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

Project Overview

Host institution: INPE – Instituto Nacional de Pesquisas Espaciais

Principal investigator: Gilberto Câmara (INPE)

Research team: Ieda Sanches, Karine Ferreira, Leila Fonseca, Lúbia 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

State-of-the-art

Humanity is changing rural and urban landscapes at an unprecedented pace. Global population will increase to around 8.5 billion by mid-century. Crop and livestock demand and production will rise by around 40% between 2008 and 2030. Growing pressures on food, water, and energy threaten the planet, at the same time we need to mitigate and adapt to climate change. More and more, citizens and politicians are pressing scientists to provide qualified information that would allow wise decisions about the future of our planet.

One of the most immediate consequences of humanity's transformation the Earth's ecosystems and landscapes is *land use change*. Globalization has increased substantially the pace of land use change in developing nations. During the 1980–2000 period, more than half of the new agricultural land across the tropics came at the expense of intact forests, and another 28% came from disturbed forests (Lambin and Meyfroidt, 2011). To understand the impact and extent of global land use change, we need a new generation of information systems.

Earth Observation satellites are the only source that provides a continuous and consistent set of information about the Earth's land and oceans. Recent decisions by major space agencies have been making unprecedented amounts of imagery available for research and operations. This brings about a unique opportunity to measure the global and local changes in our environment and assess the human impacts on land and the oceans.

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*.

Project Objectives

Objective 1: Design and build an open source knowledge platform for describing, accessing and analysing big Earth observation data. Our solution combines innovative methods for scientific data management, spacetime data analysis and semantic data description.

Objective 2: Produce large-scale land use and land cover change information on tropical forests and global agricultural production using the knowledge platform, with significant better quality than current methods.

Objective 3: Promote the knowledge platform for adoption by researchers and students, emphasizing the benefits of large-scale data sharing and reproducibility of scientific results.

Expected impact

Earth observation data analysis methods lag far behind our capacity to build sophisticated satellites. Despite the inherent nature of Earth observations as systematic data collection, scientists do not organize remote sensing images as space-time data sets due to the lack to computational support. Mostly, they produce land cover maps taking either a single or at most two time references. Scientists thus overlook the time reference inherent to Earth observation data.

Current global and large-scale land cover products are not fit for operational applications. Global land cover data sets such as MODIS Land Cover, GLC2000 and GlobCover have many mismatches on the spatial distribution of their land classes (McCallum et al., 2006). To make progress on understanding land change worldwide, we need to improve the quality of the methods we use to generate these kinds of products.

Furthermore, land use practices are becoming subtler than just a transition from one cover (e.g, forest) to another (e.g., pasture). We need to capture changes associated with forest degradation and temporary or mixed agricultural regimes. Forest transition areas have become more complex to describe and measure. Also, recent research shows that much of the recent increase of agricultural productivity in Brazil is due to double cropping-practices. These results motivate us to explore large-scale time series of remote sensing data to improve global land use and land cover classification.

Key ideas

Our project is based on one key idea. Since each Earth observation satellite revisits the same place at regular intervals, its measurements can, in principle, be calibrated so that observations of the same place in different times are comparable. These observations can be organized in regular time intervals, so that each measure from sensor is mapped into a three dimensional array in spacetime (cf. Figure 1). Researchers can then address any location in the spacetime box and perform operations in arbitrary spacetime partitions.

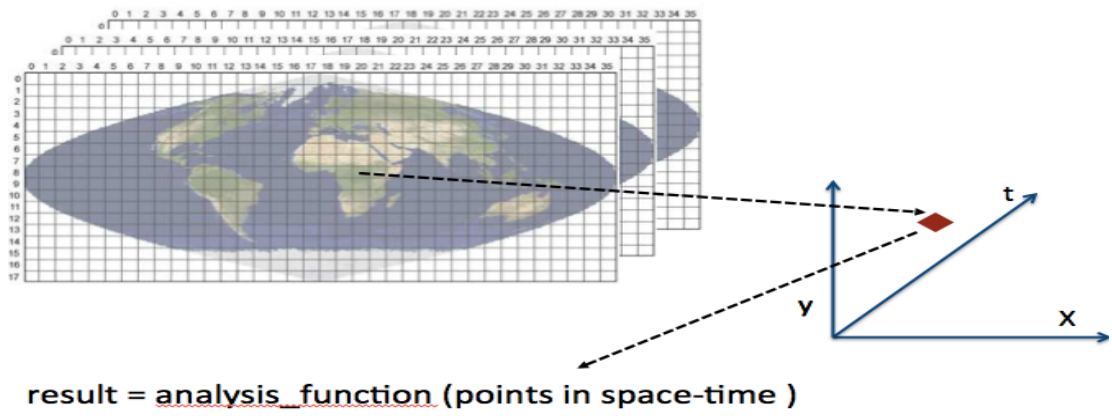

Figure 1 – Mapping data from a sensor into a position in spacetime.

To make this idea work in practice, we need to combine two key technological innovations:

(a) A scientific database based on the SciDB innovative array database management system, capable of managing large remote sensing data sets (Stonebraker et al., 2013).

(b) 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.

By organising big Earth Observation data as multidimensional arrays, we can develop new spacetime methods for information extraction. We will build a system that allows remote processing of data stored in big processing servers, thus *'bringing the user to the data'* instead of *'bringing the data to the user'*. The complete infrastructure will be open source, and thus promotes reproducible science.

We aim to make two important contributions:

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

Putting it all together

To build the proposed knowledge platform for land use and land change information, we will put the above-described building blocks together. We share the vision of the late database researcher Jim Gray: "*Today the typical scientist copies files to a local server and operates on the data sets using his own resources. Increasingly, the data sets are so large, and the application programs are so complex, that it is much more economical to move the end-user's programs to the data and only communicate questions and answers*" (Gray et al., 2005).

We envision a SDI architecture for big spatiotemporal data sets [11]. The idea of the SDI is to bring together the set of technologies that are required to manage such big data. The SDI is divided in four components: Databases, Web services, and Geographical Information System (GIS) tools. The proposed architecture is shown in Figure 2 and briefly described below.

We consider there are two main types of big spatiotemporal data sets, those in vector and raster formats. Vector data is best stored in spatial DBMS such as PostGIS that are compliant with the OGC-SFA (Open Geospatial Consortium - Simple Feature Access) specification. For raster data, such as remote sensing images, we propose the use of array databases such as SciDB. To disseminate the databases on the Internet, we propose the use of web services for raster and vector data sets as well as for their metadata. In additional to the well-established Open Geospatial Consortium standards for web services, such as WMS, WFS, WCS and CSW, we propose the WTSS – Web Time Series Service [9]. We also propose to extend the Web Coverage Service (WCS) to handle large spatiotemporal data sets stored in SciDB.

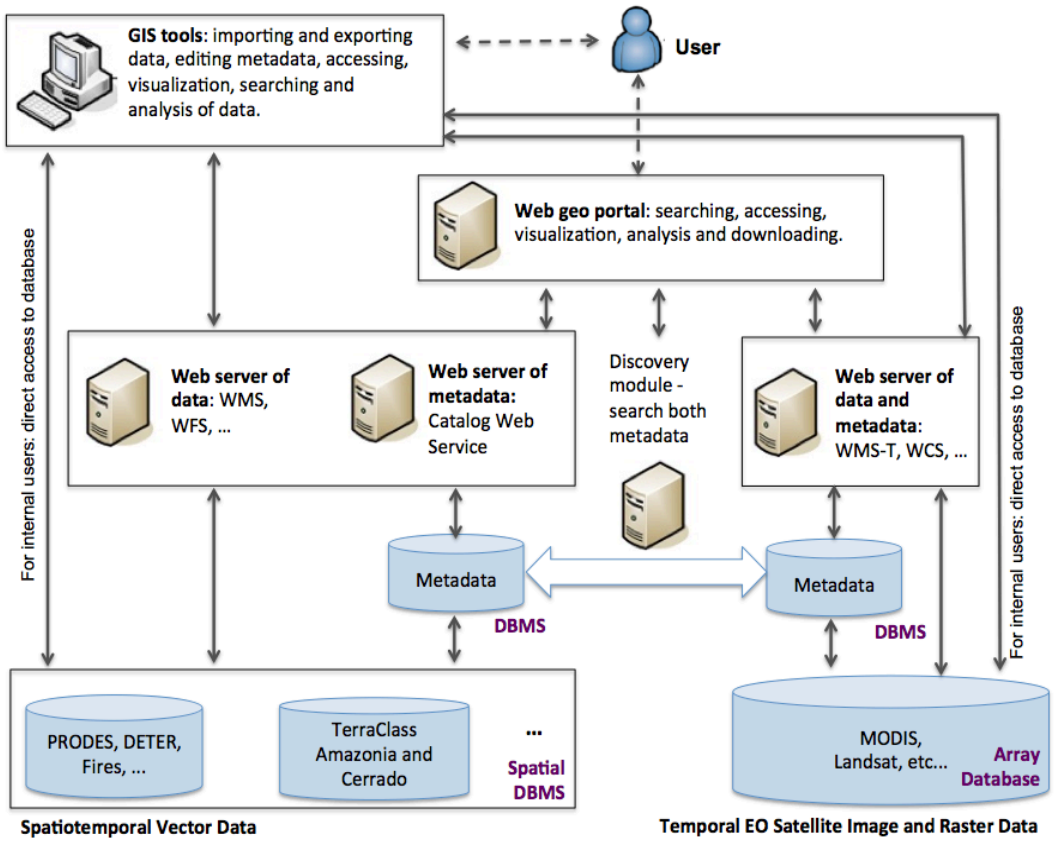

Figure 2 – Proposed spatial data infrastructure for handling big geospatial data (source: [Ferreira et al., 2015]).

The GIS tools will have to be efficient to deal with big spatiotemporal data sets. As such, they need to provide the following features:

(c) *Tools to access and combine data sets from different types of data sources*: These tools split the processing into parts. This combines spatial DBMS (e.g., PostGIS) functionalities for vector data handling with array databases that handle multidimensional raster data.

(d) *Server-side processing mechanisms:* GIS tools should allow users to process data directly on the server, avoiding the transfer of big data sets from servers to local machines.

(e) *Script language to express complex processing:* These tools allow users to express complex processing through script languages, such as R, Python and LUA.

(f) *Spatiotemporal data handling:* GIS tools should include new algorithms to analyse spatiotemporal data.

Global analysis

We aim to use the knowledge platform to produce land use and land cover information. We will approach the problem by focusing on important problems: how produce information on tropical forests and on large-scale agriculture in Brazil, a territory we know well. We will use our substantial previous experience in INPE (Brazilian Institute of Space Research) on monitoring the Amazonia tropical forest and the Brazilian crop and biofuels production.

Forest monitoring

Human use in tropical forests has been responsible for substantial GHG emissions in the last 20 years, with Brazil and Indonesia leading the amount of cleared area (Foley et al., 2005; Hougthon et al., 2012). There is still substantial uncertainty in the current estimates, partly because of the limitations of the current methods of data collection. We believe that big EO data analysis can improve information on tropical forests.

A recent work points out the challenges of using big EO data for forest monitoring. Hansen et al. (2013) produced the first global map of forest change at 30-metre resolution. Despite the important technical advances, the method used by Hansen et al. (2013) still relied on comparing two maps produced at different time instances. In this way, information about land trajectories is lost. There is no distinction between planted and native forests, and between evergreen and deciduous forests. The challenge remains on how to use big EO data to produce land trajectory information.

The multi-satellite approach is the best way of finding out land trajectories in tropical forests. MODIS land products will be the base data set for obtaining the land trajectories; LANDSAT data will be used in areas where detailed information is required. The temporal resolution of the MODIS land product provides the overall trend signal, and distinguishes the types of forests (deciduous vs. evergreen, planted vs. natural). The spatial resolution of LANDSAT and similar satellites will help to solve the problems of mixed pixels, forest degradation and detailed information. Combining MODIS-class with LANDSAT-class data at the global level is in our view the right way forward.

Agricultural monitoring

Global information on agricultural production is crucial for decision-makers. Agricultural monitoring from space is currently the only option to provide reliable and consistent information on global food production.

To improve information on global food production, we need to combine products derived from MODIS with information derived from LANDSAT-class satellites such as ESA's Sentinel-2 scale. This requires including the full temporal depth of the global archives available at such scales (Müller et al. 2015). MODIS data provides information about agricultural cycles of large-scale crop production, while LANDSAT-class data provides information on medium-scale production.

We are aware that even an intelligent combination of different satellites is not sufficient to provide complete information on agricultural production. Thus, we will focus on the major crop types that are most relevant for the global food security: wheat, maize, rice, and soy. These are the crops selected by the GEOGLAM initiative and those where improvements on land information will have the greatest immediate impact.

Vision

Combining free and big Earth Observation data, array databases, server-side processing, remote sensing time series and space time data analysis, knowledge representation, and expert knowledge on remote sensing of forests and agriculture, we will be able to build a new type of information system. This system will make good use of the large free satellite data available, while reducing the workload necessary for data organization.

Our proposed system will 'bring the user to the data'. A few data centres will be the repositories of big EO data sets, and they will make a large number of data analysis methods available to the outside users. Users will then combine these methods into workflows written in high-level open source languages (such as R and Python) so that they can make best use of big EO datasets.

References

Ferreira, K. et al. (2015). Towards a spatial data infraestructure for big spatiotemporal data sets. In: 17th Brazilian Symposium on Remote Sensing (SBSR), 2015. Proceedings, p. 7588-7594.

Foley, J. et al. (2005), “Global consequences of land use”. Science 309.5734: 570-574.

Fritz, S. et al. (2011). “Highlighting continued uncertainty in global land cover maps for the user community”. Environmental Research Letters, 6, 044005.

Fritz, S. et al. (2013), “The Need for Improved Maps of Global Cropland”. EOS Transactions American Geophysical Union, 94, 31-32.

Gray, J. et al. (2005), “Scientific data management in the coming decade”. ACM SIGMOD Record 34(4):41.

Hansen, M. et al. (2013), "High-resolution global maps of 21st-century forest cover change." Science 342.6160: 850-853.

Houghton, R. A., et al. (2012), "Carbon emissions from land use and land-cover change." Biogeosciences 9.12: 5125-5142.

Lambin, E. F., & Meyfroidt, P. (2011), “Global land use change, economic globalization, and the looming land scarcity”. PNAS, 108(9), 3465-3472.

McCallum, I. et al. (2006), "A spatial comparison of four satellite derived 1km global land cover datasets." Int Jour of Applied Earth Observation and Geoinformation, 8 (4):246-255.

Müller, H. et al. (2015), “Mining dense Landsat time series for separating cropland and pasture in a heterogeneous Brazilian savanna landscape”. Remote Sensing of Environment, 156, 490-499.

Stonebraker, M. et al. (2013). “SciDB: A database management system for applications with complex analytics”. Comp in Science & Engineering, 15(3), 54-62.