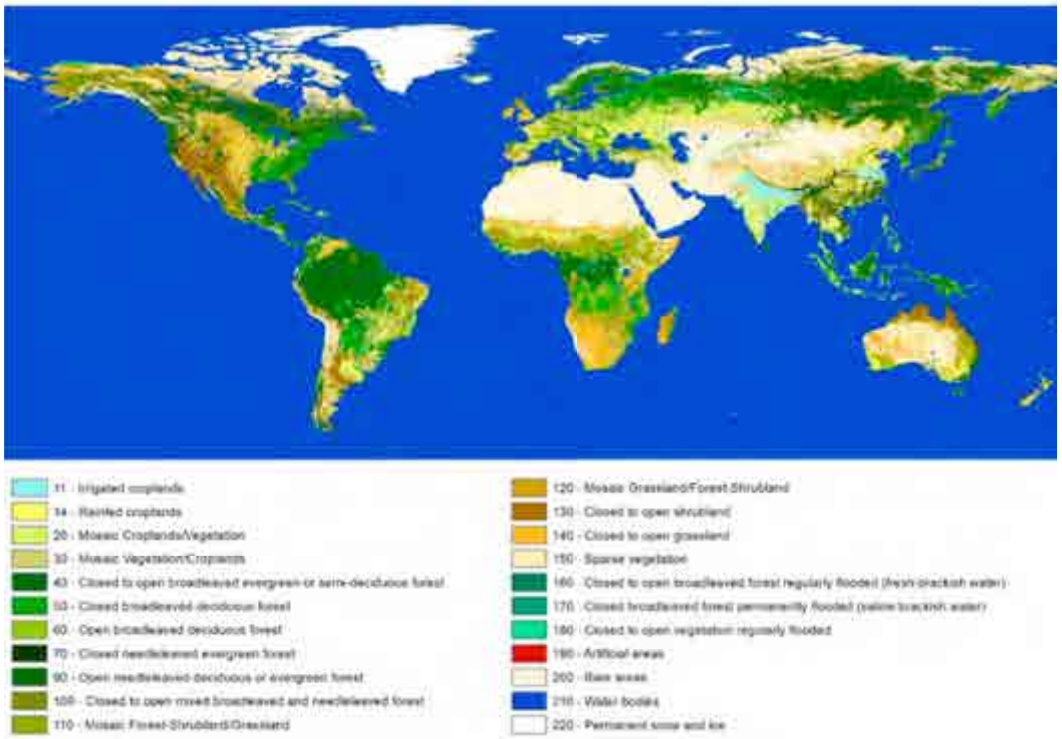


Figure 6: The GlobCover 2009 global land cover map with 22 classes legend. Figure adapted from 2

The classification module of the GlobCover processing chain consists in transforming the MERIS FR multispectral mosaics produced by the pre-processing modules into a meaningful global land cover map. The global land cover map has been produced in an automatic and global way and is associated with a legend defined and documented using the UN LCCS.

3.1.3 GLC-SHARE

The Global Land Cover-SHARE (GLC-SHARE) is a land cover database at the global level created by FAO, Land and Water Division in partnership and with contribution from various partners and institutions [11]. It provides a set of eleven thematic land cover layers resulting by a combination of available high resolution national, regional and/or sub-national land cover databases with the weighted average land cover information derived from large-scale datasets. The database is produced with a resolution of 30 arc-second (sqkm). The approach implemented is based on the utilization of the Land Cover Classification System (LCCS) and SEEA (System of Environmental-Economic Accounting) legend systems for the harmonization of the various global, regional and national land cover legends [3]. The benefit of the GLC-SHARE product is its capacity to preserve the existing and available high resolution land cover information at the regional and country level obtained by spatial and multi-temporal source data, integrating them with the best synthesis of global datasets. The database is distributed includes eleven layers, in raster format (GeoTIFF), each pixel values, the accuracy and associated information as source, date and resolution are indicated in the associated data quality indicator layer (see Figure 7).

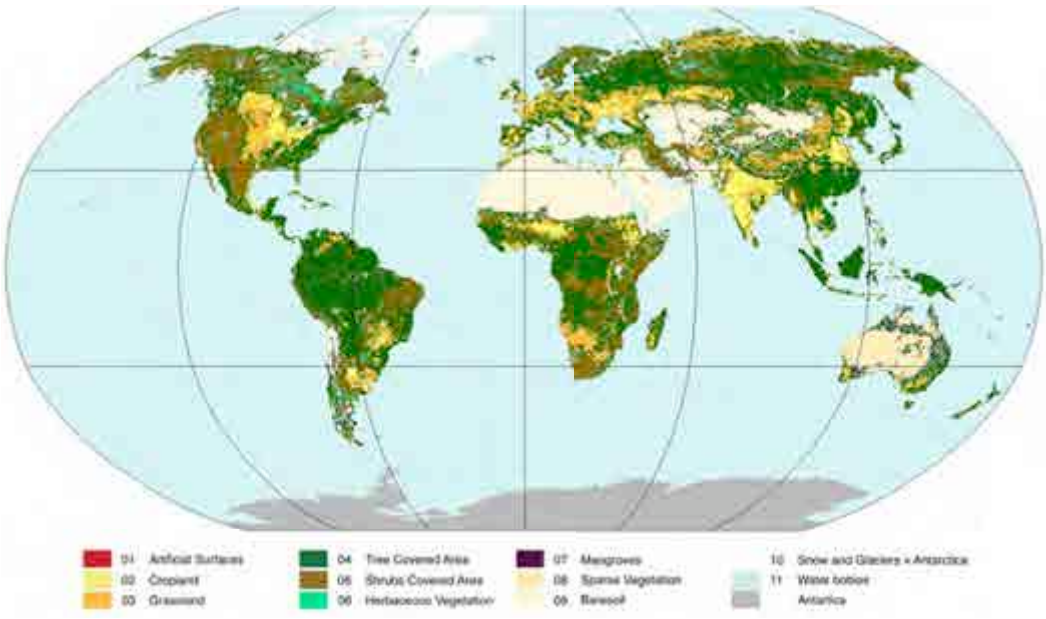


Figure 7: Distribution of dominant GLC-SHARE land cover. Adapted from 3

3.2 Applying best practices for classification analyisis

The methodology applied to the study case follows a standard workflow for land use land cover studies in supervised or semi-supervised approach, see [12] for a literature review. In these approaches we can recognize four main features: the sampling design, the classification method, the response design and the results analysis. Additionally, several relevant sub-features were included within each feature. The initial definitions include the spatial assessment unit, the sources of reference

data, the reference labeling protocol and the agreement definitions. The sampling consists of the stage for selecting a subset of the smallest spatial units that help the classification methods. It includes the interpretation and quality control, phenology characterization and the homogeneity measurement. For a while, we are just considering supervised methods for classification, specifically, the classification results provided by the TWDTW method **R** package, also known as dtwSat.

4 Application Case Study in Peru and Bolivia

Using the architecture and the methodology discussed in the previous chapters, we organized the case study accordingly to the following activities: sampling design, classification algorithm application and analysis. We considered here TWDTW algorithm to perform the classification by means of the SciDB-**R** streaming. TWDTW was implemented as an **R** extension package version of the robust and classic algorithm called Dynamic Time Warping, to measure the dissimilarity between temporal patterns and long-term time series by adding seasonal time-weight. This study was accomplished with MSc. Adeline Marinho Maciel[1] and its results are available in TerraBrasilis[2]. The results approach Bolivia and Peru from 2000 to 2016, on an yearly basis. The workflow was composed by four main features layers with the aim of improving the process of evaluating supervised algorithms such as TWDTW considering remote sensing applications. Our results showed that TWDTW **R** package in multidimensional array databases can add a substantial contribution to the iterative workflow of evaluating land use and land cover for classification satellite imagery.

4.1 Sampling Design

The sampling design considers interpretation and quality control, phenology characterization and homogeneity characteristics, and is influenced by the classification method to be used. In this study case, as we have used the TWDTW method, the sampling design is used will be used to build the temporal patterns of the land cover of interest. The study area is located in the South America and covers two countries: Peru and Bolivia. Both countries have together approximately 2,383,801km^2. Figure 8 shows a more close overview of both countries.

Usually, once the sampling method is defined, researchers perform field work campaigns, or visual interpretation of satellite images by experts in order to define the patterns of the classes. Considering the large area and the short time span for the study case, we used a different approach. We started from the GlobCover2009 global land cover classification dataset as a reference. We define as the interest classes for this study 16 classes *Barren Land, Closed Broad Deciduous Forest, Closed Broad Evergreen Forest, Herbaceous, Mosaic Cropland, Mosaic Grassland Shrubland, Mosaic Shrubland Grassland, Mosaic Vegetation, Permanently Flooded Forest, Rainfed Cropland, Regularly Flooded Forest, Shrubland, Snow, Sparse Vegetation, Urban Area* and *Salar*, this last is specific class to map areas of the Salar de Uyuni in Bolivia.

We carried on a design sampling to select samples of theses classes and their temporal patterns to be used in TWDTW. Potentially, every pixel classified as one

[1] https://www.researchgate.net/profile/Adeline_Maciel
[2] http://terrabrasilis.info/composer/E-SENSING

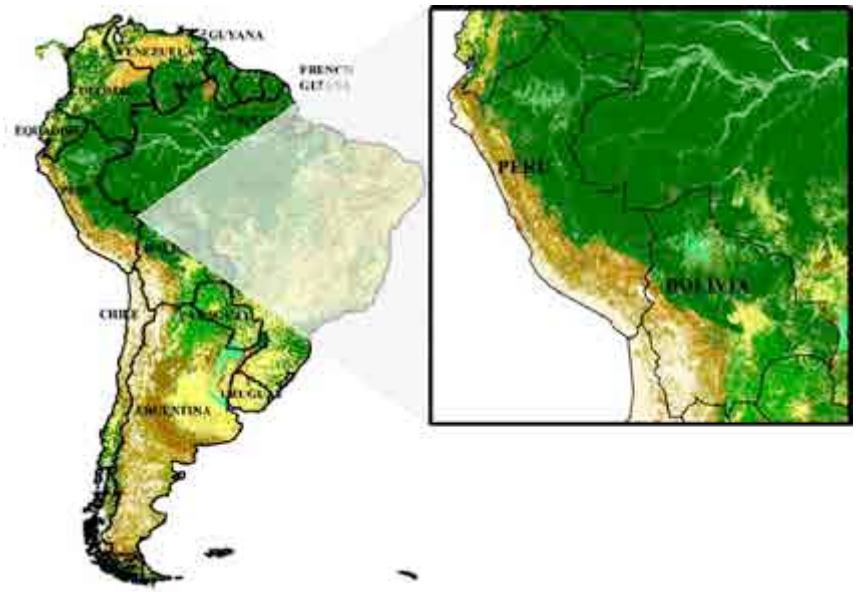


Figure 8: Peru and Bolivia. Adapted from 2

of the interest classes in the reference classification map is a sample. However, the spatial homogeneity is a important aspect to be considered. Figure 9 shows an example of a region in which a sample can not be collected and a region where a viable sample can be distinguished.

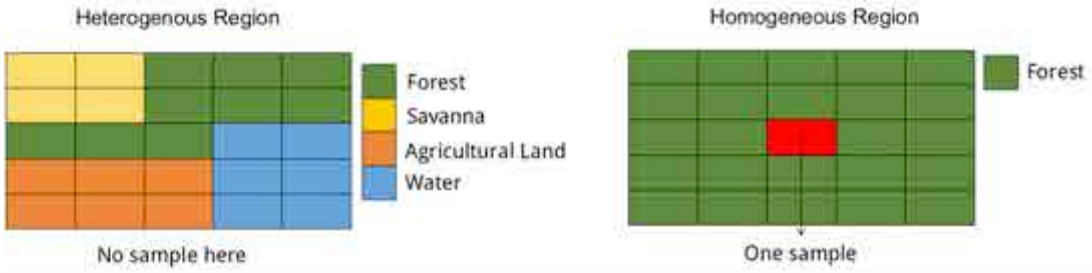


Figure 9: Automated Sampling based on existing classification maps

These spatially (considering homogeneity) and statistically (considering their distribution in size) consistent points located in Peru and BolÃ■via can be seen in Figure 10.

As the reference classification map was created using the space first approach, the spatially consistent points are not homogeneous in terms of spectral similarity along a time period. A clustering algorithm is used to divide the samples in subclasses, each subclass of a class will define a distinctive temporal pattern of the same class. Table 2 shows the number of samples for each class, and for each subclass. Figure 11 shows the 41 temporal patterns in terms of the Normalized Difference Vegetation Index (NDVI), the Enhanced Vegetation Index (EVI) and Near Infra Red (NIR) spectral attributes.

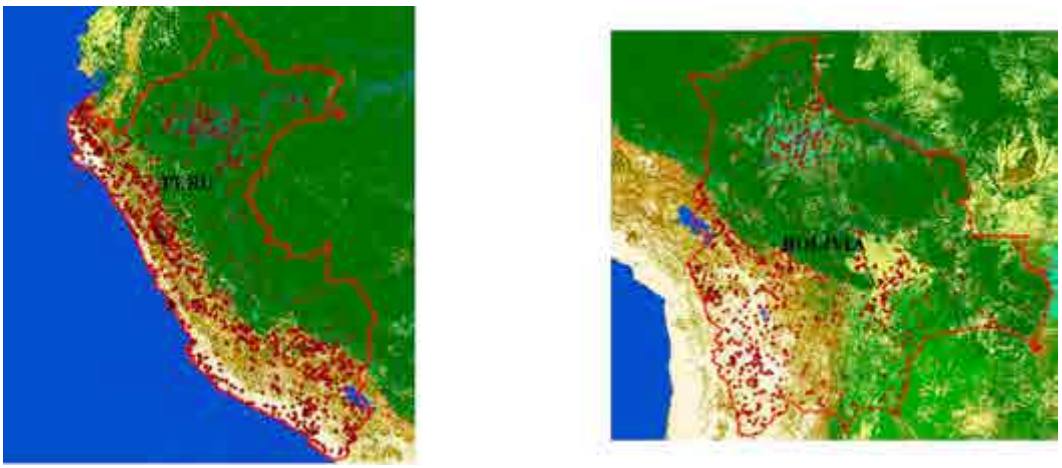


Figure 10: Sampling in Peru and Bolivia

Table 2: Number of samples per class

Class ID	Class	Number of samples
1	Barren Land	304
2	Closed Broad Deciduous Forest	295
3	Closed Broad Evergreen Forest	296
4	Herbaceous	324
5	Mosaic Cropland	330
6	Mosaic Grassland Shrubland	291
7	Mosaic Shrubland Grassland	297
8	Mosaic Vegetation	299
9	Permanently Flooded Forest	27
10	Rainfed Cropland	302
11	Regularly Flooded Forest	304
12	Salar	28
13	Shrubland	274
14	Snow	284
15	Sparse Vegetation	278
16	Urban Area	295
Total		4228

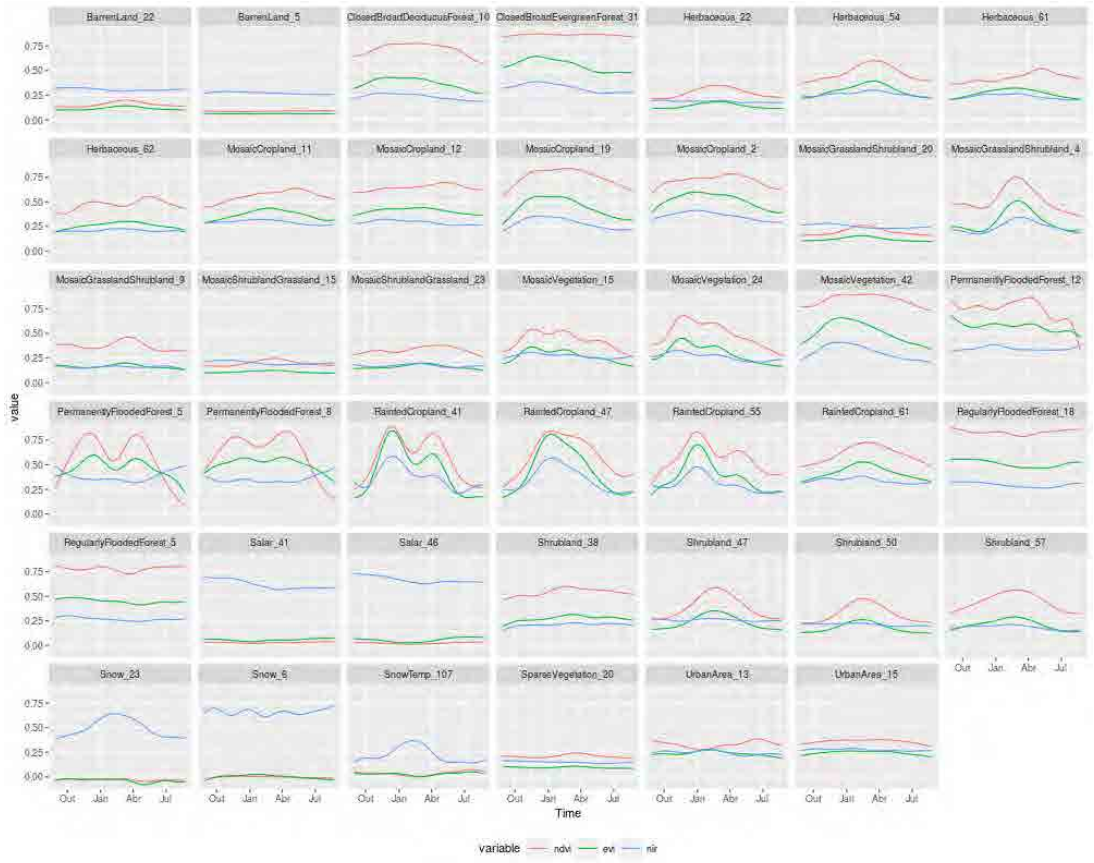


Figure 11: Temporal patterns - 16 classes, 41 subclasses in 3 attributes

Figure 12 shows a legend used to label the results of the case study.

(a) TWDTW temporal patterns classes

(b) GlobCover 2009 classes

Figure 12: Land cover data legends.

4.2 Data Model and Storage

In our SciDB, we created a 3D array containing 11 attributes: ndvi (int16), evi (int16), quality (int16), red (int16), nir (int16), blue (int16), mir (int16), view_zenith

(int16), sun_zenith (int16), relative_azimuth (int16), day_of_year (int16) and reliability (int16). The array is called mod13q1_512 and contains three dimensions: col_id (int64), row_id (int64) and time_id (int64). A few lines of which kind of information is stored in the array can be seen below:

```
{col_id,row_id,time_id} ndvi,evi,quality,red,nir,blue,mir,view_zenith,
sun_zenith,relative_azimuth,day_of_year,reliability
{60044,48622,0} 5604,4992,2062,942,3344,929,816,1373,2487,-122,58,3
{60044,48622,1} 5371,4854,3098,1244,4131,897,1299,2672,2657,-201,67,3
{60044,48622,2} 8793,5603,2120,210,3272,116,656,4630,3063,-339,85,1
{60044,48622,3} 8795,5344,2116,197,3075,106,663,3700,3282,-502,108,0
{60044,48622,4} 9209,4616,2116,103,2504,16,464,1766,3136,1165,120,0
{60044,48622,5} 8850,4549,2116,151,2477,80,435,1584,3500,1121,136,0
{60044,48622,6} 8824,4611,2116,158,2531,82,447,1640,3772,1087,152,0
{60044,48622,7} 8817,4892,2116,171,2722,95,444,222,3947,1016,161,0
{60044,48622,8} 8844,5240,2116,182,2967,103,431,1201,4078,-659,186,0
{60044,48622,9} 8946,5238,2116,164,2950,85,423,273,3946,1181,193,0
{60044,48622,10} 8848,5359,2116,186,3045,110,402,166,3727,1270,209,0
```

4.3 Applying TWDTW using SciDB-**R** Streaming

Considering an architecture that focus on the researcher needs and involves analysis of remote sensing time series, we decided to firstly apply an innovative method for the classification study, so after, we could evaluate the accuracy of our results. The preliminary results helps us to identify high demand efforts to classify and validate small areas of our region of interest instead of premature running the method for large areas. Since we are dealing with irregularly sampled and out-of-phase time series, we employ a method called Time-Weight Dynamic Time Warping (TWDTW) [13] due to its recent result studies and robustness. TWDTW takes into account a weighted extension giving an open boundary to the well known algorithm Dynamic Time Warping [14]. Figure 13 depicts an overview of the roles played by the architecture and the method deployed in this case study.

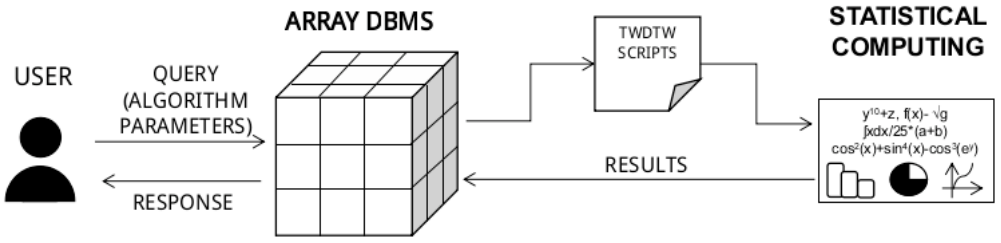


Figure 13: Architecture for big Earth Observation data analytics. Adapted from (4)

TWDTW compares temporal patterns of known vegetation index associated with land cover classes to an unknown long term time series pixel (see Figure 14). The method finds the optimal alignment within this comparison even if both time series are irregularly sampled or are out of phase in the time axis, that is, regarding the

temporal range and the phenological cycle that is relevant for this kind of classification. As a result, the method provides a dissimilarity measure.

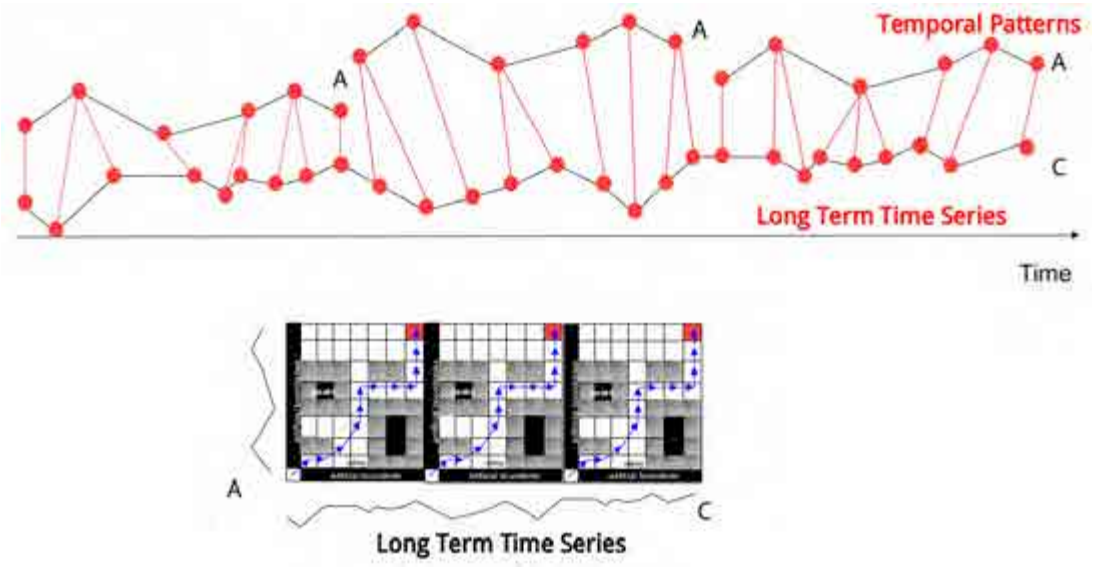


Figure 14: Open boundary TWDTW alignment from a long-term time series C divided in subintervals, and the best matching temporal patterns A in each subintervals C

The idea of SciDB-**R** streaming is very simple and is totally compatible with frameworks responsible for storing and processing large distributed data. For example, we have a big problem such as classifying satellite imagery that cover an Amazon Forest area. At first, we need to break our problem into small areas. That is exactly what SciDB does, it divides into chunks, a partitioning technique where each instance keep a subset of the array locally. This feature allows an uniformly scalable performance on large data sets. Intrinsically, the data stream then helps a fast access and complex analysis without facing errors of parallel and distributed computing (see Figure 15).

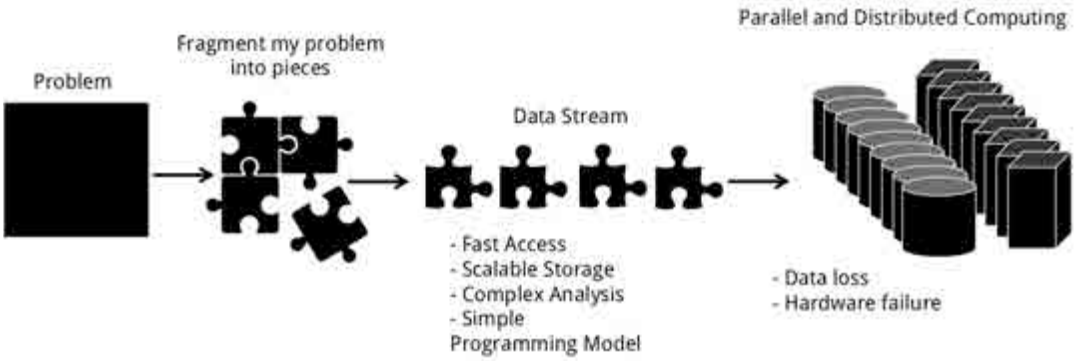


Figure 15: Big Data Streaming Analytics

After compiling and loading SciDB streaming, we can use an AFL operator call passing as input parameters: an input array (mod13q1_512), a **R** classification script, the **R** data output (an **R** binary data.frame) interface, the output column SciDB types, the output column names (the same length as used in data.frame

columns). Both types and names should be written in a comma-separated manner. Other parameters are optional, but the aforementioned parameters are the ones we used here. An exemplary call can be seen in Listing 1.

```
stream(
        scidb_array,
        'Rscript script.R',
        'format=R format output',
        'types=output column types',
        'names=output column names'
)
```

lol 1: Stream Operator Call.

The **R** classification script has standards input and output. At first, it reads a binary connection and interprets the data by means of a a simple low-level interface for serialization and unserialization. Paradigm4[3] also indicates a required different answer to SciDB when the last message is sent, that is, when ncol is equal to 0. That allows the correct written of binary chunks into SciDB in all the cases. A more detail example can be seen in Listing 2[4].

```
con_in = file("stdin", "rb")
con_out = pipe("cat", "wb")
while( TRUE )
{
  input_list = unserialize(con_in)
  ncol = length(input_list)
  if(ncol == 0)
  {
    res = list()
    writeBin(serialize(res, NULL, xdr=FALSE), con_out)
    flush(con_out)
    break
  }

  #....write algorithm

  writeBin(serialize(c(out), NULL, xdr=FALSE), con_out)
  flush(con_out)

}
close(con_in)
```

lol 2: STDIN and STDOUT - R-SciDB Streaming.

To write the algorithm, we need at the beginning load all the **R** packages and data necessary to run our classification algorithm. In our case, we need to load

[3] http://www.paradigm4.com/
[4] https://github.com/Paradigm4/stream/blob/master/examples/R_identity.R

dtwSat, *parallel*, *plyr*, the *patterns* and the *input_list* (see Listing 3). *dtwSat* is an **R** implementation of the Time-Weighted Dynamic Time Warping (TWDTW), previous mentioned method for land use and land cover mapping using satellite image time series. *parallel* is an **R** package implementation of coarse-grained parallelism in computation to split the task. Also, *plyr* is a set of clean and consistent tools that implement the split-apply-combine pattern in R. The temporal patterns and *input_list* are used as input for *dtwSat*.

```
library(dtwSat)
library(parallel)
library(plyr)
load(patterns)
attach(input_list)
```

lol 3: Load of packages, patterns and MOD13Q1 data.

As in our array in SciDB, we define the time dimension as 64 byte integer, the time instances are represented by numbers from 0 until n, where n is the last time instance gathered. That means, for MOD13Q1 if we have 391 time instances, the array is from 2000-02-18 to 2017-01-01 by 16 days. Since *dtwSat* requires a date format for a *zoo* object, we need to transform and create a time sequence from a time sequence.

```
createTimeSequence = function(year=2000:format(Sys.time(), "%Y"), frequency=16){
   res = unlist(lapply(year, function(y){
     days = seq(from = as.Date(paste0(y,"-01-01")),
                to = as.Date(paste0(y,"-12-31")),
                by = frequency)
   }))
   res = as.Date(res, origin="1970-01-01")
   res = subset(res, res >= as.Date("2000-02-18"))
   res
  }
```

lol 4: Create Time Sequence based on SciDB data.

We also need to ensure col and row identifiers are unique, since when they come from SciDB they repeat for each time instance.

```
I = unique(irow_id)
J = unique(icol_id)
indexArray = list()
k = 1
for(i in I)
for(j in J){
indexArray[[k]] = c(j=j, i=i)
 k = k + 1
}
```

lol 5: Get unique columns and rows values.

During the **R** classification script, we build MODIS timeline considering SciDB indexes starts in 0. We also need to set the breaks, a vector of class Dates with from, to and by arguments, and the time-weight function for the algorithm, responsible for computing the TWDTW local cost matrix. The time-weight function It is necessary to get the labels too. After all, we multiply all the stored values by the scale_factor described in the array metadata.

```
dates = createTimeSequence()[itime_id+1]

breaks = seq(from = as.Date("2000-09-01"),
             to = as.Date("2017-09-01"),
             by = "12 month")
label_names = as.character(labels(patterns))

weight.fun = logisticWeight(alpha=-0.1, beta=100)

scale_factor = 0.0001
dndvi = ndvi * scale_factor
devi  = evi  * scale_factor
dnir  = nir  * scale_factor
```

lol 6: Define initial TWDTW parameters.

At the end, we define a function receiving the columns and rows values as an input in order to facilitate the classification images parallelism. In this function, we select time series by array indices, remove the duplicated dates, avoid processing for timeseries smaller than 9 time instances, build the time series using zoo objects, apply TWDTW method, generate the classification values and create the data.frame output. All of these operations are performed in parallel using **R** and SciDB.

```r
fun = function(p){


  idx = which(icol_id==p["j"] & irow_id==p["i"])


  idx = idx[!duplicated(dates[idx])]
  if( length(idx) < 9 )
     return(NULL)


  x = twdtwTimeSeries(zoo(data.frame(ndvi = dndvi[idx],
                                     evi  = devi[idx],
                                     nir  = dnir[idx]),
                          dates[idx]))


  matches3 = twdtwApply(x = x,
                            y = patterns,
                            weight.fun = weight.fun,
                            keep = FALSE,
                            theta = 0.5,
                            span = 250)


  ts_classification = twdtwClassify(x = matches3,
                                        breaks = breaks,
                                        overlap = 0.5)

  aligs = ts_classification[[1]]
  k = nrow(aligs)
  data.frame(
     colid = as.double(rep(p["j"], k)),
     rowid = as.double(rep(p["i"], k)),
     time  = as.double(seq_len(k)),
     from  = as.double(as.integer(aligs$from)),
     to    = as.double(as.integer(aligs$to)),
     label = as.double(match(aligs$label, label_names)),
     dist  = as.double(aligs$distance)
  )
}

out = do.call("rbind", mclapply(indexArray, mc.cores = 3, FUN=fun))
```

lol 7: Applying TWDTW.

5 Complementary Activities

In this chapter, we discuss in more details a set of complementary activities in which the technical training fellow executed to achieve the project goals and improve his personal and professional background. These activities involve mainly weekly and monthly meetings, and the attendance of summer schools, workshops and symposiums.

5.1 Weekly and Monthly Meetings (Image Processing Division): The Research Group Integration

For the project progress, there have been weekly meetings among the project researchers and developers, where each member presented in which activities he/she was involved in the week before and planning to do the week after. In these meetings, everyone was responsible for contributing with ideas using his/her own experience to solve anyone's problem. Furthermore, Image Processing Division (DPI) members were committed to explain very deeply about a subject he/she was investigating that time.

DPI is part of the General Coordination of Earth Observation (OBT) of the Brazilian National Institute for Space Research (INPE). Its main activities involve scientific and technological research and development on digital processing of satellite images and remote sensing. This division aims to specify, design and develop systems for image processing. The head of this research group is Prof. Dr. Gilberto Câmara, who is also the supervisor and project coordinator of e-Sensing project. His main research topics are Geoinformatics, GIScience, Spatial Analysis, Land Use Change and Applied Ontology. He established a free and open access policy for INPE's data and guided INPE's team on big advances in forest monitoring by satellite.

The group also has plenty of experience with Geoinformatics. Lúbia Vinhas is an Associate Professor on Geoinformatics with a PhD in Computer Science. She currently heads INPE's Image Processing Division and is one of the manager of the TerraLib project. In the e-Sensing project she is involved with the database design, access and processing of information in SciDB. Karine Ferreira is an Associate Professor of Geoinformatics at INPE, working in the Image Processing Division. She has a PhD in Computer Science. Her research topics include data models, algebras and databases for spatiotemporal data and spatiotemporal GIS development. Gilberto Ribeiro de Queiroz has been an Associate Professor since 2005 of spatial databases at INPE, focused on the development of geotechnologies.

5.2 I Workshop de Aplicações do LuccME

The LuccME (http://luccme.ccst.inpe.br/) is an explicit spatially land-use modeling framework developed by the Earth System Science Center (CCST/INPE) and its

collaborators as an extension of the model TerraME. The main objective of the workshop was to promote the debate on the possible applications of LuccME by discussing success stories at different scales - including models of deforestation, expansion of agriculture, desertification, urban growth and other processes of change of use and coverage from the Earth. As a result, the workshope allows participants to build a collaborative networking between users and LuccME for specific topics, and survey of subsidies for future versions of the tool. The thematic sessions was really interested to this project and future collaborations could be seen.

6 Conclusions

The technical training fellowship offered many academic and professional gains. Attending to a high level research institute such as INPE, gives not only good future opportunities, but also provide a chance to exchange ideas with different researchers, and consequently, gain knowledge to solve problems using solutions that would not be possible without this experience. The e-Sensing team [1] also helped the fellow to undertake his activities, under the supervision of Prof. Dr. Gilberto Câmara, which had a great importance and had provided significant contributions to the project.

All the experiences had directly and indirectly benefited the project entitled *Integration among SciDB, TerraLib and R* because GIScience and e-Sensing team members have experts in big data, spatio-temporal analysis and database. Therefore, their support was critical during the software development. Regarding the project objectives and the obtained results, we developed an interface between the R data analysis language, the SciDB arrays database, and the TerraLib library. This because, we implemented analysis methods for big Earth Observation data developed by the project team in the **R** environment that can be efficiently executed in the SciDB environment.

Although the SciDB environment has features for complex analysis, an interface with **R** is still important because many applications require the combination of matrix and vector data. In our land use classification case study, we have seen an enhanced performance of the SciDB-**R** compared to our initial experiments. For this, it was necessary to run by means of a remote execution of the **R** scripts in a distributed environment server, minimizing in this way, the transfer of data between client and server.

[1] http://esensing.org

A STILF - A spatiotemporal interval logic formalism for reasoning about events in remote sensing data

XVIII SBSR Simpósio Brasileiro de Sensoriamento Remoto
ISBN: 978-85-17-00088-1
Anais do XVIII Simpósio Brasileiro de Sensoriamento Remoto - SBSR
28 a 31 de maio de 2017
INPE Santos - SP, Brasil

STILF - A spatiotemporal interval logic formalism for reasoning about events in remote sensing data

Adeline Marinho Maciel[1]
Lubia Vinhas[1]
Gilberto Câmara[1]
Victor Wegner Maus[2]
Luiz Fernando Ferreira Gomes de Assis[1]

[1] National Institute for Space Research – INPE
Caixa Postal 515 – 12227-010 – São José dos Campos - SP, Brasil
[2] International Institute for Applied Systems Analysis – IIASA
{adeline.maciel, lubia.vinhas, gilberto.camara}@inpe.br, {luizffga, vwmaus1}@gmail.com

Abstract. Although several studies perform time series analysis using remote sensing data provided by Earth observation satellites, few have been explored concerning the reasoning about land use change using these data. Besides, exists the challenge of make the best use of big Earth observation data sets to represent change. In this context, this work presents a new formalism - STILF (Spatiotemporal Interval Logic Formalism), and shows how to use it for reasoning about land use change using big Earth observation data. Extending the ideas from Allen's interval temporal logic, we introduce predicates $holds(o, p, t)$ and $occur(o, p, T_e)$ to build a general framework to reason about events. Events can be defined as complete entities on their respective time intervals and their lifetime is limited while objects persist in time, with a defined begin and end. Since events are intrinsically related to the objects they modify, a geospatial event formalism should specify not only what happens, but also which objects are affected by such changes. The formalism proposed and predicates extended from Allen's ideas can model and capture changes using big Earth observation data, and also allows reasoning about land use trajectories in regional or global areas. Examples for tropical forest area application is presented to better understand our proposal using STILF. For the future, the proposed formalism will be include other temporal analysis tools to thinking about events related the land use and cover change.

Keywords: land use and land cover, spatiotemporal representation, Allen's interval, events, logic formalis, remote sensing

1. Introduction

One of the recent trends in applications of remote sensing data is the use of big data sets for obtaining information about land use and land cover. Using long-term time series, scientists can obtain new information to understand how mankind is using natural resources. Satellite image time series data provides a new perspective in remote sensing data analysis (CAMARA et al., 2016a).

An example of big Earth Observation data analysis is the work by Hansen et al. (2013). Using more than 650,000 LANDSAT images and processing more than 140 billion pixels, the authors compared data from 2000 to 2010 to produce maps of global forest loss. The results for 2000 and 2010 were compared to account for forest loss during the 2000-2010 decade. The method classifies each 2D image one by one.

By contrast, methods such as the time-weighted dynamic time warping (TWDTW) (MAUS et al., 2016) and TIMESTAT (JöNSSON; EKLUNDH, 2004) work on remote sensing time series to extract long-term information for each pixel. These algorithms work on individual time series and combine the results for selected periods to generate classified maps.

The benefits of remote sensing time series analysis arise when the temporal resolution of the big data set is able to capture the most important changes. Here, the temporal autocorrelation of the data can be stronger than the spatial autocorrelation. Given data with adequate repeatability, a pixel will be more related to its temporal neighbours than to its spatial ones. In this case, *time-first, space-later* methods lead to better results than the *space-first, time-later* approach (CAMARA et al., 2016a).

Given the possible new results that can be obtained with big remote sensing data, the scientific challenge is how to best represent and detect change. Issues about representation, reasoning, modelling of changes have been researched in GIScience (PEUQUET; DUAN, 1995; GALTON, 2004). In general, these studies show the usefulness of using the concept of "events" to represent changes in spatiotemporal data. The objective of this paper is to apply the concept of "events" for representing change in big remote sensing data sets, following the ideas from Galton (2015). Additionally, we extend the interval temporal logic proposed by Allen (1984) to build a logic formalism which allows reasoning about events of change in land use and land cover data. This paper extends and improves on earlier work by our research group (CAMARA et al., 2016b).

2. A Spatiotemporal Interval Logic Formalism - STILF

To describe land use and land cover changes, we consider an approach based in time intervals. We extend the interval temporal logic from Allen (1984) to build a logic formalism for reasoning about events. Allen (1983) defines a set of thirteen relationships between two time intervals: *before*, *meets*, *during*, *starts*, *finishes*, *overlap*, with inverse relationship, and *equal*.

To extend the predicates from Allen (1984) to spatiotemporal data, we aggregate the notion of geo-objects, which are related to space. This way, the formalism is composed for a set of elements: (1) discrete geo-objects ($O = o_1, o_2, ..., o_n$); (2) properties of geo-objectos ($P = p_1, p_2, ..., p_n$); and (3) time intervals ($T = t_1, t_2, ..., t_n$). We also use the predicates: (1) $holds(o, p, t) \rightarrow bool$, which asserts that a properties p from geo-object o holds during a interval t; and (2) $occur(o, p, t_e) \rightarrow bool$, given a interval $T_e \subset T$, the properties p from geo-object o will be true during all sub-interval T_e.

The start point of the spatiotemporal interval logic formalism (STILF) is a set of time series data classified from remote sensing images. This images were previously classified by means of data mining algorithm, such as TWDTW (MAUS et al., 2016), and stored in a array database. This is a important stage for the application of the formalism, once allows that Earth observation data were stored in a database which support a large amount of remote sensing data, and subsequently, can will be used for different applications. Next, this data set will be processed, for extraction of the set of elements. Each element is composed of a discrete geo-object, its properties of geo-object and time intervals for which these properties hold.

The set of elements will be used as input data for our formalism. Combining the $holds(o, p, t)$ and $occur(o, p, t_e)$ predicates with Allen's relations, we can ask questions about trajectories of land use and land cover change. The answers will be data sets with the events that have occurred during the whole interval for which we have data. In the last stage, individual events will be combined in terms of their characteristics into recurring events or transition events. Figure 1 shows a overview of our proposed formalism.

3. Application: Examples of Reasoning About Events from Classified Land Use and Land Cover Time Series

In this section we show three examples of application using remote sensing data. The formalism presented was developed in a R programming language and applied on sample

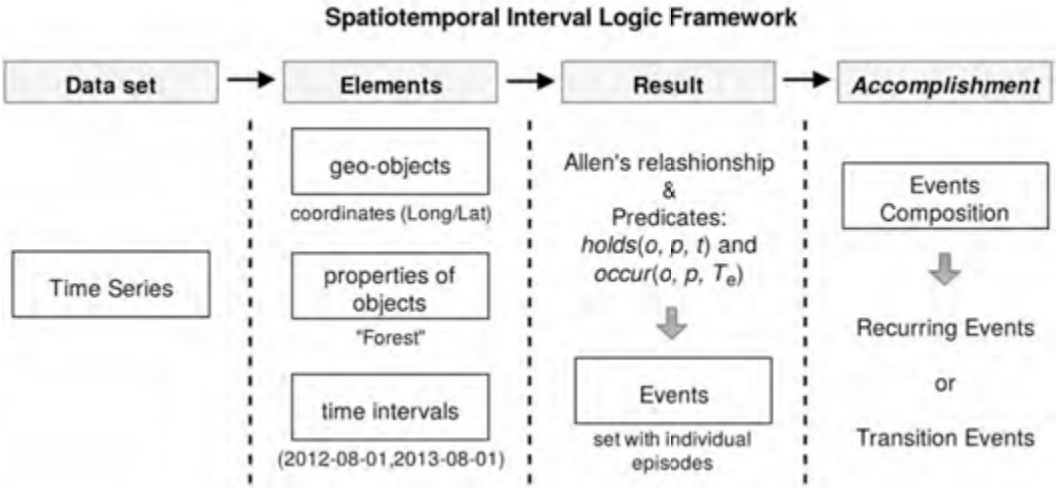


Figure 1: Spatiotemporal interval logic formalism (STILF) design

regions. The input data is composed for a set of time series classified to the municipality of Porto dos Gaúchos, located in northwest Mato Grosso (MT) state, Brazil. With territorial area 6,862.118 km^2, geographical coordinates, latitude 11° 31' 31" South and longitude 57° 24' 50" West and population of 5,400 inhabitant in 2010, according to IBGE statistics (IBGE, 2016). Porto dos Gaúchos is an area into Amazon biome.

Figure 2: Municipality of Porto dos Gaúchos with highlight for three sample regions selected to application of the formalism.

We extracted information from each sample region to discover what events had happened. These events allow us to establish the trajectories of land use and land cover. For example, the results may indicate the increase of deforestation in the municipality after earlier expansion of areas. We can also detect the conversions from pasture to soybean and from soybean to double cropping (soybean-corn or soybean-cotton).

After we classify the time series, we apply a post-processing rule to distinguish natural, intact forests from areas that had been deforested and then were allowed to regrow. This is required to be able to differentiate primary forest, without degradation, from secondary vegetation, forest areas that happened after other land use or land cover classes. The new classes generated after this stage were called "Secondary vegetation".

In the first sample region, with an area of 50.23 km^2, located in Northeast of the

municipality, we explored the ability of the formalism to detect events composition. Our query searched for events preceding and following the year of 2008, associated to the "Soybean-Millet" crop areas Question 1. The result were three graphics with information for analysing land use trajectories: (1) a custom map that highlights events that show the transition from "Pasture" to"Soybean-Millet", Figure 3; (2) a bar graph which counts the total area (km^2) for each event by year; and (3) a graph which represents the temporal sequence of the events for each pixel in the time. This type of graph show what pasture areas were transformed into crop areas (Figure 4(b)).

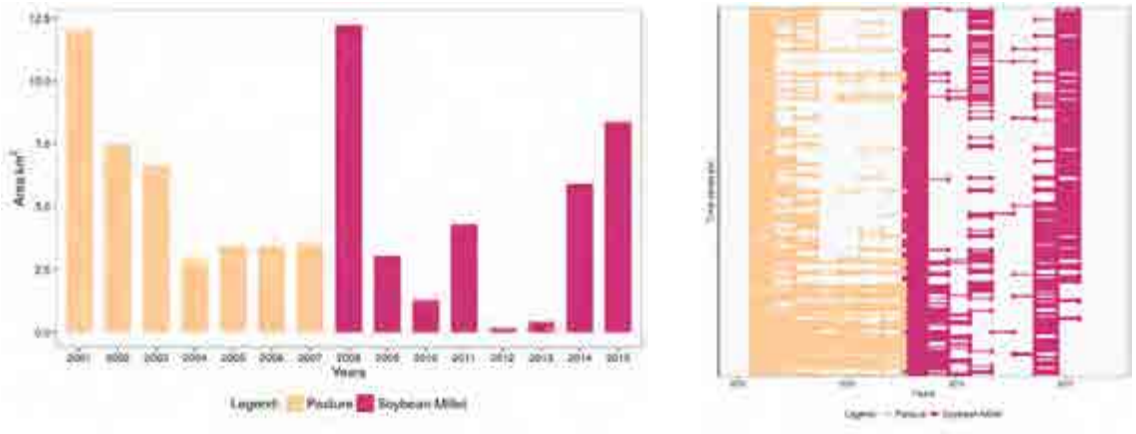
Question 1: Example of application of the spatiotemporal interval logic formalism for mapping of changes in land use and cover for the first sample region

a) *Which "Pasture" areas before 2008 have been turned in "Soybean-Millet" cropping areas?*

$$\forall o \in O, occur(o, \text{“}Pasture\text{”}, t_1) \wedge occur(o, \text{“}Soybean - Millet\text{”}, t_2) \wedge$$
$$occur(o, \text{“}Soybean - Millet\text{”}, t_3) \wedge$$
$$preceding(t_2, t_1) \Leftrightarrow (metBy(t_1, t_2) \vee after(t_1, t_2)) \wedge$$
$$next(t_2, t_3) \Leftrightarrow (meets(t_1, t_2) \vee before(t_1, t_2))$$

where $t_1 = \{2000, ..., 2007\}, t_2 = 2008, t_3 = \{2009, ..., 2015\}$

Figure 3: Sample region 1, events highlighted.



(a) Area in (km^2) with "Pasture" events turned into "Soybean-Millet" from 2008.

(b) Temporal sequence of the events.

Figure 4: Graphics to analyses of events composition - sample region 1

In the second sample region, located to the south of the municipality and area of 59.568 km^2, we investigated which "Forest" areas that have been turned into "Pasture" or " Low vegetation (a second type of pasture)" after 2001. The formal representation of the question is shown in Question 2. Three output plots were generated with information about events: a map that highlights events that happened yearly, Figure 5; a bar graph with total area for each event by year (Figure 6(a)), and a temporal representation for each pixel over time, which shows the transitions from forest to pasture (Figure 6(b)).

Question 2: Example of application of the STILF for mapping of changes in land use and cover for the second sample region

b) *Which "Forest" areas have been turned into "Pasture" or "Low-vegetation" after the year of 2001?*

$$\forall o \in O, occur(o, \text{“}Forest\text{”}, t_1) \wedge$$
$$(occur(o, \text{“}Pasture\text{”}, t_2) \vee occur(o, \text{“}Low - vegetation\text{”}, t_2)) \wedge$$
$$next(t_1, t_2) \text{ where } t_1 = 2001, t_2 = \{2002, ..., 2015\}$$

In a third sample region, located northwest of Porto dos Gaúchos municipality and area of 101.963 km^2, we wanted to know which "Forest" areas have not undergone degradation over years (Question 3) In a similar way to the PRODES system, forest areas that from regrowth after fire or deforestation are called "Secondary vegetation" by our post-processing rule and are not computed. Figure 7 shows a map that highlights the events. Figure 8 displays the amount of forest grouped by year. We can see the expansion of deforestation until 2006, when there was a significant reduction.

Question 3: Example of application of the STILF for mapping of changes in land use and cover for the third sample region

c) *Which "Forest" areas have not undergone degradation during interval of 15 years?*

$$\forall o \in O, occur(o, \text{“}Forest\text{”}, t_1) \text{ where } t_1 = \{2001, ..., 2015\}$$

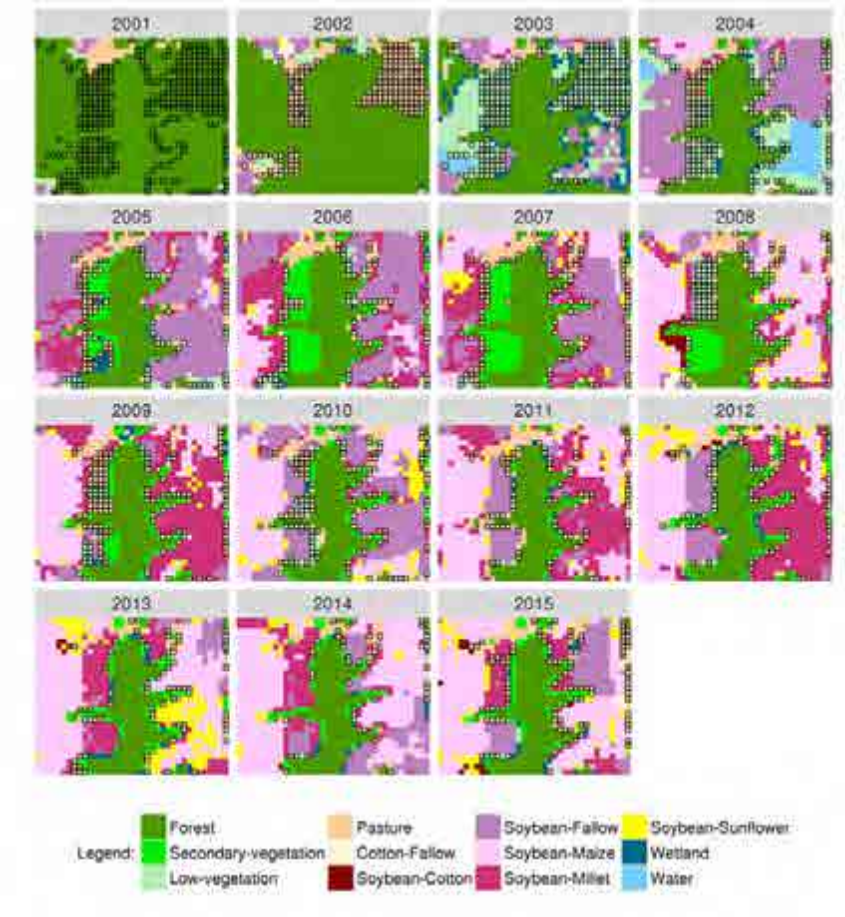
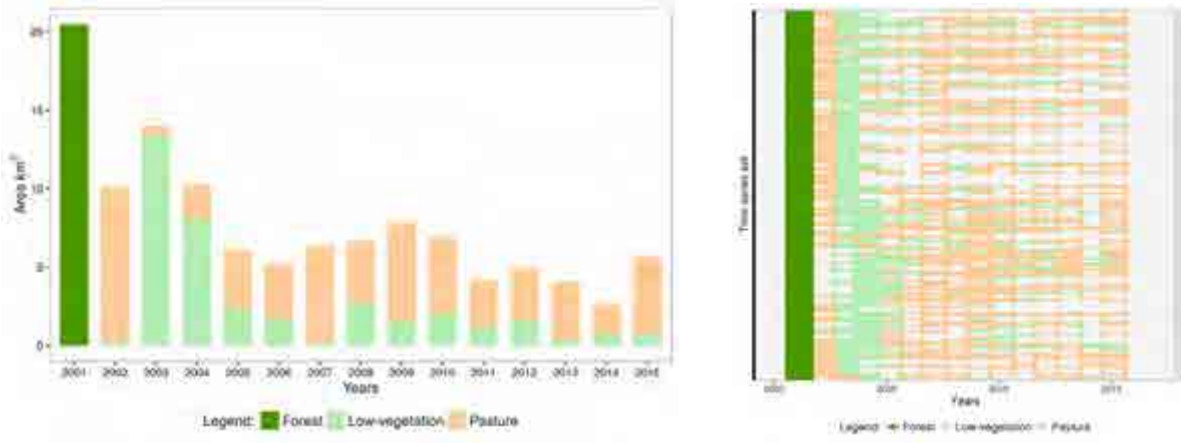


Figure 5: Sample region 2, events highlighted



(a) Area in (km^2) with "Forest" events turned into "Pasture" and/or "Low-vegetation" from 2001.

(b) Temporal sequence of the events to second sample region.

Figure 6: Graphics to analyses of events composition - sample region 2

This spatiotemporal interval logic formalism makes it easy to build questions in a logic representation in order to reason about changes in land use and land cover. We can show the trajectories of change in different perspectives. This makes it easier to understand changes in an environment. The formalism is robust. It allow different logical queries combining Allen's relations and also predicates of geo-objects.

ISBN: 978-85-17-00088-1
Anais do XVIII Simpósio Brasileiro de Sensoriamento Remoto - SBSR
28 a 31 de maio de 2017
INPE Santos - SP, Brasil

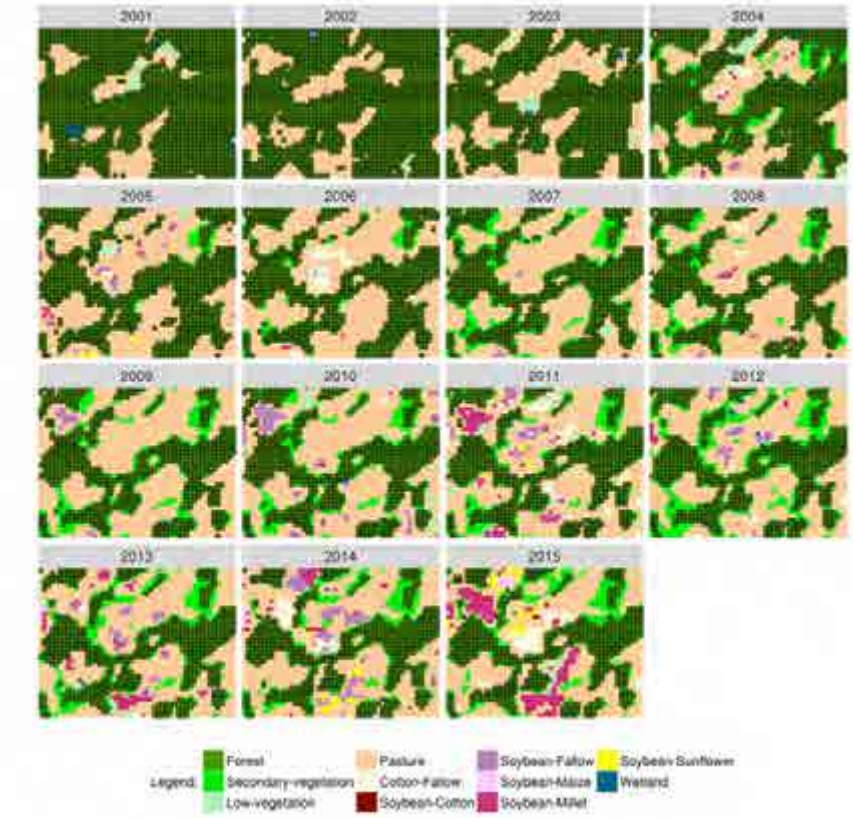


Figure 7: Sample region 3, with events highlighted

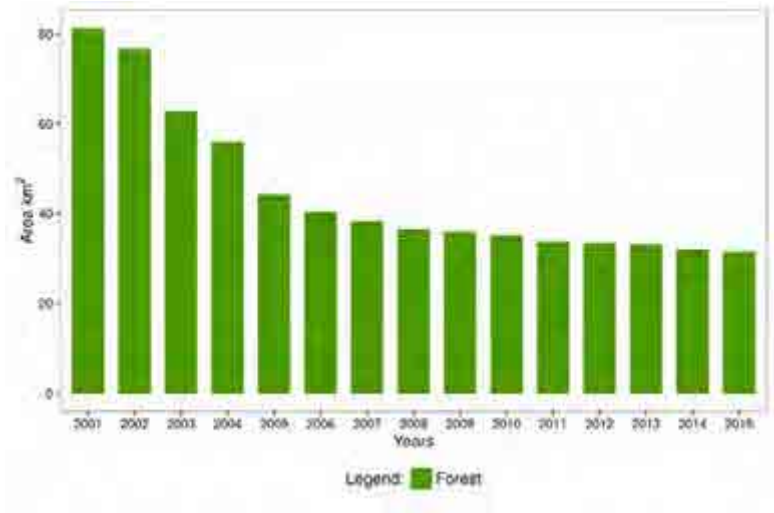


Figure 8: Area in (km^2) with "Forest" events which have not undergone degradation during the period of 15 years.

4. Final Considerations

A spatiotemporal interval logic formalism to reasoning about changes in land use and land cover was presented in this paper. This formalism is an extension from predicates defined by Allen (1984). We introduce geo-objects as new elements for analyses involving spatial data. We show three examples of application where the formalism was implemented in a programming language, take advantaging of the resources for data visualisation and results.

Acknowledgements

The authors thank CAPES and FAPESP e-science program (grants 2014-08398-6 and 2016-03397-7) by financial support.

References

ALLEN, J. F. Maintaining knowledge about temporal intervals. *Communications of the ACM*, ACM, New York, NY, USA, v. 26, n. 11, p. 832–843, 1983.

ALLEN, J. F. Towards a general theory of action and time. *Artificial Intelligence*, Elsevier Science Publishers Ltd., Essex, UK, v. 23, n. 2, p. 123–154, 1984.

CAMARA, G. et al. Big earth observation data analytics: Matching requirements to system architectures. In: *Proceedings of the 5th ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data*. New York, NY, USA: ACM, 2016. (BigSpatial '16), p. 1–6.

CAMARA, G. et al. Using dynamic geospatial ontologies to support information extraction from big earth observation data sets. In: *Ninth International Conference on Geographic Information Science (GIScience 2016)*. Montreal, CA: AAG, 2016.

GALTON, A. Fields and Objects in Space, Time, and Space-time. *Spatial Cognition & Computation*, v. 4, n. 1, p. 39–68, 2004.

GALTON, A. Outline of a formal theory of processes and events, and why giscience needs one. In: *Conference on Spatial Information Theory - COSIT 2015*. Santa Fe, NM, USA: Springer International Publishing, 2015. p. 3–22.

HANSEN, M. et al. High-resolution global maps of 21st-century forest cover change. *Science (New York, N.Y.)*, v. 342, n. 2013, p. 850–3, 2013.

IBGE. *The Brazilian Institute of Geography and Statistics*. 2016. Available on `http://www.ibge.gov.br`.

JöNSSON, P.; EKLUNDH, L. Timesat–a program for analyzing time-series of satellite sensor data. *Computers & Geosciences*, v. 30, n. 8, p. 833 – 845, 2004.

MAUS, V. et al. 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*, IEEE, PP, n. 99, p. 1–11, 2016.

PEUQUET, D. J.; DUAN, N. An event-based spatiotemporal data model (ESTDM) for temporal analysis of geographical data. *International journal of geographical information systems*, Taylor & Francis, v. 9, n. 1, p. 7–24, 1995.

B The E-Sensing Architecture for Big Earth Observation Data Analysis

THE E-SENSING ARCHITECTURE FOR BIG EARTH OBSERVATION DATA ANALYSIS

Gilberto Camara, Gilberto Queiroz, Lubia Vinhas, Karine Ferreira, Ricardo Cartaxo, Rolf Simoes, Eduardo Llapa, Luiz Assis, Alber Sanchez

National Institute for Space Research (INPE), Earth Observation Directorate
São José dos Campos, SP, Brazil

ABSTRACT

This work presents an architecture for big Earth Observation data analytics. It uses array databases to support storage and management of large volumes of satellite image time series. The analysis methods are developed in R and enable using the full depth of satellite image time series with advanced statistical learning algorithms. New kinds of web services allow data access and remote data processing of time series. The *e-sensing* architecture has been designed with a focus on land use and land cover classification using SITS, an area of Earth observation where much progress is required. This architecture is fully implemented and has already allowed innovative results in land use and land cover mapping. The method works with big data sets with a minimal set of assumptions to increase its generality. Our work promotes reproducibility and reuse of the methods and results.

Index Terms— Earth observation, web services, satellite image time series, array databases, science reproducibility, open source.

1. INTRODUCTION

The data deluge resulting from the open access policies for Earth observation (EO) data has brought about a major challenge: *How to design and build technologies that allow the EO community to analyse big data sets?*. Developing such a solution is hard because current technologies for big data management are quite different and incompatible. Alternatives include using flat files [1], MapReduce-based solutions such as Google Earth Engine [2], and distributed multidimensional array databases such as Rasdaman [3] and SciDB [4]. Each choice has its advantages and drawbacks, and fits certain needs better than others.

The first option of an infrastructure for big EO data is to store EO data as flat files and use file management systems. This is the approach taken by the Australian Data Cube [1]. This choice makes it easy to preprocess images from different sources so that they become geometrically and radiometrically compatible. Data merging and cross-calibration tasks are simple to perform. Existing pixel-based image analysis methods can be applied to big data sets. However, these simple infrastructures have a high management cost. Data analysis proceeds by searching all the relevant files. The programs open each file, extract the relevant data and then move onto the next file. When all the relevant data has been gathered in memory, the program can begin its analysis. Working with time series becomes specially burdensome because of the number of files that must be opened for a single time series to be retrieved. Managing 10,000 - 100,000 files at once can lead to scalability and performance bottlenecks.

An alternative is to take a mainstream solution used for other big data applications and adapt it to EO data. This is the case of MapReduce-based solutions such as Google Earth Engine [2]. The MapReduce model has been motivated by highly parallel applications such as text queries and there are open source implementations such as Spark. MapReduce architectures are very efficient for problems where each pixel is processed independently. They lack flexibility for big EO analytics, since they use an excessive granularity when breaking the data into parts. Region-based methods such as image segmentation are not supported, nor large-scale time series analysis are possible.

A third option is to use array databases such as Rasdaman [3] and SciDB [4]. Array DBMS reduce the impedance mismatch between the data model (raster), the storage model (arrays) and analysis functions such as linear algebra and image processing. These databases split large volumes of data in distributed servers in a "shared nothing" way. Each server controls its local data storage. Arrays are multidimensional and uniform, as each array cell holds the same user-defined number of attributes. Array databases allow organising EO data to meet the needs of different applications. Comparative studies show the SciDB architecture to be more efficient and more flexible for processing remote sensing data than MapReduce [5]. However, since array databases are designed for scientific data management, there is much less experience with them. Developers using SciDB have to spend significant effort for system configuration and performance tuning. Despite these problems, we consider array databases to be the best choice for support innovative big EO data analytics.

This work is supported by the São Paulo Research Foundation (FAPESP) e-science program (grant 2014–08398–6) and by Germany's International Climate Initiative (IKI/BMUB) under grant 17-III-084-Global-A-RESTORE+. Gilberto Camara is also supported by CNPq (grant 312151–2014–4).

One of the areas where array DBMS allow advances on big EO data analytics is when processing dense satellite image time series (SITS). Using SITS is a leading research trends in Remote Sensing [6], [7]. One of the more promising applications of SITS is measuring land use change. Land use change is important for Brazil, one of the world's largest agricultural producers with one of Earth's richest biodiversities. Many researchers have also pointed out the need for improving future global land cover products [8], [9]. Given this motivation, the *e-sensing* architecture has been designed with a focus on land use and land cover classification using SITS.

This work presents innovative methods for using the full depth of satellite image time series for extracting information from big Earth observation data. We have developed a full open source architecture that allows efficient processing of large-scale data sets, coupled with advanced data analytic methods. Our focus is on extracting the most information from dense time series of remote sensing satellites such as MODIS, LANDSAT, and SENTINEL, or combinations of those.

2. DESIGN DECISIONS

The *e-sensing* architecture has been designed with a different perspective than other proposals for Earth Observation Data Cubes [1]. We believe the gains of using big EO data will come from new analytical methods, and our design reflects such aim. A key decision for big EO architectures is the choice of programming environment. We chose R, which has more than 11,000 packages for statistical computing and graphics, including spatial analysis, time-series analysis, classification, clustering, and machine learning. Using R, it is easier for researchers to develop new methods and to collaborate with their peers. SciDB has a streaming interface that runs R scripts in parallel directly on each server (Figure 1). Combining array DBMS with R statistical computing is a natural solution for EO applications, allowing a good balance between massive parallel data processing and maximum flexibility in algorithm design.

Scientists also need tools for small-scale testing and for scaling up their work. We developed two web services to support these tasks [10]. The Web Time Series Service (WTSS) retrieves time series of Earth observation data for specific locations. The Web Time Series Processing Service (WTSPS) enables users to run R scripts on data cubes of Earth Observation data. These Web Services enable scientists to test their analysis methods first on their desktops and then move them to big EO data cubes.

Based on these considerations, the *e-sensing* architecture uses the following building blocks:

1) The SciDB open source array database [4] that allows easy mapping of big EO data to its data structure.

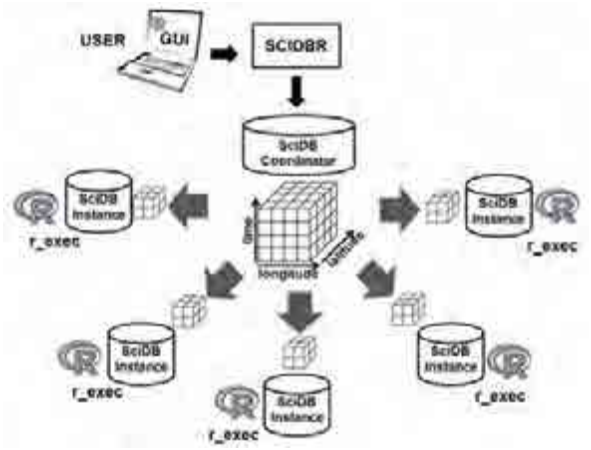


Fig. 1: Remote execution of R scripts in SciDB

2) R as the tool for big data analytics, so that researchers can thus scale up their methods, reuse previous work, and collaborate with their peers.
3) The R packages SITS [11] and dtwSat [12], for big EO analytics on satellite image time series.
4) Web services (WTSS and WTSPS) for big EO data, adapted to the needs of satellite image time series [10].
5) The architecture is fully open source, being made available online at https://github.com/e-sensing/.

3. MATCHING DATA INFRASTRUCTURES TO ANALYTICAL NEEDS

Most studies on time series for land cover classification in the literature use classical remote sensing methods [6]. For multiyear studies, researchers derive ''best-fit" yearly composites and then classify each composite image separately. The results from different periods are compared to detect change. We denote these works as taking a *space-first, time-later* approach.

Space-first, time-later methods do not use the full potential of remote sensing time series. The benefits of SITS increase when the temporal resolution of the big data set captures the most important changes. In these cases, the temporal autocorrelation of the data will be stronger than the spatial autocorrelation. Given data with adequate repeatability, a pixel is more related to its temporal neighbours than to its spatial ones. In these cases, *time-first, space-later* methods lead to better results than the *space-first, time-later* approach.

There has been much recent interest in the Earth observation community on using advanced statistical learning methods such as support vector machines [13] and random forests [14]. However, most researchers still use a *space-first, time-later* approach in connection with these methods. The dimensions of the decision space are limited to the number of spectral bands or their transformations. These approaches do

not use the power of advanced statistical learning techniques to work on high-dimensional spaces and with big training data sets [15].

The analytical methods of the *e-sensing* architecture combine data from image time series with statistical learning, using a *time-first, space later* approach. These methods use the full depth of dense time series to train advanced predictive models. These model include linear and quadratic discrimination analysis, support vector machines, random forests and neural networks. In a typical classification problem, we use time series with known land cover labels to derive measures that capture class attributes. Based on these measures, referred as training data, we provide support to select a predictive model that allows inferring classes of a larger data set.

Our proposal uses the full depth of satellite image time series to create large dimensional spaces. 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.

To illustrate the approach, Figure 2 shows the plot of the NDVI values of 370 time series for land cover class "Pasture", based on ground samples. Each thin line is one time series. The darker lines are the median and first and third quartile values. By visualizing the data, the challenge of distinguishing noise from natural variation becomes clear. The data shows natural variability due to different climate regimes and shows noise associated to cloud cover. To avoid losing information, we use the raw data such as this one to train a support vector machine, a classifier which is robust to noisy data sets.

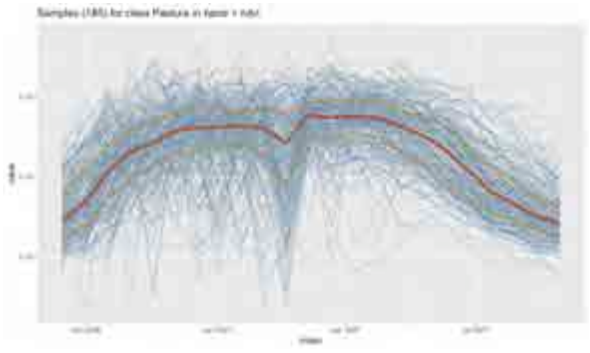


Fig. 2: Time series of 370 ground samples for land cover class "Pasture" in the state pf Mato Grosso, Brazil (source: authors).

As a case study, we developed a detailed land use change map of the state of Mato Grosso, Brazil, an area of 900,000 km^2, which has about 20 billion time series measures. We used the MODIS MOD13Q1 product from 2001 to 2016, provided every 16 days at 250-meter resolution, with 23 samples per year. By taking samples of labelled time series with 4 bands, we feed the statistical inference model with a 92-dimensional attribute space. For the analysis, we used the Normalized Difference Vegetation Index (NDVI) and the Enhanced Vegetation Index (EVI), and the near infrared (NIR) and middle infrared (MIR) bands. We defined nine classes (see Table 1 that include the most important crops and production systems in Mato Grosso. Based on a 5-fold cross validation, we estimate an overall accuracy of 94% and the Kappa index was 0.92. Producer's and user's accuracies of all classes were close to or better than 90%. This confirms the applicability of the proposed method in classify agricultural areas. In general, results show good discrimination between different crops, which improves on previous work [16], [17], [18].

Table 1: Confusion matrix of MODIS time series images, obtained by 5-fold cross validation of classification of field data, and values of producer's accuracy (PA) and user's accuracy (UA) for each class.

	1	2	3	4	5	6	7	8	9	UA
1 Cerrado	393	0	0	12	0	0	0	0	0	0.97
2 Fallow-Cotton	0	33	0	0	1	2	0	0	0	0.92
3 Forest	1	0	136	0	0	0	0	0	0	0.99
4 Pasture	6	0	1	357	3	1	0	5	0	0.96
5 Soy-Corn	0	1	1	1	352	18	0	26	4	0.87
6 Soy-Cotton	0	0	0	0	13	376	0	4	0	0.96
7 Soy-Fallow	0	0	0	0	0	0	88	0	0	1.00
8 Soy-Millet	0	0	0	0	25	2	0	199	2	0.87
9 Soy-Sunflower	0	0	0	0	4	0	0	1	47	0.90
PA	0.98	0.97	0.99	0.96	0.88	0.94	1.00	0.85	0.89	

4. COMPUTING PERFORMANCE

The architecture has been implemented operationally at Brazil's National Institute for Space Research. In terms of hardware, our architecture uses 2 clusters. Each cluster has 5 servers with 2 CPUs with 6-cores each, operating at 2.4GHz with a 15MB cache. Each server has 96 GB of RAM, and 16 TB of data storage. This gives 60 cores per cluster that can work in parallel in a "shared-nothing" data storage. The array database SciDB includes the full set of MODIS MOD09Q1 images at 250 meter resolution for South America, with 13,800 images associated to 317 billion data series. It also include selected datasets of mixed LANDSAT-8 and MODIS data sets, at 30 meter resolution.

In terms of performance, the classification scales up almost linearly. The full processing of all time series to classify 16 years of data in Mato Grosso state (900,000 km^2) takes about 6 hours using the R-SciDB interface. We also processed all of the area of Brazil's Cerrado biome (2,050,000 km^2) in about 13 hours. This shows that distributed processing with a right degree of granularity can compensate for the slower

performance of R scripts, compared with compiled languages. By using R, researchers have much flexibility when designing data analysis methods. Given these results, we argue that using SciDB combined with R is an adequate solution for big Earth Observation data analytics.

Table 2: Performance time for selected case studies

Case Study	Area (km^2)	Decision dimensions	Measures (GB)	Proc time (hours)
Mato Grosso	900,000	92	135	6
Cerrado	2,050,000	92	308	13

5. FINAL REMARKS

This paper discusses the design of an architecture that allows using satellite image time series with advanced statistical learning. Its results indicate that solutions based on array DBMS, R algorithms, and dedicated web services are well suited for satellite image time series analysis. This knowledge platform expands what can be done with big EO data, allowing scalability and reproducibility, without major compromises in performance. In the long run, it shows that the *time-first, space later* approach is an important complement of more traditional image analysis methods.

Combining array databases with R statistical computing is not an universal solution for big Earth observation data analysis. Alternative designs such as the Australian Data Cube (flat files) and Google Earth Engine (MapReduce) provide support for important studies is cases where the analysis methods are established and the novelty comes from applying them to big data. In areas where the current methods are not adequate and progress is required, such as global land cover, it is important to design new architectures such as the one proposed in the paper. We hope that our results encourage further work on the use of satellite image time series for land cover classification.

6. REFERENCES

[1] A. Lewis, S. Oliver *et al.*, "The Australian Geoscience Data Cube — Foundations and lessons learned," *Remote Sensing of Environment (online)*, 2017.

[2] N. Gorelick, M. Hancher *et al.*, "Google Earth Engine: Planetary-scale geospatial analysis for everyone," *Remote Sensing of Environment*, 2017.

[3] P. Baumann, A. Dehmel *et al.*, "The multidimensional database system RasDaMan," *ACM SIGMOD Record*, vol. 27, no. 2, pp. 575–577, 1998.

[4] M. Stonebraker, P. Brown *et al.*, "SciDB: A database management system for applications with complex analytics," *Computing in Science & Engineering*, vol. 15, no. 3, pp. 54–62, 2013.

[5] K. Doan, A. O. Oloso *et al.*, "Evaluating the impact of data placement to Spark and SciDB with an Earth Science use case," in *2016 IEEE International Conference on Big Data*, 2016, pp. 341–346.

[6] C. Gomez, J. C. White, and M. A. Wulder, "Optical remotely sensed time series data for land cover classification: A review," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 116, pp. 55 – 72, 2016.

[7] V. J. Pasquarella, C. E. Holden *et al.*, "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*, vol. 2, no. 3, pp. 152–170, 2016.

[8] S. Fritz, L. See *et al.*, "Highlighting continued uncertainty in global land cover maps for the user community," *Environmental Research Letters*, vol. 6, no. 4, p. 044005, 2011.

[9] N. Tsendbazar, S. de Bruin, and M. Herold, "Assessing global land cover reference datasets for different user communities," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 103, no. Sup C, pp. 93 – 114, 2015.

[10] L. Vinhas, G. Ribeiro *et al.*, "Web services for big Earth observation data," in *Proceedings of the 17th Brazilian Symposium on GeoInformatics*. Campos do Jordão, SP, Brazil: INPE, 2016, pp. 26–35.

[11] R. Simoes, G. Camara *et al.*, *SITS: Satellite Image Time Series Analysis*, 2017, r package version 0.9.30. [Online]. Available: https://github.com/e-sensing/sits/

[12] V. Maus, G. Camara *et al.*, "dtwSat: Time-Weighted Dynamic Time Warping for Satellite Image Time Series Analysis in R," *Journal of Statistical Software (accepted)*, 2017.

[13] G. Mountrakis, J. Im, and C. Ogole, "Support vector machines in remote sensing: A review," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 66, no. 3, pp. 247–259, 2011.

[14] M. Belgiu and L. Dragut, "Random forest in remote sensing: A review of applications and future directions," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 114, pp. 24–31, 2016.

[15] G. James, D. Witten *et al.*, *An Introduction to Statistical Learning: with Applications in R*. New York, EUA: Springer, 2013.

[16] J. Kastens, J. Brown *et al.*, "Soy moratorium impacts on soybean and deforestation dynamics in Mato Grosso, Brazil," *PLOS ONE*, vol. 12, no. 4, p. e0176168, 2017.

[17] M. N. Macedo, R. S. DeFries *et al.*, "Decoupling of deforestation and soy production in the southern Amazon during the late 2000s," *PNAS*, vol. 109, no. 4, pp. 1341–1346, 2012.

[18] D. Arvor, M. Jonathan *et al.*, "Classification of MODIS EVI time series for crop mapping in the state of Mato Grosso, Brazil," *International Journal of Remote Sensing*, vol. 32, no. 22, pp. 7847–7871, 2011.

C Reproducible geospatial data science: Exploratory Data Analysis using collaborative analysis environments

Reproducible geospatial data science: Exploratory Data Analysis using collaborative analysis environments

Alber Sánchez[1], Lubia Vinhas[1], Gilberto Ribeiro de Queiroz[1], Rolf Simoes[1], Vitor Gomes[1], Luiz Fernando F. G. de Assis[1], Eduardo Llapa, Gilberto Câmara[1]

[1]National Institute for Space Research (INPE)
Av. dos Astronautas 1758 – 12227-010
São José dos Campos – SP – Brazil

{alber.ipia, lubia.vinhas, gilberto.queiroz}@inpe.br

{rolf.simoes, gilberto.camara}@inpe.br

vitor@ieav.cta.br

{luizffga, edullapa}@dpi.inpe.br

Abstract. *The answers to current our planet's problems could be hidden in gigabytes of satellite imagery of the last 40 years, but scientists lack the means for processing such amount of data. To answer this challenge, we are building a scientific platform for handling big Earth observation data. We organized decades of satellite images into data cubes in order to put together data and analysis. Our platform allows to scale-up analysis to larger areas and longer periods of time. However, we need to provide scientists with tools and mechanisms to test and refine their routines before interacting with the Big data hosted in our platform.*
We believe that web services along collaborative analysis environments fit the hypothesis-test pattern followed by researchers while writing scientific computer code. Web services enable us to embed our platform's data and algorithms into collaborative analysis environments such as Jupyter notebooks.
To make our case, we prepared a Jupyter notebook where Earth observation scientists can interact with our platform through web services and the analytic capabilities of the programming language Python.

Resumo. *As respostas aos problemas globais atuais podem estar ocultas em gigabytes de imagens de satélite de observação da Terra adquiridas nos últimos 40 anos, mas nem sempre os cientistas possuem os meios para processá-las e transformá-las em informação. Para responder a esse desafio, estamos construindo uma plataforma científica para processar grandes volumes de dados de observação da Terra. Para isso, nós organizamos décadas de imagens de satélite em cubos de dados, a fim de juntar dados e análises. Nossa plataforma, está sendo concebida para permitir a anĺise de grandes ŕeas com dados de longos períodos de tempo mais longos. No entanto, precisamos fornecer aos cientistas ferramentas e mecanismos para testar e refinar suas rotinas antes de interagir com os dados hospedados em nossa plataforma.*
Acreditamos que os serviços Web e os ambientes de análise colaborativos encaixam com o padrão de hipótese-teste seguido pelos pesquisadores. Os serviços

da Web nos permitem incorporar os dados e algoritmos da nossa plataforma em ambientes de análise colaborativa, como os Jupyter notebooks.
Para testar nossa hipótese, nós preparamos um Jupyter notebook onde cientistas da observaço da Terra podem interagir com a nossa plataforma através de serviços web e as capacidades analíticas da linguagem de programação Python.

1. Introduction

Earth observation scientist are unable to use all the available images in their analyses because processing such volume of data demands large hardware resources, new software tools, and sound analysis techniques. These issues and requirements associated to large amounts of data are commonly addressed as the *data deluge* or *big data* [Bell et al. 2009, Boyd and Crawford 2012, Li et al. 2016]. Besides, the current satellite image distribution model is based on files. These files have their own formats and access interfaces. This distribution model had led to problems such as data duplication and the inability to track the files used or required for each analysis. The data used for Earth Observation analysis are either unavailable or just too large for independent result validation which in turn, boosts the scientific reproducibility crisis [Baker 2016, Nature 2016]. For these reasons, we are putting together data and analysis by means of a platform for handling big geospatial data. We are using our platform to research land use and land cover change.

As the amount of data increases, it is more efficient to move the algorithms to the data than the other way around [Borthakur 2007]. However, the conditions and mechanisms by which scientists move their algorithms to our platform is unknown; we would like scientist to focus on analysis and to forget about data structures and computing scalability.

We acknowledge how troublesome is the process of writing computerized scientific analysis routines and we are committed to make easier for scientists to scale up their analysis from the desktop to our platform. We believe the best moment to make our data and analysis available to scientist is at the earliest stages of their analysis. This approach can diminish the amount of rework implied while scaling up analysis.

Unfortunately, each scientist writes analysis routines on its own way. However, it is known they keep notebooks with descriptions, data and results of their experiments. Apart from this, Donald Knuth introduced *literate programming* as a way to develop, document, and publish scientific algorithms relying in both natural and machine language. Furthermore, Jim Gray proposed *Overlay Journals* as means to share, manage, and improved scientists' notebooks [Knuth 1984, Gray 2009]. These ideas are being taken to the web in the form of electronic scientific notebooks which are on-line, collaborative documents that mix code, data, descriptions, and tables to summarize the results of scientific research[Pérez and Granger 2007].

We believe that web services along collaborative analysis environments fit the hypothesis-test pattern followed by researchers while writing scientific computer code. Web services enable us to embed our platform's data and algorithms into collaborative analysis environments which are electronic approximations to the scientists' notebooks and laboratory journals.

In this paper, we examine how our platform can be integrated into the analysis workflow of Earth observation data. To achieve this, we briefly introduce our computing platform and its web services(section 2 and 3). Then, we describe analysis environments and how the into into the scientists' workflow (section 4). Finally, we test our approach by setting up Jupyter notebook — a collaborative analysis environment — in which we mixture the web services provided by our platform and the analysis analytical tools provided by the Python programming language.

2. The e-sensing platform

The *e-sensing* [1] project aims to build a platform for handling big geospatial data in order to help scientists to research land use and land cover change. We are organizing decades of satellite images into cubes — tridimensional space-time arrays —- inside our platform and finding the best way to put together data and analysis. The *e-sensing project* is ran by the Brazilian National Institute for Space Research (INPE).

The main requirements to these platforms are *analytical scaling*, *software reuse*, *collaborative work*, and *replication*. Analytical scaling is about allowing users to move their data and code between platforms of increasing processing capacities with little or no modifications at all. Software reuse means the platform must be able to use code from different origins. Collaborative work and replication are about enabling scientists to share and replicate their results [Câmara et al. 2016, Stonebraker et al. 2009]. We are addressing the software reuse, collaborative work, and replication by using open source and open access software and data. For example, inside our platform, we are only using open source software and open access data provided by NASA. But in this document we are addressing only the first step in the analytical scaling requirement.

Our platform is hosting an array database with both MODIS and LANDSAT images. We have been classifying time series of vegetation indexes of the Amazon forest into classes of Land Use and Land Cover Change (LUCC). In post-processing stages, we analyze the trajectories of LUCC over time [Assis et al. 2016, Camara et al. 2016, Lu et al. 2016, Maciel et al. 2017, Maus et al. 2016]. But the data workflow inside our platform relies on a mixture of technologies such as scripting languages (R, Python, Bash), distributed storage (SciDB, Hadoop), and operating system tools. As a result, it is hard for scientists to reproduce our results or to run their own [Câmara et al. 2016]. As mentioned earlier, we chose web services as the way to expose our platform computing capabilities while hiding its internal complexities.

On the other hand, the *CEOS Data Cube Platform* (CEOS-ODC) is a platform for storing, accessing, and managing metadata of remotely sensed data. CEOS-ODC is build on top of the *Australian Geoscience Data Cube*. Both platforms — *e-sensing* and CEOS-ODC — are interested in processing large amounts of satellite imagery and using open source tools. However, they use different type of analysis and architectures. While *e-sensing* is focused on time series analysis, the analysis supported by CEOS-ODC puts spatial before temporal analysis. Regarding architectures, *e-sensing* is built on top of array databases while CEOS-ODC is built around the programming language python and data files; this difference is subtle but important since databases are independent of program-

[1] e-sensing project `http://www.esensing.org/`

ming languages. As a consequence, the *e-sensing* platform is able to run analysis written in different languages while CEOS-ODC is constrained to python scripts [CEOS 2016].

3. A web service for retrieving time series

Sharing and re-using computer resources has been important since the 90s because writing software is error-prone and high performance hardware is expensive. Nowadays, *Web services* are the most common way to address this matter. Web services are the standardized way to access software and data over the World Wide Web independently of operating systems and programming languages. Through them, scientists can access the data and algorithms available in our platform and at the same time, web services hide complexities — such as mixed technologies, and distributed storage — behind an uniform interface.

The Web Time Series Service (WTSS) retrieves time series of Earth Observation data for specific locations. WTSS reduces the gap between data and remote-sensing time-series clients through a simple JSON representation. Traditionally, assembling time series of Earth Observation imagery is a time-consuming task because users need to sequentially open several image files, extract some pixels, and then store them. Instead, WTSS connects to an multidimensional array database and makes temporal queries on behalf of the client. WTSS exposes three main operations *list_coverages*, *describe_coverage*, and *time_series*. *list_coverages* returns a JSON list of the available coverages in the service. *describe_coverage* retrieves metadata of a specific coverage. Finally, the *time_series* operation retrieves specific time series [Vinhas et al. 2016]. WTSS implementation is publicly available on-line [2].

Moreover, WTSS has clients for the QGIS software and for the scripting languages R and Python. These WTSS clients enable scientists to access our data from on-line analysis environments.

4. Interactive and collaborative analysis environments

Literate programming is an style of coding software in which programs are treated as pieces of literature. That is, natural and machine languages are weaved together into a document where thought order prevails over code optimizations. Its goal is to create programs easier to understand and maintain and to achieve this, literate programming makes explicit the reasoning behind the code [Knuth 1984].

Note how literate programming fits the way scientists analyses their data. Once data is collected, scientists make research questions, then formulate hypotheses for later testing them on the data. The question making and hypothesis formulating is better described using natural language while data processing and hypothesis testing are automated using code.

The modern realization of literate programming are the on-line analysis environments. Using modern technologies, they add collaboration and interactivity to the traditional scientific notebooks and laboratory journals. Some examples are the R^3 and Jupyter[4] notebooks. It is worth noticing that R notebooks are focused in *R* while Jupyter

[2]e-sensing code repository `https://github.com/e-sensing/`

[3]R Notebooks `http://rmarkdown.rstudio.com/r_notebooks.html`

[4]The Jupyter Notebook `https://ipython.org/notebook.html`

notebooks support various programming languages. For this reason, we preferred the latter in this paper.

Statistical data analysis is crucial to science. From the computing perspective, the most popular and powerful computing tools for statistical analysis are R and Python. R is a computing environment designed for statistical analysis while Python is a general purpose programming language focused on readability and extensibility. Both support numerical processing, statistical data structures; the former natively while the latter trough code libraries such as SciPy [Ihaka 1998, Jones et al. 01 , OGrady 2016]. Both R and Python are supported by large communities of users coming from either the field of statistics or computer science. In this paper we preferred python because most of the author come from computer science field.

IPython adds facilities to Python for scientific computing. IPython has an interactive command with tailor-made features for scientists, such as code completion, plotting, and parallel and distributed processing. These characteristics are taken to the web in the form of Jupyter notebooks [Kluyver et al. 2016]. For example, the data and algorithms regarding the recent astronomic discovery of gravitational waves are available as Jupyter notebooks [Dal Canton et al. 2014, Usman et al. 2016, Nitz et al. 2017].

5. Analysis of time series of vegetation indexes

To test our approach, we setup up a Jupyter notebook for the exploratory analysis of time series of vegetation indexes. The time series are provided through a WTSS server attached to a cube hosted in the e-sensing platform. Our notebook is publicly available[5]. In this notebook, we mix the web services provided by our platform and the analysis analytical tools provided by the Python programming language. Our notebook presents three common jobs regarding time series of vegetation indexes: Exploratory analysis, filtering or smoothing, and classification. Figure 1 is an screen-shot of our notebook running on a web browser.

In the exploratory analysis, we get the data and then plot the time series and its location on a map. Figure 2 shows how to retrieve MODIS data into a data frame which is a table-like data structure.

Once the time series is formated as a data frame, it is simple to apply on it functions that receive and return data frame's columns as parameters. In this way, we smoothed our time series using the Kalman filter, the Fourier decomposition and the Whittaker smoother. The Kalman filter is well known in aeronautics while Fourier and Whitaker are known as good estimators of vegetation phenology [Atkinson et al. 2012, Grewal and Andrews 2010]. For example, Figure 1 shows the code and the application the Whitaker smoother to time series of vegetation indexes in a web browser.

The last example in our Jupyter notebook is classification. We used Dynamic Time Warping (DTW) to classify time series of vegetation indexes [Berndt and Clifford 1994]. DTW is an algorithm that computes a similarity measure — a distance — between two time series. Given a set of time series of known land coverages (the patterns), we compute the DTW distances to a time series of an unknown land cover (the samples). The samples

[5]Python for Data Science in Earth Observation Analysis `http://github.com/e-sensing/wgiss-py-webinar`

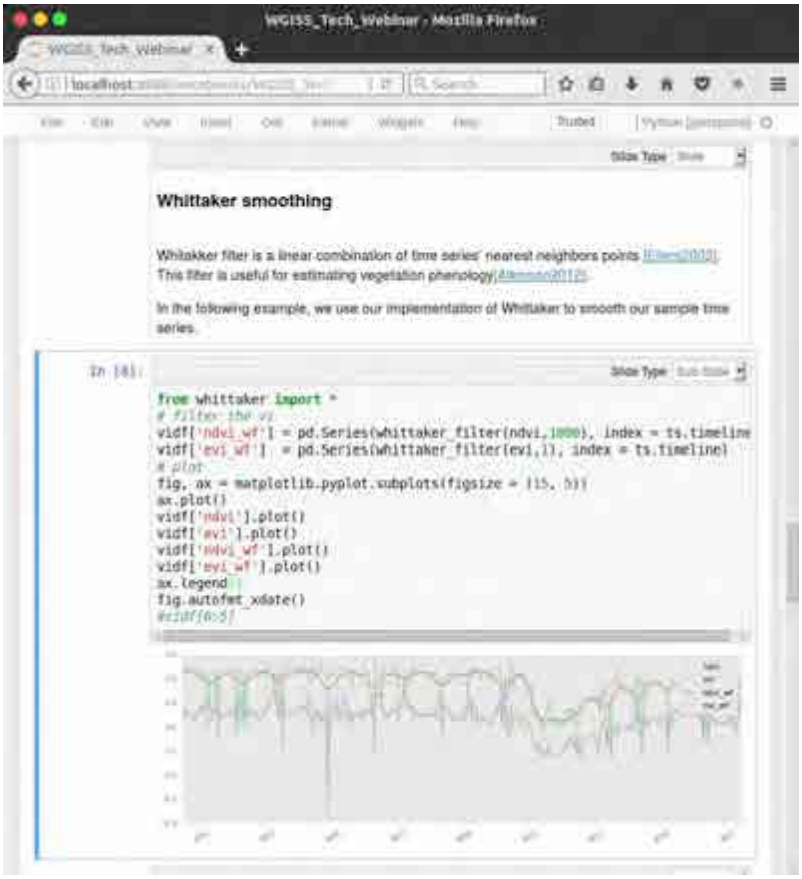


Figure 1. An on-line analysis environment for time series of Earth observation data. This environment displays a textual description of the Whitaker smoother along its Python implementation and its results when applied to a time series of vegetation indexes.

are assigned to the labels of the patterns with the shortest DTW distance.

We prepared a set of pattern time series corresponding to the land covers *cerrado* and *forest*. We also collected a set of sample points from which we know the latitude, the longitude and the land cover over a specific time interval; then we retrieved the time series of these points using WTSS. Figure 3 shows the time series of both pattern and samples. Figure 4 shows the code required to read the prepared files, retrieve the time series and to do the classification.

In summary, we joined data and analysis environments in order to plot, filter, and classify time series of Earth observation data by means of Jupyter notebooks and web services. This approach is flexible as users can use the same data and web services over different programming languages and analysis environments. For example, we setup another notebook using *R*, which is an statistical programming language. We do not describe this *R* notebook here because of lack of room, but the code is available on-line. [6]

[6]e-Sensing: Big Earth observation data analytics for land use and land cover change information `https://github.com/e-sensing/SITS_R_notebook`

```
import pandas as pd
from wtss import wtss
from tsmap import *
w = wtss("http://www.dpi.inpe.br/tws")
latitude  = -14.919100049
longitude = -59.11781088
ts = w.time_series("mod13q1_512", ("ndvi", "evi"), \
  latitude, longitude)
ndvi = pd.Series(ts["ndvi"], index = ts.timeline) * \
  cv_scheme['attributes']['ndvi']['scale_factor']
evi  = pd.Series(ts["evi"],  index = ts.timeline) * \
  cv_scheme['attributes']['evi']['scale_factor']
vidf = pd.DataFrame({'ndvi': ndvi, 'evi': evi})
```

Figure 2. Get a time series into a Python pandas data frame.

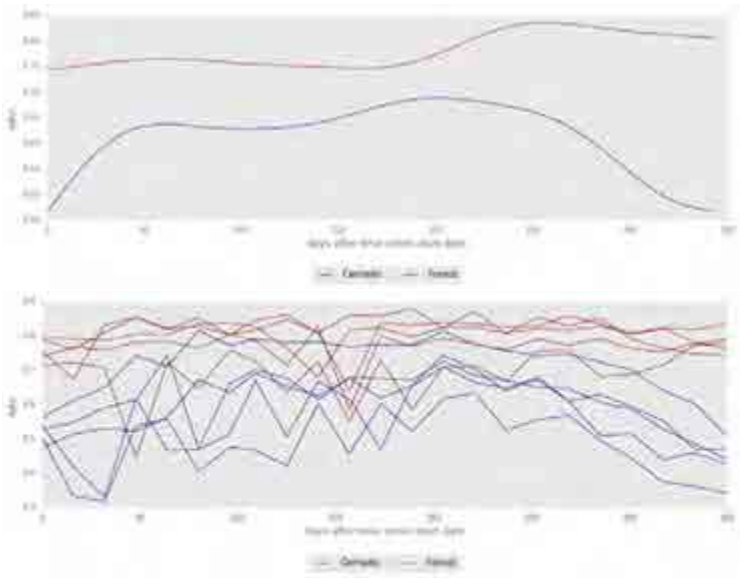


Figure 3. Patterns (top) and samples (bottom) of NDVI time series for classification.

```
from dtw import *
from tools import *
patterns_ts = pd.read_json("examples/patterns.json", orient='records')
patterns_ts["timeline"] = pd.to_datetime(patterns_ts["timeline"])
samples = pd.read_csv("examples/samples.csv")
samples_ts = wtss_get_time_series(samples)
classification = classifier_1nn(patterns_ts, samples_ts)
```

Figure 4. Python code for classifying time series using Dynamic Time Warping.

6. Conclusions

In this paper, we discussed how literate programming is being taking to the Web as interactive and collaborative analysis environments. We also showed how this environments are enhanced with web services and how both — environments and services —-help scientists to prepare their analysis routines. We set up a Jupyter notebook in which we analyzed data retrieved by the Web Time Series Service. In this way, we showed how to display, filter, smooth and classify time series of vegetation indexes. This is a convenient for scientists not only to interact with time series of Earth observation data but also to prepare their analysis routines before running them on big Earth observation data platforms such as *e-sensing*.

Web services close the gap between big Earth observation data and analysis tools by means of collaborative environments for small amounts of data. As the amount of data to be processed increases, it is better to send the analysis routine to the data which is an ongoing effort at the *e-sensing* project.

Finally, we would like to remark that the aforementioned the Jupyter notebook, the Web Time Series Service, and the analysis routine are available on-line to everyone at `http://github.com/e-sensing/wgiss-py-webinar`.

7. Acknowledgements

The authors are supported by the São Paulo Research Foundation (FAPESP) e-science program (grant 2014-08398-6). Gilberto Camara is also supported by CNPq (grant 312151-2014-4).

References

Assis, L. F., Ribeiro, G., Ferreira, K. R., Vinhas, L., Llapa, E., Sanchez, A., Maus, V., and Camara, G. (2016). Big data streaming for remote sensing time series analytics using MapReduce. In *Proceedings of the XVII Brazilian Symposium on GeoInformatics*, number November.

Atkinson, P. M., Jeganathan, C., Dash, J., and Atzberger, C. (2012). Inter-comparison of four models for smoothing satellite sensor time-series data to estimate vegetation phenology. *Remote Sensing of Environment*, 123:400–417.

Baker, M. (2016). Is there a reproducibility crisis? *Nature*, 533(7604):452–454.

Bell, G., Hey, T., and Szalay, A. (2009). Computer science. Beyond the data deluge. *Science (New York, N.Y.)*, 323(5919):1297–1298.

Berndt, D. J. and Clifford, J. (1994). Using dynamic time warping to find patterns in time series. In Fayyad, U. M. and Uthurusamy, R., editors, *KDD Workshop*, pages 359–370. AAAI Press.

Borthakur, D. (2007). The hadoop distributed file system: Architecture and design. *Hadoop Project Website*, 11(2007):21.

Boyd, D. and Crawford, K. (2012). Critical Questions for Big Data. *Information, Communication & Society*, 15(5):662–679.

Câmara, G., Assis, L. F., Ribeiro, G., Ferreira, K. R., Llapa, E., and Vinhas, L. (2016). Big earth observation data analytics: matching requirements to system architectures. In *Proceedings of the 5th ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data*, pages 1–6, Burlingname, CA, USA. ACM.

Camara, G., Maciel, A., Maus, V., Vinhas, L., and Sanchez, A. (2016). Using dynamic geospatial ontologies to support information extraction from big earth observation data sets. In *Ninth International Conference on Geographic Information Science (GI-Science 2016)*, Montreal, CA. AAG.

CEOS (2016). The CEOS Data Cube. Three-year work plan 2016-2018.

Dal Canton, T. et al. (2014). Implementing a search for aligned-spin neutron star-black hole systems with advanced ground based gravitational wave detectors. *Phys. Rev.*, D90(8):082004.

Gray, J. (2009). Jim gray on escience: A transformed scientific method. *The fourth paradigm: Data-intensive scientific discovery*, 1.

Grewal, M. and Andrews, A. (2010). Applications of Kalman Filtering in Aerospace 1960 to the Present [Historical Perspectives. *IEEE Control Systems Magazine*, 30(3):69–78.

Ihaka, R. (1998). R: Past and future history. *Computing Science and Statistics*, 392396.

Jones, E., Oliphant, T., Peterson, P., et al. (2001–). SciPy: Open source scientific tools for Python. [Online; accessed 2011/11/09].

Kluyver, T., Ragan-kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J., Grout, J., Corlay, S., Ivanov, P., Avila, D., Abdalla, S., and Willing, C. (2016). Jupyter Notebooks—a publishing format for reproducible computational workflows. *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, pages 87–90.

Knuth, D. E. (1984). Literate programming. *The Computer Journal*, 27(2):97–111.

Li, S., Dragicevic, S., Castro, F. A., Sester, M., Winter, S., Coltekin, A., Pettit, C., Jiang, B., Haworth, J., Stein, A., and Cheng, T. (2016). Geospatial big data handling theory and methods: A review and research challenges. *ISPRS Journal of Photogrammetry and Remote Sensing*, 115:119–133.

Lu, M., Pebesma, E., Sanchez, A., and Verbesselt, J. (2016). Spatio-temporal change detection from multidimensional arrays: Detecting deforestation from MODIS time series. *ISPRS Journal of Photogrammetry and Remote Sensing*, 117:227–236.

Maciel, A. M., Vinhas, L., Câmara, G., Maus, V. W., and Assis, L. F. F. G. (2017). STILF - A spatiotemporal interval logic formalism for reasoning about events in remote sensing data. In *Proceedings...*, pages 4558–4565, São José dos Campos. Brazilian Symposium on Remote Sensing, 18. (SBSR), National Institute for Space Research (INPE).

Maus, V., Camara, G., Cartaxo, R., Sanchez, A., Ramos, F. M., and de Queiroz, G. R. (2016). A time-weighted dynamic time warping method for land-use and land-cover mapping. 9(8):3729 – 3739.

Nature (2016). Reality check on reproducibility. *Nature*, 533(7604):437–437.

Nitz, A., Harry, I., Brown, D., Biwer, C. M., Willis, J., Canton, T. D., Pekowsky, L., Dent, T., Williamson, A. R., Capano, C., De, S., Cabero, M., Machenschalk, B., Kumar, P., Reyes, S., Massinger, T., Lenon, A., Fairhurst, S., Nielsen, A., shasvath, Pannarale, F., Singer, L., Macleod, D., Babak, S., Gabbard, H., Veitch, J., Sugar, C., Zertuche, L. M., Couvares, P., and Bockelman, B. (2017). ligo-cbc/pycbc: O2 production release 19.

OGrady, S. (2016). The redmonk programming language rankings: January 2016. 2016. *URL: http://redmonk. com/sogrady/2015/07/01/language-rankings-6-15/(visited on 2017/11/09)*.

Pérez, F. and Granger, B. E. (2007). IPython: a system for interactive scientific computing. *Computing in Science and Engineering*, 9(3):21–29.

Stonebraker, M., Becla, J., DeWitt, D. J., Lim, K.-t., Maier, D., Ratzesberger, O., and Zdonik, S. B. (2009). Requirements for Science Data Bases and SciDB. In {*CIDR*} *2009, Fourth Biennial Conference on Innovative Data Systems Research, Asilomar, CA, USA, January 4-7, 2009, Online Proceedings*.

Usman, S. A. et al. (2016). The PyCBC search for gravitational waves from compact binary coalescence. *Class. Quant. Grav.*, 33(21):215004.

Vinhas, L., de Queiroz, G. R., Ferreira, K. R., and Câmara, G. (2016). Web services for big earth observation data. In *GeoInfo*, pages 166–177.

Bibliography

[1] L. Assis, G. Queiroz, K. Ferreira, L. Vinhas, E. Llapa, A. Sanchez, V. Maus, and G. Camara, "Big data streaming for remote sensing time series analytics using MapReduce," in *Proceedings of the XVII Brazilian Symposium on GeoInformatics*, Campos do Jordão, SP, Brazil: Brazilian Journal of Cartography, 2016.

[2] GlobCover, "ESA Data User Element," 2017. Available at: `http://due.esrin.esa.int/page_globcover.php`.

[3] J. Latham, R. Cumani, I. Rosati, and M. Bloise, "Global land cover share (GLC-SHARE) database beta-release version 1.0-2014," *FAO: Rome, Italy*, 2014.

[4] G. Câmara, L. F. Assis, G. Ribeiro, K. R. Ferreira, E. Llapa, and L. Vinhas, "Big earth observation data analytics: matching requirements to system architectures," in *Proceedings of the 5th ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data*, (Burlingname, CA, USA), pp. 1–6, ACM, 2016.

[5] M. Stonebraker, P. Brown, D. Zhang, and J. Becla, "Scidb: A database management system for applications with complex analytics," *Computing in Science & Engineering*, vol. 15, no. 3, pp. 54–62, 2013.

[6] R Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015.

[7] G. Grekousis, G. Mountrakis, and M. Kavouras, "An overview of 21 global and 43 regional land-cover mapping products," *International Journal of Remote Sensing*, vol. 36, no. 21, pp. 5309–5335, 2015.

[8] M. A. Friedl, D. Sulla-Menashe, B. Tan, A. Schneider, N. Ramankutty, A. Sibley, and X. Huang, "Modis collection 5 global land cover: Algorithm refinements and characterization of new datasets," *Remote sensing of Environment*, vol. 114, no. 1, pp. 168–182, 2010.

[9] LP DAAC, "NASA Land Data Products and Services," 2017. Available at: `https://lpdaac.usgs.gov/dataset_discovery/modis/modis_products_table/mcd12q1`.

[10] S. Bontemps, P. Defourny, E. V. Bogaert, O. Arino, V. Kalogirou, and J. R. Perez, "Globcover 2009-products description and validation report," tech. rep., ESA (European Space Agency) Report, 2011.

[11] GeoNetwork Team, "GeoNetwork opensource portal to spatial data and information," 2017. Available at: `http://www.fao.org/geonetwork/srv/en/main.home?uuid=ba4526fd-cdbf-4028-a1bd-5a559c4bff38`.

[12] P. Olofsson, G. M. Foody, M. Herold, S. V. Stehman, C. E. Woodcock, and M. A. Wulder, “Good practices for estimating area and assessing accuracy of land change,” *Remote Sensing of Environment*, vol. 148, pp. 42–57, 2014.

[13] V. Maus, G. Câmara, R. Cartaxo, A. Sanchez, F. M. Ramos, and G. R. de Queiroz, “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 (J-STARS)*, 2016.

[14] V. Maus, G. Câmara, R. Cartaxo, A. Sanchez, F. M. Ramos, and G. R. de Queiroz, “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 (J-STARS)*, 2015.

National Institute for Space Research

Image processing division

Tools for satellite image management in array databases

Yearly Scientific Report of the Regular-Support Research Project, funded by the São Paulo Research Foundation.

Projeto FAPESP #2016/03397-7

Responsible Researcher: Alber Hamersson Sánchez Ipia

São José dos Campos, 30 de Dezembro de 2017

Project information

- Project's title:

  Tools for satellite image management in array databases

- Name of the responsible researcher:

  Alber Hamersson Sánchez Ipia

- Project's hosting institution:

  Image processing division da National Institute for Space Research

- Research team:

  Alber Hamersson Sánchez Ipia

- Research project's number:

  2016/03397-7

- Duration:

  From 01/04/2016 to 31/03/2018

- Term covered in this report:

  From 01/01/2017 to 30/12/2017

Abstract

This document presents the achievements of the sub-project *Tools for satellite image management in array databases* (FAPESP process 2016/03397-7) which is part of the project *e-Sensing: Big Earth observation data analytics for land use and land cover change information* (FAPESP 2014/08398-6). The sub-project consists on developing tools for the array database SciDB to load large amounts of satellite images.

In the last 12 months we continued the project development by designing new schemata to include more Earth observation data in our database. We also reported our results in international articles, posters, and presentations. We kept our database updated to the newest releases and we improved the data loading scripts. Finally, we answered the data requirements of our users.

Contents

Summary of the project

This project is part of the FAPESP project *e-sensing* (grant 2014/08398-6). Specifically, it corresponds to the task *Building and deployment of big Earth observation databases to support data analysis and use cases* [1].

The e-sensing project addresses the scientific question *How can we use e-science methods and techniques to substantially improve the extraction of land use and land cover change information from big Earth Observation data sets in an open and reproducible way?* Currently, scientists do not take advantage of the full potential of the freely available satellite images. Instead, they produce land cover maps taking either a single or at most two time references. As a result, the big data sets produced by remote sensing satellite are underemployed. The e-sensing project is about conceiving, building, and deploying a new type of knowledge platform for organizing, accessing, processing and analyzing big Earth observation data [1].

The e-sensing project is framed by the fast land cover and land use change and its global consequences on the Earth's systems. As the human population grows, it also grows the demand of resources from the environment. These demands are satisfied by changing the use — and in consequence, the cover — of the Earth. As the surface of the Earth is finite, mankind development is jeopardized [2, 3]

Human beings have been changing Earth surface to satisfy their needs for millennia. However, it is just until recently when we acquire the capacity to collect massive amounts of data to study the changes of Earth's surface. Satellite images date back to the 70s; they contain detailed traces of human development for the last 50 years and they are now publicly available [4].

However, scientist are unable to use all the available images in their analyzes because processing such volume of data demands large hardware resources, new software tools, and sound analysis techniques. These issues and requirements are known as the data deluge or more commonly as big data [5, 6, 7].

The current satellite image distribution model is based on files. These files have their own formats and access interfaces. This distribution model had led to problems such as data duplication and the inability to track the files used for specific analysis. This contributes to the already existent reproducibility crisis in science; specifically, the data used for Earth Observation analysis are either unavailable or just too large for independent

result validation [8, 9].

Independently of format or interface, images are stored and manipulated using an array pattern. Arrays are well-known structures for scientists, take for example the Network Common Data Form (NetCDF) and Hierarchical Data Format (HDF). Recently, computer scientists have mixed the array abstraction with the features of relational databases into what has been called array databases. Array databases add versioning, scalability, and fail-tolerance capabilities to the array abstraction [10, 11, 12, 13, 14].

As a solution to the aforementioned issues, the e-sensing project proposes an open-source knowledge platform for big Earth Observation data. Such platform would provide a homogeneous interface to organize, access, process, and analyze spatiotemporal data using by means of array databases. In this way, scientists will analyze and test their hypotheses using data with a larger extents and finer resolutions than before. Likewise, this platform will enable reproducibility as any scientist can reproduce anyone else results using the same data and interfaces [1].

The e-sensing project team chose SciDB as the array database to support the proposed knowledge platform. SciDB is an open source array database optimized for big data and analytics. SciDB is developed and maintained by the Massachusetts Institute of Technology. SciDB splits and distributes data among several servers following a shared nothing architecture paradigm [15, 16].

To achieve its goals, the e-sensing project proposed three work packages (WP) [1]:

1. Databases: This WP is about researching and developing array databases to store large Earth Observation data sets. It also develops work flows and methods for efficient storage, access and processing of large data sets.

2. Data analysis: This WP is about researching and developing spatiotemporal techniques for extracting change information on large Earth Observation data sets. This is relevant, for example, for forestry applications. This WP includes finding novel applications of remote sensing time series, and combining time series with multitemporal image processing.

3. Use case development: This WP comprises the development of applications for forestry and agriculture management where large Earth Observation data sets are useful. The use cases derived from these applications will validate the methods and data developed by the other work packages.

The first work package, databases, is composed of the tasks *Building and deployment of big Earth observation databases to support data analysis and use cases* and also *Extend*

SciDB for geographical data handling. The first one is concerned with the data required to perform analysis and the second one deals with the semantics and interoperability of spatial data. Together, this two tasks provide the foundations for the remaining work packages as they provide the e-sensing platform users with data sets and operations required by Earth Observation scientists. This report is concerned with the former task, which is split in three parts:

1. Database building. This task consists on loading satellite images to an array database.

2. Radiometric correction module. Radiometric correction allows the comparison of satellites images. As satellite images differ on spatiotemporal coverage, radiometric correction removes unwanted influences such as atmospheric effect (such as haze) or sensor errors.

3. Geometric correction module. Geometric correction is concerned with adjusting images in such a way that features of interest overlap.

The sub-task database building includes the following data sets:

- The MODIS MOD09Q1 and MOD13Q1 images at 250 meter and weekly resolutions with temporal extent from 2000 to 2014 and the spatial extent of South America. This data set has a size of 15 Terabytes.

- Part of INPE's LANDSAT-5 data collection. Its temporal extent goes from 1984 to 2012, the spatial extent covers Brazil with temporal resolution of 6 coverages a year. This data set has a size of 30 Terabytes.

- A data selection of satellites SENTINEL-2A, CBERS-4 and LANDSAT-8. The size of this data set is 10 Terabytes.

Achievements of the period

Database building

This sub-task can be divided into the database scheme, the tools to upload data, and the data uploaded. The details are below.

Database schema

Satellite images are two-dimensional numeric representation of certain properties of some segment of Earth's surface. Their two-dimensional nature is easily represented using arrays and, as a consequence, they nicely fit in array databases.

SciDB is a database able to handle large amounts of data through array abstractions and query languages. SciDB splits arrays into chunks which are distributed across the database's servers. Users control the way SciDB arranges data through array schemata. Schemata are defined in terms of dimensions, chunk sizes, overlaps, and attributes. The dimensions are the coordinate system used to create, read, and update data while the chunk sizes shape the units of storage on each dimension. The overlap parameters state the amount of data duplicated in the chunks' boundaries and the attributes are the actual contents of the arrays [16].

In our last report, we described the SciDB schema for MODIS data. It consisted on dimensions (*col_id*, *row_id*, and *time_id*), chunksizes (75, 75, 400), and no overlap between chunks. The attributes of each array correspond to each band in the image. In this schema, the spatial and temporal dimensions are absolute, meaning that every pixel in any MODIS image is uniquely referenced. We've made a slight change to this schema, we switched the chunk sizes to (40, 40, 512) as they are easier to aggregate and divide in order to match and join to other schemata.

However, Landsat and MODIS images are different regarding spatial, temporal and spectral resolution (see Figure 2.1). Their spatial reference system are different and when overlapped, their axis are oblique. Landsat spatial resolution is 10 times finer than MODIS and the number of pixel on each is also different: MODIS 4800 x 4800 pixels while a Lansat 8 has 7600 x 7700. Both Landsat and MODIS provide images each 16 days, but they differ on pre-processing characteristics (i.e MODIS uses the best

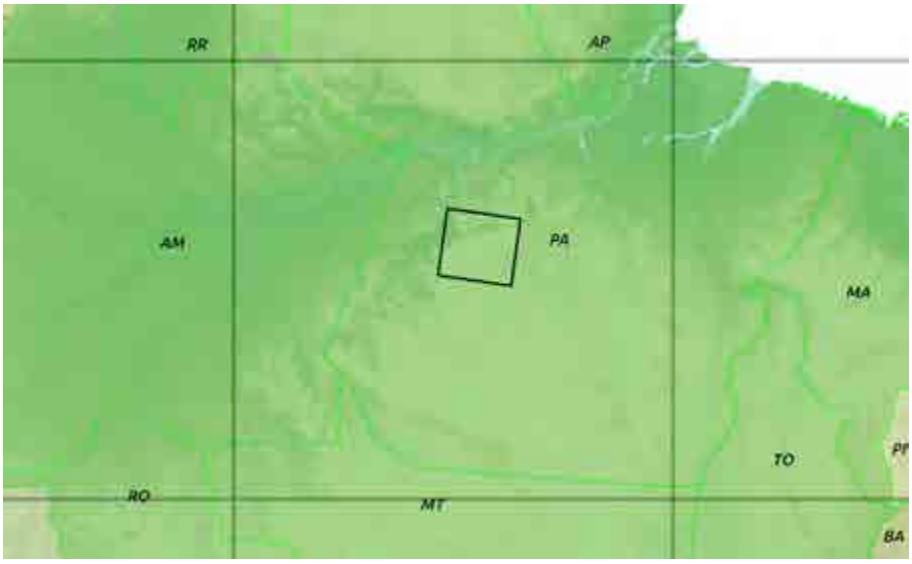


Figura 2.1: Coverage comparison between Landsat 8 (black square) and MODIS images (gray squares). The coverages correspond tho those of path 227 and row 63 (Landsat) and tile H12V09 (MODIS), over the Brazilian state of Pará. Note their obliquity due to their spatial reference systems.

pixel in the sampling period) and Landsat's overlapped zones are covered on half the time [17, 18]. One particularity of Landsat images is the overlap zones between neighbors (see Figure 2.2); the spatially overlapped regions are also sampled more frequently. In consequence, our Landsat schema has some modification in comparison to MODIS.

These differences highlight the fact that Landsat data were originally meant as images and not as arrays or data cubes. However, that perspective is changing due to the introduction of Landsat's collections and analysis ready data products.

Landsat collections are a new classification of imagery in order to simplify its usage. The collections include new and old images and all images are calibrated consistently in order to determine individual pixel quality. Landsat collections divide images into tiers according to their spatial precision using as a threshold a root mean square error of 12 meters. Tier 1 — the most precise — contains data which guarantees pixel-level spatial overlap and it is safe to use for time-series analysis[1].

On the other hand, Landsat Analysis Ready Data (ARD) are pre-processed images in such a way that users do not need to apply further processing. So far they are available for the United States only. However, it is a matter of time until they're made globally available [2].

[1]Landsat Collections `https://landsat.usgs.gov/landsat-collections`
[2]U.S. Landsat Analysis Ready Data (ARD) `https://landsat.usgs.gov/ard`

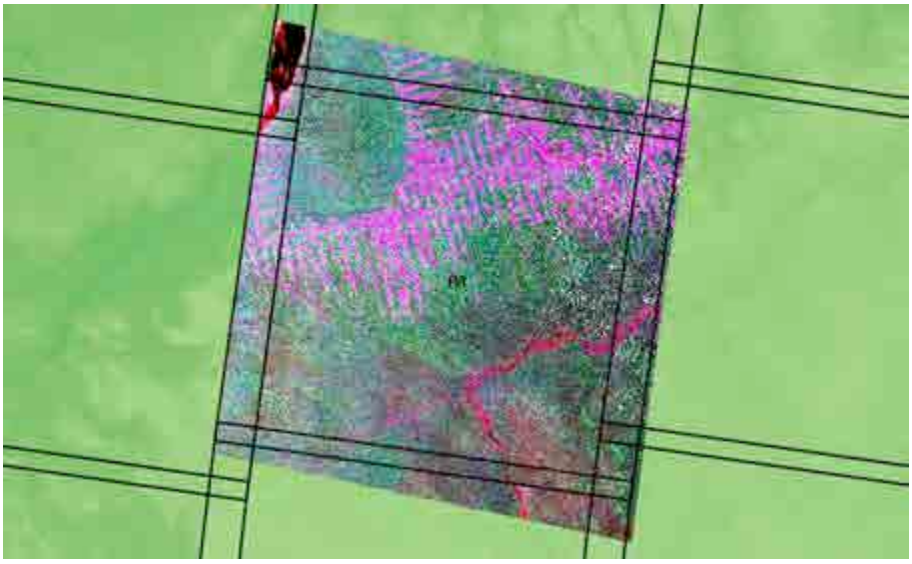


Figura 2.2: A Landsat 8 image and its neighbors' coverage (black lines). Note the overlap between the images.

Consequently, we designed a Landsat schema able to handle spatial and temporal overlap. In comparison to MODIS schema, we added two spatial dimensions to reference each image, in this way, the remaining spatial dimensions store pixel positions relative to each image. Regarding the temporal dimension, we make it ordinal, that is, instead of representing an absolute time position, it represents a relative position regarding other images in the same path and row. We also added a date attribute to the array, so we can keep track of the date of each pixel individually.

This schema allows dense arrays — that is, there are data for each combination of dimensions — while keeping the overlapped data. Besides, this is convertible into a MODIS-like schema by aggregating the overlapped data and dropping the first two dimensions (e.g. using the arithmetic mean). Next we describe the details of our Landsat schema:

- **Dimensions**. The array schema has 5 dimensions: Four for space and one for time that are respectively called *path*, *row*, *col_id*, *row_id*, and *time_id*. The first two dimensions place images relative to its neighbors; the following two dimensions number the pixels inside each image, starting at the top-left with (*0,0*). From there, the dimension value increments to the right and down until the down bottom-right pixel is reached. Lastly, the temporal dimension starts at 0 and it increases by one along the temporal resolution of the images.

- **Chunk Size**. The chunk sizes for data loading are *path=1*, *row=40*, *col_id=40*, *row_id=40*, and *time_id=512*. Since SciDB only stores the attribute data (and not the dimensions), the Landsat and MODIS chunks have the same size in disk despite the fact their schemata are different. As a consequence, both have similar performance.

- **Overlap**. The chunk overlap for Landsat and MODIS is zero, since these arrays are meant to mainly serve time series. As mentioned before, Landsat images are spatially overlapped and their overlap is not aligned to the images borders but to the satellite's flight direction. Fort this reason, the overlap property of SciDB's schemata cannot address Landsat overlaps.

Database loading tools

In our last report, we introduced the *modis2scidb-loader* [3] tools. These python scripts orchestrate the Extraction, Transformation, and Loading (ETL) processes of data.

Now, we are using the lessons learned from those scripts to write new ones — gdal2scidb [4] — focused on two things: The different satellite image formats and the parallel loading capabilities of SciDB. MODIS and Landsat data are delivered differently, the first uses a Hierarchical Data Format (HDF) where a single files stores all the image bands while the second uses one GeoTIFF file for each band on a single image; likewise, other satellites and data products deliver data using other formats.

These new scripts are built on top of Python modules which eases code reuse and adaptability to new imagery formats and to new SciDB versions. The new scripts hide the format complexities using the *ImageSeries* abstraction. An *ImageSeries* is a set of images of the same satellite, sensor, path and row but different acquisition time. The scripts are able to build *ImageSeries* objects from imagery and load the data to SciDB, independently if the imagery is MODIS, Landsat or Sentinel.

As the old scripts, the new ones use GDAL to read and transform the data into SciDB's binary format. Then, using the data is loaded using the Operating System and SciDB parallel processing capabilities [16, 19]

[3] SciETL - Extract, Transform and Load of Geospatial Data for SciDB `https://github.com/e-sensing/scietl`

[4] gdal2scidb - Python scripts for exporting a raster (supported by GDAL) to SciDB binary format and CSV `https://github.com/albhasan/gdal2scidb/tree/dev`

Uploaded data

- MODIS collection 6. Vegetation index product MOD13Q1 (in progress). These data are being uploaded because of the update from SciDB version 15 to version 16.
- Landsat 8. Collection 1, surface reflectance (in progress).
- Tropical Rainfall Measuring Mission. Satellite estimations of rainfall in the tropics.

Radiometric and geometric correction modules

The radiometric module is under development. Using the database language, we have tested basic radiometric corrections.

However, as mentioned earlier, MODIS and Landsat imagery are being adapted to time-series analysis and data cubes by implementing uniform radiometric and geometric corrections on their archives. These changes match the guidelines provided by organization such as the Committee on Earth Observation Satellites [5], and the Open Data Cube [6], which are coordinating international efforts in order to manage and organize satellite imagery. As a result, it is expected other sources of imagery will also deliver their images as Analysis Ready Data (e.g. the European Sentinel program).

As the imagery is made available with geometric and radiometric corrections (e.g. MODIS, Landsat collection 1 surface reflectance), the relevance of this module diminishes and for that reason is no longer a priority.

Other activities

Here we introduce other activities related to the fulfillment of this research project:

- Development of bash scripts to create time-series of MODIS vegetation indexes using the GeoTIFF format. These scripts create a single Geotiff image that includes the complete time series of a MODIS band. The GeoTIFF format is easier to consume by client-side analysis applications such as those provided by the programming language *R*. These scripts are available on-line at `https://github.com/albhasan/hdf2tif`

[5] CEOS `http://ceos.org/`
[6] Open Data Cube `https://www.opendatacube.org`

- Development of R scripts for processing time series using SciDB's stream. These scripts allow SciDB to run user-provided routines of analysis of vegetation indexes. Available on-line `https://github.com/albhasan/sdbStreamR4ts`

Impact assessment

The impact of the *e-sensing platform* can be assessed by the data and services provided by it as well their associated publications (see Chapter 3). [7]

These achievements are supported by the SciDB array database and its schemata, which enable the development of analysis tools. The database loading tools enable building spatiotemporal arrays made of satellites images taken by different sensors. The results can be seen as the uploaded data allowed the publications of several articles.

The experience and lessons learned by the *e-sensing* project staff are valuable for international organizations such as the Committee on Earth Observation Satellites, which whom we are sharing knowledge regarding big Earth data management.

[7] For a publication list, visit the *e-sensing* website `http://esensing.org/`

Participation in scientific events

The author of this report took part of the following:

1. Presentation of "Reproducible geospatial data science: Exploratory Data Analysis using collaborative analysis environments". In this presentation, we introduced how to use the data and services provided by the *e-sensing platform* in order to analyze time series of vegetation data, using on-line, open-source tools. Presented at the XVIII Brazilian Symposium on Geoinformatics on December 2017. Available online `http://mtc-m16c.sid.inpe.br/col/sid.inpe.br/mtc-m16c/2017/12.01.19.04/doc/1sanchez_camara.pdf`

2. Presentation of the paper "The e-sensing architecture for big Earth observation data analysis" and the poster "Big Earth observation data for fast detection of deforestation using adaptative filtering" at the 2017 Conference on Big Data from Space (BiDS'17). Here we presented the *e-sensing platform* and some partial results of the application of filtering to time-series of vegetation indexes to detect changes associated with deforestation. BiDS'17 took place in Toulouse, France on November 2017.

3. Short presentation about analysis of time series of vegetation indexes using the Python programming language. This is part of the Webinars from the CEOS Working Group on Information Systems & Services (WGISS) Technology Exploration Subgroup on August 2017. This presentation enables other communities to popularize the *e-sensing platform* and allowed us to meet other scientists working on related topics. Available online: `https://youtu.be/92Yg-57zkE4`

4. Co-author of "Climate drivers of the Amazon forest greening". This article exposes the relation between seasonal leaf production and increments in insolation and precipitation by using satellite and field observations [20].

5. Co-author of poster "Carbon Monoxide Measurements as a Biomass Burning Tracer at the Amazon Basin" presented at the 19th WMO/IAEA Meeting on Carbon Dioxide, Other Greenhouse Gases, and Related Measurement Techniques (GGMT-2017) at Duebendorf, Switzerland, on August 2017.

References

[1] E-SENSING: Big Earth observation data analytics for land use and land cover change information (proposal). 2014. `http://esensing.org/docs/e-sensing_proposal.pdf`. Accessed: 2012-12-20.

[2] FOLEY, J. A. et al. Global consequences of land use. v. 309, n. 5734, p. 570–574, 2005.

[3] ROCKSTRÖM, J. et al. Planetary boundaries: exploring the safe operating space for humanity. *Ecology and Society*, v. 14, n. 2, 2009.

[4] BELWARD, A. S.; SKOIEN, J. O. Who launched what, when and why; trends in global land-cover observation capacity from civilian earth observation satellites. *ISPRS Journal of Photogrammetry and Remote Sensing*, International Society for Photogrammetry and Remote Sensing, Inc. (ISPRS), v. 103, p. 115–128, 2015. ISSN 09242716. Disponível em: <http://dx.doi.org/10.1016/j.isprsjprs.2014.03.009>.

[5] BELL, G.; HEY, T.; SZALAY, A. Computer science. Beyond the data deluge. *Science (New York, N.Y.)*, v. 323, n. 5919, p. 1297–1298, 2009. ISSN 0036-8075. Disponível em: <http://www.sciencemag.org/content/323/5919/1297.full.pdf>.

[6] BOYD, D.; CRAWFORD, K. Critical Questions for Big Data. *Information, Communication & Society*, v. 15, n. 5, p. 662–679, 2012. ISSN 1369-118X.

[7] LI, S. et al. Geospatial big data handling theory and methods: A review and research challenges. *ISPRS journal of Photogrammetry and Remote Sensing*, Elsevier, v. 115, p. 119–133, 2016.

[8] BAKER, M. Is there a reproducibility crisis? *Nature*, v. 533, n. 7604, p. 452–454, may 2016. ISSN 0028-0836. Disponível em: <http://www.nature.com/doifinder/10.1038/533452a>.

[9] NEWS, N. Reality check on reproducibility. *Nature*, v. 533, n. 7604, p. 437–437, may 2016. ISSN 0028-0836. Disponível em: <http://www.nature.com/doifinder/10.1038/533437a>.

[10] BAUMANN, P. A Database Array Algebra for Spatio-Temporal Data and Beyond. *Proceedings of the 4th International Workshop on Next Generation Information Technologies and Systems*, n. Mdd, p. 76–93, 1999. Disponível em: <http://dl.acm.org/citation.cfm?id=646411.692530>.

[11] BRENNAN, J. et al. Working with NASA's HDF and HDF-EOS earth science data formats. *Earth Observer Newsletter*, v. 25, n. April 2013, p. 16–19, 2013. Disponível em: <https://scholar.google.nl/scholar?hl=nl&q=Working+with+NASA's+HDF+and+HDF-EOS+Earth+Science+Data+Formats+&btnG=&lr=#0>.

[12] REW, R.; DAVIS, G. NetCDF: An Interface for Scientific Data Access. *IEEE Computer Graphics and Applications*, v. 10, n. 4, p. 76–82, 1990. ISSN 02721716.

[13] GRAY, J. et al. Scientific data management in the coming decade. *SIGMOD Rec.*, v. 34, n. 4, p. 34–41, 2005. ISSN 01635808. Disponível em: <http://portal.acm.org.library.capella.edu/citation.cfm?doid=1107499.1107503%5Cnhttp://portal.a

[14] STONEBRAKER, M. et al. Requirements for Science Data Bases and SciDB. In: *{CIDR} 2009, Fourth Biennial Conference on Innovative Data Systems Research, Asilomar, CA, USA, January 4-7, 2009, Online Proceedings*. [s.n.], 2009. Disponível em: <http://www-db.cs.wisc.edu/cidr/cidr2009/Paper_26.pdf>.

[15] STONEBRAKER, M. The Case for Shared Nothing. *{IEEE} Database Eng. Bull.*, v. 9, n. 1, p. 4–9, 1986. Disponível em: <http://sites.computer.org/debull/86MAR-CD.pdf>.

[16] STONEBRAKER, M. et al. SciDB: A Database Management System for Applications with Complex Analytics. *Computing in Science Engineering*, v. 15, n. 3, p. 54–62, 2013. ISSN 1521-9615.

[17] ROY, D. et al. Landsat-8: Science and product vision for terrestrial global change research. *Remote Sensing of Environment*, Elsevier B.V., v. 145, p. 154–172, apr 2014. ISSN 00344257. Disponível em: <http://dx.doi.org/10.1016/j.rse.2014.02.001 http://linkinghub.elsevier.com/retrieve/pii/S003442571400042X>.

[18] SOLANO, R. et al. *MODIS vegetation indices (MOD13) C5 user's guide*. [S.l.], 2010. v. 2, 2010 p.

[19] TANGE, O. GNU Parallel: the command-line power tool. *The USENIX Magazine*, v. 36, n. 1, p. 42–47, 2011. Disponível em:

<https://www.usenix.org/publications/login/february-2011-volume-36-number-1/gnu-parallel-command-line-power-tool>.

[20] FABIEN, H. W. et al. Climate drivers of the Amazon forest greening. *PLOS ONE*, Public Library of Science, v. 12, n. 7, p. e0180932, jul 2017. ISSN 1932-6203. Disponível em: <https://doi.org/10.1371/journal.pone.0180932 http://dx.plos.org/10.1371/journal.pone.0180932>.

National Institute for Space Research (INPE)

Image Processing Division (DPI) and Remote Sensing Division (DSR)

Use of Remote Sensing Time Series for Brazilian Agriculture Monitoring

Researcher: Michelle Cristina Araujo Picoli

Supervisor: Prof. Dr. Gilberto Câmara

FAPESP project number: 2016/23750-3

Yearly Scientific Report (from January, 2017 to December, 2017)

Abstract

Brazil is one of the largest agricultural producers in the world and a leading producer of biofuels. However, the use of remote sensing images to provide estimates of crop yield is still limited. This is due to the limitations of current data analysis methods, which focuses on processing a single image. The expectation of this project is that methods of "big analytics data" can significantly improve the use of satellite image in the generation of information about crop yield in Brazil. This project will focus in the specification and validation activities on methods for agricultural monitoring using big Earth observation data. These methods should be based on analysis of satellite image time series. The tasks to be performed are: (a) Detection of planted area of soybeans, maize and sugarcane, rice and wheat crops in selected areas, using methods that process large scale satellite image time series; (b) Detailed assessment of big Earth observation data analytics for agricultural mapping. In this project will development analytical methods for detecting large agricultural areas in Brasil, with the specific tasks of mapping land cover associated to soybeans, maize and sugarcane. The project results will be compare with ground truth data, that will be acquired, and with results from IBGE (Brazil's Census Bureau). The methods for agricultural monitoring should be developed in the R language and work with data stored in the SciDB array database.

Contents

1. Summary of the activities developed of the period

The project entitled "Use of Remote Sensing Time Series for Brazilian Agriculture Monitoring" is part of the thematic project "E-Sensing: Big Earth observation data analytics for land use and land cover change information" (FAPESP grant 2014/08398-6), coordinated by Prof. Dr. Gilberto Camara.

This project, whose purpose is to specify and validate activities on methods for agricultural monitoring using big Earth observation data. These methods should be based on analysis of satellite image time series.

In the first year of this project the main activities developed were:

(1) A bibliographical survey about the state of the art regarding the theme of the project and the main problems to be solved (Chapter 2);

(2) The development analytical methods for detecting large agricultural area, with the specific tasks of mapping land cover associated to soybeans, maize, cotton, millet and sunflower (Chapter 3);

(3) Comparison of our results with ground truth data and with IBGE statics (Chapter 4);

(4) The creation of an article explaining the main results achieved in this first year, participation in scientific events, and contribution in other papers (Chapter 5)

These activities are detailed in the next chapters.

2. Introduction

Since the 1980s, Brazil has become one of the world's largest agricultural exporters. Brazil is the world's largest producer of sugarcane, coffee, orange juice, and the second producer of soybeans, beef and chicken meat. Brazilian crop and livestock producers face a major challenge. While producing food for a growing world demand, Brazilian agriculture has to contribute to the country's commitments to reduce its deforestation rates and GHG emissions (Garnett, 2015). In the next decades Brazil needs to balance economic gains with sustainable practices in agriculture. To achieve this aim, Brazilian needs to increase its agricultural productivity, using cultivated land in more productive way (Nepstad et al., 2014).

Brazil's federal government has acted to reduce deforestation and resulting emissions. Combining rapid assessment of new forest cuts with strong law enforcement, Brazil cut tropical deforestation by 80% from 2005 until 2010 (Assunção et al., 2015). These initiatives were complemented by actions from the private sector, such as the Soy Moratorium. The moratorium is an agreement signed by the major soybean traders pledging not to buy soy grown in Amazon forest areas cleared after July 2008. During 2004 and 2005, 30% of soy expansion in this region occurred through deforestation.

In 2014, only 1% of the new soy expansion in the Amazon biome resulted from direct conversion from forest (Gibbs et al., 2015). Despite these advances, the environmental impacts of crop production and cattle-raising in Brazil's Amazonia and Cerrado biomes continue to raise concerns (Nepstad et al., 2014). To develop adequate public policies that balance production with protection, Brazil needs comprehensive information on land change dynamics.

Previous studies in Brazilian agricultural dynamics have focused in the state of Mato Grosso, one of the world's fast moving agricultural frontiers. Spera et al. (2014) use satellite remote sensing to examine patterns of cropland expansion in Mato Grosso from 2001 to 2011. They use the MODIS EVI time series, coupled with a decision-tree algorithm. Data from crop specific growing season lengths and maximum EVI thresholds was used to classify large-scale croplands in five classes: soy, cotton, soy-maize, soy-cotton, and irrigated. The paper describes how crop expansion depends on land attributes such as soil, climate and topography. The authors found out that most suitable areas for cropland expansion in Mato

Grosso had been occupied by 2006. As a consequence, farmers increased double cropping systems to make up for the scarcity of high quality remaining agricultural land. Since the paper deals on how land quality affects farmers' decision-making, it does not include accuracy assessments of the classification results.

Arvor et al. (2011) use MODIS EVI time series to identify five crop classes: soybean, maize and cotton crops planted in single or double cropping systems. They assume that maize is only planted in consortium with soybeans. The authors collected ground data sets in 50 farms in Mato Grosso on 2005–2006 and 2006-2007. The study uses a two-step classification method, first creating a cropland mask and then discriminating the crop varieties of interest inside the mask. To create the mask, they assumed that crop EVI profiles are identifiable as having one of two cycles with high maximum values and low minimum values. To classify crop types inside the cropland mask, they use the Jeffries–Matsushita (JM) distance to rank the 23 dates of each MODIS EVI series. The authors use the best subset of these dates as inputs to a supervised classifier, followed by post-processing using segmentation to produce more homogeneous results. Reported accuracy is 85% for the agricultural mask and 74% for the crop classification, using validation data not included in the training set.

To describe the spatial dynamics of crop production in Mato Grosso from 2001 to 2014, Kastens et al. (2017) use MODIS NDVI time series. They take ground reference data from 2009 to 2016 to train and validate a random forest classification model. Reported accuracy was 79% for distinguishing five crop classes (soybean-fallow, fallow-cotton, soybean-cotton, and soybean crop). The soybean-crop class includes maize, millet, sorghum and sunflower, which the authors stated they could not distinguish well.

Studies covering the whole Amazonia biome focus on deforestation and its relation to pasturelands. The PRODES system by Brazil's National Institute for Space Research (INPE) maps clear cuts in the Amazon forest yearly, producing a forest/non-forest mask (Hansen et al., 2008). Hansen et al. (2013) produce global maps of forest cover change using LANDSAT-class data. Parente et al. (2017) present maps of pastureland areas in Brazil using LANDSAT-8 images. INPE, together with the Brazilian Agriculture Research Corporation (EMBRAPA), produced TerraClass, a map of land cover change in the Amazonia biome (Almeida et al., 2016). TerraClass produces a cropland mask, and does not distinguish between different crops. These efforts are relevant and produce important data sets, but none provides a complete assessment of land cover change and its relation to different cropping systems.

There are no previous works in the literature that map both the dynamics of crop expansion and the land changes due to pasture expansion in Brazil's agricultural frontiers. To address this challenge, we have developed new methods to produce consistent multi-year maps of the different types of land cover in Brazil. These maps provide information on crop production systems and pasture expansion into natural vegetation. The results enable an informed assessment of the interplay between production and protection in the Brazilian Amazonian and Cerrado biomes.

This project proposes innovative methods for using satellite image time series to produce land use and land cover classification over large areas in Brazil. Using the full depth of the MODIS time series data to classify natural and human-transformed land areas in state of Mato Grosso, Brazil's agricultural frontier. Our hypothesis is that building high dimensional spaces using all values of the time series, coupled with advanced statistical learning methods, is a robust and efficient approach for land cover classification of large data sets.

The method improves on work by Arvor et al. (2011), Spera et al. (2014), and Kastens et al. (2017), by including non-crop cover types and providing a more detailed distinction between crops.

3. Material and Method

3.1. Study area

Mato Grosso (MT) has 903,357 km2 of extension, being the third largest state of Brazil. It includes three of Brazil's biomes: Amazonia, Cerrado and Pantanal. The Cerrado biome covers 40% of the whole territory and is a important biome related a animals species (more than 1,500 species), birds (837 species), amphibians (150 species) and reptiles (120 species). The Pantanal, which occupies 7% of the state, is a bio-diversity rich biome, and is an UNESCO World Natural Heritage and Biosphere Reserve. In the Amazonia biome in Mato Grosso there are two types of forest: the Amazon Forest and the Seasonal Forest, which together occupy about 53% of the territory of Mato Grosso.

3.2. Data

We used the MOD13Q1 product from NASA from 2001 to 2016, provided every 16 days at 250-meter spatial resolution in the Sinusoidal projection (Didan, 2015)[1]. To do the analysis, we selected the Normalized Difference Vegetation Index (NDVI) and the Enhanced Vegetation Index (EVI), and the original near infrared (NIR) and middle infrared (MIR) bands. We defined nine classes: (1) forest, (2) cerrado, (3) pasture, (4) soybean-fallow (single cropping), (5) fallow-cotton (single cropping), (6) soybean-cotton (double cropping), (7) soybean-maize (double cropping), (8) soybean-millet (double cropping), (9) soybean-sunflower (double cropping). According to the Brazilian Institute of Geography and Statistics (IBGE), crop classes (4)-(9) accounted for more than 93% of MT agricultural land area in 2015. Crop and pasture ground data was collected by co-authors Coutinho, Esquerdo and Antunes through farmer interviews in October 2009 and in October 2013. Samples for cerrado and forest classes were provided by co-author Bergotti. Ground samples for soybean-fallow class were provided by co-author Arvor, based on his previous work (Arvor et al., 2012). Table 1 lists the distribution of the ground samples.

Table 1: Ground samples used as training data for Mato Grosso.

[1] Since the 2004 MODIS image presented high amount of noise in the MIR band, results from 2004 were not used in the analysis.

Class label	Count	Freq
Cerrado	400	18.90%
Fallow-Cotton	34	1.60%
Forest	138	6.50%
Pasture	370	17.50%
Soy-Maize	398	18.80%
Soy-Cotton	399	18.90%
Soy-Fallow	88	4.20%
Soy-Millet	235	11.10%
Soy-Sunflower	53	2.50%

To get an overall view of the temporal signatures of the ground samples, we use a generalized additive model (GAM) to estimate the joint distribution the set of samples for each class (Maus et al., 2016). The GAM estimates use a smoothing function that approximates the idealized temporal patterns. One can observe that the temporal signatures of classes soy-maize, soy-millet and soy-sunflower are similar, leading to some possible confusion. As our experiments show, these are the classes which are harder to distinguish.

3.3. Method

3.3.1 Combining satellite image time series with statistical learning methods

This work combines SITS with statistical learning. 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.

There has been much recent interest in using classifiers such as support vector machines (Mountrakis et al., 2011) and random forests (Belgiu and Dragut, 2016). 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 (Brown et al., 2013; Kastens et al., 2017). Then, using software such as TIMESAT (Jönsson and Eklundh, 2004), they derive a small set of phenological parameters from vegetation indexes, like beginning, peak, and length of growing season (Estel et al., 2015; Pelletier et al., 2016).

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). 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 project, 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 Maus et al. (2016), 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. 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.

As an example, Figure 1 shows the plot of the NDVI values of 370 time series for land cover class ”Pasture”, based on ground samples. Each thin line is one time series. The darker lines are the median and first and third quartile values. By visualizing the data, the challenge of distinguishing noise from natural variation becomes clear. The data shows natural variability due to different climate regimes and shows noise associated to cloud cover. To avoid losing information, we use the raw data to train a support vector machine, a classifier which is robust to noisy data sets (Hastie et al., 2009).

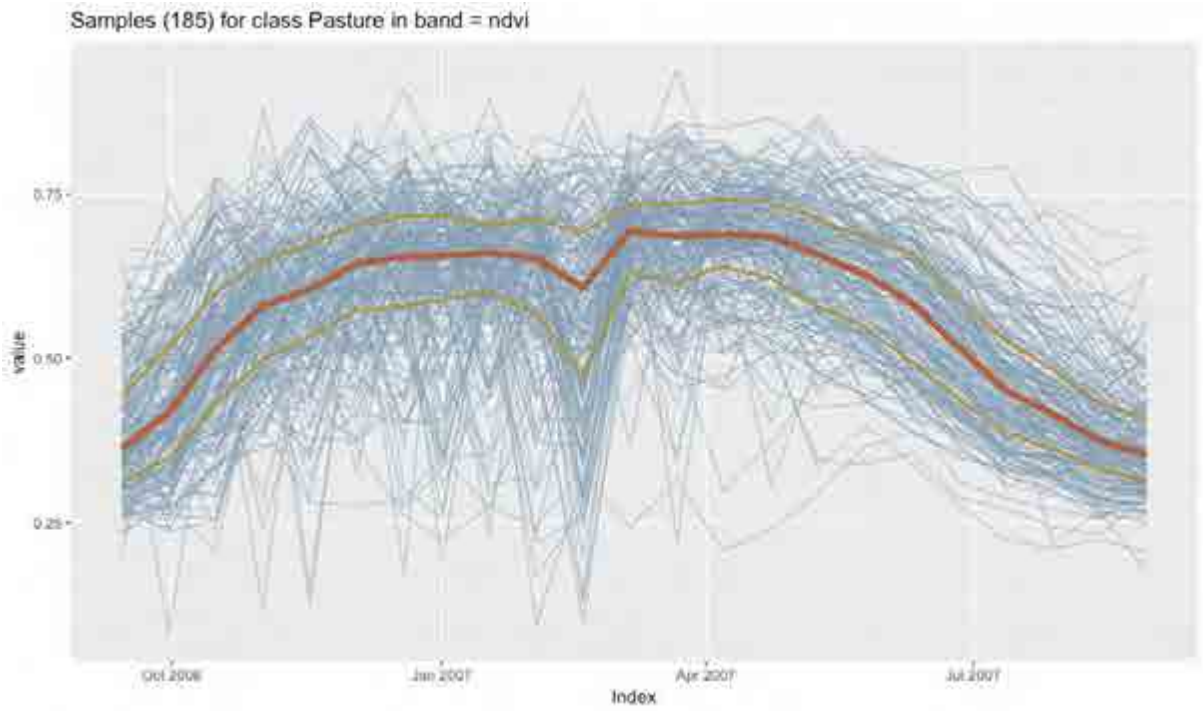


Figure 1: Time series of 370 ground samples for land cover class ”Pasture” in the state of Mato Grosso, Brazil (source: authors).

The support vector machine is a classifier which considers that the boundary between two classes is non-linear. In its simplest form, an SVM implements a linear classifier by defining boundaries in an n-dimensional space to distinguish two classes. SVMs build hyperplanes that represent the largest separation between the two classes. The hyperplanes maximize the distance from them to the nearest data point on each side. The training samples that define the hyperplane of maximum margin are called support vectors. There are many cases where the classes cannot be correctly distinguished by linear hyperplanes. In these situations, the SVM algorithm uses non-linear mappings to project the input vectors to a very high-dimension feature space. In this new feature space, the SVM builds a linear decision surface (Cortes and Vapnik, 1995). SVM implementations include polynomial and radial kernels to deal with non-linear class boundaries. In the case of noisy satellite image time series, we found out that using an SVM with radial kernels improves classification accuracy relative to the simpler linear kernel.

3.3.2 Computational infrastructure

Progress on big EO data analytics depends on researchers developing and sharing new methods. Thus, an architecture for big EO data analytics should meet the needs of the researchers. Results should be shared with the scientific community, enabling collaborative work. Researchers should be able to replicate best practices and build their own infrastructure. To achieve these goals, our architecture uses the following building blocks:

1. The SciDB open source array database (Stonebraker et al., 2013) that allows easy mapping of big EO data to its data structure.

2. R as the tool for big data analytics, so that researchers can thus scale up their methods, reuse previous work, and collaborate with their peers.

3. The R packages SITS (Simoes et al., 2017) and dtwSat (Maus et al., 2017), for big EO analytics on satellite image time series.

4. A set of web services for big EO data, adapted to the needs of satellite image time series (Vinhas et al., 2016).

Array databases split large volumes of data in distributed servers in a “shared nothing” way. A big array is broken into “chunks” that are distributed among different servers. Array DBMS such as SciDB (Stonebraker et al., 2013) reduce the impedance mismatch between the data model (raster), the storage model (arrays) and the analysis functions. Entire collections of image data fit into single spatiotemporal arrays. Using array DBMS with statistical computing is a natural solution for EO applications. SciDB has an R interface that allows researchers to parallelising complex analysis and run algorithms on large remote sensing data sets (Figure 2). This solution is a suitable compromise between the needs for massive paralel data processing and maximum flexibility in algorithm design.

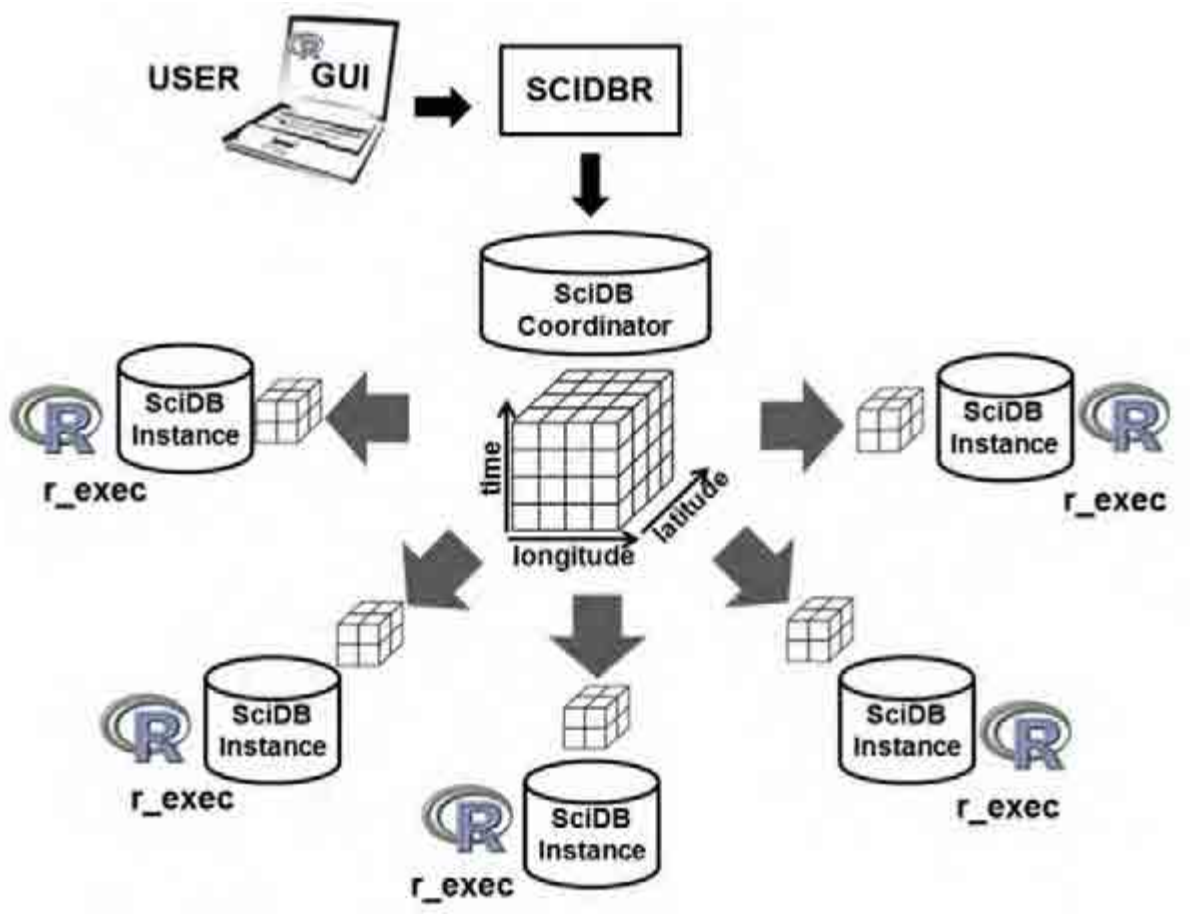


Figure 2: Remote execution of R scripts in SciDB

In terms of hardware, our architecture has 5 servers with 2 CPUs with 6-cores each, operating at 2.4GHz with a 15MB cache. Each server has 96 GB of RAM, and 16 TB of data storage. This gives 60 cores that can work in parallel in a “shared-nothing” data storage. The array database SciDB includes the full set of MODIS MOD09Q1 images at 250 meter resolution for South America, with 13,800 images associated to 317 billion data series. The case study described in the project covers the state of Mato Grosso, Brazil, an area of 900,000 km2, which has about 20 billion measures. The full processing of all time series to classify 16 years of data in Mato Grosso takes about 6 hours using the R-SciDB interface. Given these results, we argue that using SciDB combined with R is an adequate solution for big Earth Observation data analytics.

3.4. Post-processing Masks

We applied three masks to the final classified maps. The sugarcane masks from 2003 to 2016 come from the Canasat project (www.dsr.inpe.br/canasat/). This project maps sugarcane areas in the South-Central region of Brazil using LANDSAT images (Adami et al., 2012). Sparovek et al. (2015) provided the urban area mask. The water mask comes from Pekel et al. (2016), who used three million LANDSAT satellite images to quantify changes in global surface water over the past 32 years (1984 to 2015).

4. Results and Discussion

To estimate the classification accuracy, we ran a 5-fold cross-validation procedure (Wiens et al., 2008). In this validation, we run 5 different assessments. For each assessment, 80% of the samples are used for training and 20% for prediction. The accuracy of all 5 classifications is averaged to produce a single estimation. Using a 5-fold validation has some advantages to other validation methods. The goal of cross-validation is to find out how well a given statistical learning procedure can be expected to perform on independent data (James et al., 2013). Increasing the number of folds reduces the bias of the estimate of the model performance on independent data, at the cost of increasing its variance. Given the number of samples of each class (see Table 1), we consider that a 5-fold cross validation is adequate for our training set.

The 5-fold cross validation estimates an overall accuracy of 94% and the Kappa index was 0.92. Producer's and user's accuracies of all classes were close to or better than 90% (Table 2). This confirms the applicability of the proposed method in classify agricultural areas. As expected, the matrix shows some confusion between the classes' soybean-maize and soybean-millet. Since maize and millet have similar physical characteristics, they can be spectrally confused (Figure 5). Both are grasses, with lanceolate leaves; the height of maize can reach up to 3.5 meters, while millet varies between 1.5 and 3 meters, and can reach more than 5 meters. In general, results show a good discrimination between different crops, which improves on previous work (Kastens et al., 2017; Macedo et al., 2012; Arvor et al., 2012, 2011).

Measured deforestation in Mato Grosso from 2005 to 2016 was 4.1 million hectares, a decrease of 12% of the total forest area, considering both Amazonia and Cerrado biomes. The areas classified as forest were compared with the Hansen et al. (2013) mapping for the year 2000. These authors used LANDSAT images to map the percent of tree crown cover densities. Trees were defined as all vegetation taller than 5 meters height. In order to separate the forest areas, we took the areas with more than 25% tree cover from the Hansen map. We found out that 99% of the pixels classified as forest match the pixels indicated by Hansen et al. (2013) as having more than 25% tree cover. For the cerrado class, 62% of the pixels match the pixels indicated by Hansen et al. (2013) as having more than 25% tree cover. This

difference occurs because our cerrado class includes both wooded and wooded-herbaceous physiognomies.

Table 2: Confusion matrix of MODIS time series images, obtained by 5-fold cross validation of classification of field data, and values of producer's accuracy (PA) and user's accuracy (UA) for each class.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	UA
1 Cerrado	393	0	0	12	0	0	0	0	0	0.97
2 Fallow-Cotton	0	33	0	0	1	2	0	0	0	0.92
3 Forest	1	0	136	0	0	0	0	0	0	0.99
4 Pasture	6	0	1	357	3	1	0	5	0	0.96
5 Soy-Maize	0	1	1	1	352	18	0	26	4	0.87
6 Soy-Cotton	0	0	0	0	13	376	0	4	0	0.96
7 Soy-Fallow	0	0	0	0	0	0	88	0	0	1
8 Soy-Millet	0	0	0	0	25	2	0	199	2	0.87
9 Soy-Sunflower	0	0	0	0	4	0	0	1	47	0.9
PA	**0.98**	**0.97**	**0.99**	**0.96**	**0.88**	**0.94**	**1**	**0.85**	**0.89**	

The pixels labelled as pasture were compared to the pasture mapping done by Parente et al. (2017), who produced a pasture mask for Brazil in 2015 using LANDSAT-8 images and a random forest classifier. The difference between the total pasture area in our work and that mapped by Parente et al. (2017) for the state of Mato Grosso was 4%. Correlation between the individual pasture pixels in both works was 89%. Part of this difference can be explained by the fact that the map by Parente et al. (2017) uses additional masks to exclude indigenous areas and national parks. An additional factor is that Parente et al. (2017) use LANDSAT images, while we use MODIS. Note that user's accuracy of the pasture class on the map made by Parente et al. (2017) is 83%, while user's accuracy of the pasture class of SVM classification is 96%. Further detailed studies are required to assess the quality of these approaches and to improve pasture assessments in the Amazonia and Cerrado biomes.

Figure 3 shows two of the resulting maps with the spatial distribution of land cover classes, for the years 2005 and 2016. The full data set, including all resulting maps and the ground sample data, as well as the software used to produce the maps, are openly available in the internet.

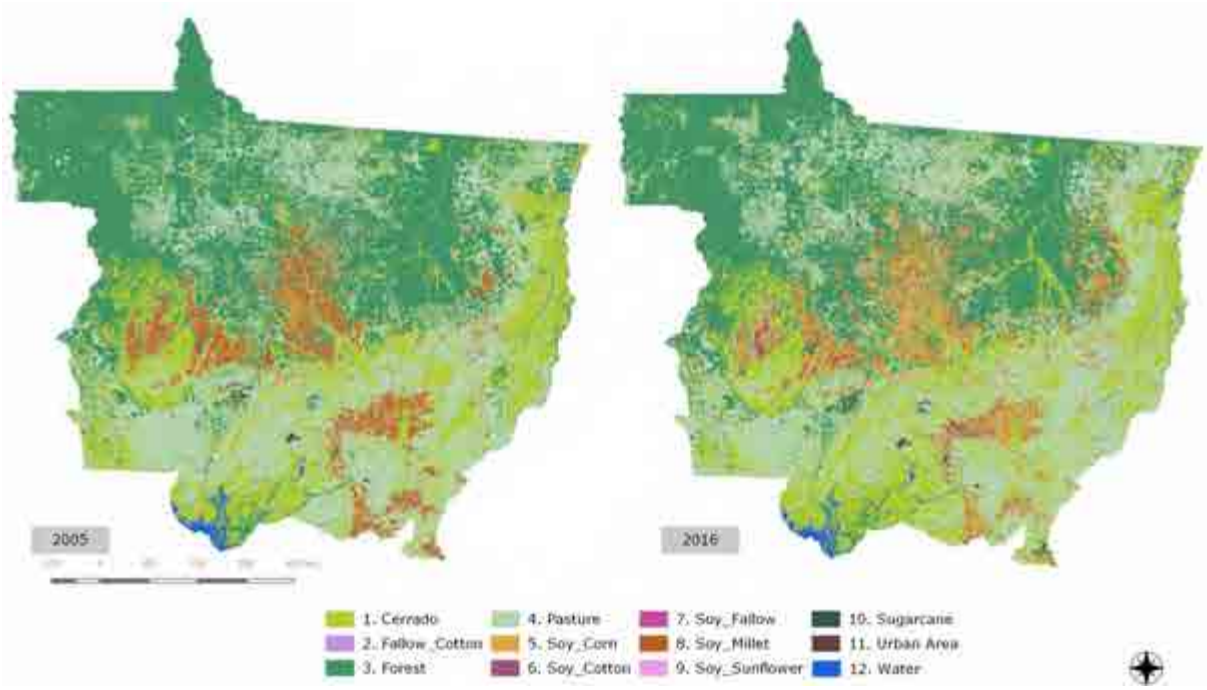


Figure 3: SVM classification for state of Mato Grosso in 2005 and 2016, with sugarcane, urban area and water masks.

We compared our crop classification to IBGE official crop statistics (IBGE, 2017). IBGE hold yearly sample surveys of agricultural production at the municipal level, the so-called PAM ("Pesquisa Agrícola Municipal"). At state level, the soybean, cotton, maize and sunflower areas mapped by our work had a correlation of 98%, 96%, 73%, and 80%, respectively, to the state level results of the IBGE PAM (Figure 4). Compared to the IBGE PAM, the classification overestimated the soybean and the maize areas, and underestimated cotton and sunflower areas. These differences may have been caused by the spatial resolution of the MODIS images (250 meters), which generates spectral mixing due to different land uses within a single pixel (Friedl et al., 2002). However, the lack of a reliable reference data set precludes an objective assessment. The IBGE PAM results are based on samples and not on surveys. Thus, they contain uncertainties as well, and should not take as absolute references. To produce the PAM, IBGE staff do not go to the field.

They contact large producers and also rely on subjective estimates of the local IBGE staff. Therefore, comparing our results with data from the PAM does not entail an accuracy estimate of our work. Correlation between the sum of agricultural areas classified in this study and the estimates by IBGE for the harvests from 2005 to 2016, are equal to 98%. Thus, we

consider that the proposed methodology is effective for mapping agricultural crops in Mato Grosso.

In Mato Grosso, the cropland area increased by 1.83 million hectares (26.5%) from 2005 to 2016. The greatest expansion occurred between 2007- 2008 and 2012-2013, with a growth rate of 11.9% and 11.6%, respectively. The expansion of agricultural areas occurred mainly around the BR163 highway, which has its starting point in the center of the state and goes as far north as Mato Grosso, as if dividing the state in half. At the edge of this road it is possible to observe the expansion of agriculture in the northern direction on the Amazonia biome. Municipalities such as Querencia and Tabaporã, where there was almost no presence of agriculture in the early 2000s, today are expressive producers of soybeans. Arvor et al. (2012) also observed the same expansion trend around the BR163 highway. This area has the highest soybean yields in Mato Grosso due to its soil, topography and climate (Spera et al., 2014). It is area is also the one with largest proportion of double cropping due to a longer rainy season (Arvor et al., 2013) Furthermore, the Brazilian government the plans to asphalt the BR163 highway until it connects with the Mirituba and Santarem harbors in the state of Para on the Amazon river. This would decrease the transportation costs for soybean exports

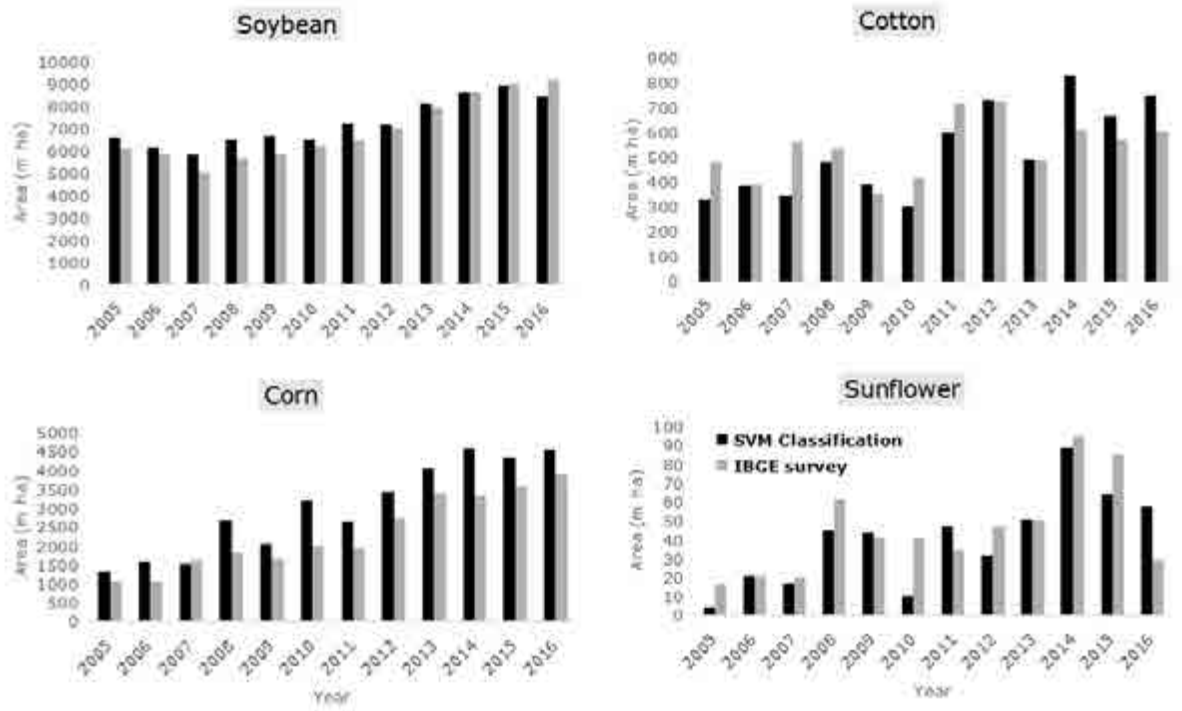


Figure 4: Total area of soybean, cotton, maize and sunflower in state of Mato Grosso estimated by SVM classification and the IBGE cropland survey.

The soybean class had also decreased in area from 2005 to 2006 and from 2006 to 2007 due to the economic crisis, when the Brazilian currency was devalued compared with the US dollar. The unfavorable exchange rate affected soybean production from 2005 to 2007 (Arvor et al., 2012). However, soybeans had a significant increase of 0.93 million hectares (12.9%) between 2012 and 2013. According to the Brazilian National Supply Company (CONAB, July 2013), this growth is due to better prices for soybean in the international market, and its repercussions in the domestic market. New commercial arrangements, such as advance commercialization, also contributed to this increase.

Due to the increased demand for food and biofuels, producers in the state of Mato Grosso intensified agricultural production by adopting double cropping systems. Area cultivated with double cropping systems, involving soybeans (first cycle) + some other crop (second cycle) or some other crop (first cycle) + cotton (second cycle), increased from 6.58 to 8.43 million hectares during 2005 to 2016, an increase of 28%. Double-cropping systems are currently predominant in Mato Grosso. Maize area also grew due by replacing millet as the crop of choice for planting in consortium with soybeans. Millet lost an area of 1.87 million hectares (61%) to maize. Due to improvements in maize varieties, and the increase in Brazil's maize exports, maize replaced millet as a more profitable option. In the municipality of Campo Verde (located in the cerrado biome, in the southeast of the state of Mato Grosso), it is possible to observe this transition from single to double crop systems, with the replacement of fallow cotton by soybean cotton, from 2005 to 2015 (Figure 5).

The double cropping system is more profitable for the producer, and represents a better use of agricultural areas, allowing the increase of production at the same time that reduces the pressure of expansion over native vegetation. Double cropping also enables to adopt no tillage practice (apart for cotton) which is better from an ecological point of view. Previous authors (Kastens et al., 2017; Spera et al., 2014; Arvor et al., 2012) had already pointed out the increase in double cropping production associated with soybeans.

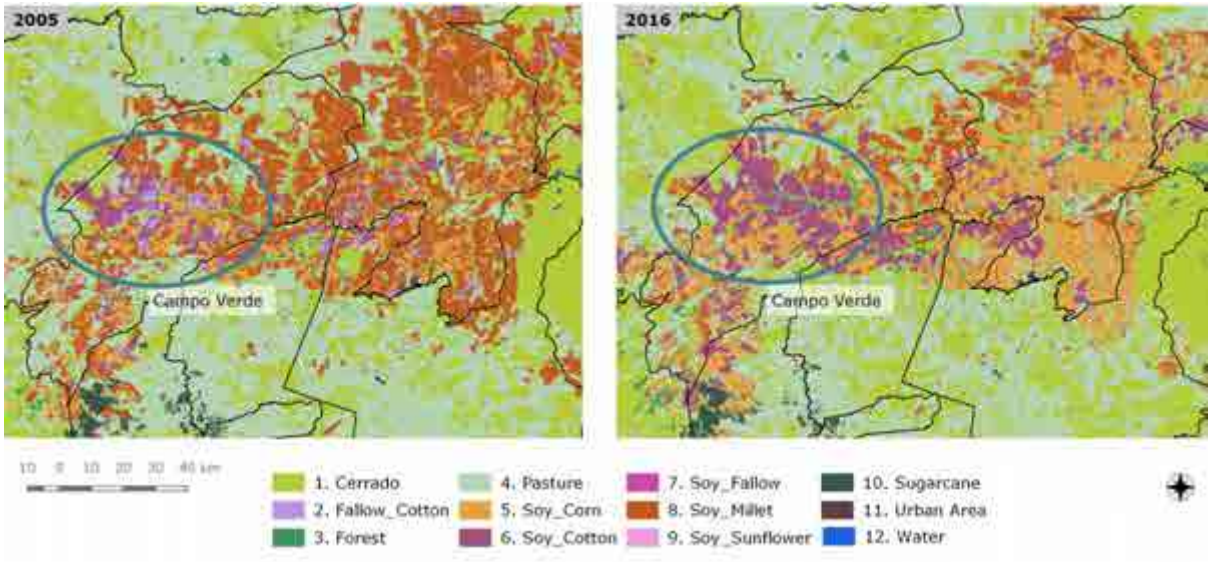


Figure 5: Change in the production system in the municipality of Campo Verde - MT from single crop in 2005 to double crop in 2016. In the gray highlight it is possible to observe the change from the fallow cotton class in 2005 to the soybean cotton class in 2016.

Pasture area in Mato Grosso between 2005 and 2015 declined 4.6 million hectares, from 28.1 to 23.5 million hectares. Our results for 2016 point to a total to 28.9 million hectares. We consider this result to be an outlier that needs to be checked by producing the 2017 estimates in due time. According to IBGE, the number of cattle heads in the state has increased from 26.7 in 2005 to 29.3 million in 2015, a growth of 10% (IBGE, 2017). In Figure 6, we show that the stocking rate in Mato Grosso has grown steadily. The cattle heads grew by 10%, while pasture decreased by 16% between 2005 and 2015. In general, there is a trend towards pasture intensification coupled with abandonment of frontier areas, especially those at the most Northern part of the state.

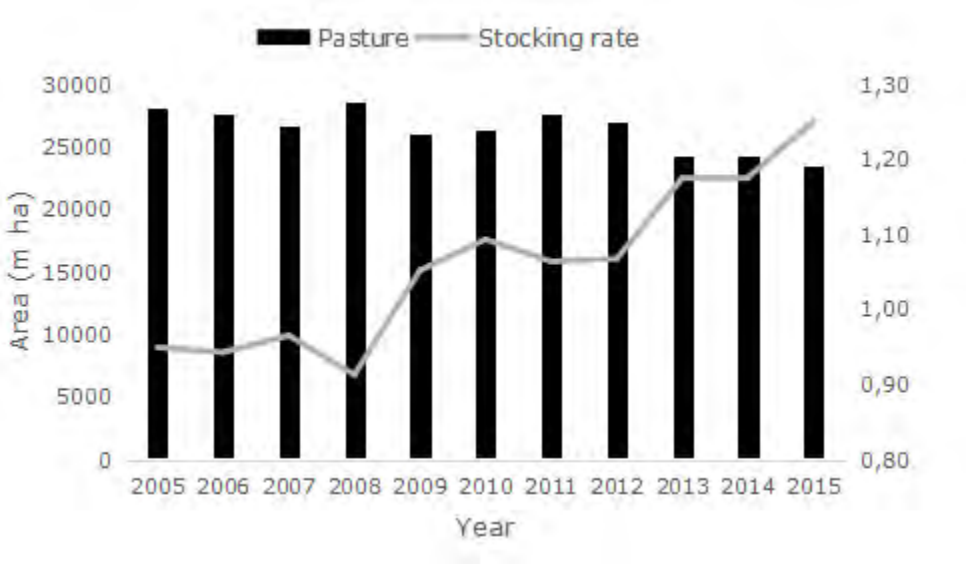


Figure 6: Pasture area, provide by SVM classification, and stoking rate, calculate by IBGE cattle head value and pasture areas, in Mato Grosso state, from 2005 to 2016.

The results point out to important trends in agricultural intensificationin Mato Grosso. Double cropping systems are now the most common production system in the state, thus increasing the potential for land sparing. As pointed out by other authors (Spera et al., 2014; Gibbs et al., 2015; Kastens et al., 2017) the impact of crop production in deforestation has decreased since 2005. Arguably, this is due to a combination of factors, including the Soy Moratorium, increased law enforcement, and the occupation of the best farming areas in the state's Amazonia biome (Spera et al., 2014). A less studied issue is the increase in pasture productivity. Pasture expansion and intensification has been less studied than crop expansion, although it has a stronger impact on deforestation and GHG emissions. Our data points to a significant increase in stocking rate in Mato Grosso, and to the possible abandonment of pasture areas opened in the state's frontier. Further studies, that couple fieldwork, mapping and economic models, are required for better understanding of the underlying driving forces for the cattle-growing sector.

Our results point out the conflicting forces at play in the agricultural expansion in Mato Grosso. In some segments (such as crop production), there is a consolidation in place. The best producing areas have been occupied, and emphasis now is on increasing productivity by adoption of double-cropping systems. In the case of cattle-raising, one can see mixed signs.

On one hand, there is a modest, but significant, increase in stocking rate. However, there is still expansion going on in the Northern frontiers of the state, which need to be better

studied. Many factors could be at play, including land speculation, and indirect land use due to crop expansion. This situation poses important challenges. The large scale mapping that we produced for Mato Grosso needs to be expanded to the whole Amazonia and Cerrado biomes, and also needs to be supported by economic analysis. There is a need for continuous improvement of land cover classification using remote sensing time series, by using LANDSAT-class satellites to increase spatial resolution and classification accuracy.

5. Publications Submitted and Participation in scientific events

Papers submitted to international journals

I. **Michelle Picoli**, Gilberto Camara, Ieda Sanches, Rolf Simões, Alexandre Carvalho, Adeline Maciel, Alexandre Coutinho, Julio Esquerdo, João Antunes, Rodrigo Begotti, Damien Arvor, Claudio Almeida. Big Earth Observation Time Series Analysis for Monitoring Brazilian Agriculture. Submitted to ISPRS Journal of Photogrammetry and Remote Sensing (under review).

II. Adeline Maciel, Gilberto Câmara, Lúbia Vinhas, **Michelle Picoli**, Rodrigo Begotti, Luiz Assis. Spatiotemporal interval logic for reasoning about land use change dynamics. Submitted to Inter. Journal of Geographical Information Science (2nd revision).

Data sets submitted to public repositories

I. Gilberto Camara, **Michelle Picoli**, Rolf Simoes, Adeline Maciel, Alexandre Carvalho, Alexandre Coutinho, Julio Esquerdo, João Antunes, Rodrigo Begotti, Damien Arvor (2017): Land cover change maps for Mato Grosso State in Brazil: 2001-2016, links to files. PANGAEA, https://doi.org/10.1594/PANGAEA.881291

Papers submitted to scientific conferences

I. Adeline Maciel, Lubia Vinhas, Gilberto Camara, **Michelle Picoli**, Rodrigo Begotti. An interval-based approach for reasoning about land use change trajectories. Submitted to International Geoscience and Remote Sensing Symposium, IGARSS 2018.

Participation in scientific events

I. 18o. Simpósio Brasileiro de Sensoriamento Remoto, SBSR 2017. Santos - SP, 28-31 May 2017.

II. 5th Sampling and Research Methodology School and 4th International Workshop on Surveys for Evaluation of Public Policies, V ESAMP e IV WIPAPP. Cuiabá-MT, 17-20 October 2017.

References

Adami, M., Rudorff, B. F. T., Freitas, R. M., Aguiar, D. A., Sugawara, L. M., Mello, M. P., 2012. Remote sensing time series to evaluate direct land use change of recent expanded sugarcane crop in Brazil. Sustainability 4 (4),574–585.

Almeida, C., Coutinho, A., Esquerdo, J., Adami, M., Venturieri, A., Diniz, C., Dessay, N., Durieux, L., Gomes, A., 09 2016. High spatial resolution land use and land cover mapping of the Brazilian Legal Amazon in 2008 using Landsat-5/TM and MODIS data. Acta Amazonica 46, 291 – 302.

Arvor, D., Dubreuil, V., Simões, M., Bégué, A., 2013. Mapping and spatial analysis of the soybean agricultural frontier in mato grosso, Brazil, using remote sensing data. GeoJournal 78 (5), 833–850.

Arvor, D., Jonathan, M., Meirelles, M., Dubreuil, V., Durieux, L., 2011. Classification of MODIS EVI time series for crop mapping in the state of Mato Grosso, Brazil. International Journal of Remote Sensing 32 (22), 7847–7871.

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

Assunção, J., Gandour, C., Rocha, R., 2015. Deforestation slowdown in the brazilian amazon: Prices or policies? Environment and Development Economics 20 (6), 697–722.

Belgiu, M., Dragut, L., 2016. Random forest in remote sensing: A review of applications and future directions. ISPRS Journal of Photogrammetry and Remote Sensing 114, 24–31.

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.

Câmara, G., Assis, L. F., Ribeiro, G., Ferreira, K. R., Llapa, E., Vinhas, L., 2016. Big earth observation data analytics: matching requirements to system architectures. In: Proceedings of the 5th ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data. ACM, Burlingname, CA, USA, pp. 1–6.

CONAB, July 2013. Follow up of brazilian crop: grains (ninth survey). Tech. rep., National Supply Company (CONAB).

Cortes, C., Vapnik, V., 1995. Support-vector networks. Machine Learning 20 (3), 273–297.

Didan, K., 2015. MOD13Q1 MODIS/Terra Vegetation Indices 16-Day L3 Global 250m SIN Grid V006. NASA EOSDIS Land Processes DAAC. Available at: https://doi.org/10.5067/modis/mod13q1.006.

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.

Friedl, M., McIver, D., Hodges, J., Zhang, X., Muchoney, D., Strahler, A., Woodcock, C., Gopal, S., Schneider, A., Cooper, A., Baccini, A., Gao, F., Schaaf, C., 2002. Global land cover mapping from MODIS: algorithms and early results. Remote Sensing of Environment 83 (1–2), 287 – 302.

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.

Garnett, T., 2015. Where are the best opportunities for reducing greenhouse gas emissions in the food system (including the food chain)? Food Policy 36, S23–S32.

Gibbs, H. K., Rausch, L., Munger, J., et al., 2015. Brazil's soy moratorium. Science 347 (6220), 377–378.

Gibbs, H. K., Ruesch, A. S., Achard, F., Clayton, M. K., Holmgren, P., Ramankutty, N., Foley, J. A., 2010. Tropical forests were the primary sources of new agricultural land in the 1980s and 1990s. Proceedings of the National Academy of Sciences 107 (38), 16732–16737.

Gomez, C., White, J. C., Wulder, M. A., 2016. Optical remotely sensed time series data for land cover classification: A review. ISPRS Journal of Photogrammetry and Remote Sensing 116, 55 – 72.

Hansen, M., Shimabukuro, Y., Potapov, P., Pittman, K., 2008. Comparing annual MODIS and PRODES forest cover change data for advancing monitoring of Brazilian forest cover. Remote Sensing of Environment 112 (10), 3784–3793.

Hansen, M. C., Potapov, P. V., Moore, R., Hancher, M., Turubanova, S. A., Tyukavina, A., Thau, D., Stehman, S. V., Goetz, S. J., Loveland, T. R., Kommareddy, A., Egorov, A., Chini, L., Justice, C. O., Townshend, J. R. G., 2013. High-resolution global maps of 21st-century forest cover change. Science 342 (6160), 850–853.

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

IBGE, 2017. Brazilian institute of geography and statistics. municipal agricultural production. Available at: http://www2.sidra.ibge.gov.br/.

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 timeseries 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., Andréfouët, S., Cohen, W. B., Gómez, 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., 2014. Bringing an ecological view of change to landsat-based remote sensing. Frontiers in Ecology and the Environment 12 (6), 339–346.

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., Helmut, G., 2006. Land-Use and Land-Cover Change. Global Change – The IGBP Series. Springer.

Macedo, M. N., DeFries, R. S., Morton, D. C., Stickler, C. M., Galford, G. L., Shimabukuro, Y. E., 2012. Decoupling of deforestation and soy production in the southern Amazon during the late 2000s. Proceedings of the National Academy of Sciences of the United States of America 109 (4), 1341–1346.

Maus, V., Camara, G., Appel, M., Pebesma, E., 2017. dtwSat: Time-Weighted Dynamic Time Warping for Satellite Image Time Series Analysis in R. Journal of Statistical Software (accepted).

Maus, V., Camara, G., Cartaxo, R., Sanchez, A., Ramos, F. M., de Queiroz, G. R., 2016. A time-weighted dynamic time warping method for land-use and land-cover mapping. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 9 (8), 3729 – 3739.

Mountrakis, G., Im, J., Ogole, C., 2011. Support vector machines in remote sensing: A review. ISPRS Journal of Photogrammetry and Remote Sensing 66 (3), 247–259.

Nepstad, D., McGrath, D., Stickler, C., Alencar, A., Azevedo, A., Swette, B., Bezerra, T., DiGiano, M., Shimada, J., Seroa da Motta, R., Armijo, E., Castello, L., Brando, P., Hansen, M. C., McGrath-Horn, M., Carvalho, O., Hess, L., 2014. Slowing amazon deforestation through public policy and interventions in beef and soy supply chains. Science 344 (6188), 1118–1123.

Parente, L., Ferreira, L., Faria, A., Nogueira, S., Araújo, F., Teixeira, L., Hagen, S., 2017. Monitoring the Brazilian pasturelands: A new mapping approach based on the Landsat 8 spectral and temporal domains. International Journal of Applied Earth Observation and Geoinformation 62, 135–143.

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.

Pekel, J. F., Cottam, A., Gorelick, N., Belward, A. S., 2016. High-resolution mapping of global surface water and its long-term changes. Nature 540, 418–422.

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.

Petitjean, F., Inglada, J., Gancarski, P., 2012. Satellite image time series analysis under time warping. IEEE Transactions on Geoscience and Remote Sensing 50 (8), 3081–3095.

Simoes, R., Camara, G., Carvalho, A., Maus, V., Assis, L., Maciel, A., Queiroz, G., 2017. SITS: Satellite Image Time Series Analysis. R package version 0.9.30. URL https://github.com/e-sensing/sits/

Sparovek, G., Barreto, A. G. O. P., Matsumoto, M., Berndes, G., 2015. Effects of governance on availability of land for agriculture and conservation in Brazil. Environmental Science & Technology 49, 10285–10293.

Spera, S. A., Cohn, A. S., VanWey, L. K., Mustard, J. F., Rudorff, B. F., Risso, J., Adami, M., 2014. Recent cropping frequency, expansion, and abandonment in Mato Grosso, Brazil had selective land characteristics. Environmental Research Letters 9 (6), 064010.

Stonebraker, M., Brown, P., Zhang, D., Becla, J., 2013. Scidb: A database management system for applications with complex analytics. Computing in Science & Engineering 15 (3), 54–62.

Verbesselt, J., Hyndman, R., Newnham, G., Culvenor, D., 2010. Detecting trend and seasonal changes in satellite image time series. Remote sensing of Environment 114 (1), 106–115.

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.

Wiens, T. S., Dale, B. C., Boyce, M. S., Kershaw, G. P., 2008. Three way kfold cross-validation of resource selection functions. Ecological Modelling 212 (3), 244–255.

National Institute for Space Research (INPE)

Image Processing Division (DPI)

Report of Post-Doc Activities

Post-Doc: Rodrigo Anzolin Begotti

FAPESP grant: 2016/16968-2

Para o período de referência do presente relatório, as atividades de pesquisa do bolsista de pós-doutorado Rodrigo Anzolin Begotti (processo FAPESP n° 16/16968-2) contemplam a validação e aprimoramento de métodos de alerta de desmatamento em "big data" para o programa DETER do INPE e a elaboração de um artigo científico. Para tal, o bolsista realizou a tarefa de qualificar parte do acervo de fotos do INPE (Fototeca e Projeto Geoma) com o objetivo de gerar amostras espectrais georeferenciadas para toda a Amazônia brasileira. As atividades relacionadas à qualificação do acervo de fotos foram descritas no relatório científico anterior.

Para analisar a detectabilidade e verificar a existência de padrões temporais relacionados ao processo de conversão da floresta em corte raso e floresta degradada, foram utilizadas séries temporais de imagens do sensor MODIS, coleção MOD13Q1 de acesso livre, disponível a partir do ano 2000. Três bandas dessa coleção em particular foram utilizadas: i) *EVI* (*Enhanced Vegetation Index*); ii) *NDVI* (*Normalized Difference Vegetation Index*); e iii) *NIR* (Infra-Vermelho Próximo). Das 7013 fotografias qualificadas que compõem o nosso conjunto de amostras, foram utilizadas 2509 amostras agrupadas em sete classes de cobertura distintas (Tabela 1). As amostras foram analisadas em ambiente *R* (http://www.r-project.org) utilizando o pacote *SITS* (*Satellite Image Time Series*; disponível em https://github.com/e-sensing/sits).

Tabela 1: Amostras espectrais utilizadas na análise da série temporal MODIS.

Classe	Número de amostras
Floresta primária	1337
Corte raso	41
Cicatriz de fogo florestal	43
Corte seletivo	130
Degradação florestal leve	165
Degradação florestal moderada	433
Degradação florestal alta	360

O intervalo das séries temporais foi estabelecido seguindo o calendário agrícola, iniciando-se no primeiro dia do mês de setembro e terminando no último dia do mês de

agosto do ano seguinte. De um modo geral, início do calendário agrícola coincide com o final da estação seca e início da estação chuvosa. Para a detecção de padrões temporais, foi utilizado o modelo *SVM* (*Support Vector Machine*), implementado no pacote *SITS*. O desempenho do classificador foi avaliado utilizando-se validação cruzada com cinco repetições. A cada repetição, o algoritmo separa aleatoriamente 20% das amostras para validação e com a parte restante dos dados, realiza os procedimentos computacionais para o aprendizado de máquina (*machine learning*). Após cinco repetições, a acurácia geral da classificação foi igual a 63%. A Tabela 2 apresenta a matriz de confusão para as classes utilizadas. Há considerável confusão entre as classes de degradação florestal e de floresta primária uma vez que proporcionalmente o número de amostras das classes de degradação florestal é menor do que a quantidade de amostras de floresta primária. Há também grande confusão entre as classes de degradação florestal e de corte seletivo. Além disso, a presença de nuvens pode ter alterado os padrões temporais.

Tabela 2: Matriz de confusão da classificação das amostras do acervo de fotos por meio do modelo *Support Vector Machine*.

	Corte raso	Cicatriz de fogo florestal	Degradação florestal alta	Degradação florestal leve	Degradação florestal moderada	Floresta primária	Corte seletivo
Corte raso	26	0	4	0	1	0	0
Cicatriz de fogo florestal	0	14	6	0	2	2	1
Degradação florestal alta	7	12	144	18	71	34	9
Degradação florestal leve	0	0	13	21	25	24	6
Degradação florestal moderada	3	4	91	46	165	64	30
Floresta primária	5	13	98	65	147	1184	52
Corte seletivo	0	0	4	15	22	16	32

Com base nesse primeiro resultado, as classes Degradação florestal alta e Degradação florestal moderada foram fundidas em uma única classe chamada Degradação florestal. Da mesma forma, a classe Degradação florestal leve foi incorporada à classe Corte seletivo. Com o novo arranjo das amostras em cinco classes o modelo *SVM* foi utilizado para uma nova classificação. A acurácia geral obtida foi de 71%. A matriz de confusão das cinco classes é descrita na Tabela 3. A confusão entre as classes Degradação florestal e Floresta primária diminuiu, ao contrário da classe Corte seletivo. A classe Cicatriz de fogo florestal em ambos os resultados apresenta grande confusão. Isso se deve à grande variação na severidade dos danos provocados pelo fogo na vegetação. Há também grande variação na ocorrência e duração dos incêndios florestais ao longo do tempo, mesmo que eles ocorram predominantemente no período seco do ano. Os próximos passos serão testar a classificação utilizando novos conjuntos de amostras em regiões específicas.

Tabela 3: Matriz de confusão da classificação das amostras agrupadas do acervo de fotos por meio do modelo *Support Vector Machine*.

	Corte raso	Degradação florestal	Cicatriz de fogo florestal	Floresta primária	Corte seletivo
Corte raso	25	4	0	1	0
Degradação florestal	10	470	14	103	94
Cicatriz de fogo florestal	0	8	16	1	0
Floresta primária	6	246	13	1190	119
Corte seletivo	0	65	0	29	82

Entre os dias 17 e 29 de setembro de 2017 o bolsista realizou trabalho de campo nos estados de Mato Grosso e Pará, na região da Rodovia BR-163 entre os municípios

de Sinop-MT e Novo Progresso-PA. O objetivo foi incorporar novas amostras àquelas já existentes e obter um novo conjunto de amostras de degradação florestal, particularmente de corte seletivo. Aproximadamente 2858 km foram percorridos de automóvel (Figura 1). Foram visitados 13 pontos de degradação florestal obtidos a partir dos dados do Programa Mapeamento da Degradação Florestal na Amazônia Brasileira (DEGRAD) ocorridos no ano de 2017. Foram visitados também 59 pontos de desmatamento ocorrido no ano 2016/2017 a partir de dados do Programa de Monitoramento da Floresta Amazônica Brasileira por Satélite (PRODES). Foram realizadas visitas a áreas de floresta pertencentes a três propriedades rurais que possuíam licença concedida para a execução de Plano de Manejo Florestal. A Figura 2 mostra a localização dos pontos de degradação e desmatamento visitados. Ocasionalmente, foram coletadas informações a respeito da localização e da cultura plantada em áreas agrícolas.

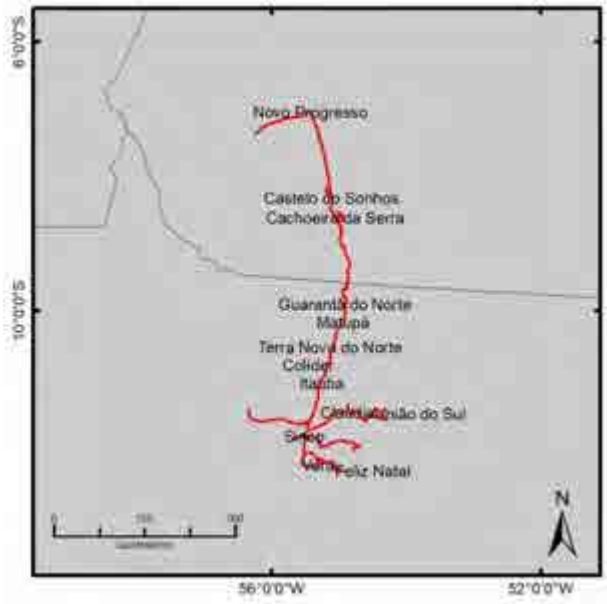


Figura 1: Trajetos rodoviários percorridos durante o trabalho de campo.

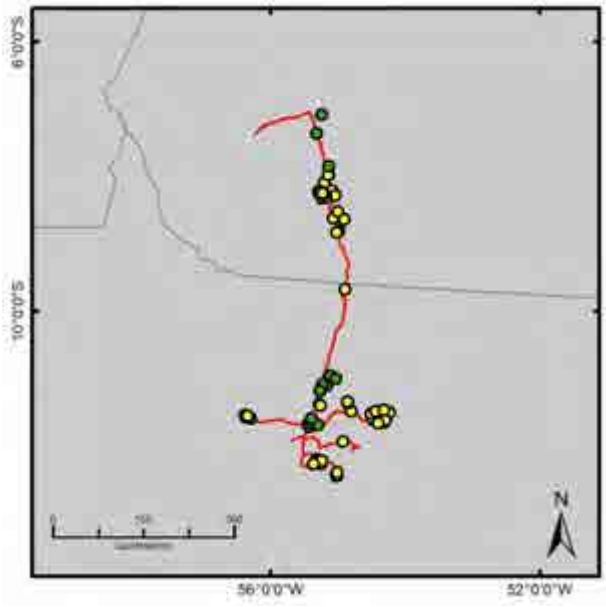


Figura 2: pontos visitados de degradação florestal (verde) e desmatamento (amarelo).

Nas áreas sob licença de Plano de Manejo Florestal, foram registradas a localização de cada UPA (Unidade de Produção Anual) e o ano em que ocorreu ou que estava planejado ocorrer a retirada de árvores de valor madeireiro. Esses dados serão utilizados para analisar de forma mais robusta o comportamento espectral da floresta submetida à degradação florestal por corte seletivo, uma vez que com esses dados validados em campo, sabe-se quando a atividade madeireira ocorreu ou irá ocorrer. Nesse último caso, a visita em campo foi importante para avaliar a estrutura da floresta antes da retirada das árvores. A Figura 3 mostra a distribuição das amostras de corte seletivo em uma das áreas visitadas de acordo com o ano no qual a degradação pela atividade ocorreu.

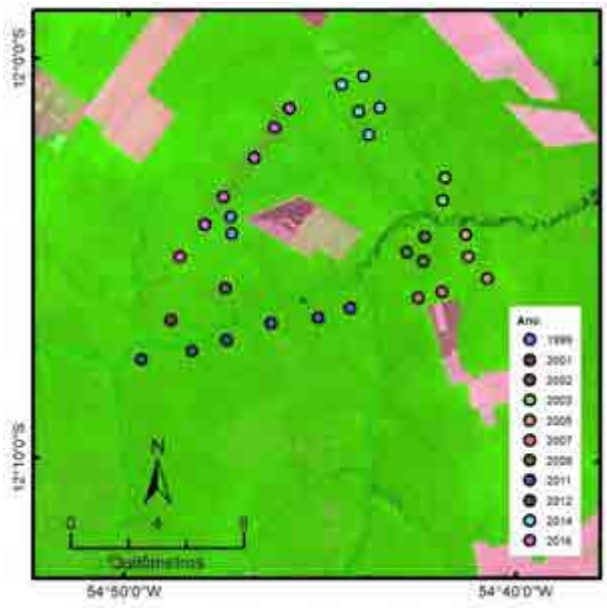


Figura 3: Degradação florestal por corte seletivo em Planos de Manejo Florestal.

O bolsista participou da elaboração como co-autor de três manuscritos e um resumo expandido submetido para congresso científico internacional. No manuscrito entitulado "An interval-based approach for reasoning about land use change trajectories", cujos autores são Adeline M. Maciel, Lubia Vinhas, Gilberto Camara, Luiz F. Assis, Michele C. A. Picoli e Rodrigo A. Begotti foi submetido para o periódico *International Journal of Geographical Information Science*. No momento estão sendo feitas as correções sugeridas pelos revisores. A contribuição do bolsista para esse manuscrito se deu na obtenção e processamento dos dados, na escrita e revisão do texto. O manuscrito entitulado "Big Earth observation time series analysis for monitoring Brazilian agriculture" foi submetido para o periodico *ISPRS Journal of Photogrammetry and Remote Sensing* e se encontra em revisão. O bolsista participou da obtenção e seleção das amostras espectrais e do processamento das imagens resultantes, além de auxiliar na elaboração e revisão do texto e das figuras que compõem o manuscrito. No momento o bolsista está participando da preparação do manuscrito entitulado "Spatio-temporal patterns of forest cover in Brazilian Amazonia: Different landscapes generated by different land use (or human occupation) contexts". Os autores são Adriana Afonso, Rodrigo A. Begotti e Maria I. S. Escada. O bolsista está contribuindo com a elaboração e revisão do texto e das figuras que compõem o

manuscrito. O objetivo é submetê-lo para o periódico *PNAS Proceedings of National Academy of Sciences of the United States of America*.

Para o evento *International Geoscience and Remote Sensing Symposium* que será realizado em julho de 2018 na cidade de Valencia, Espanha, o bolsista participou da elaboração do resumo expandido submetido ao comitê científico sob o título "An interval-based approach for reasoning about land use change trajectories". O bolsista contribuiu com a obtenção e processamento dos dados, na escrita e revisão do texto. Os autores desse trabalho são Adeline M. Maciel, Lubia Vinhas, Gilberto Camara, Michele C. A. Picoli e Rodrigo A. Begotti.

Instituto Nacional de Pesquisas Espaciais

Programa de Pós Graduação em Computação Aplicada

Relatório científico 02

FAPESP – Fundação de Amparo à Pesquisa do Estado de São Paulo

Pesquisa de doutorado:

Métodos de análise de dados espaço-temporais

Projeto

Esensing: Big Earth observation data analytics for land use and land cover change information

FAPESP Research Program in e-science - Grant 2014/08398-6

Bolsista – Rennan de Freitas Bezerra Marujo

Orientadora – Prof. Leila Maria Garcia Fonseca

Processo n. 2016/08719-2

São José dos Campos – SP

Dezembro 2017

1 Introdução

Este é o segundo relatório referente à bolsa de Doutorado fluxo contínuo - processo número 2016/08719-2 - outorgada à Rennan de Freitas Bezerra Marujo, aluno regular do Programa de Pós-Graduação em Computação Aplicada do Instituto Nacional de Pesquisas Espaciais. A bolsa financia a pesquisa “Métodos de análise de dados espaço-temporais” desenvolvida no projeto e-sensing - processo 2014/08398-6 - sob a orientação da Profa. Dra. Leila Maria Garcia Fonseca.

A Terra está em constante mudança, sendo a caracterização e mapeamento da cobertura terrestre essenciais para planejar e gerenciar seus recursos naturais. Entender os processos ativos, como o desmatamento, a expansão urbana e os fenômenos naturais, é vital para a preservação dos ecossistemas. Portanto, é importante desenvolver ferramentas capazes de detectar tais variações (KUENZER et al., 2015).

O processo de detecção de mudanças é uma tarefa difícil de realizar, sendo comumente utilizado nesta operação sensores orbitais multi-espectrais (COPPIN et al., 2004). Sensores orbitais de alta resolução espacial captam informações da superfície terrestre com mais detalhes do que os sensores de baixa resolução espacial. Entretanto, há um compromisso entre as características de resoluções temporal, espacial e radiométrica que podem limitar o desempenho do sensor em algumas aplicações (LEFSKY, COHEN, 2003). Sensores de média resolução espacial (10 a 50 metros) podem preencher a lacuna entre o detalhamento provido por imagens de alta resolução espacial e a frequência de aquisição de imagens obtidas por sensores de baixa resolução espacial (EHLERS et al., 2002).

O INPE foi pioneiro na distribuição livre de imagens orbitais de média resolução espacial de imagens com o segundo Satélite Sino-Brasileiro de Recursos Terrestres (CBERS-2) gratuitamente na internet (BANSKOTA et al., 2014). Esta politica de dados livres encorajou o Serviço Geológico dos Estados Unidos (United States Geological Survey - USGS) a disponibilizar os dados Landsat em 2008 (WOODCOCK et al., 2008, BANSKOTA et al., 2014), o que resultou numa maior quantidade de acessos e aplicação destas imagens (WULDER et al., 2012).

Os métodos para a detecção de mudança normalmente utilizam séries temporais curtas, variando de duas a cinco imagens, não utilizando o potencial completo das séries históricas (COPPIN et al., 2004). Neste contexto, as séries temporais podem fornecer observações de padrões, não encontradas em observações de data única, como tendências, periodicidades e modelos de previsão (EHLERS, 2009).

Séries temporais de imagens orbitais são um conjunto contínuo e consistente de informações sobre a Terra (EHLERS, 2009), integrando a informação espectral e espacial com a componente temporal, proporcionando informação rica para detalhar as variações espaciais ao longo do tempo (PETITJEAN et al., 2012). No entanto, a ausência de imagens de boa qualidade devido à presença de nuvens, baixa resolução temporal, bem como defeitos do sensor (por exemplo, as lacunas em imagens Landsat-7) exigem a sua correção e, em muitas aplicações, há a necessidade do uso de mais de um sensor (LEFSKY; COHEN, 2003; SHEN et al., 2016).

Atualmente, devido a maior quantidade de dados de sensoriamento remoto disponível (WULDER et al., 2012), as abordagens que utilizam imagens de múltiplas fontes tornaram-se promissoras. Isso ocorre devido à melhora dos mapeamento e do monitoramento das variáveis da vegetação ao longo do tempo quando utiliza-se aquisições mais frequentes (MOUSIVAND et al., 2015).

Por outro lado, são necessárias técnicas de processamento de imagens para unificar esses dados de forma que eles possam ser integrados em uma mesma base de dados prontos para análise (EHLERS, 2009), pois estes dados possuem diferentes resoluções espaciais, espectrais, temporais e angulares (MOUSIVAND et al., 2015).

Neste contexto, este relatório descreve as atividades acadêmicas e de pesquisa realizadas pelo bolsista desde janeiro de 2017 até dezembro de 2017, segundo ano do projeto de pesquisa, que envolvem participação em eventos, requisitos do programa de pós-graduação, pesquisa e redação de artigos científicos. Neste relatório também são apresentados o resumo do projeto de pesquisa, resultados e plano de trabalho para as próximas etapas da pesquisa.

2 Resumo do projeto de pesquisa

2.1. Resumo

O Objetivo da pesquisa é desenvolver algoritmos de análise espaço-temporal para extrair informações de grandes bancos de imagens de observação da Terra. Este trabalho foca no desenvolvimento de métodos de análise espaço-temporal para detecção de mudanças de uso e cobertura da terra em grandes conjuntos de dados. Para isto serão desenvolvidos (1) técnicas de unificação dos dados obtidos de sensores diferentes que envolvem várias etapas: calibração radiométrica e geométrica, tratamento da cobertura de nuvem, compatibilização das bandas espectrais dos sensores, etc; (2) geração das séries temporais; (3) análise das séries temporais para detecção de mudanças do uso e cobertura da Terra. Durante este período, foram desenvolvidas técnicas de unificação dos dados dos satélites Landsat-8, Landsat-7 e CBERS-4. Além disso, foram também estudadas técnicas de reconstrução de dados de observação da Terra.

2.2. Objetivo

Desenvolver algoritmos de análise espaço-temporal para extrair informações de grandes bancos de imagens de observação da Terra.

2.3. Metas da tese de doutorado:

- Conceber, implementar e validar métodos de detecção de mudanças de uso e cobertura da terra para grandes bancos de dados com séries temporais extraídas de imagens de sensoriamento remoto multisensores;

- Publicar dois artigos em congresso internacional e dois artigos em revista científicas.

2.4. Cronograma proposto no relatório 01 (Dezembro 2016)

Cronograma

Tarefa	1°Ano		2°Ano		3°Ano	
Estudar métodos de análise de séries temporais de sensoriamento remoto	X	X				
Desenvolvimento métodos para séries temporais multi-sensor			X	X		
Artigos científicos		X		X		
Defesa de tese						

2.5. **Situação atual dos objetivos**

Os objetivos traçados para o segundo ano de pesquisa foram alcançados, de modo que o desenvolvimento dos métodos constam na proposta de tese do bolsista (Anexo 2). Dentre as abordagens utilizadas estão: a compatibilização dos dados dos diferentes sensores, Landsat-8/OLI, Landsat-7/ETM+ e CBERS-4/MUX, por meio de regressão linear; o preenchimento de dados nulos e redução de nuvens com uma adaptação do método de Maxwell et al. (2007) e o preenchimento de dados nulos com uma adaptação da metodologia de casamento de templates de séries temporais proposta por Vuolo et al. (2017). A adaptação do método de Maxwell et al. (2007) encontra-se com uma implementação prévia e começará a ser testada. A adaptação da metodologia de Vuolo et al. (2017) será implementado no decorrer do próximo ano.

No que tange a publicação de artigos científicos, o bolsista teve como aceito para publicação o artigo “Raster Data Processing with TerraLib for Lua: an application to fill Landsat-7 SLC-off gaps” na revista Journal of Computational Interdisciplinary Sciences (JCIS). O bolsista apresentou o trabalho “CBERS-4/MUX automatic detection of clouds and cloud shadows using decision trees” no XVIII Simpósio Brasileiro de Sensoriamento Remoto. O bolsista, submeteu e apresentou um resumo intitulado “Spectral compatibilization between Landsat-8/OLI, Landsat-7/ETM+ and CBERS-4/MUX bands through linear regression and linear mixture model” no 17º Workshop de Computação Aplicada do INPE (WORCAP 2017). A continuação deste trabalho foi submetida na categoria full paper ao Brazilian Symposium on Geoinformatics (Geoinfo 2017), sendo aceita para publicação nos anais do evento.

3 Resumo das atividades desenvolvidas

Como parte das obrigações do curso de pós graduação o bolsista realizou o estágio docência na Universidade Federal de São Paulo (Unifesp), auxiliando na disciplina de Lógica de Programação. O bolsista frequentou os seminários promovidos pela pós-graduação em computação aplicada, atingindo os 40 créditos exigidos pelo programa. O bolsista apresentou e foi aprovado no exame de qualificação em maio de 2017 (Anexo 1), no qual apresentou temas relacionados a séries temporais e abordagem multisensor. O bolsista também apresentou e foi aprovado no exame de proposta de tese, ocorrido em 12 de Dezembro de 2017 (Anexo 2), no qual propôs a unificação dos dados dos sensores Landsat-8/OLI, Landsat-7/ETM+ e CBERS-4/MUX por meio de regressão linear e adaptações de métodos para preencher dados nulos das observações e, assim, gerar um cubo de dados para análise espaço-temporal.

O bolsista testou abordagem por modelo linear de mistura e por regressão linear para unificação dos dados, tendo esta ultima apresentado melhores resultados. Devido à presença de nuvens e falhas no sensor Landsat-7/ETM+, alguns métodos de reconstrução de dados de observação da Terra foram pesquisados para preencher estas lacunas. Na proposta de tese o bolsista sugeriu modificações em dois métodos de reconstrução para gerar um cubo de imagens de sensores óticos. O primeiro método, alteração da abordagem de Maxwell et al. (2007), foi implementado e está sendo testado.

O bolsista também apresentou o trabalho “CBERS-4/MUX automatic detection of clouds and cloud shadows using decision trees” no XVIII Simpósio Brasileiro de Sensoriamento Remoto. Apresentou também o pôster “Spectral compatibilization between Landsat-8/OLI, Landsat-7/ETM+ and CBERS-4/MUX bands through linear regression and linear mixture model” no 17º Workshop de Computação Aplicada do INPE (WORCAP 2017) e deu continuidade a este trabalho com um full paper no Brazilian Symposium on Geoinformatics (GEOINFO 2017).

3.1. **Pesquisa**

O bolsista fez um levantamento sobre métodos de normalização de imagens provenientes de diferentes sensores sendo as mais adequadas: regressão linear e modelo linear de mistura espectral. Com base nisso o bolsista realizou testes de compatibilização envolvendo os sensores Landsat-8/OLI, Landsat-7/ETM+ e CBERS-4/MUX por meio das duas abordagens. Os resultados indicaram que a abordagem utilizando regressão linear é melhor para compatibilização destes dados. Devido à presença de nuvens no momento da aquisição das imagens, bem como falhas no sensor Landsat-7/ETM+, algumas regiões das imagens apresentam valores nulos. Com base nisso, o bolsista pesquisou métodos de reconstrução de dados de observação da Terra. Deste modo, na proposta de tese foram sugestionadas alterações em dois métodos para suprimir estas lacunas. Ambas as adaptações

encontram-se descritas no documento de proposta (Anexo 2). A primeira alteração consiste em uma adaptação do método de Maxwell et al. (2007) para preenchimento das falhas do Landsat-7/ETM+. Basicamente o método original utiliza regiões obtidas por segmentação de imagens de data próxima, para estimar valores faltantes utilizando a média dos segmentos. Na alteração proposta, será utilizada a variância dos segmentos, possibilitando assim um preenchimento mais próximo da realidade. A segunda alteração consiste em adaptar o método de Vuolo et al. (2017) para geração de séries temporais sem lacunas. No método original os autores fazem uso de distância euclidiana para casar templates e estimar os valores faltantes de acordo com o template mais semelhante. Na alteração proposta, será utilizada a distância DTW, que é mais apta a notar diferenças em séries temporais (PETITJEAN et al., 2012).

3.2. **Submissão e apresentação de artigos:**

Baseado nos estudos realizados no ano anterior, o bolsista apresentou o trabalho citado no último relatório intitulado “CBERS-4/MUX automatic detection of clouds and cloud shadows using decision trees” no simpósio brasileiro de sensoriamento. O artigo “Raster Data Processing with TerraLib for Lua: an application to fill Landsat-7 SLC-off gaps” foi aceito para publicação revisão na revista “Journal of Computational Interdisciplinary Science (JCIS)”.

O bolsista apresentou o trabalho “Spectral compatibilization between Landsat-8/OLI, Landsat-7/ETM+ and CBERS-4/MUX bands through linear regression and linear mixture model” no 17º Workshop de Computação Aplicada do INPE (WORCAP 2017) com. A continuação deste trabalho foi submetida e aceita na categoria full paper no evento Brazilian Symposium on Geoinformatics (Geoinfo 2017).

4 Atividades desenvolvidas

4.1. **Pesquisa**

Inicialmente, o bolsista fez um levantamento de sensores de média resolução espacial disponibilizados sem custo para o usuário. Dentre esses sensores optou-se por trabalhar com os dados do Landsat-7/ETM+, Landsat-8/OLI e CBERS-4/MUX.

Os dados do programa Landsat foram selecionados devido ao seu intenso uso no monitoramento do uso e cobertura da terra e por ter a maior série histórica de dados orbitais (WULDER et al., 2012. As bandas espectrais dos sensores e resolução espacial do satélite CBERS-4 são similares as do Landsat. No futuro, pode-se incluir os dados do Sentinel, que também estão sendo disponibilizados sem custo.

4.1.1. Compatibilização de dados de sensores diferentes

Duas abordagens de normalização espectral foram testadas: abordagem por modelo linear de mistura e por regressão linear.

A abordagem por regressão linear assume que a relação das bandas de diferentes sensores depende da iluminação e da geometria de observação. Baseia-se no principio de que imagens de sensores semelhantes, calibradas e atmosfericamente corrigidas são consistentes e comparáveis, apresentando pequena diferença. Assim, uma referência de valores de reflectância é utilizada em uma regressão com os valores de reflectância de um alvo, resultando em uma equação de ganho e offset para cada banda espectral. Adotando os dados do Landsat-8/OLI como referência, foram normalizados dados dos sensores Landsat-7/ETM+ e do sensor CBERS-4/MUX.

A normalização por modelo linear de mistura espectral consiste em obter respostas espectrais referência, também chamada de endmember, para diversas classes, por exemplo: vegetação, solo e água. Adotando os dados do Landsat-8/OLI como referência, coleta-se endmembers para as classes vegetação, solo e água/sombra. Seguindo o mesmo principio, esses endmembers são coletados em imagens dos sensores Landsat-7/ETM+ e do sensor CBERS-4/MUX, obtendo assim para cada classe de cada sensor uma imagem proporção. As imagens de proporção são, então, utilizadas no processo inverso utilizando os endmembers referência, gerando assim imagens sintéticas do sensor Landsat-8/OLI.

A comparação desses dois métodos foi utilizada para preparação do artigo “Spectral normalization between Landsat-8/OLI, Landsat-7/ETM+ and CBERS-4/MUX bands through linear regression and spectral unmixing” (Anexo 4). A normalização por regressão linear apresentou resultados melhores do que a do modelo linear de mistura espectral, de modo que este método será utilizado para integrar as imagens destes sensores. As bandas com maior comprimento de onda apresentaram maior correlação do que as de comprimento de onda mais curto. Esse resultado pode ser justificado pelo fato das bandas de menor comprimento de onda sofrerem mais alteração devido aos efeitos atmosféricos (JENSEN, 2007).

4.1.2. Preenchimento de lacunas em cubo de dados de sensoriamento remoto

Devido à presença de nuvens e defeitos dos sensores, outro tópico de pesquisa estudado foi a reconstrução dos dados. O primeiro método é baseado na abordagem de Maxwell et al. (2007), que estima os valores faltantes pela média das regiões da imagem segmentada, como pode ser observado na Figura 1. Na adaptação do método proposta na tese, os valores são estimados pela variância dos pixels das regiões para preencher as lacunas, realizando um preenchimento mais próximo da realidade. Este método está sendo testado.

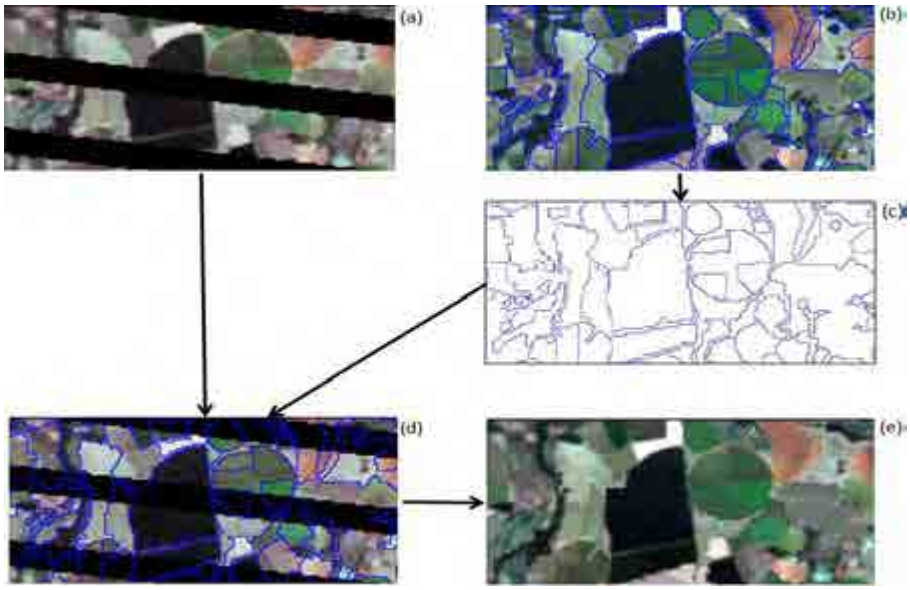


Figura 1: Abordagem de Maxwell et al., (2007) para preenchimentos de lacunas em imagens do satélite Landsat-7 (a) utilizando imagens de data próxima obtidas por outro sensor (b), segmentando-a (c), de modo a usar o valor médio deste segmento (d) para preencher as lacunas (e).

O segundo método estudado, de Vuolo et al. (2017), utiliza séries temporais como templates para preencher as lacunas de séries semelhantes, por meio de distância euclidiana. Na adaptação do método proposta, será usada uma série gerada a partir de diversos sensores, possibilitando mais observações. Além disso, será utilizada a distância DTW, que é mais adequada para detectar diferenças em séries temporais (PETITJEAN et al., 2012).

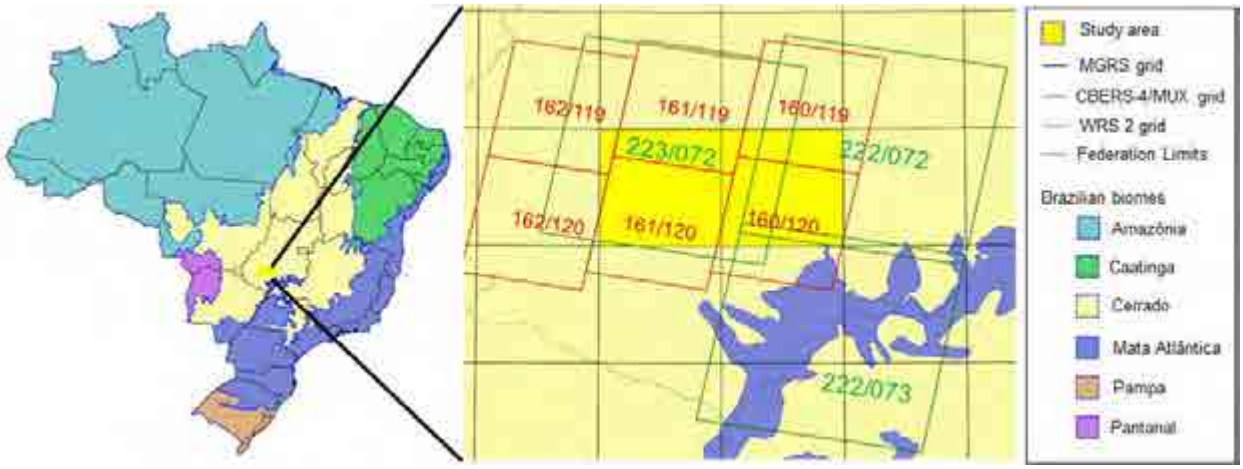
4.1.3. Casos de estudo

Dois casos de estudo serão usados nas regiões no sul do estado de Goiás, no bioma cerrado. Para este estudo, imagens de Agosto de 2015 à Outubro de 2017 foram selecionadas. Essas datas foram selecionados devido ao inicio da operação do satélite CBERS-4 e porque as plantações na região, normalmente, se iniciam entre Setembro e Outubro.

As áreas de estudo consistem em duas grades do sistema Military Grid Reference System (MGRS), identificadas por 22KDF e 22KEF. Essas áreas intersectam com a área útil do Path/Row Landsat 223/072, 222/072 e 222/073 (WRS 2 - Worldwide Reference System 2) e com 6 Path/Rows CBERS-4/MUX: 162/119, 162/120, 161/119, 161/120, 160/119 e 160/120 (CBERS WRS Path Row), como pode ser observado na Figura 2. Todas as imagens, dos sensores em estudo, que apresentem cobertura de nuvens inferior a 50%, no período de agosto de 2015 e outubro de 2017 serão adquiridas, processadas para reflectância de superfície, registradas, normalizadas e terão suas lacunas preenchidas. A seleção das grades selecionadas representam duas condições extremas, a primeira na qual as imagens do sensor referência estão totalmente contidas na área de estudo e a segunda onde a área de estudo esta localizada nas bordas do sensor referência.

Para inspeções futuras do cubo de imagens gerado, haverá uma banda contendo informação sobre o histórico do pixel. Essa banda conterá informação referente ao sensor de origem e sobre todos os processamento realizados naquele ponto, incluindo sensores utilizados para preencher lacuna e datas utilizadas. Para validar o cubo de imagens, serão realizadas duas classificações de cobertura da Terra, a primeira baseada no método proposto por Vuolo et al. (2017) e a segunda utilizando a metodologia proposta. Inicialmente, a classificação será realizada com um único

sensor e posteriormente usará os 3 sensores para verificar a melhoria da acurácia no mapeamento. Como referência serão utilizados dados do TerraClass Cerrado (INPE et al., 2012) e os resultados obtidos serão avaliados por meio de matriz de confusão.



4.2. **Submissão de trabalhos**

4.2.1. Artigo aceito para publicação pela revista Journal of Computational Interdisciplinary Sciences (JCIS)

Em setembro de 2016 como resultado da disciplina de Banco de dados Geográficos desenvolveu-se, em colaboração com os professores da disciplina Dr. Gilberto Ribeiro de Queiroz, Dra. Lúbia Vinhas e Dra. Karine Reis Ferreira o trabalho “Raster Data Processing with TerraLib for Lua: an application to fill Landsat-7 SLC-off gaps”.

Conforme anexo 3, o objetivo desse artigo é utilizar conceitos de banco de dados de imagens para realizar um Binding, usando a linguagem de programação Lua para acessar métodos de processamento raster implementados em linguagem C++ na biblioteca geoespacial TerraLib. Após o Binding, os métodos da TerraLib foram utilizados para preencher as lacunas existentes em imagens do sensor ETM+ a bordo do satélite Landsat-7.O artigo explora a utilização da interface SWIG para construir a comunicação entre Lua e C++.

4.2.2. Artigo aceito para publicação no Symposium on Geoinformatics (Geoinfo)

Em Dezembro de 2017 os resultados da normalização das imagens Landsat-7/ETM+, Landsat-8/OLI e CBERS-4/MUX foram publicados no artigo “Spectral

normalization between Landsat-8/OLI, Landsat-7/ETM+ and CBERS-4/MUX bands through linear regression and spectral unmixing". Este trabalho foi primeiramente apresentado como forma de resumo no 17º Workshop de Computação Aplicada do INPE (WORCAP 2017) e posteriormente no formato full paper nos anais do evento Symposium on Geoinformatics (Geoinfo 2017).

Conforme anexo 4, o objetivo desse artigo consiste em comparar a normalização espectral de três sensores ópticos de média resolução espacial: Landsat-7/ETM+, Landsat-8/OLI e CBERS-4/MUX. Comparou-se duas abordagens, a normalização por regressão linear e a normalização por modelo linear de mistura espectral.

Como resultado foi possível observar que as bandas espectrais dos canais com menor comprimento de onda estão menos correlacionadas. Foi observado também que existe uma correlação maior entre as imagens Landsat-7/ETM+ com Landsat-8/OLI do que CBERS-4/MUX com Landsat-8/OLI. A abordagem utilizando regressão linear apresentou resultados mais consistentes que a do modelo linear de mistura espectral.

5 Próximas etapas do trabalho

Seguindo o cronograma da proposta de tese, os métodos propostos para análise de séries temporais serão implementados, testados e avaliados, podendo gerar uma publicação em revista.

5.1. Cronograma da proposta de tese

	2018											
↓ Tarefa e Mês →	J	F	M	A	M	J	J	A	S	O	N	D
Tarefa 1	•	•										
Tarefa 2	•	•										
Tarefa 3		•	•	•								
Tarefa 4					•	•						
Tarefa 5							•	•	•			
Tarefa 6										•	•	
Tarefa 7						•	•				•	•
Tarefa 8						•	•	•	•	•	•	•

• Tarefa 1 – Revisão Bibliográfica;

• Tarefa 2 – Obter e preprocessar dados;

• Tarefa 3 – Compatibilização de dados de diferentes sensores;

• Tarefa 4 – Preenchimento de lacunas por meio de variância em segmentação multiescala;

• Tarefa 5 – Preenchimento de lacunas por casamento de template utilizando distância DTW;

• Tarefa 6 – Validar resultados com um caso de estudo;

• Tarefa 7 – Publicação dos resultados em conferências ou revistas internacionais;

• Tarefa 8 – Escrita da tese.

Referências Bibliográficas:

BANSKOTA, A., KAYASTHA, N., FALKOWSKI, M. J., WULDER, M. A., FROESE, R. E., WHITE, J. C. Forest monitoring using Landsat time series data: A review. **Canadian Journal of Remote Sensing**, v. 40, n. 5, p. 362–384, 2014.

COPPIN, P.; JONCKHEERE, I.; NACKAERTS, K.; MUYS, B.; LAMBIN, E. Digital change detection methods in ecosystem monitoring: a review. **International Journal of Remote Sensing**, v. 25, n. 9, p. 1565–1596, 2004.

EHLERS, M.; JANOWSKY, R.; GAEHLER, M. New remote sensing concepts for environmental monitoring. **Remote Sensing for Environmental Monitoring**, v. 4545, p. 1–12, 2002.

EHLERS, R. S. Análise de séries temporais., 2009. http://www.icmc.usp.br/pessoas/ehlers/stemp/.

JENSEN, J. Remote Sensing of the Environment: An Earth Resource Perspective. New Jersey: Pearson Prentice Hall, 2007.

KUENZER, C.; DECH, S.; WAGNER, W. Remote Sensing Time Series. Cham: Springer International Publishing, 2015. 458p. (Remote Sensing and Digital Image Processing, v. 22).

LEFSKY, M. a.; COHEN, W. B. Selection of Remotely Sensed Data. **Remote Sensing of Forest Environments**., 2003. v. 90, p. 13–46.

MAUS, V.; CAMARA, G.; CARTAXO, R.; SANCHEZ, A.; RAMOS, F. M.;QUEIROZ, G. R. de. 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,aug 2016.

MAXWELL, S. K.; SCHMIDT, G. L.; STOREY, J. C.; MAXWELL; S., K.; SCHMIDT; G., L.; STOREY; J., C. A multi-scale segmentation approach to filling gaps in Landsat ETM+ SLC-off images. **International Journal of Remote Sensing**, v. 28, n. 23, p. 5339–5356, 2007.

MOUSIVAND, A.; MENENTI, M.; GORTE, B.; VERHOEF, W. Multi-temporal, multi-sensor retrieval of terrestrial vegetation properties from spectral-directional radiometric data. **Remote Sensing of Environment**, Elsevier Inc., v. 158, p.311–330, 2015. ISSN 00344257.

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

SAMAIN, O.; GEIGER, B.; ROUJEAN, J. L. Spectral normalization and fusion of optical sensors for the retrieval of BRDF and albedo: Application to VEGETATION, MODIS, and MERIS data sets. **IEEE Transactions on Geoscience and Remote Sensing**, v. 44, n. 11, p. 3166–3178, 2006.

SHEN, H.; HUANG, L.; ZHANG, L.; WU, P.; ZENG, C. Long-term and fine-scale satellite monitoring of the urban heat island effect by the fusion of multi-temporal and multi-sensor remote sensed data: A 26-year case study of the city of Wuhan in China. **Remote Sensing of Environment**, Elsevier Inc., v. 172, p. 109–125, 2016.

VUOLO, F.; NG, W.-T.; ATZBERGER, C. Smoothing and gap-filling of high resolution multi-spectral time series: Example of Landsat data. **International Journal of Applied Earth Observation and Geoinformation**, Elsevier B.V., v. 57, p. 202–213, 2017.

WOODCOCK, C. E.; ALLEN, R.; ANDERSON, M.; BELWARD, A.;BINDSCHADLER, R.; COHEN, W.; GAO, F.; GOWARD, S. N.; HELDER, D.;HELMER, E. H.; NEMANI, R.; OREOPOULOS, L.; SCHOTT, J.;THENKABAIL, P. S.; VERMOTE, E. F.; VOGELMANN, J. E.; WULDER,M. A.; WYNNE, R. H. Free Access to Landsat Imagery. **Science**, v. 320, n. May, p. 1011–1012, 2008.

WULDER, M. A.; MASEK, J. G.; COHEN, W. B.; LOVELAND, T. R.;WOODCOCK, C. E. Opening the archive: How free data has enabled the science and monitoring promise of Landsat. **Remote Sensing of Environment**, v. 122,p. 2–10, 2012.

PLOS ONE

RESEARCH ARTICLE

Climate drivers of the Amazon forest greening

Fabien Hubert Wagner[1]*, Bruno Hérault[2], Vivien Rossi[3], Thomas Hilker[4†], Eduardo Eiji Maeda[5], Alber Sanchez[6], Alexei I. Lyapustin[7], Lênio Soares Galvão[1], Yujie Wang[7], Luiz E. O. C. Aragão[1,8]

1 Remote Sensing Division, National Institute for Space Research - INPE, São José dos Campos 12227-010, SP, Brazil, 2 CIRAD, UMR Ecologie des Forêts de Guyane, Kourou 97379, France, 3 UR B&SEF Biens et services des écosystèmes forestiers tropicaux, CIRAD, Yaoundé BP 2572, Cameroon, 4 Department of Geography and Environment, University of Southampton, Southampton SO17 1BJ, United Kingdom, 5 Department of Environmental Sciences, University of Helsinki, Helsinki, FI-00014, Finland, 6 Earth System Science Center, National Institute for Space Research - INPE, São José dos Campos 12227-010, SP, Brazil, 7 Goddard Space Flight Center, NASA, Greenbelt, MD 20771, United States of America, 8 College of Life and Environmental Sciences, University of Exeter, Exeter, EX4 4RJ, United Kingdom

† Deceased.
* wagner.h.fabien@gmail.com

Citation: Wagner FH, Hérault B, Rossi V, Hilker T, Maeda EE, Sanchez A, et al. (2017) Climate drivers of the Amazon forest greening. PLoS ONE 12(7): e0180932. https://doi.org/10.1371/journal.pone.0180932

Editor: Benjamin Poulter, Montana State University Bozeman, UNITED STATES

Received: January 31, 2017

Accepted: June 24, 2017

Published: July 14, 2017

Data Availability Statement: Data and codes are available from Figshare. EVI data: https://doi.org/10.6084/m9.figshare.5151622.v1 Litterfall data: https://doi.org/10.6084/m9.figshare.5153824.v1.

Funding: This project and F.H.W. have been funded by the Fapesp (Fundação de Amparo a Pesquisa do Estado de São Paulo, processo 13/14520-6, processo 15/50484-0 and processo 16/17652-9). L.E.O.C.A. is thankful for the support of FAPESP (grant 50533-5) and CNPQ (grant 304425/2013-3). E.E.M. was funded by the Academy of Finland (project: 266393). A.S. has been funded by the

Abstract

Our limited understanding of the climate controls on tropical forest seasonality is one of the biggest sources of uncertainty in modeling climate change impacts on terrestrial ecosystems. Combining leaf production, litterfall and climate observations from satellite and ground data in the Amazon forest, we show that seasonal variation in leaf production is largely triggered by climate signals, specifically, insolation increase (70.4% of the total area) and precipitation increase (29.6%). Increase of insolation drives leaf growth in the absence of water limitation. For these non-water-limited forests, the simultaneous leaf flush occurs in a sufficient proportion of the trees to be observed from space. While tropical cycles are generally defined in terms of dry or wet season, we show that for a large part of Amazonia the increase in insolation triggers the visible progress of leaf growth, just like during spring in temperate forests. The dependence of leaf growth initiation on climate seasonality may result in a higher sensitivity of these ecosystems to changes in climate than previously thought.

Introduction

The Amazonian forests account for 14% of the global net primary production (NPP) and are a major component (66%) of the inter-annual variation in global NPP [1]. While large seasonal swings in leaf area have been reported at least in parts of the Amazon basin [2–4], the environmental controls that trigger the synchronous development of new leaves are not well understood [5–7]. As a result, current earth system models inadequately represent the dynamics of leaf development, despite its major role for photosynthesis of tropical vegetation [8]. In equatorial forests, leaf flushing correlates with increased light availability and photosynthetically active radiation during the dry season [4, 9], and is theoretically driven by a change in daily insolation [10]. However, water availability constrains leaf phenology in southern Amazonia and most of the Congo basin, impeding the maintenance of the evergreen state during the dry season [11].

remote sensing

MDPI

Article

Examining Multi-Legend Change Detection in Amazon with Pixel and Region Based Methods

Mariane S. Reis *, Luciano V. Dutra, Sidnei J. S. Sant'Anna and Maria Isabel S. Escada

Brazilian National Institute for Space Research—INPE, São José dos Campos 12245, SP, Brazil; dutra@dpi.inpe.br (L.V.D.); sidnei@dpi.inpe.br (S.J.S.S.); isabel@dpi.inpe.br (M.I.S.E.)
* Correspondence: reis@dpi.inpe.br; Tel.: +55-12-3208-6781

Academic Editors: Guangxing Wang, Erkki Tomppo, Dengsheng Lu, Huaiqing Zhang, Qi Chen, Arko Lucieer and Prasad S. Thenkabail
Received: 1 October 2016; Accepted: 9 January 2017; Published: 15 January 2017

Abstract: Post-classification comparison is one of the most widely used change detection methods. However, it presents several operational problems that are often ignored, such as the occurrence of impossible transitions, difficulties in accuracy assessment and results not accurate enough for the purpose. This work aims to evaluate post-classification comparison change detection results obtained from LANDSAT5/TM data in a region of the Brazilian Amazon, using three legends in different levels of detail and both pixel wise and region based classifiers. A distinctive characteristic of the used approach is that each change mapping is the result of the combination of 100 land cover classifications for each date, obtained using varied training samples. This approach allowed to account for the training samples choice into the methodology, as well as the construction of confidence mappings. We presented and discussed different approaches for evaluating change results, such as the likelihood of land cover transitions occurring within the study area and time gap, the use of rectangular matrices to incorporate the occurrence of impossible or non evaluable changes and classification uncertainty. In general, change mappings obtained from region based classifications showed better results than the ones obtained from pixel based classifications. Globally, the use of region based approaches, in contrast to pixel based ones, led to an increase in accuracy of 15.5% for the change mapping from the most detailed legend, 7.8% for the one with the legend with intermediate level of detail and 3.6% for the less detailed one. In addition, individual transitions between land cover classes were better identified using region based approaches, with the exception of transitions from a non agriculture class to an agricultural one. The proposed quality mappings are useful to help to evaluate the change mappings, mainly in legend levels with higher level of detail and if reference samples are unreliable or unavailable. It was possible to access, in a spatially explicit way, that at least 29.0% of the pixel based change mapping and 21.9% of the region based one from the most detailed legend were erroneous classified, without ground truth information on the evaluated date. These values decreased to 0.5% and 1.4% (respectively the pixel and region based approaches) for results with the legend with the intermediate level of detail and are non existent in the results from the less detailed legend. The more generalized the legend (lower number of classes), the most similar are the accuracy of region and pixel based change mappings. These accuracy values also increase as fewer classes are considered in the legend, since similar classes are assembled during clustering, which reduces the overlap between groups. However, this accuracy is still low for operational purposes in areas with few changes, even considering the very high accuracy of the land cover classifications used to generate the change mappings (land cover classification with Overall Accuracy higher than 0.98 resulted in change mappings with Overall Accuracy around 0.83).

Keywords: change detection; multi-legend; pixel based classification; region based classification; Amazon

Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II. *http://www.jstatsoft.org/*

dtwSat: Time-Weighted Dynamic Time Warping for Satellite Image Time Series Analysis in R

Victor Maus
University of Münster
INPE, IIASA

Gilberto Câmara
INPE
University of Münster

Marius Appel
University of Münster

Edzer Pebesma
University of Münster

Abstract

The opening of large archives of satellite data such as LANDSAT, MODIS and the SENTINELs has given researchers unprecedented access to data, allowing them to better quantify and understand local and global land change. The need to analyse such large data sets has lead to the development of automated and semi-automated methods for satellite image time series analysis. However, few of the proposed methods for remote sensing time series analysis are available as open source software. In this paper we present the R package **dtwSat**. This package provides an implementation of the Time-Weighted Dynamic Time Warping method for land cover mapping using sequence of multi-band satellite images. Methods based on dynamic time warping are flexible to handle irregular sampling and out-of-phase time series, and they have achieved significant results in time series analysis. **dtwSat** is available from the Comprehensive R Archive Network and contributes to making methods for satellite time series analysis available to a larger audience. The package supports the full cycle of land cover classification using image time series, ranging from selecting temporal patterns to visualising and assessing the results.

Keywords: dynamic programming, MODIS time series, land cover changes, crop monitoring.

1. Introduction

Remote sensing images are the most widely used data source for measuring land use and land cover change (LUCC). In many areas, remote sensing images are the only data available for this purpose (Lambin and Linderman 2006; Fritz *et al.* 2013). Recently, the opening of large archives of satellite data such as LANDSAT, MODIS and the SENTINELs has given researchers unprecedented access to data, allowing them to better quantify and understand local and global land change. The need to analyse such large data sets has lead to the development of automated and semi-automated methods for satellite image time series analysis. These

Campo Verde Database: Seeking to Improve Agricultural Remote Sensing of Tropical Areas

Ieda Del'Arco Sanches, Raul Queiroz Feitosa, *Senior Member, IEEE*, Pedro Marco Achanccaray Diaz, *Student Member, IEEE*, Marinalva Dias Soares, Alfredo José Barreto Luiz, Bruno Schultz, and Luis Eduardo Pinheiro Maurano

***Abstract*— In tropical/subtropical regions, the favorable climate associated with the use of agricultural technologies, such as no tillage, minimum cultivation, irrigation, early varieties, desiccants, flowering inducing, and crop rotation, makes agriculture highly dynamic. In this letter, we present the Campo Verde agricultural database. The purpose of creating and sharing these data is to foster advancement of remote sensing technology in areas of tropical agriculture, primarily the development and testing of methods for crop recognition and agricultural mapping. Campo Verde is a municipality of Mato Grosso state, localized in the Cerrado (Brazilian Savanna) biome, in central west Brazil. Soybean, maize, and cotton are the primary crops cultivated in this region. Double cropping systems are widely adopted in this area. There is also livestock and forestry production. Our database provides the land-use classes for 513 fields by month for one Brazilian crop year (between October 2015 and July 2016). This information was gathered during two field campaigns in Campo Verde (December 2015 and May 2016) and by visual interpretation of a time series of Landsat-8/Operational Land Imager (OLI) images using an experienced interpreter. A set of 14 preprocessed synthetic aperture radar Sentinel-1 and 15 Landsat-8/OLI mosaic images is also made available. It is important to promote the use of radar data for tropical agricultural applications, especially because the use of optical remote sensing in these regions is hindered by the high frequency of cloud cover. To demonstrate the utility of our database, results of an experiment conducted using the Sentinel-1 data set are presented.**

***Index Terms*— Agricultural mapping/monitoring, double cropping systems, free available database, remote sensing, synthetic aperture radar (SAR), tropical agriculture.**

I. Introduction

FOOD security is a major concern worldwide and faces the challenge of a continuously increasing global population and limited availability of natural resources. Consequently, agriculture is a key economic activity worldwide, primarily for food but also for fiber and energy (biofuel) production.

Tropical areas have an important position in global food production. Brazil, for instance, is one of the largest global producers and exporters of sugar, coffee, orange juice, soybean, maize, and beef. Brazil is also the lead producer of sugarcane ethanol, an alcohol-based biofuel. Much of this progress is the result of intense research in tropical agriculture. The Brazilian Cerrado biome, for example, was previously considered an area unsuitable for cultivation but has become an agricultural frontier in recent decades and is currently one of the top grain and beef-producing regions in the world [1].

To assure that food production meets the world demands and its environmental impacts are minimized, it is necessary to monitor agriculture activities regularly. Compared to temperate regions, this mission is considerably more challenging for tropical agricultural areas because of the favorable climate associated with the different cultivation systems adopted (e.g., no tillage, minimum cultivation, irrigation, crop rotation, and early varieties) cause intense dynamism and demand year-round monitoring. For this purpose, satellite remote sensing technology can contribute significantly, since it offers repetitive, timely, and accurate information regarding agricultural activity over large areas at relatively low cost [2].

Currently, a variety of high-quality remote sensing data are available free of charge that can be used to monitor agriculture, such as Moderate Resolution Imaging Spectroradiometer (MODIS) products and images from the Landsat series. Several studies have been conducted in this field using these data [2]–[4], but there is still considerable room for advancement, especially in tropical areas. For example, efficient methodologies to identify areas of double cropping (i.e., two consecutive crops cultivated on the same land within a single growing season) using multitemporal remote sensing data have been developed [5], but it remains difficult to identify which two crops are cultivated [4]. Moreover, most crop pattern recognition research has been conducted using a database of temperate regions [6], [7].

In optical remote sensing, cloud cover represents a major constraint, especially in tropical countries [8]–[10]. Alternatives to overcome or at least minimize this problem might be the combined use of data from different sensors,

Manuscript received October 23, 2017; revised November 22, 2017; accepted December 29, 2017. This work was supported by the "Science without Borders" CNPq/CAPES Program under Project 402.597/2012-5. *(Corresponding author: Raul Queiroz Feitosa.)*

I. Del'Arco Sanches and L. E. P. Maurano are with the National Institute for Space Research, São José dos Campos 12227-010, Brazil (e-mail: ieda.sanches@inpe.br; luis.maurano@inpe.br).

R. Q. Feitosa, P. M. A. Diaz, and M. D. Soares are with the Pontifical Catholic University of Rio de Janeiro, Rio de Janeiro 22753801, Brazil (e-mail: raul@ele.puc-rio.br; pmad9589@ele.puc-rio.br; mdiasoares@gmail.com).

A. J. B. Luiz is with the Brazilian Agricultural Research Corporation (Embrapa), Jaguariuna 13820-000, Brazil (e-mail: alfredo.luiz@embrapa.br).

B. Schultz is with Geoambiente, São José dos Campos 12244-000, Brazil (e-mail: schultz.florestal@gmail.com).

Color versions of one or more of the figures in this letter are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LGRS.2017.2789120

MAPEAMENTO DE PADRÕES DE INTENSIDADE DA DEGRADAÇÃO FLORESTAL: ESTUDO DE CASO NA REGIÃO DE SINOP, ESTADO DO MATO GROSSO.

Vinicius do Prado Capanema
Taise Farias Pinheiro
Maria Isabel Sobral Escada
Sidnei J.S. Sant'Anna

Instituto Nacional de Pesquisas Espaciais - INPE
Av. dos Astronautas 1758 - 12227-010 - São José dos Campos – SP, Brasil
{vinicius.capanema, taise.pinheiro}@inpe.br, {isabel, sidnei }@dpi.inpe.br.

RESUMO

Neste trabalho é apresentada uma metodologia para mapear e classificar a partir de imagens OLI/Landsat, níveis de intensidade de degradação florestal de forma semiautomática, e tem como área de estudo a região de Sinop, situada no estado do Mato Grosso. A abordagem metodológica constou de duas etapas: i) classificação espectral da imagem por meio da técnica de Modelo Linear de Mistura Espectral, para a geração de uma imagem-índice, combinando as frações solo e vegetação. Nessa etapa, a imagem resultante foi fatiada e os elementos indicadores de degradação florestal especificamente decorrentes de exploração madeireira, tais como, presença de pátios de estocagem, carreadores para derrubada e escoamento da madeira, e cicatrizes de fogo, foram identificados e mapeados; ii) classificação estrutural dos padrões de intensidade de degradação florestal considerado células de 1 km^2. Técnicas que exploram as métricas de paisagem e de mineração de dados foram empregadas para classificação dos padrões de degradação. O desempenho da classificação, que teve como suporte informações coletadas em campo, apresentou exatidão global e índice Kappa de 96% e 91%, respectivamente. Os resultados obtidos mostraram que essa abordagem, por considerar a intensidade da degradação, pode ser replicada em estudos temporais de análise das condições da paisagem florestal, pois a célula, sendo uma unidade fixa no tempo e no espaço, possibilita mensurar a direção e magnitude da estrutura dos elementos associados à degradação e analisar os seus efeitos colaterais no espaço e no tempo. A metodologia proposta possibilitou gerar gradientes espaciais de intensidade de degradação florestal, cujas informações podem subsidiar o planejamento de políticas e de ações de controle e de fiscalização em áreas florestais.

Palavras chaves: Degradação Florestal. Exploração Madeireira**.** Classificação Espectral, Mineração de dados Classificação Estrutural

ABSTRACT

Keywords:Forest Degradation, selective Logging, Spectral Classification, Data Mining, Structural Classification

1. INTRODUÇÃO

A degradação florestal pode ser entendida como um processo que resulta na alteração das condições biofísicas e estruturais originais dos sistemas florestais. Na literatura, contudo, não há um consenso quanto à sua definição que pode assumir, de acordo com o objetivo do estudo, um enfoque mais específico, geral ou operacional (LUND, 2009; PINHEIRO, 2015). No escopo do Sistema de Mapeamento da Degradação Florestal (DEGRAD), que faz parte do Programa de Monitoramento da Floresta Amazônia Brasileira por satélite, desenvolvido pelo INPE (2008), a degradação florestal é definida como o *processo gradual e de longo prazo da perda da cobertura florestal por meio da extração seletiva de madeira e/ou da ocorrência de incêndios florestais*. Esse conceito se aproxima mais da perspectiva operacional, pois especifica elementos que descrevem a degradação florestal e que são passíveis de serem detectadas com imagens de satélite e técnicas de processamento de imagens. Uma das vantagens do uso do conceito operacional de degradação, que é o utilizado neste trabalho, é que a revisita dos sensores orbitais que geram imagens periódicas de uma mesma área, possibilita ao usuário acompanhar todos os estágios de um determinado fenômeno que, no caso deste estudo, é o processo de degradação florestal.

International Journal of Geographical Information Science
Vol. 00, No. 00, July 2017, 1–18

RESEARCH ARTICLE

A spatiotemporal calculus for reasoning about land use change dynamics

Adeline Marinho Maciel[a*], Gilberto Camara[a], Lubia Vinhas[a], Michelle Cristina Araujo Picoli[a], Rodrigo Anzolin Begotti[a] and Luiz Fernando Ferreira Gomes de Assis[a]

[a]*Image Processing Division, National Institute for Space Research, Av. dos Astronautas 1758, So Jos dos Campos, SP, 12227-001, Brazil*

(*v0.0 released July 2017*)

*Corresponding author. Email: adeline.maciel@inpe.br

ISSN: 1365-8816 print/ISSN 1362-3087 online

DOI: 10.1080/1365881YYxxxxxxxx
http://www.informaworld.com

Big Earth Observation Time Series Analysis for Monitoring Brazilian Agriculture

Michelle Picoli[a,*], Gilberto Camara[a], Ieda Sanches[a], Rolf Simões[a], Alexandre Carvalho[b], Adeline Maciel[a], Alexandre Coutinho[c], Julio Esquerdo[c], João Antunes[c], Rodrigo Begotti[a], Damien Arvor[d], Claudio Almeida[a]

[a]*National Institute for Space Research (INPE), São José dos Campos, Brazil*
[b]*Institute of Applied Economic Research (IPEA), Brasilia, Brazil*
[c]*Embrapa Agricultural Informatics, Brazilian Agricultural Research Corporation (Embrapa), Campinas, Brazil*
[d]*Universite de Rennes 2, Rennes, France*

Abstract

This paper presents innovative methods for using satellite image time series to produce land use and land cover classification over large areas in Brazil from 2001 to 2016. We use MODIS time series data to classify natural and human-transformed land areas in state of Mato Grosso, Brazil's agricultural frontier. Our hypothesis is that building high-dimensional spaces using all values of the time series, coupled with advanced statistical learning methods, is a robust and efficient approach for land cover classification of large data sets. We use the full depth of satellite image time series to create large dimensional spaces for statistical classification. Data consists of MODIS MOD13Q1 time series 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 a support vector machine model with a 92-dimensional attribute space. Using a 5-fold cross validation, we obtained an overall accuracy of

*Corresponding author

Email addresses: mipicoli@gmail.com (Michelle Picoli), gilberto.camara@inpe.br (Gilberto Camara), ieda.sanches@inpe.br (Ieda Sanches), rolf.simoes@inpe.br (Rolf Simões), alexandre.ywata@ipea.gov.br (Alexandre Carvalho), adeline.maciel@inpe.br (Adeline Maciel), alex.coutinho@embrapa.br (Alexandre Coutinho), julio.esquerdo@embrapa.br (Julio Esquerdo), joao.antunes@embrapa.br (João Antunes), rodrigo.begotti@inpe.br (Rodrigo Begotti), damien.arvor@gmail.com (Damien Arvor), claudio.almeida@inpe.br (Claudio Almeida)