A discussion of crowdsourced geographic information initiatives and big Earth observation data architectures for land-use and land-cover monitoring

Luiz Fernando F. G. de Assis, Karine Reis Ferreira, Lúbia Vinhas
Brazilian National Institute for Space Research (INPE)
São José dos Campos, Brazil
luizffga@dpi.inpe.br,
karine.ferreira@inpe.br,
lubia.vinhas@inpe.br

Téssio Novack, Alexander Zipf
GIScience Chair, Heidelberg University
Heidelberg, Germany
novack@uni-heidelberg.de,
zipf@uni-heidelberg.de

Abstract

The effective monitoring of land-use and land-cover changes (LULCC) is a basic requirement for understanding socio-enviromental processes of local to global scales. Remote sensing data and methods have long been established as the most effective approach for monitoring LULCC. The potential for further increasing the effectiveness of this approach is proportional to the astonishingly large amount of satellite imagery provided, many times free-of-cost, by space agencies worldwide. However, scientists still lack of ways of organizing, structuring and analyzing this colossal amount of remote sensing data in a way that leverages administrative and scientific LULCC monitoring. Hence, an efficient image data storage, query and processing architecture that manages different satellite specifications and climatic conditions is required for generating and sharing updated and area-extensive LULCC information. Furthermore, because reliable LULCC monitoring with remote sensing data requires extensive training and validation analysis performed by humans, the potential of big Earth Observation (EO) data for LULCC monitoring is still limited by the amount and time-availability of the analysts involved in the project. In this paper, we discuss the potential of Citizen Science for improving the feasibility and effectiveness of LULCC monitoring supported by big EO data architectures. We put forward general ideas on how to promote and stimulate an active involvement of citizens in EO data analytics for LULCC monitoring. For that, we briefly present and critically evaluate how existing approaches that allow citizens to contribute with up-to-date and detailed LULCC information mitigate the issue of exhaustive sampling required in LULCC monitoring with automatic remote sensing image classification.
Keywords: Citizen Science, Land-Use/Land-Cover, Remote Sensing, Big Earth Observation Data.

1 Introduction

Land-use and land-cover changes (LULCC) have many complex interrelations with different socio-economic and environmental processes. This interplay occurs from the local to the global scales thus affecting individuals and the local environment as well as society and the Earth system as a whole. Researchers and decision-makers need reliable, up-to-date and spatially extensive LULCC information at different scales in order to better understand these interrelations and thus conceive effective policies for environment protection and the sustainable management of natural resources.

For a few decades now, due to their many advantages, remote sensing data and methods, coupled with field observations and other geographic data, have been used for monitoring LULCC. More recently, the open data policy adopted by many national space agencies has opened access to petabytes of remote sensing data of different spatial, spectral and temporal resolutions. This offers new possibilities for improving scientific and governmental LULCC monitoring activities, for example, deforestation detection, wildfires mapping, crop monitoring.

However, exploring this massive amount of data requires the understanding of an appropriate software architecture in order to ensure the interoperability and reproducibility of their analysis. Such investigations cannot be compared to small-scale ones that usually take place in different processing and development environments. By enabling their common configuration parameters, this software architecture leads to an unbiased and thus more consistent data analysis, which, in its turn, leads ultimately to more efficient decision making.

A few architectures for big EO data handling have emerged with recent advances in distributed systems, cloud computing and geospatial services (Lewis et al., 2017; Camara et al., 2016; Wang et al., 2018). These advances have enabled the optimization of queries, the access to different remote sensing data, the down and upscaling of analyses in real-time as well as the possibility that more people

collaboratively work on the same analysis. Applications supported by big EO data architectures frequently require exhaustive experimentations, high skilled people and more extensive sampling. These requirements could be partially fulfilled with the support of crowdsourced geographic information. Human visual image interpretation and the collection of in situ data are simple and reliable ways of providing up-to-date and detailed LULCC information (Fritz et al., 2009) as well as validating existing ones (See et al., 2015), thus inserting society's perception in the pipeline of LULCC analysis (Jokar et al., 2013; Comber et al., 2015; Moorthy et al., 2017). In other words, citizens could be part of the analysis process by cleaning, creating, enriching, improving, labeling, organizing, sampling, tuning and validating LULCC datasets inside big EO data architectures. However, the question on how could the knowledge of the crowd be concretely explored for improving LULCC monitoring with big EO data analysis remains to this date largely open. The main challenge is to support and promote an active scientific citizenship within big EO data architectures for LULCC monitoring.

To achieve this goal, it is crucial to understand the variety of opportunities and potentials (as well as the limitations) of crowdsourced geographic information initiatives (Assis, et al., 2018). Only then, the following tasks can be approached: 1) to engineer the data and business rules that could benefit big EO data analytics; 2) to enable time-effective, low-cost and user-friendly LULCC participatory monitoring and validation; and, 3) to evaluate the volunteers contributions and compare them to those from in-house experts within complex architectures for LULCC monitoring.

In this short paper, we briefly discuss the few existing architectures for big EO image data handling and the initiatives for leveraging LULCC monitoring with citizen science. Furthermore, we elaborate on the potential sinergies of combining big EO data analytics and citizen science for improving LULCC monitoring. Lastly, we mention possibly promising future research avenues.

2 Big Earth Observation Data Architecture

Due to the growing volume of data produced by new satellites, LULCC applications have triggered the need for a set of requirements designed to enable the developement of complex analyses. These architecture requirements include, but are not limited to, distributed storage and processing components, high performance computing, usability to other domain specialists, consistent and reliable analysis results, maintainability, testability, and interoperability. They revolve around interdisciplinary groups, heterogeneous data specification, and the core operations for the analysis (Pressman, 2005).

Currently, Google Earth Engine (GEE) is considered one of the most efficient user-friendly platform for high-performance analysis in a planetary-scale (Gorelick et al., 2017). GEE allows users to manipulate scripts and share their results without the technical capacities required for dealing with the complexity involved in this type of processing. It provides a set of intrinsically parallel functions on large geospatial datasets, based on the MapReduce paradigm (Dean and Ghemawat, 2010). However, GEE requires additional efforts to adapt computational tasks that don't fit into the MapReduce paradigm or to use data that is not already available. Alternatives to GEE include the MapReduce main open source implementation, Hadoop, and its spatial variations, namely, Hadoop-GIS (Aji et al., 2013) and SpatialHadoop (Eldawy and Mokbel, 2015). Although both platforms achieve high performance in batch processing jobs by supporting a set of spatial index structures, queries, data types and operations, they have the drawbacks of either being a black box or of producing an overwhelming amount of reads and writes on disk.

There are also solutions such as GeoMesa (a suite of big data tools built on top of Hadoop and column-family databases) that enables indexing, managing and analyzing both vector and raster data (Hughes et al., 2015). It integrates with Spark, but it does not deal with big spatial data in streaming mode as GeoSpark does (Yu, Wu and Sarwat, 2015). Although Geospark provides in-memory processing that outperforms most of the MapReduce-based platforms, preparing data as streams for processing burdens the users in comparison to other solutions. This is because users need to download the imagery via a pre-order request and the access is limited to a small number of files per request. Then, it is necessary to wait for the order to become available for download. Finally, the imagery should be organized and provided in a way that they can be analyzed as a ordered sequence of files (Assis, et al., 2016).

Another emergent solution aiming to facilitate the organization of the data are the multidimensional array databases. Exemplary databases such as Rasdaman (Baumann et al., 1998) and SciDB (Stonebraker, et al., 2013) highlight the efficiency of relying on the mathematical concept of 3-D array to manage EO satellite images. In this case, the images are arranged chronologically with the coordinates $<x, y, t>$ of a pixel corresponding respectively to its latitude, longitude and time (Camara, et al., 2016;

Planthaber, Stonebraker and Frew, 2012). This 3D array concept allows innovative spatio-temporal analyses. Data Cubes also represent multidimensional arrays in the geospatial context (Lewis et al., 2017). It aims to prepare large, heterogeneous, and complex remote sensing data directly for the analyses. The term ''analysis ready data'' was created based on the motivation to provide data with a minimum set of requirements to start the analysis (Giuliani et al., 2017).

3 Participatory Land-Use and Land-Cover Monitoring

Big EO data combined with the appropriate software architecture and the effective analysis methods are still not enough to achieve the desired improvement in LULCC monitoring. Subtle class differences (e.g. pastures and savannas), high class variability as well as classes typical of some regions are factors that frequently require more extensive and minucius sampling as well as local knowledge that remote sensing experts sometimes lack.

Manually collected data by non-specialists but local experts can provide valuable and reliable information for LULCC monitoring. Furthermore, this would reduce the burden on the staff of experts regarding sampling and the validation of LULCC products (Lillesand, Kiefer and Chipman, 2014). Such a channel for the mutual exchange of LULCC knowledge between citizens and scientists can benefit both parts by generating detailed and up-to-date LULCC information (Newman et al., 2012; Bonney et al., 2016). The development of such a channel requires taking into account the issues of data quality and consistency as well as the issues involved in environmental monitoring (Irwin, 1995). Open Street Map (OSM) is the most prominent example of a platform for the participatory production of geographic information. It enables local experts and engaged volunteers to contribute with geographic information that is frequently not available elsewhere and would be otherwise difficult to obtain. These volunteers understand that the task relates not only to feed the application, but also to improve both the application efficacy and their learning curve about the topic with experience.

More specifically regarding LULCC information, OSM LULC, LACO-Wiki, LandSense, Geopedia, AgroSense and Collect Earth are standard Web-GIS applications for managing LULCC information based on volunteers. Land-Cover Geo-Wiki proposes a collaborative manner to extend, validate and update LULCC information (Fritz et al., 2009; Fritz et al., 2011). They include citizen engagement requirements (Fritz, et al., 2012) and quality assurance methods (Comber et al., 2013; Foody et al, 2013).

Unfortunately, this and other similar platforms have limited data availability. This relates to the *want-to* vs. *have-to* dialetic. An user might be aware of what to do but not motivated for doing so. As a consequence, data availability varies a lot from place to place even in consolidated platforms such as OSM. Younger volunteers might feel stimulated by game-design elements and principles such as rewards, prestige and competition. Gamification concepts combined with high-tech volunteer-oriented deployments (e.g., Foldit, Mozak) can encourage more participants due to their intrinsic driven motivation. Furthermore, specific knowledge marathons are another way of stimulating a group of people for hours, days or even weeks to explore data, discuss new ideas and develop new designs. Social media, workshops and events offer an alternative way to share our own awareness about the fact, therefore, they play an important role for these approaches.

4 Citizen Science applied to Big Earth Observation Data Analytics

Big EO data architectures have made it possible that citizens contribute to LULCC monitoring through user-friendly and intuitive tools. More than providing valuable information for the understanding of the issues involved in environmental monitoring (Irwin, 1995; Miller-Rushing, Primack and Booney, 2012; See et al., 2016), citizens might create together with scientists a sustainable and democratic analysis process. With the appropriate training, accessibility and architecture design, they are able to collaborate with reliable local LULCC information and thus help the staff of experts in remote sensing applications which might even involve complex computational tasks and statistical analysis (Lillesand, Kiefer and Chipman, 2014; Schultz et al., 2017; Wan et al., 2017).

Crowdsourcers have indeed proven to be able to perform tasks of different degrees of complexity, from tagging images (Herfort et al., 2017) to collaboratively build websites (Paolacci, Chandler and Iperotis, 2010). Thus, the assistance of the crowd has been resourted to in a large variety of applications both in academia and in outsourced services. The potentially large amount of data produced by crowdsourcers is of particular relevance for data-driven analyses of more dynamic phenomena such as LULCC (Lary et al., 2016). Besides, the collection of some types of data by experts might be time and cost excessive, they sometimes possess a certain bias

towards the interests of specific campaigns and communities of experts. In the domain of LULCC, this is reflected by the existence of different classification systems (Friedl et al., 2010; Team, 2018; GlobCover, 2018). It can thus be argued that crowdsourced labeled LULCC data would be more in accordance with the local perceptions and therefore more usable to the local communities. This also enables experts to focus more on computational, structuring and gate-keeping tasks (Lease, 2011).

New big EO data architectures should enable, i.e., contain a mechanism for, the collection of data by individuals with less related expertise. They should strive to optimize the arrange of independent citizens with varying experiences and skill levels. In this regard, machine learning methods could be effectively applied. More specifically, a semi-supervised category of learning methods called active learning is particularly relevant. In active learning, the algorithm receives as input a large amount of data without labels. It then attempts to learn how to classify them by interactively requesting labeled samples from the least amount of (and most skilled) individuals or so-called "human annotators" (Schohn and Cohn, 2000). Active learning offers an adequate trade-off between computational resources and the most useful citizen-provided information. Thus, in the context of big EO data and citizen-assisted LULCC monitoring programs, regions with limited or lack of samples could be more accurately classified by means of this optimized and effective sampling.

5 Final Remarks and Future Work

We have briefly discussed in the previous sections the recent advances in crowdsourced geographic information production and big EO data architectures. Furthermore, we provided some initial ideas on how to combine these two worlds. A deeper discussion of these and other related ideas is the aim of our future work and should throw light on questions of very distinct nature: from engaging citizens (e.g., how to engage and motivate citizens to contribute as a way to ensure the sustainability of the project?), to project sustainability (e.g., how can citizens be assigned to tasks and their contributions managed in an intelligent manner as a way to optimize the effectiveness/data input relation?), to reliability of information (e.g., to which extent appropriate architecture designs should help citizens to contribute with reliable LULCC information?), to the use of an specific "technique" such as active learning (e.g., to what extent crowdsourced data may improve the accuracy of machine learning algorithms within big EO data architectures?), and passing by decision-making support (e.g., what are the possibilities and constraints in terms of remote sensing and auxiliary data inputs to a crowd-assisted LULCC monitoring system?).

Researchers and developers focusing on the above questions and more generally on combining big EO data architectures and crowdsourced geographic information for LULCC monitoring will face many challenges. However, this is a fresh ground for new ideas and therefore an exciting topic for a deeper discussion and a deeper literature review.

References

Aji, A., Wang, F., Vo, H., Lee, R., Liu, Q., Zhang, X. & Saltz, J. (2013), ‘Hadoop gis: a high performance spatial data warehousing system over mapreduce’, Proceedings of the VLDB Endowment 6(11), 1009–1020.

Assis, L., Queiroz, G., Ferreira, K., Vinhas, L., Llapa, E., Sanchez, A., Maus, V. & Camara, G. (2017), Big data streaming for remote sensing time series analytics using MapReduce, in ‘Proceedings of the XVII Brazilian Symposium on GeoInformatics’, Brazilian Journal of Cartography, Campos do Jordão, SP, Brazil.

Assis, L., Novack, T., Ferreira, K., Vinhas, L. & Zipf, A. (2018), Citizen science for big earth observation data analytics in land use and land cover change monitoring: From scope to future directions,in‘EGU General Assembly Conference Abstracts’, Vol. 20, p. 14638.

Baumann, P., Dehmel, A., Furtado, P., Ritsch, R. & Widmann, N. (1998), The multidimensional database system rasdaman, in ‘Acm Sigmod Record’, Vol. 27, ACM, pp. 575–577.

Bonney, R., Phillips, T. B., Ballard, H. L. & Enck, J. W. (2016), ‘Can citizen science enhance public understanding of science?’, Public Understanding of Science 25(1), 2–16.

Camara, 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, pp. 1–6.

Comber, A., Mooney, P., Purves, R., Rocchini, D. & Walz, A. (2015), ‘Comparing national differences in what people perceive to be there: mapping variations in crowdsourced land cover’, The

International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences 40(3), 71.

Comber, A., See, L., Fritz, S., Van der Velde, M., Perger, C. & Foody, G. (2013), 'Using control data to determine the reliability of volunteered geographic information about land cover', International Journal of Applied Earth Observation and Geoinformation 23, 37–48. 4

Dean, J. & Ghemawat, S. (2010), 'Mapreduce: a flexible data processing tool', Communications of the ACM 53(1), 72–77.

Eldawy, A. & Mokbel, M. F. (2015), Spatialhadoop: A mapreduce framework for spatial data, in 'Data Engineering (ICDE), 2015 IEEE 31st International Conference on', IEEE, pp. 1352–1363.

Foody, G. M., See, L., Fritz, S., Van der Velde, M., Perger, C., Schill, C. & Boyd, D. S. (2013), 'Assessing the accuracy of volunteered geographic information arising from multiple contributors to an internet based collaborative project', Transactions in GIS 17(6), 847–860.

Friedl, M. A., Sulla-Menashe, D., Tan, B., Schneider, A., Ramankutty, N., Sibley, A. & Huang, X. (2010), 'Modis collection 5 global land cover: Algorithm refinements and characterization of new datasets', Remote sensing of Environment 114(1), 168–182.

Fritz, S., McCallum, I., Schill, C., Perger, C., Grillmayer, R., Achard, F., Kraxner, F. & Obersteiner, M. (2009), 'Geo-wiki. org: The use of crowdsourcing to improve global land cover', Remote Sensing 1(3), 345–354.

Fritz, S., McCallum, I., Schill, C., Perger, C., See, L., Schepaschenko, D., Van der Velde, M., Kraxner, F. & Obersteiner, M. (2012), 'Geo-wiki: An online platform for improving global land cover', Environmental Modelling & Software 31, 110–123.

Fritz, S., See, L., McCallum, I., Schill, C., Obersteiner, M., Van der Velde, M., Boettcher, H., Havlík, P. & Achard, F. (2011), 'Highlighting continued uncertainty in global land cover maps for the user community', Environmental Research Letters 6(4), 044005.

Giuliani, G., Chatenoux, B., De Bono, A., Rodila, D., Richard, J.-P., Allenbach, K., Dao, H. & Peduzzi, P. (2017), 'Building an earth observations data cube: Lessons learned from the swiss data cube (sdc) on generating analysis ready data (ard)', Big Earth Data 1(1-2), 100–117.

GlobCover (2017), 'ESA Data User Element'. Available at: http://due.esrin.esa.int/page globcover.php.

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D. & Moore, R. (2017), 'Google earth engine: Planetary-scale geospatial analysis for everyone', Remote Sensing of Environment 202, 18–27.

Herfort, B., Reinmuth, M., de Albuquerque, J. & Zipf, A. (2017), Towards evaluating crowdsourced image classification on mobile devices to generate geographic information about human settlements, in 'Proceedings of the 20th AGILE'.

Hughes, J. N., Annex, A., Eichelberger, C. N., Fox, A., Hulbert, A. & Ronquest, M. (2015), Geomesa: a distributed architecture for spatio-temporal fusion, in 'Geospatial Informatics, Fusion, and Motion Video Analytics V', Vol. 9473, International Society for Optics and Photonics, p. 94730F.

Irwin, A. (1995), Citizen science: A study of people, expertise and sustainable development, Psychology Press.

Jokar Arsanjani, J., Helbich, M., Bakillah, M., Hagenauer,J. & Zipf, A. (2013), 'Toward mapping land-use pat-terns from volunteered geographic information', International Journal of Geographical Information Science27(12), 2264–2278.

Lary, D. J., Alavi, A. H., Gandomi, A. H. & Walker, A. L.(2016), 'Machine learning in geosciences and remotesensing', Geoscience Frontiers7(1), 3–10.

Lease, M. (2011), 'On quality control and machine learning in crowdsourcing.',Human Computation11(11).

Lewis, A., Oliver, S., Lymburner, L., Evans, B., Wyborn,L., Mueller, N., Raevksi, G., Hooke, J., Woodcock, R., Sixsmith, J. et al. (2017), 'The australian geoscience datacube—foundations and lessons learned', Remote Sensing of Environment202, 276–292.

Lillesand, T., Kiefer, R. W. & Chipman, J. (2014),Remote sensing and image interpretation, John Wiley & Sons.

Miller-Rushing, A., Primack, R. & Bonney, R. (2012), 'The history of public participation in ecological research', Frontiers in Ecology and the Environment 10 (6), 285–290.

Moorthy, I., Fritz, S., See, L. & McCallum, I. (2017), Land-sense: A citizen observatory and innovation marketplace for land use and land cover monitoring, in 'EGU General Assembly Conference Abstracts', Vol. 19, p. 8562.

Newman, G., Wiggins, A., Crall, A., Graham, E., Newman, S. & Crowston, K. (2012), 'The future of citizen science:emerging technologies and shifting paradigms',Frontiers in Ecology and the Environment 10 (6), 298–304.

Paolacci, G., Chandler, J. & Ipeirotis, P. G. (2010), 'Running experiments on amazon mechanical turk', Judgment and Decision Making5(5).

Planthaber, G., Stonebraker, M. & Frew, J. (2012), Earthdb: scalable analysis of modis data using scidb, in 'Proceedings of the 1st ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data', ACM,pp. 11–19.

Pressman, R. S. (2005), Software engineering: a practitioner's approach, Palgrave Macmillan.

Schohn, G. & Cohn, D. (2000), Less is more: Active learning with support vector machines, in 'Proceedings of the Seventeenth International Conference on Machine Learning', Morgan Kaufmann Publishers Inc., pp. 839–846.

Schultz, M., Voss, J., Auer, M., Carter, S. & Zipf, A. (2017),'Open land cover from openstreetmap and remote sensing', International Journal of Applied Earth Observation and Geoinformation 63, 206–213.

See, L., Mooney, P., Foody, G., Bastin, L., Comber, A.,Estima, J., Fritz, S., Kerle, N., Jiang, B., Laakso, M. et al.(2016), 'Crowdsourcing, citizen science or volunteered geographic information? The current state of crowdsourced geographic information', ISPRS International Journal of Geo-Information 5 (5), 55.

See, L., Perger, C., Hofer, M., Weichselbaum, J., Dresel,C. & Fritz, S. (2015), 'Laco-wiki: an open access on-line portal for land cover validation. ',ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences 2.

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.

Team, G.(2017), 'GeoNetwork open source portal to spatial data and information'. Available at: http://www.fao.org/geonetwork/srv/en/main.home?uuid=ba4526fd-cdbf-4028-a1bd-5a559c4bff38.

Wan, T., Lu, H., Lu, Q. & Luo, N. (2017), 'Classification of high-resolution remote-sensing image using open-street map information', IEEE Geoscience and Remote Sensing Letters.

Wang, L., Ma, Y., Yan, J., Chang, V. & Zomaya, A. Y.(2018), 'pipscloud: High performance cloud computing for remote sensing big data management and processing',Future Generation Computer Systems 78, 353–368.

Yu, J., Wu, J. & Sarwat, M. (2015), Geospark: A cluster computing framework for processing large-scale spatial data, in 'Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems', ACM, p. 70.