GEOMATICS AND ENVIRONMENTAL ENGINEERING • Volume 16 • Number 3 • 2022

https://doi.org/10.7494/geom.2022.16.3.131

Paloma Merodio Gómez[1], Andrea Ramírez Santiago[2],
Olivia Jimena Juárez Carrillo[3], Francisco Javier Jiménez Nava[4]

The Potential Contribution of Earth Observation Data Cubes for the Production of Information for Sustainable Development in Emerging Countries

Abstract: One of the great challenges of achieving the shared vision of the 2030 Agenda for Sustainable Development is having high-quality, timely, comparable, and accessible data that allows to measure and report progress on the Sustainable Development Goals (SDG). Hence, in many countries, geospatial information (including Earth observation) and algorithms implemented in cloud computing platforms have become important tools to monitor indicators of the SDG thanks to their broad accessibility and global coverage. However, emerging countries still face barriers to the implementation of technologies to manage the large amounts of EO data. This article aims to show the advantages of satellite-based EO in the measurement of SDG indicators, as well as challenges emerging countries face in the use of these technological tools. It addresses why the open-source tool Open Data Cube (ODC) should be seen as a response to the said challenges. Finally, there is a description regarding the experience of Mexico with the use and application of this tool for the measurement of SDG indicators, from the development and implementation of the Mexican Geospatial Data Cube (MGDC) to the results obtained from its application in the support for the measurement of SDG indicators 6.6.1 *Change in the extent of water-related ecosystems over time* and 15.1.1 *Forest area as a proportion of total land area*.

Keywords: Earth observation, Geospatial Data Cube, satellite imagery, Sustainable Development Goals, emerging countries

Received: 2 March 2022; accepted: 17 May 2022

[1] National Institute of Statistic and Geography, Aguascalientes, Mexico, email: paloma.merodio@inegi.org.mx, https://orcid.org/0000-0002-6086-7023
[2] National Institute of Statistic and Geography, Aguascalientes, Mexico, email: andrea.santiago@inegi.org.mx, https://orcid.org/0000-0002-6201-1907
[3] National Institute of Statistic and Geography, Aguascalientes, Mexico, email: jimena.juarez@inegi.org.mx
[4] Independent researcher, Aguascalientes, Mexico, email: fjjn1960@gmail.com

1. Introduction: Satellite-based Earth Observation in Support of the SDG

In 2015, the leaders of 193 member countries of the United Nations agreed on the 17 Sustainable Development Goals (SDG), to eradicate poverty, protect the planet and ensure prosperity for all people by the year 2030. These 17 goals aim to reach 169 targets, which are monitored and evaluated through 231 unique indicators.

At the national level, the National Statistics Offices (NSOs) usually have custody of the production of the indicators. However, the diversity, scope and scale of the SDG require the participation of many stakeholders, in addition to NSOs [1, 2], which contribute to the development of tools and technological platforms that are accessible and easy to implement in all countries, including emerging countries [3], as well as the production and analysis of timely, accessible, reliable, and high-quality data that will guarantees the principle of "leaving no one behind" [1].

The *Voluntary National Reviews Synthesis Report* [4], on the key findings of the progress of the SDG, concluded that countries in all regions are exploring technology solutions to meet their needs and address the issues of accessibility, the sharing and integration of data, as well as data breakdown at a high level [2].

Reliance on geospatial information, including Earth Observation (EO), that is the gathering of information about Earth's physical, chemical, and biological systems via satellite-based remote sensing technologies, is gaining momentum as countries have begun implementation of the 2030 Agenda at the local, subnational, national, regional, and global levels [2].

These new technologies aim to replace costly traditional approaches in the countries that use them [5], as well as being easy to include in emerging countries where this type of technology is not commonly included. Therefore, EO provides a substantial contribution to the achievements of the SDG by enabling informed decision-making and monitoring expected results [6].

In order to show, promote, and implement the potential of this type of data in monitoring and evaluating the SDG, the Group on Earth Observations (GEO) launched an initiative called Earth Observations for the Sustainable Development Goals (EO4SDG) in 2015, with the participation of members of GEO and other organizations and initiatives [7]. Figure 1, produced by GEO, indicates the targets and indicators that Earth Observation supports, both directly and indirectly.

In addition to this, the Working Group on Geospatial Information (WGGI) of the Inter-Agency and Expert Group on the SDG Indicators (IAEG-SDG) has estimated that approximately 20% of the SDG indicators can be assessed and measured directly using geospatial information or through its integration with statistical data [9].

While EO are a necessary data source for monitoring and driving progress on the SDG, UN member states cannot always harness this value as there are different types of challenges identified in their adoption, such as: institutional; financial; policy and regulation related (lack of political commitment, coordination, institutional

alliances and arrangements); technological (lack of standardization of data processing methods, complexity in access to data, lack of relevant data that fit the purpose, lack of development of Information and communications technology application); related to technical and human capacities (i.e., skills and knowledge); as well as sufficient use cases and good practice examples [2, 10].

Target Contribute to progress on the Target, not necessarily the Indicator									Goal	Indicator Direct measure or indirect support to the Indicator					
							1.4	1.5	1 No poverty	1.4.2					
						2.3	2.4	2.c	2 Zero hunger	2.4.1					
					3.3	3.4	3.9	3.d	3 Good health and well-being	3.9.1					
									4 Quality education						
								5.a	5 Gender equality	5.a.1					
		6.1	6.3	6.4	6.5	6.6	6.a	6.b	6 Clean water and sanitation	6.3.1	6.3.2	6.4.2	6.5.1	6.6.1	
					7.2	7.3	7.a	7.b	7 Affordable and clean energy	7.1.1					
								8.4	8 Decent work and economic growth						
					9.1	9.4	9.5	9.a	9 Industry, innovation and infrastructure	9.1.1	9.4.1				
						10.6	10.7	10.a	10 Reduced inequalities						
	11.1	11.3	11.4	11.5	11.6	11.7	11.b	11.c	11 Sustainable cities and communities	11.1.1	11.2.1	11.3.1	11.6.2	11.7.1	
				12.2	12.4	12.8	12.a	12.b	12 Responsible consumption and production	12.a.1					
					13.1	13.2	13.3	13.b	13 Climate action	13.1.1					
		14.1	14.2	14.3	14.4	14.6	14.7	14.a	14 Life below water	14.3.1	14.4.1	14.5.1			
	15.1	15.2	15.3	15.4	15.5	15.7	15.8	15.9	15 Life on land	15.1.1	15.2.1	15.3.1	15.4.1	15.4.2	
								16.8	16 Peace, justice and strong institutions						
17.2	17.3	17.6	17.7	17.8	17.9	17.16	17.17	17.18	17 Partnerships for the goals	17.6.1	17.18.1				

Fig. 1. Earth observations and geospatial information linkages to SDG, targets, and indicators

Source: [8]

This article aims to show the advantages of EO in the measurement of the SDG indicators. To this end we first explain the challenges in the use of EO for emerging countries before explaining how the open-source tool ODC responds to these challenges, as well as the development and implementation of the Mexican Geospatial Data Cube. Subsequently, we describe the experience of Mexico regarding the use and application of this tool for the measurement of SDG indicators 6.6.1 *Change in the extent of water-related ecosystems over time* and 15.1.1 *Forest area as a proportion of total land area*. Finally, we present a discussion of these matters and some conclusions.

2. Satellite-based Earth Observation Challenges in Emerging Countries

Geospatial information (including EO), together with new tools and technologies, offer opportunities to emerging countries' capacities to monitor all facets of sustainable development as they mitigate *in situ* data availability shortages and provide reliable, up-to-date, cost-effective, and scalable data [11–13].

Freely available data from satellite constellations such as MODIS (Moderate Resolution Imaging Spectroradiometer), Landsat, and Sentinel have increasingly democratized access to timely satellite imagery from around the world. However, this growing availability of satellite imagery reveals different challenges, as traditional remote sensing analysis and data management techniques are not sufficient to handle this large amount of data

The data from EO can be described as Big Data, since it is characterized by massive amounts of data, multiple data sources, a multi-temporal and multi-dimensional heterogeneous structure [13]. Therefore, powerful technological capabilities are required for its handling and analysis together with advanced analytical methods that support multiple data models and reduce data transfer, as well as advanced visualization techniques that can be easily integrated into different graphical user interfaces, including web systems and mobile devices [11].

New methods and algorithms, research infrastructures, and computational resources are useful for preserving, compressing, grouping, and modeling EO data during analysis, interpretation, and visualization in a variety of applications [11]. Likewise, the need to establish national reference data has motivated countries to strengthen their capacities to collect and analyze data, and they have tested proposals aligned with their existing planning instruments, their sustainable development strategies, and their procedures for information production.

EO technologies are evolving at a fast pace which makes it harder for the non-specialists to remain up to date with new tools and techniques, and the new skills these might imply. Not acknowledging the benefits that spatial technologies bring to sustainable development could impede them being used to their full potential [14, 15], therefore, the development of geospatial tools that are easy to implement and use are essential so that non-specialists can obtain information ready for analysis, without the need to participate in its processing.

In emerging countries, the lack of trained and experienced personnel to produce information derived from satellite data with local resources, provide user support, and generate applications from space technologies can be a barrier to expanding the use of satellite technologies.

Other obstacles to the wider use of satellite technologies include restricted access to data, lack of standardization, data that is not fit for purpose, lack of data ready for analysis, and insufficient frequency of observations [14]. However, leveraging satellite data has helped many developing and emerging countries address

some of their most pressing challenges. The 2019 survey conducted under the auspices of the African Association for Remote Sensing of the Environment showed that the success of geospatial information and EO depends on three factors [15]:

1) a well-informed public sector to develop a strategy and an architecture for geospatial data,
2) highly developed academic institutions to support capacity development in EO and geoinformation sciences, in engineering and space technology,
3) a prosperous private sector that serves as an engine for economic growth.

3. Data Cubes as a Response to Challenges in the Use of Satellite-based Earth Observation

In recent years, there has been a global move from satellite operators towards producing more usable datasets, to reduce the work required before mining and analyzing. However, the large amount of data that is available requires to be migrated from the traditional approach of users who downloaded data and did local processing towards high-performance computing data centers (local or cloud-based), using Big Data processing tools [16].

Various implementations of platforms capable of analyzing EO data reflect the growing interest in developing large-scale analytical tools allowing effective and efficient information retrieval [17]. A specific sub-group of such tools are referred to as data cubes or EO data cubes. Data cubes are revolutionizing the way users can work with EO data. To reduce the processing load on users, the generation of high-quality multidimensional arrays of satellite information is a fundamental requirement, as they minimize the time and scientific knowledge required to access and use satellite information. They reduce the barrier to generating geospatial information products ready to be analyzed by automating pre-processing steps to support the utilization of the growing volume of EO data, thus expanding the number of potential users [18].

The term EO data cube is a novel and often unknown concept; Strobl [19, p. 32.] defined a data cube as a data structure that "is based on regularly and irregularly gridded, spatial and/or temporal data with n dimensions (or axes) and characterized by the presence of the 6 faces". In order to describe EO data cubes into meaningful and manageable parts, Strobl [19] identified six different aspects (Fig. 2); each of which regards one well-established data science domain. Strobl emphasized that to enable and facilitate the full interoperability of EO data cubes it is important to make sure all the views are adequately addressed and kept technology neutral [19].

Furthermore, a data cube is often compared to (or confused with) cloud-based processing platforms; to better characterize data cubes, it is important to differentiate them from cloud-based platforms such as the Copernicus Data and Information Access Services (DIAS), the Google Earth Engine (GEE) or Earth on Amazon Web Services. Cloud-based platforms often provide free and open access to

global EO datasets together with powerful analysis tools, yield fast results and allow to avoid the burden of data preparation; however, they lock users into a platform dependency [17]. Giuliani et al. [17] listed some of the identified challenges; most of these potential drawbacks can be tackled by utilizing EO data cubes, as Table 1 describes.

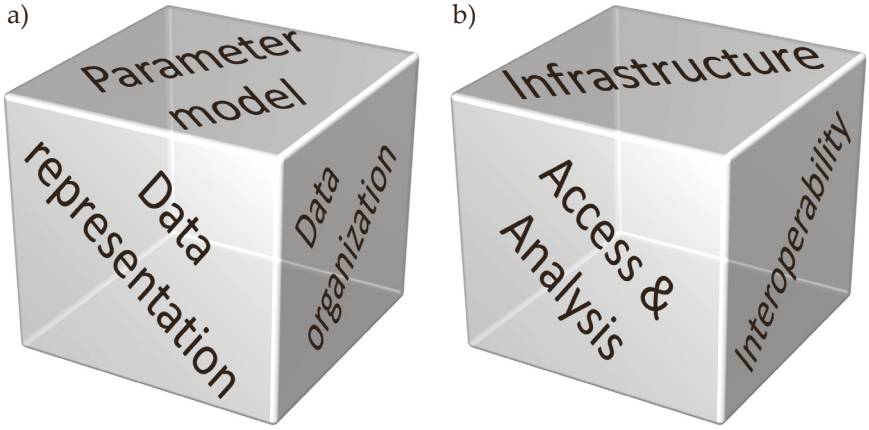


Fig. 2. The data-oriented faces of the EO data cube (a); the functionality-oriented faces of the EO data cube (b)

Source: [19]

Table 1. EO data cube advantages over cloud-based processing platforms

Identified concern in cloud-based processing platforms	How this concern may be tackled by EO data cubes
Users do not know whether a given platform will be sustained or evolved in the future	Users can install and maintain an open-source solution on their own computing infrastructure, they may also keep any version they wish to, with no need to update
Limited time and spatial scale for analyses is provided	Data cubes provide processing scalability (depending on the computational infrastructure available)
Only cloud-based computing is provided (no options for hubs or local computing solutions)	Data cubes provide hardware flexibility
Users are requested to upload their analytical processing and even local data, while data download is discouraged or not even allowed	Users can install and use on their own computing infrastructure and develop solutions that allow to work in closed environments, helping to guarantee data confidentiality and providing a further sense of ownership. (Furthermore, avoiding commercial and internet dependence can help to fulfill national data security standards)
Platform providers require the right to "own" all the data utilized on the platform	

Table 1. cont.

Users get only those datasets that providers offer, limiting data interoperability	Users are allowed storing different type of data. Data cubes support an efficient and joint use of multiple datasets, enhancing their interoperability and complementarity
Data are often not ready to be analyzed	Data cubes design favors the use of higher quality datasets (analysis ready data or ARD), which can enhance interoperability and generate better results

In particular, the ODC aims to meet the challenge of Big Data as a new approach to store, organize, manage, and analyze EO data, therefore it is now considered a promising technology for conducting time series analysis of large satellite datasets. Several countries and even regions (Africa, Pacific) are implementing the ODC open-source solution where the open-source nature of the ODC has been an important factor in the selection of this tool, as well as the ability of the ODC to be deployed on various computing infrastructures ranging from national supercomputing facilities to numerous commercial cloud infrastructures, which has allowed the establishment of sovereign operational capabilities that can be controlled and managed in the country [15, 20].

4. Implementation of the Open Data Cube in Mexico

In Mexico, the National Institute of Statistic and Geography (Instituto Nacional de Estadística y Geografía – INEGI) has the task of monitoring the environmental, socioeconomic, and demographic phenomena that occur throughout the country. Particularly, the National Subsystem of Geographic and Environmental Information (Sistema Nacional de Información Estadística y Geográfica – SNIEG), in its geographic component, generates data on natural resources, among which are the National Land Use and Vegetation Map (Fig. 3) and the National Water Bodies Map (Fig. 4), which are part of the Spatial Data Infrastructure of Mexico [21].

The land use and vegetation information has been generated through various methodologies that use inputs with multi-year temporal resolutions and mostly manual and exhaustive processes in the territory, classifying satellite images from various sources. Since 1968, INEGI has produced and disseminated the different series (versions) of the Land Use and Vegetation Map, scale 1 : 250,000 which contain the location and distribution of agricultural land use, of natural and induced vegetation in the country, of livestock and forestry use, and other uses that occur in the territory related to plant cover [22]. This map is updated every 5 years, and to date there have been seven series (or versions).

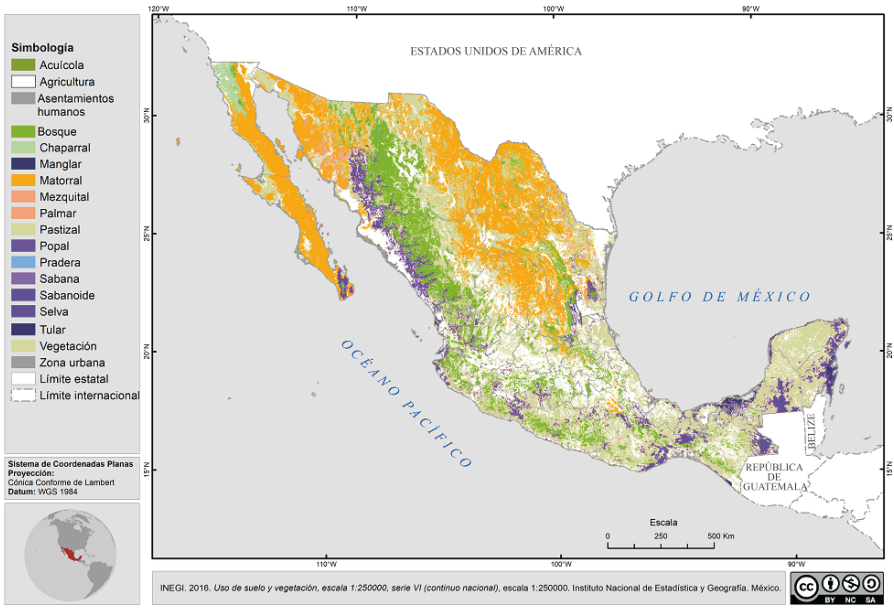


Fig. 3. Example of the national Land Use and Vegetation Map (series VI), scale 1 : 250,000

Source: CONABIO (http://www.conabio.gob.mx/informacion/gis/)

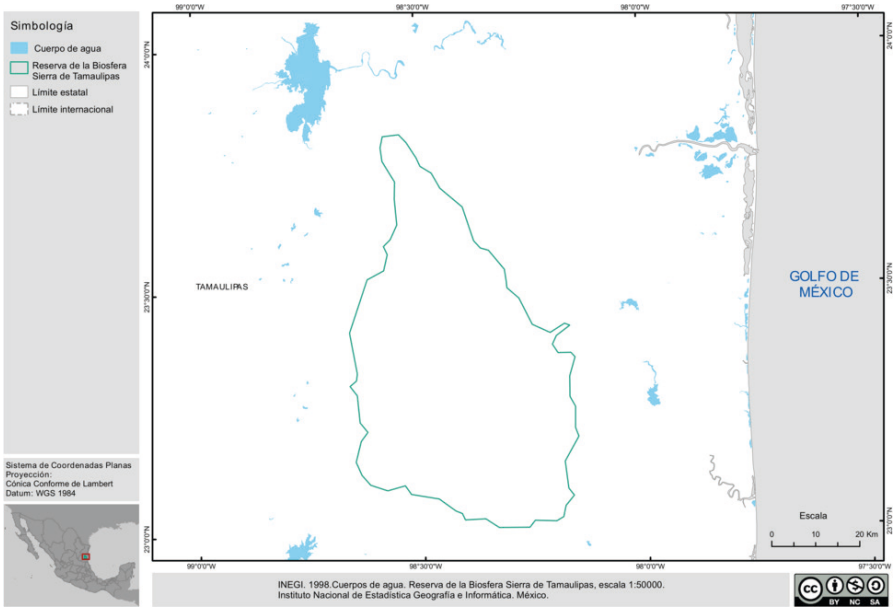


Fig. 4. Example of the local Water Bodies Map, scale 1 : 50,000

Source: CONABIO (http://www.conabio.gob.mx/informacion/gis/)

The spatial resolution varies in each series since they have been made with satellite images from different satellites. Series I was generated with aerial photographs and *in situ* data collected with a completely cartographic approach with output from printed maps. For series II, the methodology was changed for the use of printed space maps. From series III to series VI, the entire process was carried out under a Geographic Information System environment and the use of satellite images was implemented [22, 23].

Series III, was carried out through the photointerpretation of Landsat ETM satellite images of 30 m resolution (of the year 2002); series IV used Spot 5 satellite images of 10 m resolution (of the years 2007 and 2008); series V used Landsat TM5 satellite images of 30 m resolution (of the year 2011) were used; series VI (Fig. 3) used Landsat TM8 satellite images of 30 m resolution (of the year 2014) [22]; and finally, series VII used images from Mexican Geospatial Data Cube (MGDC) together with MGDC products, such as the Landsat geomedian. All these versions were supported by previous information (bibliographic and cartographic) and information obtained *in situ* [23].

Traditional methodologies present various areas of opportunity such as the automation of the satellite image classification processes, as well as the use of timely information that allows the creation of national continuums with temporal homogeneity. However, incorporating new methodologies that allow the generation of large volumes of highly structured data entails technological challenges and technical capabilities for analysts. For this reason, the adoption of new technologies such as the ODC is essential, which aims to solve both needs, guaranteeing an intelligent management of satellite data that allows maximizing computational capacities and bringing information to the end user in a format ready for analysis, thus allowing the generation of information in a timelier manner [24, 25].

INEGI, in collaboration with Geoscience Australia, implemented the Mexican Geospatial Data Cube (MGDC) using ODC, which facilitates access, use and processing of time series satellite images, pixel by pixel, throughout the national territory, thus allowing the selection of a region or period of interest specifying location and date of collection. Currently, MGDC imagery collection consists of more than 132,000 Landsat scenes with their metadata. There is a collection of images of the entire national territory from 1984 to 2021 (and part of 2022) with monthly updates.

The images that are incorporated into a data cube have been corrected and provide comparability in time of their observations at the pixel level, these corrected images referred to as Analysis Ready Data (ARD) [26]. For the specific case of the MGDC, around 86.4% of the images in the collection correspond to NASA's ARD, that is, Quality Level 1 (T1) images from NASA Collection 1, 13.2% correspond to Quality Level 2 (T2) and 0.4% to real time (RT) quality [26]. Each pixel in an image represents a 30 m × 30 m region; the projection parameters (Albers) are defined to better fit INEGI's information production goals [20, 24].

The MGDC architecture consists of a set of Python libraries, a PostgreSQL database, and a collection of organized (indexed) Landsat images; all these resources are

hosted locally on INEGI's servers. The functions of the libraries allow cataloging, indexing, and processing of the thousands of images ordered from the collection based on the format they have when they are downloaded from the space agency. The database does not store the images but rather the path where each one is located within the designated local storage unit (to which the servers have access). In addition, the database also records the metadata of each image (source, date of observation, geographic region it covers, spectral bands, quality level, etc.), which is relevant when generating derivative products. Thus, MGDC facilitates the handling and access to large volumes of satellite images through a programming interface based on Python [21].

In practical terms, the data structure that is generated by sorting a set of satellite images with this tool is a massive dense (no void cells) array of raster data. This means that, in terms of access, all the images in MGDC are a single multidimensional structure, described by several axes; coordinates in these axes are the mechanism that allows access to the data; this coordinate system (x, y, date, band) is homologated for all data; hence, the location of the pixels becomes independent of the source image. In order to include different sensors (like Sentinel-2), more complex design decisions need to be made to either combine both data sources into a single cube (sacrificing information as Sentinel-2 provides more bands), or to keep them apart in different cube instances [21]. Either way, the spatial resolution is compatible (10 m) in theory: each Landsat pixel represents 9 Sentinel-2 pixels.

As of March 2022, the MGDC imagery collection contains more than 132,000 scenes from all over the national territory. And around five hundred images more are collected every month, which together are approximately 500 GB in size; this represents around 6 TB of annual storage. The volume of the complete image collection amounts to 90 TB in its unzipped version. However, MGDC optimizes its storage (converting GeoTIFF images to netCDF), reducing its volume by one third without losing information.

MGDC is designed to produce an ever-expanding range of analysis- and/or decision-ready derivative products. The implementation of the systematic generation of products like the Normalized Difference Vegetation Index (NDVI) and the Modified Normalized Difference Water Index (MNDWI) to the context of Mexico is currently being tested [27]. That is, with the implementation of this platform, INEGI processes (essential for national methodologies) can evolve to include the analysis of a region of interest or the entire territory, not only using an observation from one point in time, but with historical data that can allow time-series analysis to observe, monitor and understand its behavior over time. In addition, it will help develop capabilities in remote sensing, data science, process engineering and software architecture, as well as infrastructure capabilities that include efficient access to satellite imagery, broadband to collect large and frequent data, and Big Data technology to store and process satellite information [28].

Since MGDC is a recent and disruptive implementation in INEGI, a few obstacles have been encountered by the technical team. The objective of this article is not

to provide a harsh assessment of the tool, as the personnel involved are currently in a ramp-up stage to build and strengthen technical skills.

However, for the sake of completeness, some of the challenges found in the use and implementation of the MGDC need to be listed:

- MGDC is difficult to install. The flexibility of the tool results in very particular installation pipelines.
- The use of the tool through the API to analyze the existing data is easy enough, however, the integration of new raster data products (Sentinel, MODIS) requires expert knowledge, so far there is great documentation on the indexing of Landsat imagery, but to expand to and leverage the potential of different data sources there is a need to "define" the products; very different skills are needed to achieve this.
- Another challenge is that despite the nature of this tool in facilitating the use of large collections of images, the lack of (or insufficient) interoperability of the software with Big Data ecosystems of common use to data scientists and often required for handling large volumes of data such as Apache Spark.

5. Mexican Geospatial Data Cube Application for Measuring SDG Indicators 6.6.1 and 15.1.1

In Mexico, the National Statistical and Geographic Information System (SNIEG), represented in the Specialized Technical Committee for the SDG, has the production of information to measure progress in public policies associated with the SDG as one of its objectives. SNIEG, through INEGI and its Specialized Executive and Technical Committees, establishes the objectives and strategies for the generation of information [29, 30]. In this way, the indicators (6.6.1 and 15.1.1) analyzed in this article were selected based on the coincidence of both the shortlist of indicators developed by WGGI and the list of indicators developed by SNIEG, specifically, those corresponding to the National Subsystem of Geographic and Environmental Information, developed in the Department of Natural Resources and Environment of INEGI.

Indicator 6.6.1:
Change in the extent of water-related ecosystems over time

Target 6.6 aims to protect and restore water-related ecosystems. Particularly, indicator 6.6.1 aims to understand how and why these ecosystems are changing in extent over time.

This indicator includes five ecosystem categories:

1) vegetated wetlands,
2) rivers and estuaries,
3) lakes,
4) aquifers,
5) artificial waterbodies.

The measurement of all the components of Indicator 6.6.1 is important to allow informed decisions towards the protection and restoration of water-related ecosystems. However, due to a lack of data within countries to support monitoring of the Indicator, UN Environment Programme proposed a global methodology, one which is internationally recognized, and which combines national data from ground sampling and global data based on EO, resulting in global datasets with extensive spatial and temporal scale which are internationally comparable [31, 32].

Statistical and geospatial data on permanent water, seasonal water, reservoirs, wetlands, mangroves, and lake water quality are available on the SDG 6.6.1[5] data portal. It is important to mention that not all the data series represented on the site use the same reference period, because the recorded observations are captured by different satellites, according to the type of ecosystem related to the water that is being observed (Tab. 2) [32].

Table 2. Data from Earth observations

Ecosystem	Unit	Features
Lakes & Rivers (permanents)	surface area	Annual and multi-annual changes in permanent water area (1984–present) statistics for new and lost permanent water (2000–2019) statistics aggregated at national, sub-national & basin scales
Lakes & Rivers (seasonal)	surface area	Annual and multi-annual changes in seasonal water area (1984–present) statistics for new and lost seasonal water (2000–2019) annual seasonality statistics for periods: 0–1, 3–6, 7–11 months statistics aggregated at national, sub-national & basin scales
Reservoirs	surface area	Annual and multi-annual changes in reservoir surface area (1984–present) statistics for new and lost reservoir area (2000–2019) statistics aggregated at national, sub-national & basin scales
Mangroves	surface area	Annual and multi-annual changes in mangrove area (2000–2016) statistics aggregated at national, sub-national & basin scales
Wetlands	surface area	Wetlands area (baseline area comprised of data btw 2016–2018) statistics aggregated at national, sub-national & basin scales wetlands area changes will be included starting in 2021/2022
Lakes	water quality	Monthly, annual, and multi-annual measurements of trophic state and turbidity for 4200 lakes globally (at 300 m resolution)

Source: [32]

[5] www.sdg661.appp.

Some satellites, such as Landsat, have data from the early 1970s, allowing the measurement of changes in open water bodies (lakes). However, there are satellites that have been incorporated more recently, such as the European Sentinel and several Japanese satellites, which allow the capture of images and data for other types of ecosystems and parameters related to water.

Global data for rivers and groundwater are not yet available at useful spatial and temporal resolutions to be incorporated into the indicator 6.6.1 methodology, so these data should continue to be provided from models or ground measurements.

The global results for indicator 6.6.1, are available on the UN Water portal, which show the gain or loss in the extent of the water body compared to the established baseline period (2001–2005) [33].

On the other hand, the national methodology based on the MGDC, enabled the generation of national-level products based solely on the analysis of time series of images [34]. To measure this indicator, two data sources were used: the National Water Bodies Map [34] and MGDC-based product called Landsat Surface Water Classification Index from Space (ICASE), an adaptation of WOfS (Water Observations from Space) methodology. The National Water Bodies Map is derived from the set of national topographic data at scale 1 : 50,000 made with SPOT 5, SPOT 6, Geoeye and WorldView satellite images (spatial resolution from 1 m to 10 m). The Landsat ICASE is a geospatial analysis product that provides information on the presence of surface water [27].

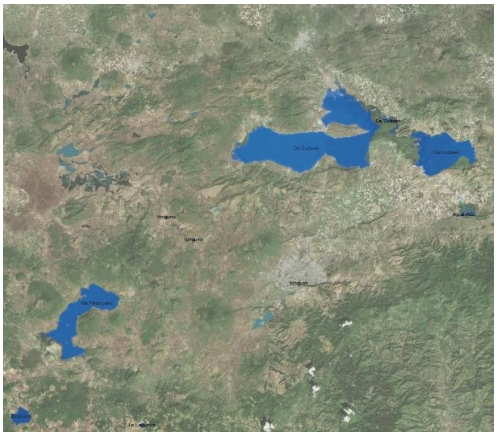
The ICASE consists of annual national mosaics, from 1982 to 2020; each ICASE mosaic maps the level of the presence of water identified in Landsat satellite images, through a regression tree that considers values from the spectral bands, both as individual bands and in combination [27].

Through this process, surface water is detected using an automated water mapping algorithm. The number of times water is detected for each location is summed through time and then compared to the number of clear observations of that location. The result is a percentage value of the number of times water was observed at that location, providing a nationally consistent tool to complement the studies of surface water dynamics, both spatially and temporally [35].

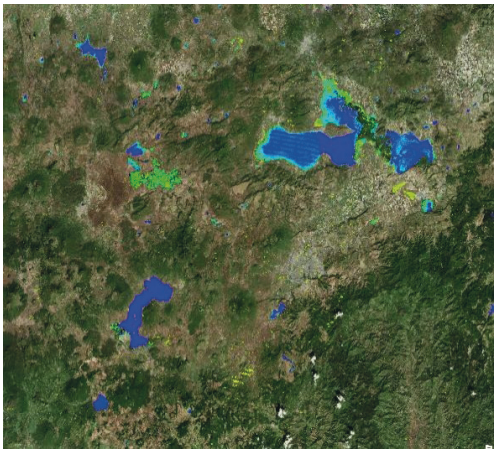
The 2015 ICASE mosaic (derived from Landsat 7 and 8) was obtained by processing all the 2015 images available on MGDC. Quality indicators considered in this method are saturation of pixels, contiguity of bands, the clouds or cloud shadow and the shadow of the terrain [27].

The 2015 mosaic was one of the first ICASE mosaics generated. To observe the possible uses of this mosaic (or an annual series of it) for measuring indicator 6.6.1, four lakes were selected in central western Mexico (Yuriria, Cuitzeo, Pátzcuaro and Zirahuén), where it was determined that the water surface is 10% less than the 2010 reference data in the National Water Bodies Map (Fig. 5). Experts analyzed these outcomes and resolved that the information was consistent with the status of the water bodies known from different sources. Hence, the 2015 mosaic and the ICASE algorithm was considered fit for further research.

a)



b)



Water not detected
Water detected in 1% of the observations
Water detected in 5% of the observations
Water detected in 20% of the observations
Water detected in 50% of the observations
Water detected in 80% of the observations
Water detected always

Fig. 5. National Water Bodies Map for 2010, scale 1 : 50 000 (area: 44 798 ha) (2010 Landsat image background) (a). ICASE Assessment for 2015 (area: 40 315.40 ha) (2015 Landsat geomedian background) (b)

Based on the positive feedback provided from thematic experts in the previous exploration of this product, more ICASE mosaics were generated. In the later study, the aim was to measure the annual change in the surface of Chapala Lake from the period 1985 to 2019, which is one of the most important lakes in Mexico. The water pixels were quantified and converted to surface values (in square kilometres) in order to showcase the value of a tool such as the MGDC to decision-makers (Fig. 6).

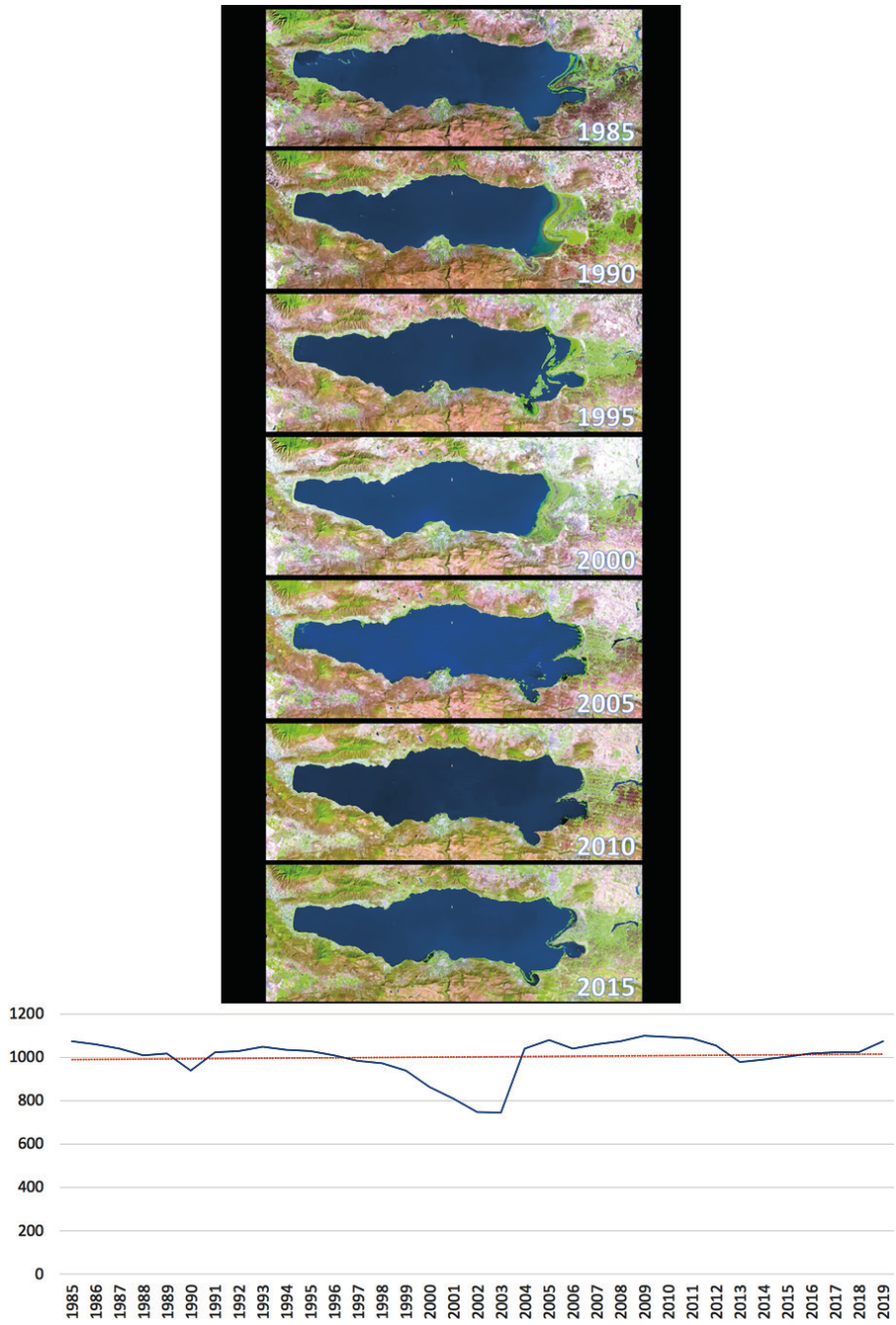


Fig. 6. Graphic of annual changes [km^2] in the surface of Chapala Lake using Landsat ICASE (1985–2019) and reference images for years 1985, 1990, 1995, 2000, 2005, 2010 and 2015. Each ICASE mosaic is derived from annual sets of Landsat images in the MGDC

Indicator 15.1.1:
Forest area as a proportion of total land area

This indicator provides a measure of the relative extent of forest in each country and is a key element for forest policy and planning in the context of sustainable development. Total forest area is the total land spanning more than 0.5 ha with trees higher than five meters and a canopy cover of more than 10%, or trees able to reach these thresholds *in situ*. And the total land area is the total country area excluding area under inland waters and coastal waters [36].

At the national level, many countries carry out their forest area assessment at infrequent intervals, which is why global data from EO helps countries update their forest area estimates more frequently [35]. The Food and Agriculture Organization of the United Nations (FAO) collects data on forest area at regular intervals (currently every 5 years) through the Global Forest Resources Assessment (FRA), as well as collecting data on land area through the annual FAO Questionnaire on Land Use, Irrigation and Agricultural Practices. These data are supplemented by national statistical data and other official government data [37, 38].

The global results for indicator 15.1.1 are available on the FAO portal on Global Forest Resources Assessments (FRA), which provides essential information for understanding the extent of forest resources, their status, management and uses [38].

The results for Mexico from the FRA 2015, indicate the total forest area from 1990 to 2015. In addition to the data from FAO, FAR 2015, the dissemination platform of the SDG Global Indicators Database shows the results for Mexico on the total forest area and the forest area as a proportion of the total land area [38].

To measure this indicator at the national level, data from the INEGI national Land Use and Vegetation maps are used. These data sets are the key to observe changes in land use, the effectiveness of the protection of natural areas, the increase in agricultural areas and the urbanization, among others.

INEGI's series of mosaics called Landsat geomedian is the result of searching for the single image that best represents all the Landsat images of the same year which capture a portion of the national territory and that are available in MGDC.

There are several methods for generating summary mosaics of national coverage. Mechanisms based in pixel time series statistics, consider the pixel sequence for each band over time and calculates some statistics that summarize the observations to create an image composite. A common practice is to estimate a one-dimensional statistic for each of the bands, however, it does not preserve spectral relationships, which is desirable in cases where the image resulting will be used as a starting point for analytical processes [39].

The geomedian is a composition-based approach at the pixel level that takes a collection of satellite images from a specific time period (defined by the user) and "compiles" them into a single image. In Figure 7 the process for obtaining the geomedian from the collection of images to the construction of mosaics is presented.

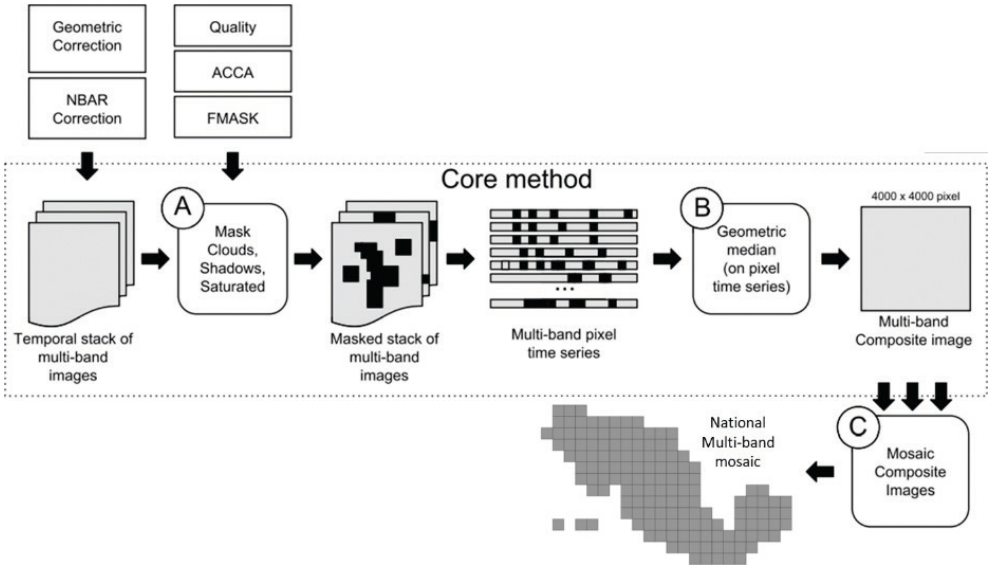


Fig. 7. Production of a Landsat geomedian mosaic in INEGI
Source: [26]

This algorithm is used to perform a multivariate statistical summary (similar to a median value) for all the observed values of the same pixel in an indicated period (Fig. 7: Temporal stack of multi-band images). Hence, the application of this algorithm produces an image composed of "summary pixels," which maintains spatial consistency, and by working with all the pixel bands at the same time, it also preserves the ratio between these values. Filtering out cloud and noisy pixels using quality mask layers (Fig. 7: Mask, Clouds, Shadows, Saturated), this image composite provides a good representation of a typical observation that lacks outliers and with reduced spatial noise and maintains spectral relationships among the bands [39].

After applying the algorithm to each pixel time-series, the "summary" pixels (Fig. 7: Geometric median (on pixel time-series)) are integrated into a single image (Fig. 7: Mosaic composite images); it is worth mentioning that the computational task was divided into smaller mosaics (Fig. 7: 4000 × 4000 pixel multi-band composite image) prior to this integration into a single image.

The final product is a mosaic of multispectral images that represent the characteristics of the surface at a specific period. INEGI's particular series has one country-wide geomedian image per year for the last three decades subdivided into 144 mosaics to facilitate distribution of the data (Fig. 7: National multi-band mosaic) [26].

Mexico has a multi-temporal vision of the distribution of vegetation and land use, which makes it possible to observe changes in uses, effectiveness of the protection of natural areas, increase in agricultural areas and urbanization, among other variables that are relevant for achieving sustainable development.

One of the advantages of using MDGC for the calculation of NDVI, is that it allows us to learn the behavior of vegetation for a whole year. For the purpose of mapping land use and vegetation, this approach provides a more representative classification for each pixel, unlike what could be obtained using a single value of an original image, which is a true value but only observed on one certain date.

An initial exploratory study of the application of the MGDC-based methodology can be seen in Figure 8, which shows the results of visually comparing two false-color images of the Montes Azules area and the Marqués de Comillas Area, which are divided by the Usumacinta River. Montes Azules area was declared a Natural Protected Area in 1978, the river marks the limit of the protected area. This study consisted in searching for an MGDC image of the region from a time close to 1978 in order to evaluate the status of the vegetation. The selected image, taken as an initial reference, is from 1986 (the MGDC collection does not have data from 1978); this image is a single observation (Landsat 5) (since there was not enough available images to produce a geomedian). The image used to compare it with was the 2017 geomedian, which is cloud-free by design (its construction used Landsat 7 and 8 data).

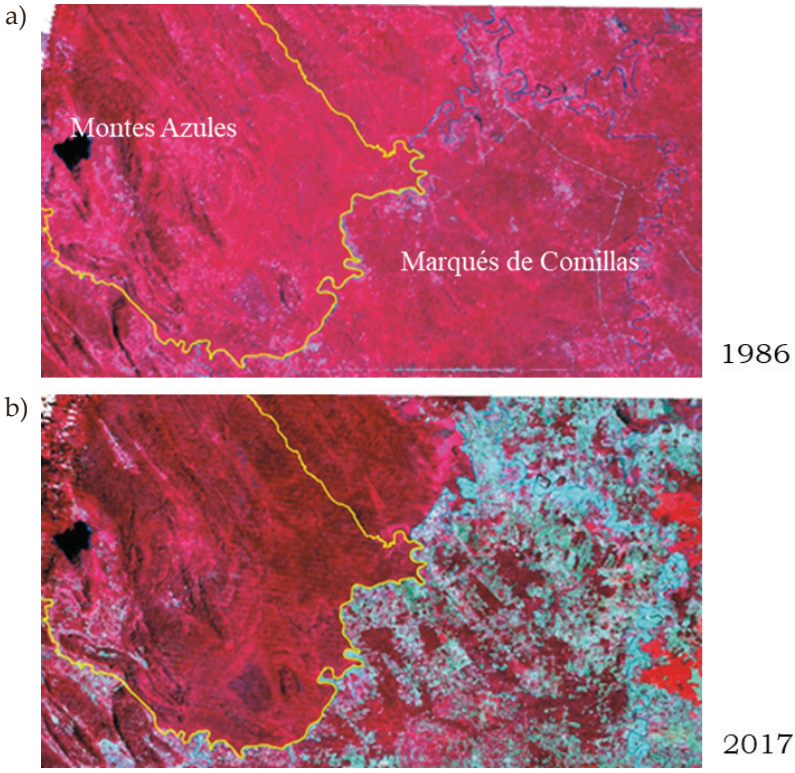


Fig. 8. Changes in vegetation cover and effectiveness of the Montes Azules Natural Protected Area in southeastern Mexico from 1986 (a) to 2017 (b), shown in "false color" spectral combination to enhance the presence of vegetation

Once this study helped to reveal the potential of the MGDC and its products to evaluate public environment policy, INEGI resolved to operationalize the platform and to publish derived information products (geomedian and ICASE so far).

6. Discussion

The data reported at the global level for monitoring the SDG are relevant and useful for making comparisons on sustainable development between countries. In addition, when a national data set is not available, global data sets offer a complementary alternative to produce information that would not otherwise exist. However, to focus on national development policies and advance the 2030 Agenda, more frequent data and higher spatial resolution are needed.

In situ information provides more detailed and high-quality data, however, frequently updating this type of information can incur significant costs in both temporal and monetary terms. In this way, satellite-based EO provide the opportunity to build consistent data sets that are constantly updated and allow for better monitoring and measurement of the SDG indicators.

One way to reduce the barriers that some countries face in the inclusion of satellite-based EO technologies is to implement open source technology solutions, which make it possible to store, organize, manage and analyze data from EO more easily, such as it is the ODC which has the ability to be deployed in various computer infrastructures, which has allowed the establishment of operational capabilities that can be controlled and managed in the country, revolutionizing the way users can work with data, since it has the potential to generate ARD, thus reducing the processing load.

Specifically, the MGDC's image collection only consists of Landsat images, which were provided directly to INEGI on hard drives by NASA and USGS personnel in order to avoid the task of downloading huge volumes (90 TB) and allow for a faster and more complete implementation. Arrangements on the future incorporation of Sentinel-2 to the cube are in place and the storage resources are considered within the Information Technology plan from INEGI. However, including a second data source involves technical and design decisions as well a huge effort to download and preprocess the data, since in this case, Sentinel-2 images are not available for download in the level of quality required. When the development of the ODC began, the most usual sensor in this sort of technology was Landsat, very few others had managed to feed Sentinel-2 data into a data cube instance.

The development of SDG indicators based on EO data processed using tools such as the ODC are strong examples of how to solve the key challenge of providing timely information as the basis for sound, evidence-based decision making. In the specific cases of the selected indicators, the overestimation of the values calculated for indicator 15.1.1 by the global methodology has been resolved thanks to the

incorporation of the knowledge of local specialists, the inclusion of field information, and the use of satellite images with higher spatial resolution.

Although INEGI's National Water Bodies Map at a 1 : 50 000 scale provides useful information for monitoring indicator 6.6.1., the integration of data derived from the MGDC in its methodology will allow the development of sub-indicators for the rainy season as well as for the dry season.

For this, a broader temporal assessment and sub-indicators of seasonality are needed due to the dynamic nature of the water. Studies can be carried out in areas of interest and annual monitoring. Although it is proposed to calculate indicator 6.6.1 every 3 or 5 years with this methodology, the product allows an annual product with an additional seasonal analysis (dual).

These initial efforts using the National Water Bodies Map to evaluate ICASE will play a significant role in showing the added value of EO technologies and in the adoption of automatic processes in official activities.

7. Conclusions

The implementation of new data sources, technologies, and processes are increasing the production and availability of information necessary for more efficient monitoring of the SDG indicators. However, it is not always easy for NSOs to analyze data from EO, as well as incorporate large volumes of data into their methodologies, due to a lack of infrastructure and capacity, hence the importance of developing and implementing tools that allow easy management and efficient analysis of data, such as the MGDC.

The recent implementation of the MGDC as an institutional and transversal platform in Mexico further recognizes the valuable contribution of satellite EO to the SDG. The adoption of the MGDC allows to further harness the potential of global data sets by providing a developing country with the ability to build custom data sets based on those globally available and thus take advantage of the local knowledge to evaluate different products.

It is important to mention that the products generated by the MGDC, in addition to their use in the measurement of SDG indicators, will be used to complement the traditional mapping of natural resources, which has been delivered prepared and updated by INEGI for 50 years. The methodological design of this activity has always included field verification, which will also be used for MGDC products. In this sense, INEGI is working to ensure that the products generated by the MGDC are complemented and integrated with a wide range of on-site validation data (field verification) to ensure that derived products are robust and improve the monitoring of public policies and the SDG.

Geospatial information (including EO) and statistical data can be integrated in support of national priorities and global goals, not only facilitating the monitoring

of SDG indicators, but also the assessment of public policies and decision-making at the national level based on evidence. In this way, this open tool has given emerging countries ownership of these global data sets and the possibility to apply them for their specific and more localized studies, and at the same time to address international commitments, such as the indicators of the 2030 Agenda that can be supported by EO.

Author Contributions

Paloma Merodio Gómez: conceptualization, writing – original draft, supervision, writing – review & editing, project administration.

Andrea Ramírez Santiago: conceptualization, methodology, validation, investigation, resources, writing – original draft, writing – review & editing, visualization, project administration.

Olivia Jimena Juárez Carrillo: methodology, software, validation, formal analysis, investigation, resources, data curation, writing – review & editing, visualization.

Francisco Javier Jiménez Nava: methodology, validation, investigation, resources, writing – review & editing, visualization.

References

[1] Marcovecchio I., Thinyane M., Estevez E., Fillottrani P.: *Capability Maturity Models as a Means to Standardize Sustainable Development Goals Indicators Data Production*. Journal of ICT Standardization, vol. 6(3), 2018, pp. 216–244. https://doi.org/10.13052/jicts2245-800X.633.

[2] Kavvada A., Metternicht G., Kerblat F., Mudau N., Haldorson M., Laldaparsad S., Friedl L. et al.: *Towards delivering on the sustainable development goals using earth observations*. Remote Sensing of Environment, vol. 247, 2020, 111930. https://doi.org/10.1016/j.rse.2020.111930.

[3] Thinyane M.: *Small data and sustainable development–Individuals at the center of data-driven societies*. [in:] *2017 ITU Kaleidoscope: Challenges for a Data-Driven Society (ITU K)*, IEEE, 2017, pp. 1–8. https://doi.org/10.23919/ITU-WT.2017.8246991.

[4] UN DESA: *2019 Voluntary National Reviews Synthesis Report*. United Nations Department of Economic and Social Affairs, New York 2019. https://sdghelpdesk.unescap.org/sites/default/files/2019-11/252302019_VNR_Synthesis_Report_DESA.pdf [access: 19.04.2022].

[5] Andries A., Morse S., Murphy R.J., Lynch J., Woolliams E.R.: *Seeing Sustainability from Space: Using Earth Observation Data to Populate the UN Sustainable Development Goal Indicators*. Sustainability, vol. 11(18), 2019, 5062. https://doi.org/10.3390/su11185062.

[6] Anderson K., Ryan B., Sonntag W., Kavvada A., Friedl L.: *Earth observation in service of the 2030 Agenda for Sustainable Development*. Geo-spatial Information Science, vol. 20(2), 2017, pp. 77–96. https://doi.org/10.1080/10095020.2017.1333230.

[7] Paganini M., Petiteville I., Ward S., Dyke G., Steventon M., Harry J., Kerblat F.: *Satellite earth observations in support of the sustainable development goals. The CEOS Earth Observation Handbook Special 2018 Edition*. European Space Agency, 2018. http://eohandbook.com/sdg/files/CEOS_EOHB_2018_SDG.pdf [access: 19.04.2022].

[8] GEO and UN-GGIM: *Earth Observations and Geospatial Information: Supporting Official Statistics in Monitoring and Achieving the 2030 Agenda*. Group on Earth Observations, United Nations Committee of Experts on Global Geospatial Information Management, 2017. https://earthobservations.org/documents/publications/201704_geo_unggim_4pager.pdf [access: 19.04.2022].

[9] Arnold S., Chen J., Eggers O.: *Global and Complementary (Non-authoritative) Geospatial Data for SDGs: Role and Utilisation*. 2019. https://ggim.un.org/documents/Report_Global_and_Complementary_Geospatial_Data_for_SDGs.pdf [access: 19.04.2022].

[10] Allen C., Metternicht G., Wiedmann T.: *Prioritising SDG targets: assessing baselines, gaps and interlinkages*. Sustainability Science, vol. 14(2), 2019, pp. 421–438. https://doi.org/10.1007/s11625-018-0596-8.

[11] Filchev L., Pashova L., Kolev V., Frye S.: *Challenges and solutions for utilizing Earth Observations in the "Big Data" era*. Paper presented at the BigSkyEarth conference: AstroGeoInformatics, Tenerife, Spain, December 17–19, 2018. https://doi.org/10.5281/zenodo.2391937.

[12] Rosa W. (ed.): *A New Era in Global Health: Nursing and the United Nations 2030 Agenda for Sustainable Development*. Springer Publishing Company, 2017.

[13] Li W., El-Askary H., Lakshmi V., Piechota T., Struppa D.: *Earth Observation and Cloud Computing in Support of Two Sustainable Development Goals for the River Nile Watershed Countries*. Remote Sensing, vol. 12(9), 2020, 1391. https://doi.org/10.3390/rs12091391.

[14] UN ECOSOC: *Exploring space technologies for sustainable development and the benefits of international research collaboration in this context*. Report of the Secretary-General, Commission on Science and Technology for Development, Twenty-third session, Geneva, 23–27 March, 2020, United Nations Economic and Social Council, 2020. https://unctad.org/system/files/official-document/ecn162020d3_en.pdf [access: 19.04.2022].

[15] Woldai T.: *The status of Earth Observation (EO) & Geo-Information Sciences in Africa – trends and challenges*. Geo-spatial Information Science, vol. 23(1), 2020, pp. 107–123. https://doi.org/10.1080/10095020.2020.1730711.

[16] Dhu T., Giuliani G., Juárez J., Kavvada A., Killough B., Merodio P., Ramage S.: *National Open Data Cubes and Their Contribution to Country-level Development Policies and Practices*. Data, vol. 4(4), 2019, 144. https://doi.org/10.3390/data4040144.

[17] Giuliani G., Masó J., Mazzetti P., Nativi S., Zabala A.: *Paving the Way to Increased Interoperability of Earth Observations Data cubes*. Data, vol. 4, 2019, 113. https://doi.org/10.3390/data4030113.

[18] Giuliani G., Chatenoux B., De Bono A., Rodila D., Richard J.P., Allenbach K., Dao H., Peduzzi P.: *Building an Earth Observations Data Cube: Lessons Learned from the Swiss Data Cube (SDC) on Generating Analysis Ready Data (ARD)*. Big Earth Data, vol. 1(1–2), 2017, pp. 100–117. https://doi.org/10.1080/20964471.2017.1398903.

[19] Strobl P., Marchetti P.G.: *The Six Faces of the Data Cube*. [in:] Marchetti P., Soille P. (eds.), *Proceedings of the 2017 conference on Big Data from Space (BIDS' 2017): 28th–30th November 2017 Toulouse (France)*, Publications Office of the European Union, 2017, pp. 32–35. https://doi.org/10.2760/383579.

[20] Asmaryan S., Muradyan V., Tepanosyan G., Hovsepyan A., Saghatelyan A., Astsatryan H., Grigoryan H. et al.: *Paving the Way towards an Armenian Data Cube*. Data, vol. 4(3), 2019, 117. https://doi.org/10.3390/data4030117.

[21] Juárez Carrillo O.J., Merodio Gómez P., Ponce Medina M.D.S., Ornelas de Anda J.L., Corona Iruegas A.A.: *Cubo de datos geoespaciales para el uso de las imágenes satelitales en la generación de información geográfica y estadística*. Realidad, Datos y Espacio Revista Internacional de Estadística y Geografía, vol. 11(3), 2020, pp. 124–139.

[22] INEGI: *Land Use and Vegetation*. Instituto Nacional de Estadística y Geografía. https://www.inegi.org.mx/temas/usosuelo/#Descargas [access: 19.04.2022].

[23] INEGI: *Uso del suelo y vegetación: Metodología*. Instituto Nacional de Estadística y Geografía. https://www.inegi.org.mx/contenidos/temas/mapas/usosuelo/metadatos/metodologia.pdf [access: 1.03.2021].

[24] Lewis A., Oliver S., Lymburner L., Evans B., Wyborn L., Mueller N., Raevksi G. et al.: *The Australian geoscience data cube – foundations and lessons learned*. Remote Sensing of Environment, vol. 202, 2017, pp. 276–292. https://doi.org/10.1016/j.rse.2017.03.015.

[25] Dhu T., Dunn B., Lewis B., Lymburner L., Mueller N., Telfer E., Lewis A. et al.: *Digital Earth Australia – unlocking new value from earth observation data*. Big Earth Data, vol. 1(1–2), 2017, pp. 64–74. https://doi.org/10.1080/20964471.2017.1402490.

[26] INEGI: *Producción y publicación de la Geomediana Nacional a Partir de imágenes del Cubo de Datos Geoespaciales de México: documento metodológico*. Instituto Nacional de Estadística y Geografía, México 2020. https://www.inegi.org.mx/contenidos/productos/prod_serv/contenidos/espanol/bvinegi/productos/nueva_estruc/702825198763.pdf [access: 19.04.2022].

[27] INEGI: *Producción del Índice de Clasificaciones de Agua Superficial desde el Espacio (ICASE) Landsat: documento metodológico*. Instituto Nacional de Estadística y Geografía, México 2021. https://www.inegi.org.mx/contenidos/productos/prod_serv/contenidos/espanol/bvinegi/productos/nueva_estruc/889463903642.pdf [access: 19.04.2022].

[28] UN ECOSOC: *In-depth review of satellite imagery / earth observation technology in official statistics: Prepared by Canada and Mexico*. Economic Commission for Europe Conference of European Statisticians, 67th plenary session, Geneva, 26–28 June 2019, United Nations Economic and Social Council, 2019. https://unece.org/DAM/stats/documents/ece/ces/2019/ECE_CES_2019_16-1906490E.pdf [access: 19.04.2022].

[29] CTEODS: *Estrategia de Indicadores ODS para el 2020*. Segunda sesión del Consejo Consultivo Nacional, Comité Técnico Especializado de los Objetivos de Desarrollo Sostenible, 2019. https://www.snieg.mx/DocumentacionPortal/Consejo/sesiones/doc_22019/ODS.pdf [access: 17.04.2021].

[30] SNIEG: *Acerca del SNIEG*. Sistema Nacional de Información Estadística y Geográfica. https://www.snieg.mx/home/acerca-de/ [access: 15.01.2022].

[31] UN STATS: *SDG Indicators Metadata repository* [Indicator 6.6.1: Change in the extent of water-related ecosystems over time]. United Nations Statistics Division. https://unstats.un.org/sdgs/metadata/files/Metadata-06-06-01a.pdf [access: 1.03.2021].

[32] UN Environment and UN Water: *Measuring change in the extent of water-related ecosystems over time: Sustainable Development Goal Monitoring Methodology Indicator 6.6.1*. 2020. https://files.habitatseven.com/unwater/SDG-Monitoring-Methodology-for-Indicator-6.6.1.pdf [access: 1.03.2021].

[33] UN Water: *Indicator 6.6.1 – Progress on Water-related Ecosystems*. https://sdg6data.org/indicator/6.6.1 [access: 5.03.2021].

[34] INEGI: *National Water Bodies Map*. Instituto Nacional de Estadística y Geografía. https://www.inegi.org.mx/temas/hidrologia/ [access: 1.03.2021].

[35] Mueller N., Lewis A., Roberts D., Ring S., Melrose R., Sixsmith J., Lymburner L. et al.*: Water observations from space: Mapping surface water from 25 years of Landsat imagery across Australia*. Remote Sensing of Environment, vol. 174, 2016, pp. 341–352. https://doi.org/10.1016/j.rse.2015.11.003 [access: 15.01.2022].

[36] UN STATS: *SDG Indicators Metadata repository* [Indicator 15.1.1: Forest area as a proportion of total land area]. United Nations Statistics Division. https://unstats.un.org/sdgs/metadata/files/Metadata-15-01-01.pdf [access: 1.03.2021].

[37] UN FAO: *Global Forest Resources Assessments*. Food and Agriculture Organization of the United Nations. http://www.fao.org/forest-resources-assessment/past-assessments/en/ [access: 15.03.2021].

[38] UN FAO: *Global Forest Resources Assessments 2015*. Food and Agriculture Organization of the United Nations. http://www.fao.org/forest-resources-assessment/past-assessments/fra-2015/en/ [access: 15.03.2021].

[39] Roberts D., Dunn B., Mueller N.: *Open Data Cube Products Using High-Dimensional Statistics of Time Series*. [in:] *IGARSS 2018: IEEE International Geoscience and Remote Sensing Symposium*, IEEE, 2018, pp. 8647–8650. https://doi.org/10.1109/igarss.2018.8518312.