Automated cloud coverage analysis with Brazil Data Cube

Thainara Lima[1], Rômulo Marques [1], Ueslei Sutil[1], Cláudia Almeida[1], Claudio Barbosa[1], Thales Körting[1], Gilberto Queiroz[1]

[1]Instituto Nacional de Pesquisas Espaciais Avenida dos Astronautas, Jardim da Granja, São José dos Campos, SP, Brazil, CEP 12227010

thainara.lima@inpe.br, ueslei.sutil@inpe.br, mr.romulomarques@gmail.com, claudia.almeida@inpe.br, claudio.barbosa@inpe.br, thales.korting@inpe.br, gilberto.queiroz@inpe.br

Abstract. *This paper approaches the topic of Data Cubes applied to the management and monitoring of the brazilian land surface. It presents the results of an instrumental work that aimed at developing a tool in Python language, capable of processing raster and vector data from cloud masks of the Brazil Data Cube STAC catalog, with the purpose of generating statistical assessments of the cloud coverage of a given place in a given period of time. The program was developed in the Jupyter Notebook environment, and hence, it is an open-source software development.*

1. Introduction

Earth Observation (EO) data retrieved from space platforms have exceeded the petabyte-scale, and nowadays some of them are freely and openly available from different data repositories, allowing a better scientific understanding and deeper knowledge of our planet´s biosphere and its limits (Giuliani et al., 2019). Handling and exploring big EO data pose a number of issues in terms of volume, velocity and variety, which requires a change of standard from traditional data-centric approaches, in order to face the challenges caused by data magnitude and management (Nativi et al., 2017).

To tackle the barriers between analyst's expectation and big data access, researchers started to develop EO Data Cubes (EODC) as a new paradigm that provides solutions for the storage, organization, management, and analysis of big EO data (Baumann et al., 2019). A data cube is generated from the collection of satellite images, properly co-registered, which are subject to a pan-sharpening operation in order to attain a better quality, and then cropped according to the user´s needs, based on the database grid system. These data, organized in space and time, can be used for various purposes of management and observation of the Earth's surface (FERREIRA, et al., 2020).

One of these applications is the monitoring of the cloud coverage in images relying on the cloud mask information. These masks, included in the Brazil Data Cube (BDC) repository, make it possible to evaluate, according to Polidorio et al. (2005), eventual interferences in the images, which reduce the radiometric responses or cause the complete occlusion of features, either by the clouds themselves or by the projected shadows.

Lucena [et al. 2020] developed a tool for visualization of cloud coverage implemented in Brazil Data Cube, called Brazil Data Cube Cloud Coverage (BDC3) viewer. This tool was developed to visualize cloud coverage information based on the Spatio Temporal Asset Catalog (STAC) specification, such as seasonal cloud coverage average, total annual cloud coverage, scene, area or period with maximum or minimum cloud coverage, and cloud coverage timeseries.

In this context, aiming to contribute to the work presented by Lucena [et al. 2020], we present a prototype of a tool for visualizing information about cloud coverage, considering a given area of interest, as well as a temporal period, such as time series of the percentage of cloud cover, the average rate of cloud occurrence, the number of cloud-free images, scene and period of images with cloud cover below a given threshold, area or period with maximum cloud coverage.

2. Brazil Data Cube (BDC)

Since January 2019, the BDC project is being developed by Brazil's National Institute for Space Research (INPE). The project's main objective is to create multidimensional data cubes of analysis-ready from EO images for all Brazilian territory, generate land use and land cover information, as well as satellite image time series analysis (BDC, 2021). In the data cube generating process, two types of cubes are created: identity data cubes and regular data cubes. The identity data cube uses all available images from a single sensor, without applying a temporal compositing function, keeping all available images with their original acquisition dates. On the other hand, the regularly spaced data cubes are created using functions for temporal compositing (monthly or every 16 days): median, average, and stack. The stack compositing function is also called the *best pixel* approach, where it consists of ranking the time step images selecting the observation from the best-ranked image (Ferreira *et al.,* 2020).

Inside BDC, the metadata about the data cubes are stored in a relational database called STAC (Spatial Temporal Asset Catalog). The language is an open specification one, based on JSON and RESTful, that was created to increase the interoperability of searching for geospatial data, including satellite imagery (Ferreira et al., 2020). The images in the STAC are organized in hierarchical levels (catalog → collections → item → assets), where the cloud mask is available accessing the assets of the cube. The Catalog Specification provides structural elements to group Items and Collections. It is important to notice that Collections are Catalogs, but with required metadata and description of a group of related Items. The Item object represents a unit of data and metadata, representing a single scene of data at a given place and time, and includes Asset links, to enable direct access or download of the asset. The Asset is any file that represents information about the Earth captured in a certain space and time.

3. Methodology

The Data Cube Collection from Landsat-8/OLI was considered in this prototype. The choice of the collection was based on the spatial coverage of the cube and information availability. Here, we considered the identity data cube instead of the regular data cube. The regular data cube uses the best pixel composition, so, in order to obtain a faithful cloud information from the satellite image, the identity cube was considered. It is important to emphasize that the work presented refers to a prototype, and therefore, other collections may be considered in the future.

Based on each specified collection, the cloud viewer tool is being implemented in the BDC project using the BDC-STAC, which allows access to information about metadata and satellite imagery. For each collection, the cloud mask presents specific pixel values to represent cloudy areas. In Landsat-8/OLI, e.g., in addition to cloud and non-cloud information, there is also shadow information in the image. Despite this, these prototype was developed focus only on cloud detection.

Based on the general specifications of the BDC-STAC, which are organized inside the Python library called stac.py, the cloud viewer tool proposes an extension for visualization of cloud coverage statistics information. The tool allows user interaction, and the cloud information can be obtained considering specific study areas, delimited through a GeoJSON file. Figure 1 shows a flowchart of our prototype.

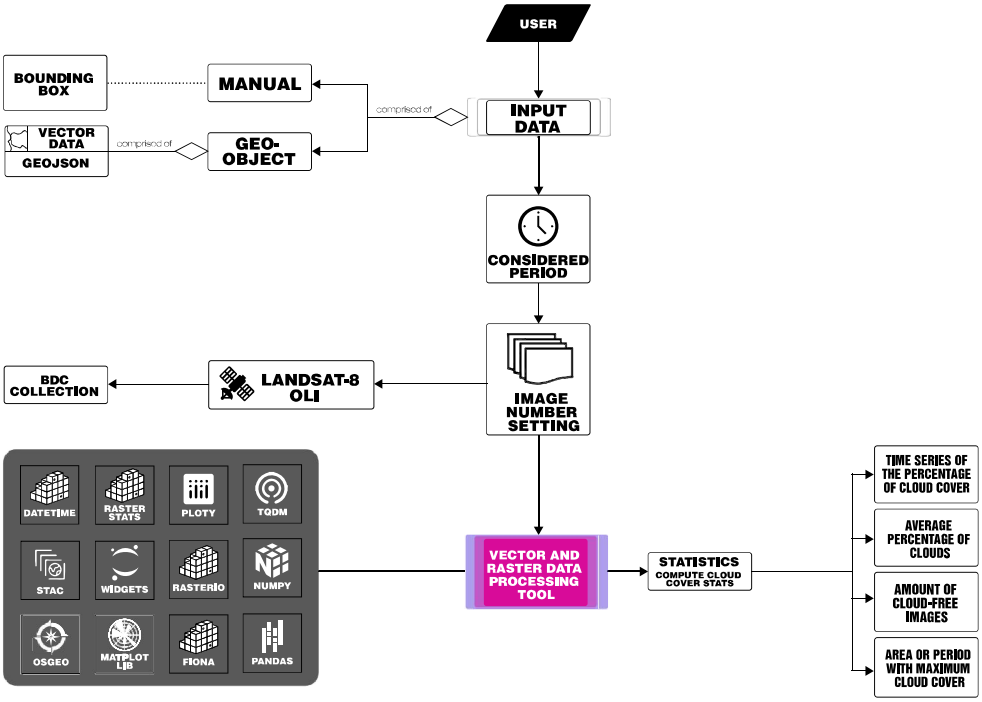

Figure 1. Architecture flowchart.

As it is shown in the flowchart, the user enters as input information: the collection, considered period, and the delimitation of the study area, which can be assessed through the coordinates of the bounding box or the GeoJSON vector data. From these specifications, the tool accesses the cloud masks through BDC-STAC by applying a series

of libraries in Python (see the flowchart in Figure 1). From the cloud masks, the user can obtain different statistical information: (a) cloud cover percentage timeseries; (b) average cloud percentage; (c) number of cloud-free images; (d) scene, area or period with maximum cloud coverage. In addition to the visual information, the tool allows the user to visualize the cloud mask (with coverage less than 20%, for instance), and the grouped generated data will be available to users for download.

4. Results and Discussion

In order to illustrate the application of the tool, a case of study is presented, considering as study area the city of São José dos Campos throughout the collection processing period (January/2016 – March/2021). We analyzed our test case from for the Landsat collection. From 417 images, 170 images were available, and 247 were invalid (with only null values). Figure 2 shows the percentage of cloud cover for all available images from the Landsat collection. The data shows a large variation in cloud cover in São Jose dos Campos, where it is possible to notice a seasonal variation considering the oscillation in rainfall, with greater intensities from September to March, and lower intensities from June to August, a period regarded as of reduced rainfall.

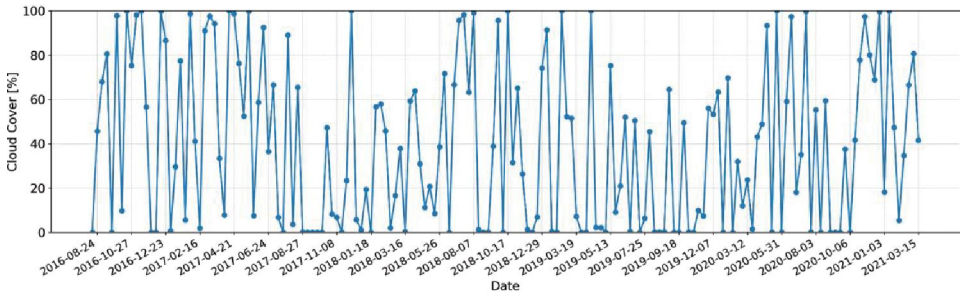

Figure 2. Cloud cover percentage from January/2016 to March/2021.

It is possible to observe in Table 2, running an average filter through the data, that the average cloud cover in São José dos Campos is 39.24%, and the widest variation in the data, from 0% to 99.99%, was observed on April 14, 2017 (Table 2). With our ptototype, using previous information, it is possible to obtain images from the Landsat collection without clouds, as well as the images with less than 20% of cloud coverage. Besides indicating the number of images with 0% or lower than 20% of clouds, it is important to notice that our prototype also presents the identification of the images and allows the user to download them.

Table 2. Statistical metrics for January/2016 to March/2021.

Average percentage of cloud coverage	Maximum cloud coverage percentage	Minimum coverage percentage	Number of images without cloud cover	Number of images with less than 20% of cloud cover
39.24%	99.99%	0%	13	74

Figure 3 shows the day with the highest percentage of clouds (99.99%). It happened on 2017-04-14 and its cube collection is named as LC8_30_v001_044054_2017-04-14. Its spatial map is displayed in Figure 3 (a), where only a small portion of the map (denoted inside the red circle) is not covered with clouds. This information was compared with the GOES satellite IR-4 image (Figure 3(b)) for the same day and it is possible to observe a cloudiness zone in the eastern portion of São Paulo State, possibly associated with the data found in the Landsat cubes.

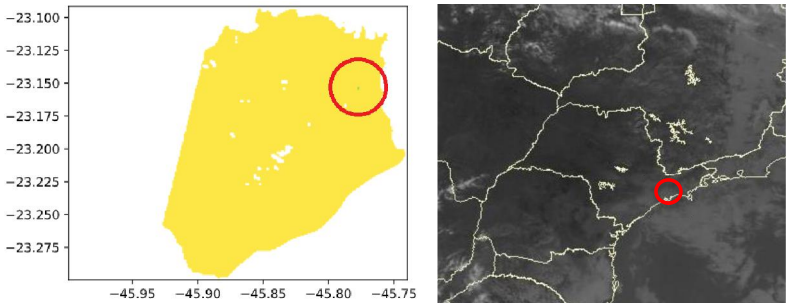

Figure 3. (a) - Cloud cover for São José dos Campos on 2017-04-14. The area in yellow represents the area covered by cloud. The red circle indicates the area in cyan, that represent area without clouds. (b) - GOES-13 IR-4 Channel over South America on 2017-04-14 (INPE-CPTEC). The red circle indicates Sao Jose dos Campos.

Filtering the dataset in months, as shown in Figure 4, our prototype allows the user to obtain monthly information of cloud cover for different years, which can be important for planning the choice of the best study period. Based on the monthly average, BDC3 allows the analysis of cloud coverage considering the different seasons of the year, where it is possible to observe that the highest percentage of coverage generally occurs during summer and autumn.

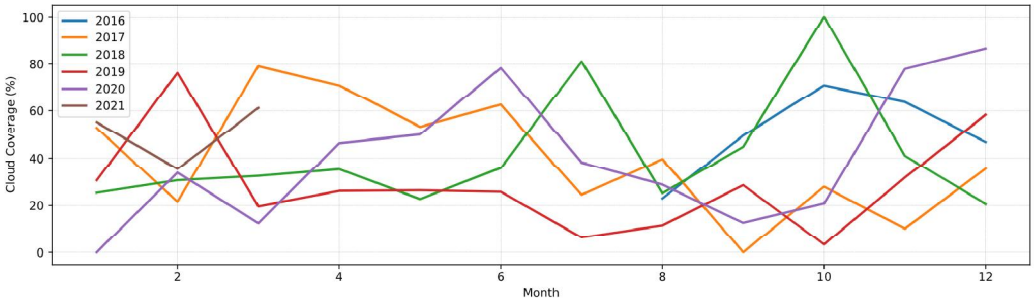

Figure 4 - Monthly average cloud cover considering the entire period of images available in the collection.

Based on the cloud mask obtained for each date, Figure 5 shows the accumulation of clouds over the last six years (2016-2021), where it is possible to generate a classification indicating the areas with the highest occurrence of clouds.

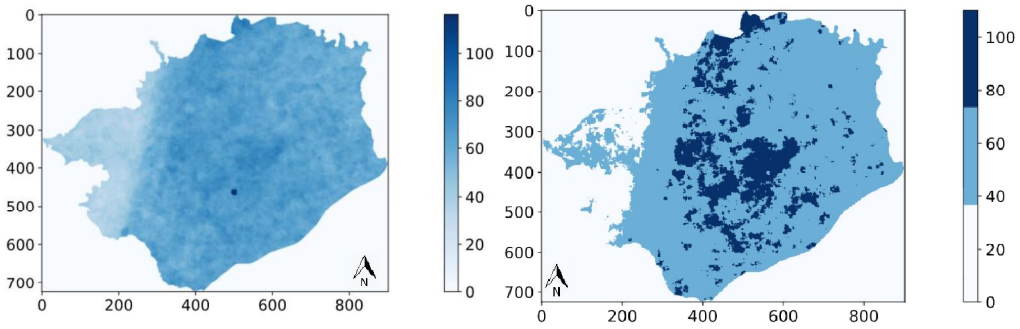

Figure 5 – Areas with a higher occurrence of cloud. The colored bars represent the frequency of cloud occurrence.

5. Conclusions

This article concerns a prototype development, and hence, is limited to presenting some applications for the cloud masks of the Brazil Data Cube project. The results presented show the potential of the tool for the evaluation of seasonal weather behaviors, such as cloud cover. Therefore, it is projected as a useful functionality for the BDC, as it will allow the performance of multicriteria analysis based on its integration with further available tools. The next steps to be developed regard improving the script and integrating the tool to the BDC Portal, as an interactive online platform.

References

BAUMANN, Peter; MISEV, Dimitar; MERTICARIU, Vlad; HUU, Bang P. (2019). Datacubes: Towards space/time analysis-ready data. In Service-oriented mapping (pp. 269-299). Springer, Cham.

Ferreira, K. R.; Queiroz, G. R. et al. (2020) "Earth Observation Data Cubes for Brazil: Requirements, Methodology and Products." Remote Sensing, 12, 4033.

Giuliani, G.; Camara, G.; Killough, B.; Minchin, S. (2019) "Earth Observation Open Science: Enhancing Reproducible Science Using Data Cubes". Data, 4, 147.

Lucena, F. R. S. M.; Escobar-Silva, E. V.; Marujo, R. F. B.; et al. (2020) "Brazil Data Cube Cloud Coverage Viewer." XXI GEOINFO. 222-227 p.

Nativi, Stefano; Mazzetti, Paolo; Craglia, Max. (2017). A view-based model of data-cube to support big earth data systems interoperability. Big Earth Data, 1(1-2), 75-99.

Polidorio, A. M. *et. al.* (2005), "Detecção automática de sombras e nuvens em imagens CBERS e Landsat 7 ETM". XII SIMPÓSIO BRASILEIRO DE SENSORIAMENTO REMOTO. 4233-4240 p.

Soille, P.; Burger, A.; Marchi, D.D.; Kempeneers, P.; Rodriguez, D.; Syrris, V.; Vasilev, V. (2018). A versatile data-intensive computing platform for information retrieval from big geospatial data. Future Gener. Comput. Syst. 81, 30–40.