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).

doi: 10.2760/383579

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.
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/.

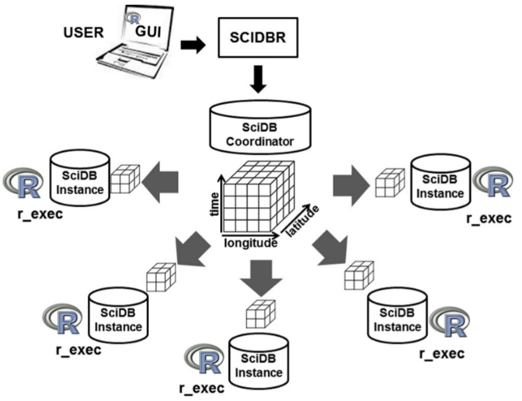

Fig. 1: Remote execution of **R** scripts in SciDB

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.

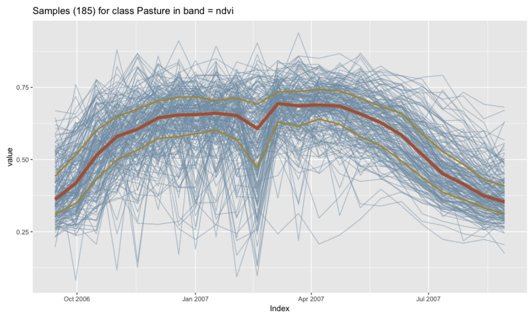

Fig. 2: Time series of 370 ground samples for land cover class ”Pasture” in the state pf Mato Grosso, Brazil (source: authors).

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)	Data dimensions	Measures (millions)	Proc time (hours)
Mato Grosso	900,000	92	20,000	6
Cerrado	2,050,000	92	50,000	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.