Big Earth Data: From Data to Information

GI_Forum 2018, Issue.1
Page: 184 - 193
Full Paper
Corresponding Author:
stefan.lang@sbg.ac.at
DOI: 10.1553/giscience2018_01_s184

Martin Sudmanns, Stefan Lang and Dirk Tiede
University of Salzburg, Austria

Abstract

While we still lack a community-agreed definition of 'big Earth data', there is clear evidence that the data's unprecedented volume, variety and velocity, as well as veracity, require changes to traditional ways of storing and analysing Earth observation (EO) data and sharing their value with the public. Alongside the challenges, opportunities also arise when continental or global-scale analyses become readily feasible, or when time series / time lapse analyses provide unprecedented insights. This contribution presents a broad overview of current trends, limitations and opportunities for an increased use of larger volumes of EO data, while having in mind that the main objective is to use big Earth data to foster the fifth 'V', namely value. The data become valuable if they can be transformed into information which matches specific contexts. The overview on big Earth data is complemented by specific implementations and use cases.

Keywords:

data cube, multi-temporal analytics, Sentinel, Copernicus, earth observation

1 Big Earth data as a recent trend in Earth observation

The ever-increasing volume, variety and velocity of satellite images, acquired by a variety of Earth observation (EO) acquisition platforms, can be valuable for numerous application areas or even extended into completely new ones. The unprecedented combination of high spatial and high temporal resolutions in particular allows for new approaches to solve spatial problems on Earth. Applications of these advances include shorter response times in reacting to land-use changes over large areas, such as (illegal) deforestation of rain forest (Maus et al., 2016); observation of unclear and confusing city structures and changes therein in Asia, Africa and South America (Bachofer, 2017); or prompt reaction to natural and humanitarian disasters such as earthquakes and floods (Lang, Schoepfer, Zeil, & Riedler, 2017; Schwarz et al., 2018). The European Copernicus programme is one of the main drivers of this expansion of EO data, illustrated in Figure 1, which shows the increase of users registered to access satellite data via the ESA portals. However, the latent message here is that not only have the numbers of users increased, but also the diversity of application domains. While arguably most of the users in the pre-Copernicus era were EO experts, EO data is increasingly being used by non-EO experts in fields such as Biology, Ecology, Geology or Marine Science.

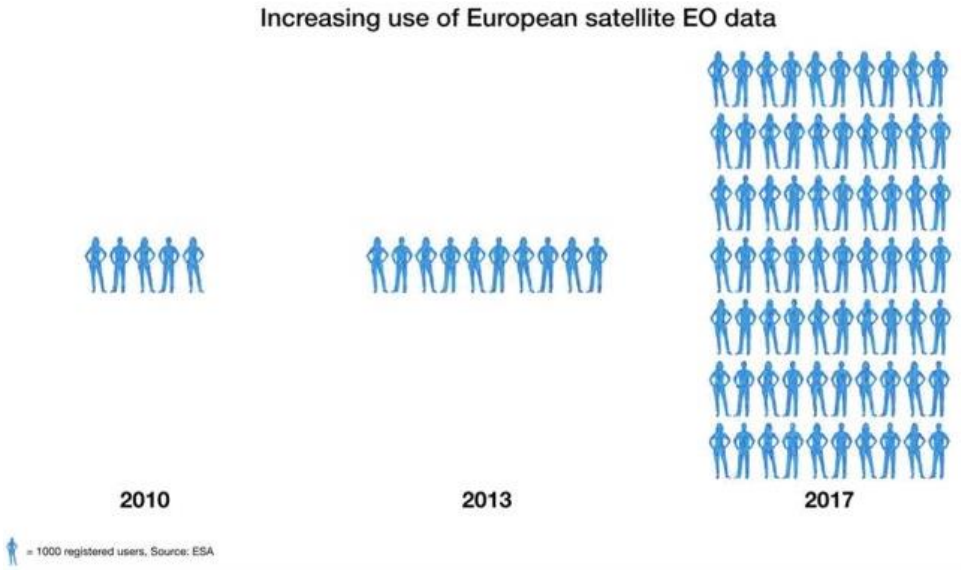

Figure 1: Increasing use of European satellite EO data, documented by registered users (Source: ESA, created and shared by isardSAT).

With wider interest in, and the increasing volume, variety and velocity of, EO data, new challenges arise. What is the best way to store and manage data, and how can it be made accessible for end-users without the need for downloading? How can we produce useful information systematically from the content of an image, or even a long time series of images? How can EO data from different sensors be combined, or how can they be combined with non-EO data? Arguably, in most analyses only a small portion of the available images are currently being used. There are several possible reasons for this. The prevalent requirement is still for data to be downloaded prior to analysis, partly because local workflows cannot be easily translated into Web-based workflows. Other reasons are the tools and computation capabilities that are simply not powerful enough yet, and the lack of interoperability and transferability of approaches.

If we take a closer look, we see that there is no community-agreed definition of what 'big Earth data' actually is and whether the term describes the phenomenon appropriately. Synonyms might be 'big EO data', 'big environmental data' or 'big data in remote sensing'. However, newly emerging trends in how data volumes are provided and processed, and the modifications to associated methods of analysis, are evident. The open access policy of the Landsat and Sentinel programmes in particular (European Commission, 2013; Wulder & Coops, 2014) seems to be a key driver for putting the term 'big Earth data' on the agenda (Baumann et al., 2016; Guo, 2017; OGC, 2017). In the optical EO domain, the Copernicus Sentinel-2 satellites are specifically designed to deliver high resolution (HR) imagery as a key information source of the European Copernicus programme, which collects significantly more data than any comparable initiative, past or present. Under the present acquisition plan, each of the two currently operational Sentinel-2 satellites produces more than 1.7 Terabytes (Tb) of data per day for the 1C processing level (ESA, 2017), as shown in Figure 2. This

translates into several hundred images every day. Since the beginning of the exploitation phase of Sentinel-2A in mid-2015, more than 3.7 million images have been acquired (as at March 2018). The Sentinel-2 constellation covers the vast majority of the Earth's land surface with dual satellites (2A / 2B, 180° apart), a broad swath-width of up to approximately 290 km, a 5-day revisiting time at the equator, which is even more frequent at higher latitudes, and 13 multispectral bands ranging from 10m to 60m spatial resolution (Drusch et al., 2012). A comparable increase of data volumes and velocities in the radar domain is provided by the Sentinel-1 satellite constellation.

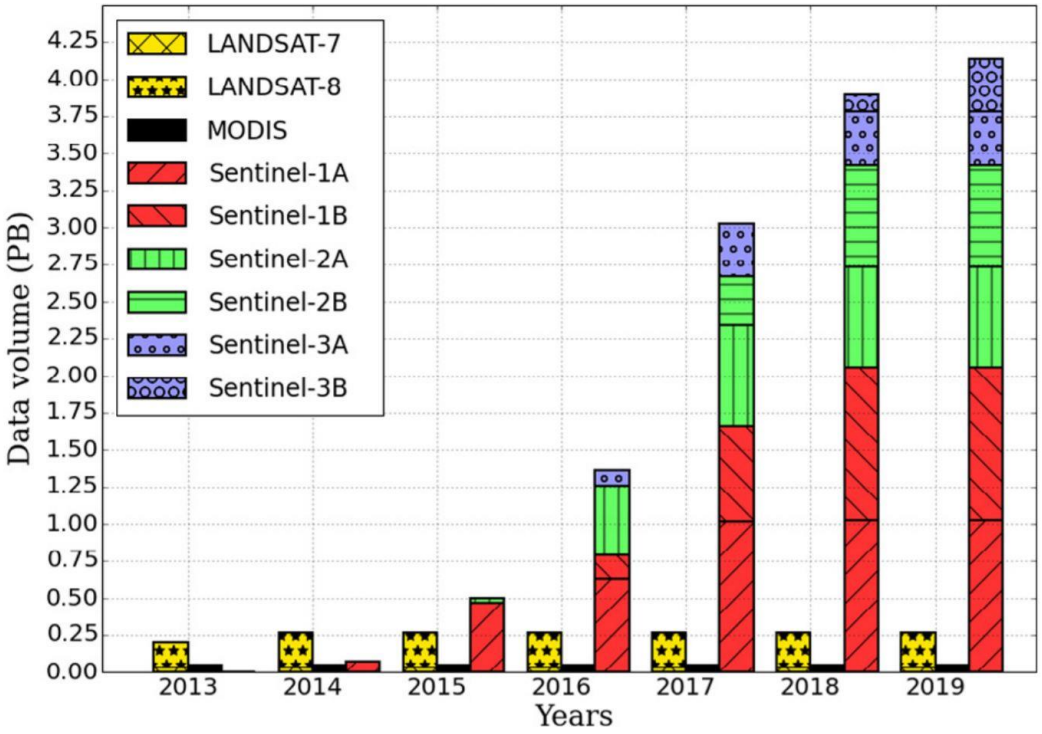

Figure 2: Yearly data volume (from Soille et al., 2018 without changes).

One of the main challenges for the community will be to systematically derive meaningful information from the data. This is indeed not an easy task in non-big Earth data domains and gets even more complicated, regarding automation and accuracy, for the large amount of data we face. We also need to define what data is and what information is. Further, this overarching goal is closely tied to technological developments, innovations in the application domains, and interdisciplinary research in the fields of remote sensing and computer vision, GIScience, computer science, and human-computer interactions.

2 Big data, big discussions

Challenges and opportunities

As for any new technology, the direction(s) and magnitude(s) that follow from the initial developments are unclear. We can infer from observations of other developments, e.g. smartphones or social networks, that there are many opportunities, but also challenges which need to be tackled. This is especially the case if expectations are high, such as using EO technology to help solve global problems. Table 1 illustrates some characteristics of big Earth data and their potential impacts, in terms of both challenges and opportunities.

Table 1: Characteristics of big Earth data and the potential impact.

Characteristics	Potential impact
The data are open access.	Continental or even global analyses are possible. Business models based on EO data can be established.
The data are not collected for a single purpose, but **are a constant stream.**	Long time-series analyses are possible. Knowledge about spatio-temporal characteristics of phenomena prior to data collection is less important. Data quality and reliability become even more important.
The volume of data is so huge that downloads **are cumbersome or even no longer possible.**	Remote processing emerges together with cloud computing and sharing economy (of data and processing capabilities). Data provider becomes platform provider for services. Questions of data security and safety arise. Smaller companies become competitive as they can rent processing capabilities and can use market places for selling their products and services.
Acquisition frequency increases up to several days for high-resolution images.	Phenomena with inherent high temporal changes can be observed. The actual date of a change can be determined more accurately. Several application domains can make better use of EO data, e.g. urban development, insurance companies or agriculture.
Time between acquisition and data provision is reduced to several hours	Near real-time analyses become possible. Requirements of new application domains, e.g. for disaster management, can be satisfied.
Image archives date back some decades (e.g. MODIS, Landsat).	Analyses of long-term trends are possible, e.g. to study the impact of climate change or human-induced changes in urban developments. New technologies emerge which can handle the temporal domain better, e.g., data cubes.
Analysts have access to more computing resources.	Larger study areas and time-series analysis can be considered. Machine-learning approaches are increasingly being used.

The main opportunity comes through the increased spatio-temporal coverage of EO data. This opens new analysis fields, e.g. monitoring specific global trends in the context of the UN Sustainable Development Goals (SDGs), or long time-series analysis with temporal resolutions of a few days. Application domains which deal with impacts of climate change, monitoring of natural resources, or human impact benefit from the additional evidence provided by EO analyses. Challenges may arise with the use of big Earth data, including questions of data storage and management, data quality and reliability, whether methods are transferable and can be applied on a continental or even global scale, or how the results can be interpreted and whether they can be trusted. The impact of the increased volume of data and the availability of massive computing power in cloud environments can as yet only be surmised vaguely, not forecast or calculated. These are just some of the many open questions, which will be looked at briefly in the sections below.

Production of information from images

In contrast to other domains, where the data volume and especially the velocity are challenges in their own right, in Earth observation, the raw data need first to be transformed into information. The process of transforming data into information includes many pre-processing steps, such as radiometric calibration and correction, or format conversions. But we need to distinguish clearly between the terms data and information, which are not as sharply differentiated as they might seem to be at first glance. In the case of optical EO, data are the raw images, delivered for example on processing Level 1C; to be considered as information, data need to be interpreted to some degree. Contrary to what the common terminology might suggest, information does not simply 'sit' inside data from which it can simply be 'extracted'. Hence, the term 'information production' seems to be more accurate than 'information extraction'.

In the context of image understanding, it needs to be understood that an image is a 2D representation of the physical 4D world (Marr, 1982). Therefore, image understanding aims at reconstructing a scene from a 2D image to allow questions to be answered and information to be produced from it. We are referring here explicitly to the arrangement of physical objects and their measurable attributes. The data interpretation process requires a framework of prior knowledge on the existence and arrangements of the physical objects that compose the image and to which the observations relate.

Advent of new technologies and applications

Alongside the term 'big Earth data', several technologies have emerged or been transferred from other domains. While an exhaustive review goes beyond the scope of this contribution, some of the most salient developments can be mentioned by way of examples. Most prominently, the data cube concept can be seen as the backbone of modern big Earth data analytics (Baumann et al., 2016; Nativi, Mazzetti, & Craglia, 2017; Wagemann, Clements, Marco Figuera, Rossi, & Mantovani, 2018), together with map-reduce-based solutions such as Google Earth Engine (Gorelick et al., 2017). Furthermore, machine-learning technologies, e.g. convolutional neural networks (CNN) (Long, Shelhamer, & Darrell, 2015), are increasingly being applied to EO images (Längkvist, Kiselev, Alirezaie, & Loutfi, 2016).

Data cubes have long been familiar from data warehouse technologies, where attributes are indexed by coordinates in different (including non-spatial) dimensions. However, in contrast to EO data cubes, other types of data cubes are sparsely populated and can be represented in relational databases. An EO data cube stores data in two or three spatial dimensions and at least one non-spatial dimension, e.g. time. A definition of a data cube is provided by the data cube manifesto (Baumann, 2017):

> *'A datacube is a massive multi-dimensional array […]; "massive" entails that we talk about sizes significantly beyond the main memory resources of the server hardware. Data values, all of the same data type, sit at grid points as defined by the d axes of the d-dimensional datacube. Coordinates along these axes allow addressing data values unambiguously.'*

Operational examples for larger implementations of data cubes are Digital Earth Australia (DEA), which evolved from the Australian Geoscience Data Cube (Lewis et al., 2017), the Swiss Data Cube (SDC) (Giuliani et al., 2017), and the EarthServer (Baumann et al., 2016). While DEA and SDC are based on the technology of the Open Data Cube Initiative (ODCI), the EarthServer uses the Rasdaman array database system (Baumann, Dehmel, Furtado, Ritsch, & Widmann, 1998) to manage and query the data.

Deep learning technologies, such as CNN, are increasingly being used to identify the content in EO images. An example can be found in Long et al. (2015). Implementations range from simply attaching labels to whole images to marking image objects using dense semantic segmentation, in which every pixel is classified. However, there are also critical voices, those of people who are not satisfied with the results or the approach and criticize it as a 'black box' (Marcus, 2018). Physical model-based approaches exist that rely not on learning but on explicit structural knowledge; they can be seen as an alternative or complementary effort (Baraldi et al., 2010).

While there is and will continue to be ongoing discussion about the best technological foundation and its implementation, there seems little doubt that big Earth data fosters new application areas and research questions. For example, Tiede, Baraldi, Sudmanns, Belgiu, & Lang (2017) presented an approach for semantic content-based image retrieval and semantic analysis of EO images directly at the database level. Pekel, Cottam, Gorelick, & Belward (2016) mapped long-term changes of surface water on a global scale, and Lewis et al. (2017) used the AGDC to map the water dynamics of the whole of Australia. Hansen et al. (2013) mapped forest cover change on a global scale. These are just a few examples of new approaches and categories of geospatial information products. They can be used in operational applications, such as monitoring the achievements of SDGs or calculating essential climate variables as inputs for climate models.

3 The potential of big Earth data applications

The concept of 'from data to information' outlined above encompasses the whole workflow but also stresses the need to produce reliable, global, multi-temporal, geospatial information from big Earth data to unfold the data's potential, e.g., in science or political decision-making. Just a few of the research domains concerned with big Earth data, then, are:

- EO big data handling and data selection
- Semantics / feature and information extraction
- Innovative concepts of time-series analysis
- Remote processing platforms and connections to public and private big data initiatives
- Machine learning, artificial intelligence and knowledge-based systems

Many contributions to the issues (1 and 2) of GI_Forum 2018 address the topic 'from data to information'. They range from mapping natural resources such as grassland (Bekkema & Eleveld) or forests (Korman et al.; Peters, Liu, Bruce, O'Hehir, & Li) or environmental changes in general (Augustin, Sudmanns, Tiede, & Baraldi), to assessing settlement damage in war zones (Braun). Specific topics of image analysis workflows are covered: data management using data cubes (Augustin et al.), transferable models for image classification (Wolfe, Jin, & Bahr), and time-series analyses (Augustin et al.; Braun; Peters et al.). Multiple data sources are considered, ranging from high-resolution optical satellite images, in particular from Sentinel-2 (Augustin et al.; Korman et al.), to multi-temporal LiDAR data (Peters et al.) and radar data (Braun).

More precisely, Peters et al. develop a framework for automatic LiDAR data processing and the prediction of timber yield. They aim to update an existing forest inventory based on airborne laser scanning (ALS) using UAV-based LiDAR. Wolfe et al. demonstrate the usefulness of visual modelling tools for complex classification tasks using Softmax Regression or a Support Vector Machine (SVM) classifier. In the Sentinel-2 HR domain, Bekkema and Eleveld have developed a new method to map and assess grassland management intensity using the C5.0 univariate decision tree (DT) algorithm and report on the suitability of Sentinel-2 data for this task. Augustin et al. report on the development and setting-up a of a semantic data cube using the Open Data Cube for semantic queries and analyses through time, presenting surface water dynamics during the Syrian conflict as a case study. Korman et al. estimate the forest above-ground biomass (AGB) and its temporal and spatial variation in Croatia in 2016 using Sentinel-2, Braun uses a time series of Sentinel-1 radar imagery for the identification of changes resulting from armed combats in the city of Raqqa, Syria.

4 Conclusion and outlook

The topic of big Earth data has not lost its actuality and has even increased constantly in the last 40 years, boosted by the free and open data policy for HR Landsat and Sentinel data. Since the launch of the first civil satellites, problems of handling large volumes of EO data have always been present for the remote sensing community: in one of the early remote sensing textbooks published forty years ago, Barrett and Curtis (1978), we find a sub-section entitled 'Problems of handling large quantities of remote sensing data' (p. 318). However, the message should not be a discouraging one, since – in tandem with the increasing computing capacity available over the years – the community has already developed and implemented a wide spectrum of algorithms and tools to tackle the challenges. Today's tools are far more sophisticated than earlier ones, which included simply reducing the data volume by lossy generalization, compression or even deleting data. Some of the new tools, like data cubes,

developments in artificial intelligence, and algorithms for efficient time series analysis, represent the cutting edge of the field.

The kaleidoscope of different topics, technologies and solutions presented in this issue of GI_Forum is impressive, providing a snapshot of what is a very dynamic and lively domain. And it may well be outdated in the near future due to the speed of technical developments. Nevertheless, when looking back we can identify a general shift in 'big Earth data'-related topics, from rather technical and practical questions about how to transmit and store data to questions about how to produce information from the available data and thus generate value from it. As the remote sensing community and those working in related domains (such as computer science) which have influenced the evolution of remote sensing have made significant progress towards providing solutions to the first set of questions, we may confidently expect the second set of questions to be tackled as well – namely, how to get from data to information.

References

Augustin, H., Sudmanns, M., Tiede, D., & Baraldi, A. (2018). A Semantic Earth Observation Data Cube for Monitoring Environmental Changes during the Syrian Conflict. GI_Forum 2018, 1.

Bachofer, F. (2017). Assessment of building heights from pléiades satellite imagery for the Nyarugenge sector, Kigali, Rwanda. Rwanda Journal, 1(1S). https://doi.org/10.4314/rj.v1i2s.6d

Baraldi, A., Durieux, L., Simonetti, D., Conchedda, G., Holecz, F., & Blonda, P. (2010). Automatic Spectral-Rule-Based Preliminary Classification of Radiometrically Calibrated SPOT-4/-5/IRS, AVHRR/MSG, AATSR, IKONOS/QuickBird/OrbView/GeoEye, and DMC/SPOT-1/-2 Imagery--Part I: System Design and Implementation. IEEE Transactions on Geoscience and Remote Sensing, 48, 1299–1325.

Barrett, E. C., & Curtis, L. F. (1978). Introduction to environmental remote sensing (Reprinted.). Science paperbacks: Vol. 117. London: Chapman and Hall.

Baumann, P. (2017). The Datacube Manifesto. Retrieved from EarthServer website: http://earthserver.eu/tech/datacube-manifesto

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

Baumann, P., Mazzetti, P., Ungar, J., Barbera, R., Barboni, D., Beccati, A., and others. (2016). Big data analytics for earth sciences: The EarthServer approach. International Journal of Digital Earth, 1–27. https://doi.org/10.1080/17538947.2014.1003106

Bekkema, M., & Eleveld, M. E. (2018). Mapping Grassland Management Intensity Using Sentinel-2 Satellite Data. GI_Forum 2018, 1.

Braun, A. (2018). Assessment of building damage in Raqqa during the Syrian civil war with time-series of radar satellite imagery. GI_Forum 2018, 1.

Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., . . . Bargellini, P. (2012). Sentinel-2: ESA's Optical High-Resolution Mission for GMES Operational Services. Remote Sensing of Environment, 120, 25–36. https://doi.org/10.1016/j.rse.2011.11.026

ESA. (2017). Sentinel High Level Operations Plan (HLOP): COPE-S1OP-EOPG-PL-15-0020 (No. 2). Frascati. Retrieved from https://earth.esa.int/documents/247904/685154/Sentinel_High_Level_Operations_Plan

European Commission. (2013). Commission Delegated Regulation (EU) No 1159/2013 of 12 July 2013 supplementing Regulation (EU) No 911/2010 of the European Parliament and of the

Council on the European Earth monitoring programme (GMES) by establishing registration and licensing conditions for GMES users and defining criteria for restricting access to GMES dedicated data and GMES service information. Retrieved from http://data.europa.eu/eli/reg_del/2013/1159/oj

Giuliani, G., Chatenoux, B., Bono, A. de, Rodila, D., Richard, J.-P., Allenbach, K., . . . 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, 14(2), 1–18. https://doi.org/10.1080/20964471.2017.1398903

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. https://doi.org/10.1016/j.rse.2017.06.031

Guo, H. (2017). Big Earth data: A new frontier in Earth and information sciences. Big Earth Data, 1(1-2), 4–20.

Hansen, M. C., Potapov, P. V., Moore, R., Hancher, M., Turubanova, S. A., Tyukavina, A., . . . Townshend, J. R. G. (2013). High-resolution global maps of 21st-century forest cover change. Science, 342(6160), 850–853. https://doi.org/10.1126/science.1244693

Korman, D., Berta, A., Mesić, Z., Ziza, I., Jantol, N., Križnjak, D., . . . Kušan, V. (2018). Handling And Extraction Of Sentinel 2 Data For The Comparison Of Seasonal And Monthly Models For Biomass Assessment - Case Study: Continental First Age Class Forests And (Sub-) Mediterranean Thickets And Maquis In Croatia. GI_Forum 2018, 1.

Lang, S., Schoepfer, E., Zeil, P., & Riedler, B. (2017). Earth Observation for Humanitarian Assistance. GI_Forum 2017, 1, 157–165.

Längkvist, M., Kiselev, A., Alirezaie, M., & Loutfi, A. (2016). Classification and Segmentation of Satellite Orthoimagery Using Convolutional Neural Networks. Remote Sensing, 8(4), 329.

Lewis, A., Oliver, S., Lymburner, L., Evans, B., Wyborn, L., Mueller, N., . . . Wang, L.-W. (2017). The Australian Geoscience Data Cube -- Foundations and lessons learned. Remote Sensing of Environment, 202, 276–292. https://doi.org/10.1016/j.rse.2017.03.015

Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 3431–3440).

Marcus, G. (2018). Deep Learning: A Critical Appraisal. arXiv preprint arXiv:1801.00631.

Marr, D. (1982). Vision: A computational investigation into the human representation and processing of visual information.

Maus, V., Câmara, G., Cartaxo, R., Sanchez, A., Ramos, F. M., & Queiroz, G. R. de. (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.

Nativi, S., Mazzetti, P., & Craglia, M. (2017). A view-based model of data-cube to support big earth data systems interoperability. Big Earth Data, 1(1-2), 75–99.

OGC. (2017). Big Geospatial Data – an OGC White Paper. Retrieved from http://docs.opengeospatial.org/wp/16-131r2/16-131r2.html

Pekel, J.-F., Cottam, A., Gorelick, N., & Belward, A. (2016). High-resolution mapping of global surface water and its long-term changes. Nature, 540(7633), 418–422. https://doi.org/10.1038/nature20584

Peters, S., Liu, J., Bruce, D., O'Hehir, j., & Li, J. (2018). Cost efficient estimates of (radiata pine) wood volumes using multi-temporal LiDAR data - an approach for operational forestry. GI_Forum 2018, 1.

Schwarz, B., Pestre, G., Tellman, B., Sullivan, J., Kuhn, C., Mahtta, R., . . . Hammett, L. (2018). Mapping Floods and Assessing Flood Vulnerability for Disaster Decision-Making: A Case Study Remote Sensing Application in Senegal. In P.-P. Mathieu & C. Aubrecht (Eds.), Earth

Observation Open Science and Innovation (pp. 293–300). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-65633-5_16

Soille, P., Burger, A., Marchi, D. de, Kempeneers, P., Rodriguez, D., Syrris, V., & Vasilev, V. (2018). A versatile data-intensive computing platform for information retrieval from big geospatial data. Future Generation Computer Systems, 81, 30–40. https://doi.org/10.1016/j.future.2017.11.007

Tiede, D., Baraldi, A., Sudmanns, M., Belgiu, M., & Lang, S. (2017). Architecture and prototypical implementation of a semantic querying system for big Earth observation image bases. European journal of remote sensing, 50(1), 452–463. https://doi.org/10.1080/22797254.2017.1357432

Wagemann, J., Clements, O., Marco Figuera, R., Rossi, A. P., & Mantovani, S. (2018). Geospatial web services pave new ways for server-based on-demand access and processing of Big Earth Data. International Journal of Digital Earth, 11(1), 7–25.

Wolfe, J., Jin, X., & Bahr, T. (2018). Creating Models Of Custom Image Classification Workflows Using Softmax Regression And Support Vector Machine. GI_Forum 2018, 1.

Wulder, M. A., & Coops, N. C. (2014). Make Earth observations open access. Nature, 513(7516), 30–31. https://doi.org/10.1038/513030a