www.editada.org

International Journal of Combinatorial Optimization Problems and Informatics, 17(3), 2026, 317-328. ISSN: 2007-1558.
https://doi.org/10.61467/2007.1558.2026.v17i3.1283

Version Generation of a Multispectral Dataset Based on Satellite Images for Poverty Prediction at Municipal Level in Mexico

Armida Castillo González[1], Vitervo Lopez Caballero[1],
Alan Hernández-Solano[2], Lucia Morales Morales[1]
[1] Centro Nacional de Investigación y Desarrollo Tecnológico, Interior Internado Palmira S/N, Col. Palmira, C.P. 62490 Cuernavaca, Morelos, México.
[2] Instituto de Investigaciones para el Desarrollo con Equidad (EQUIDE), Universidad Iberoamericana, México City, México.
Email address(es): d24ce117@cenidet.tecnm.mx, vitervo.lc@cenidet.tecnm.mx , alan.hernandez@ibero.mx, lmorales@cenidet.tecnm.mx
✉Corresponding author: Vitervo Lopez Caballero

Abstract. Poverty in Mexico has traditionally been measured biennially through socioeconomic surveys that identify people with income deficiencies and at least one of six social deprivation dimensions. This approach faces significant limitations: high costs, operational difficulties in regions with violence, and restricted capacity for timely policy responses, as evidenced during COVID-19. Emerging methods based on satellite imagery and remote sensing expand poverty analysis scope and efficiency. However, implementation faces challenges from scarce primary data, particularly satellite imagery and field measurements essential for predictive model training. This article develops a multispectral database for Mexico designed to facilitate training and validation of advanced machine learning algorithms for poverty prediction. This contribution provides a methodological tool for designing robust and scalable approaches while reducing reliance on traditional socio-economic survey methods.
Keywords: Multispectral dataset, Poverty, Mexico, Machine learning, Satellite images.

Article information
Received: Jan 7, 2026
Accepted: Jun 9, 2026

1 Introduction

Poverty is a structural challenge that affects millions of people around the world. Within the framework of the 2030 Agenda for Sustainable Development, adopted by the Member States of the United Nations (UN) in 2015, Sustainable Development Goal 1 (SDG 1) establishes the goal of "ending poverty in all its forms everywhere" (UN, 2023). This objective, essential for global progress, requires not only effective public policies and sufficient resources, but also the development of innovative tools that make it possible to measure, analyze and address poverty with greater precision and efficiency (World Bank, 2020).

Measuring poverty remains a persistent challenge for governments and international organizations. Traditional methods, such as censuses and surveys, are resource-intensive and have significant limitations, especially in rural, marginalized or hard-to-reach areas (Huang et al., 2021). In this context, emerging technologies offer promising alternatives. The use of satellite imagery, combined with machine learning techniques, has shown great potential to characterize socioeconomic conditions in large geographical areas, including those affected by conflict or with a high incidence of extreme poverty (Chitturi, 2021). Among these methodologies, convolutional neural networks (CNNs) stand out for their ability to analyze images and detect visual patterns linked to developmental indicators (Okaidat et al., 2021).

Mexico, due to its wide geographical and socioeconomic diversity, as well as its recent institutional transformations in poverty measurement, is an ideal case study to explore this type of approach. The country underwent a significant institutional change in 2025 when, through a constitutional reform published on July 17, the National Council for the Evaluation of Social Development

Policy (CONEVAL) was dissolved, and its poverty measurement functions were transferred to the National Institute of Statistics and Geography (INEGI) (INEGI, 2025a).

This institutional transition coincided with the presentation of results that show a notable reduction in multidimensional poverty: according to official INEGI figures, poverty in Mexico decreased by 6.8 percentage points during the period 2022-2024, from 36.3% to 29.6% of the national population, representing a reduction of approximately 8.3 million people in poverty (INEGI, 2025b) — figures reported at the state and national level, as municipal-level estimates have not yet been published under the new INEGI mandate. However, this apparent improvement in poverty indicators has generated intense academic and political debate around the methodological comparability and credibility of the measurements. The institutional change raises questions about methodological continuity, given that CONEVAL had developed a robust multidimensional approach that included both income and social deprivation dimensions in education, health, social security, housing, basic services, and food (La Crónica de Hoy, 2025).

The transfer of these functions to INEGI, while formally maintaining the multidimensional methodology established in the General Law of Social Development, has raised concerns among specialists about the preservation of technical independence and the temporal comparability of measurements (Bloomberg Línea, 2025; COPARMEX, 2025). Additionally, although the reduction in poverty is significant, the same report shows a paradoxical increase in the vulnerable population due to social deprivation, which went from 37.9 million in 2022 to 41.9 million in 2024, suggesting that improvements in income did not necessarily translate into proportional improvements in access to basic services (Alto Nivel, 2025).

Against this backdrop, this article presents the construction of a publicly accessible multispectral database for Mexico, designed as an input for the development and validation of municipal-level poverty detection algorithms. The database integrates vegetation indices (NDVI), nighttime luminosity, land cover, and water bodies— variables extracted for the year 2020, which corresponds to the most recent official municipal-level poverty estimates published by CONEVAL under its five-year measurement cycle — all of which have proven relevant in remote sensing-based poverty prediction studies (Hall et al., 2023; Heitmann, 2019). Together, these satellite-derived indicators capture key dimensions associated with poverty, including agricultural activity, urbanization, access to electricity, and water resource availability (Heitmann, 2019).

In Mexico, there are several public sources that provide satellite images useful for territorial and environmental analysis. These sources offer data that can be used to develop predictive models of poverty at the residential level. The National Institute of Statistics and Geography (INEGI) is one of the main entities that provides access to satellite images. Through its high-resolution imaging program, it provides data with a degree of detail and accuracy ranging from 1 meter to more than 30 cm, which is essential for the analysis of land cover and urbanization (INEGI, n.d.a). The National Commission for the Knowledge and Use of Biodiversity (CONABIO) also offers access to satellite imagery through its Satellite Image Reception System (SRIS), providing data from sensors such as MODIS, VIIRS, and AVHRR (CONABIO, n.d.). In addition, the Mexican Space Agency (AEM), in collaboration with the European Space Agency (ESA), has produced vegetation maps in Mexico using images from the constellation Sentinel-2 (INEGI, n.d.c). On the other hand, the Mexican Institute of Water Technology (IMTA, n.d.) has used satellite images to evaluate water storage in aquifers, useful data for the analysis of water bodies and their relationship with poverty (Instituto Mexicano de Tecnología del Agua, 2023).

These public sources offer a variety of data that can be used for the analysis of poverty in Mexico. However, integrating these various sources into a single, consistent and accessible dataset is non-existent. Therefore, the creation and publication of a dataset that combines NDVI, night lights, land use, and water bodies at the municipal level represents a significant contribution. This effort provides a valuable tool for researchers, policymakers, and organizations seeking to understand and address poverty in Mexico.

This work is considered relevant for three reasons. First, it offers a technical comparison of various Application Programming Interfaces (APIs) used to obtain geospatial data, identifying advantages and limitations of each one (Tamiminia et al., 2020; Soares et al., 2023; Kumar & Mutanga, 2018). Second, it proposes a replicable methodology for the extraction of satellite images and variables at the municipal level in Mexico, which can be used in future studies on poverty and territorial development (Lim et al., 2024; Martínez-Fernández et al., 2024). Finally, the free availability of both the database and the processing scripts will allow researchers, civil society organizations, and policymakers to develop predictive models and verify official measurements, contributing to an informed debate on the robustness and transparency of poverty estimates (Cardille et al., 2023; Jacobsen et al., 2020). This last point is particularly timely: given that INEGI is expected to publish its first municipal-level poverty estimates in 2025, the publicly available scripts can be readily adapted to extract equivalent satellite composites for 2025 and, combined with the 2020 dataset presented here, enable machine learning models to empirically assess whether the institutional transition from CONEVAL to INEGI affects the comparability and consistency of municipal poverty measurements.

The article is organized as follows. Section 2 presents a comparative study of various Application Programming Interfaces (APIs) for obtaining satellite data, evaluating their resolution, band availability, ease of use and access to information. Section 3 describes the methodology implemented to extract geospatial variables at the municipal level, including NDVI, land cover, night lights, and water bodies. Section 4 details the stages of preprocessing, generation of temporal compounds and calculation of metrics by municipality, as well as the final configuration of the dataset. Section 5 discusses the contribution of the dataset, its structure, public availability, potential applications in machine learning models for poverty prediction, and presents a correlation analysis between the satellite-derived variables and official municipal poverty estimates as empirical validation of the dataset's predictive capacity. Finally, section 6 presents the conclusions, highlighting the relevance of the methodological approach and the usefulness of the dataset for research and decision-making in public policies.

2 Comparative study of application programming interfaces (APIs)

In order to select the most suitable platform for the generation of the multispectral dataset, a comparative study of various geospatial APIs was carried out, considering five fundamental evaluation criteria.

First, the spatial and temporal resolution was analyzed, assessing both the level of detail of the images (ideally ≤30 m/pixel) and the availability of historical series with a minimum of five years (Lim et al., 2024). Second, the availability of spectral bands relevant to the calculation of indices, such as NDVI, detection of water bodies and analysis of nightlights, was examined (Martínez-Fernández et al., 2024). Likewise, the ease of use and integration was evaluated, considering the existence of client libraries, accessible documentation and cloud processing tools (Tamiminia et al., 2020). Another criterion was access to data, prioritizing free or low-cost solutions for academic purposes and without severe restrictions on use (Copernicus Data Space Ecosystem, n.d.). Finally, the quality of the documentation and the support of the community were included, reflected in tutorials, forums and specialized technical support (Soares et al., 2023).

Based on the application of these criteria, four APIs that meet the minimum requirements established were identified and analyzed: Google Earth Engine, NASA EARTHDATA, Copernicus Open Access Hub and Planet API. All of them are widely recognized as reference platforms in access to satellite remote sensing data for scientific applications (Soares et al., 2023). Table 1 summarizes the main results of this comparison.

Table 1. Satellite API Comparison

Criterion	Google Earth Engine	NASA Earthdata	Copernicus Open Access Hub	Planet API
Spatial and temporal resolution	10-30 m, 16 days	15-30 m, 1-16 days	10-60 m, 1-10 days	3-5 m, daily
Historical data	More than 40 years	More than 50 years	More than 10 years	Approx. 10 years
Spectral bands for NDVI and water detection	Visible, NIR, SWIR, Thermal	Visible, NIR, SWIR, Thermal	Visible, NIR, SWIR, radar	Visible, NIR
Ease of Use and Integration	Python, JavaScript, R with cloud processing	Python with local download required	Python with basic processing tools	Python, REST API with business model
Access and cost conditions	Completely free for academic use	Free access	Free access	Free limited trial
Documentation and Community Support	Extensive documentation, very active community	Full documentation, active community	Moderate documentation, basic support	Commercial documentation, paid technical support

Based on the comparative analysis carried out, Google Earth Engine (GEE) was selected as the most appropriate platform for the purposes of this study, mainly because of its cloud processing capacity, which eliminates the limitations associated with local hardware, and because of its extensive catalog with more than 40 years of satellite data (Kumar & Mutanga, 2018). In addition, the platform integrates native tools that allow calculating indices widely used in remote sensing, such as NDVI, land cover, night lights and detection of bodies of water (Soares et al., 2023).

An additional advantage is its free access for academic purposes, which distinguishes it from other commercial platforms. In addition, GEE provides libraries in Python and JavaScript, accompanied by clear documentation and an active community of users, which together facilitate the development and replicability of research projects (Kumar & Mutanga, 2018). These features consolidate Google Earth Engine as the most practical and efficient option for multispectral remote sensing studies.

3 Extraction of geospatial variables as municipal poverty proxies

The methodological strategy for the derivation of geospatial indicators was designed with the purpose of capturing key dimensions associated with municipal poverty, taking advantage of the massive data analysis capabilities of Google Earth Engine (GEE) (Tamiminia et al., 2020; Soares et al., 2023; Kumar & Mutanga, 2018; Amani et al., 2020). Four variables were selected with support in scientific literature for their significant correlation with indicators of well-being and socioeconomic development (Zhao et al., 2017). The broken-down indicators are shown below.

3.1 NDVI (Normalized Difference Vegetation Index)

NDVI is a robust indicator for assessing vegetation density and vigor (Kassa et al., 2023). This study used Sentinel-2 images with a spatial resolution of 10 m (Zanaga et al., 2022), following a processing flow that included:

1. Temporal (year 2020) and spatial (cloud cover <10%) filtering.
2. Calculation of NDVI using bands B8 (NIR) and B4 (RED).
3. Image composition from the median to reduce noise and outliers.
4. Obtaining the municipal average of the NDVI.

The NDVI reflects agricultural activity and the availability of natural resources, factors closely associated with rural poverty conditions (Gibson et al., 2021).

3.2 Land Cover

The spatial distribution of land cover is a proxy for the degree of urbanization, agricultural intensity, and the use of natural ecosystems, all of which are linked to socio-economic development (Arino et al., 2020). The ESA WorldCover 2021 product, with a resolution of 10 m, was used, which classifies the earth's surface into 11 categories (Zanaga et al., 2022). The procedure was:

1. Import of the global land cover product.
2. Reclassification of categories into three levels of socioeconomic risk.
3. Estimation of the percentage proportion of each class by municipality.
4. The proportion of urban and agricultural areas was considered an indicator of the degree of local economic development (Wang et al., 2020).

3.3 Nightlights

Night-time light emissions detected by satellite sensors have shown a high correlation with GDP, urbanization and access to services (Hu et al., 2015). For this analysis:

1. Monthly VIIRS images for 2020 were processed.
2. An annual compound was generated by averaging radiance.
3. The average intensity by municipality was calculated.

Light intensity is interpreted as an indirect indicator of infrastructure, urban dynamics, and economic activity, being a solid predictor of territorial inequality (Goldblatt et al., 2018).

3.4 Water Bodies

Access to water resources is a critical factor in the resilience and economic development of communities, particularly in rural contexts (Pekel et al., 2016). In this study:

1. The JRC Global Surface Water assembly was used.
2. Permanent water bodies were selected (presence >90% of the time).
3. The percentage of municipal areas covered by water was calculated.

This indicator allows us to approximate the availability of water resources and their potential influence on food security and community well-being.

Table 2 summarizes the technical characteristics and processing applied to each geospatial variable, as well as its link with socioeconomic indicators of municipal poverty.

Table 2. Geospatial variables used as municipal poverty proxies

Variable	Space Source	Spatial Resolution	Applied Processing	Associated socioeconomic indicator
NDVI	Sentinel-2 (B8, B4)	10m	Temporal filtering (2020), cloudiness <10%; NDVI calculation; composition by median; Municipal average	Agricultural activity and access to natural resources, associated with rural poverty (Gibson et al., 2021)
Land Cover	ESA WorldCover 2021	10 m	Import of the global product; reclassification into socioeconomic risk levels; Percentage estimate by municipality	Degree of urbanization and agricultural intensity, linked to socioeconomic development (Wang et al., 2020)
Night Lights	VIIRS DNB Monthly	500 m	Monthly Image Processing (2020); annual compound by average radiance; Average municipal intensity	Infrastructure, urban dynamics and economic activity, predictor of territorial inequality (Goldblatt et al., 2018)
Water	JRC Global Surface Water	30 m	Permanent water selection (presence >90%); Percentage calculation of municipal area covered by water	Water resource availability, food security, and community well-being (Pekel et al., 2016)

4 Methodology

The construction of the dataset was designed under a systematic approach, composed of four main stages, ranging from the acquisition of geographical boundaries to the final structuring of the dataset, as shown in Figure 1. This methodology was supported by good practices for the massive processing of satellite data (Tamiminia et al., 2020) and by consolidated geospatial analysis frameworks in Google Earth Engine (Cardille et al., 2023).

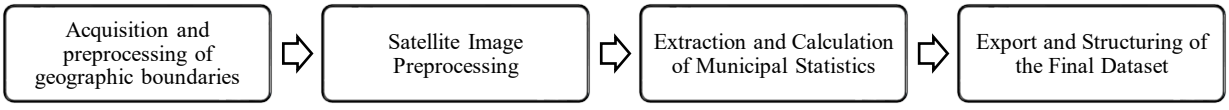


Fig. 1. Stages of the implemented methodology.

4.1 Stage 1: Acquisition and Preprocessing of Geographical Boundaries

4.1.1 Obtaining Municipal Boundaries

The municipal geographic boundaries were obtained from the National Geostatistical Framework of the INEGI, which offers the territorial division of Mexico into different levels of disaggregation to reference statistical information (INEGI, n.d.b). They were downloaded in shapefile format (.cpj, .shp, .shx, .dbf, .prj) from the official INEGI portal and uploaded as custom assets to Google Earth Engine under the identifier projects/poverty-prediction-457518/assets/Mexico (Figure 2).

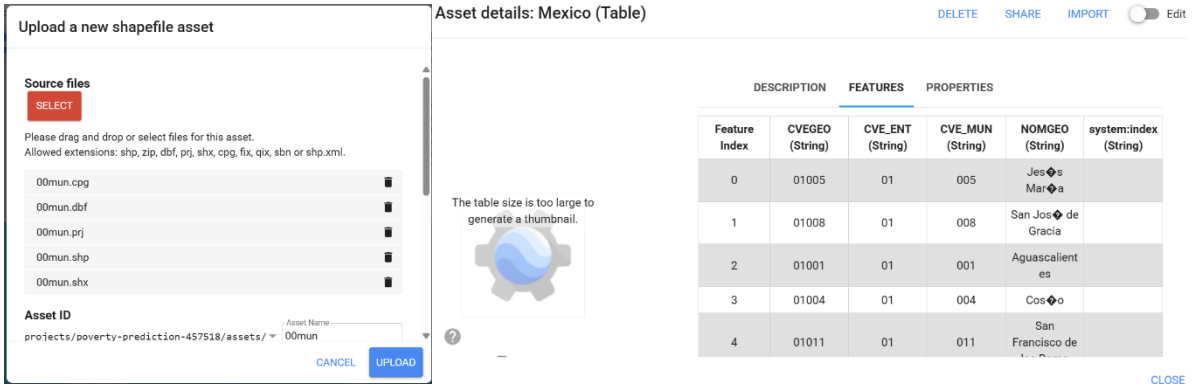


Fig. 2. Municipal boundaries loaded in GEE

4.1.2 Setting Up Assets in Google Earth Engine

Once the vector files are uploaded, the platform automatically converted them to vector table format optimized for distributed processing. The geometric integrity of the polygons was verified, and the unique identifiers were validated: CVE_ENT (federal entity code) and CVE_MUN (municipal code) for the subsequent aggregation of average statistics extraction by municipality. Processing was configured to allow filtering by state using EE. Filter.eq('CVE_ENT', federal entity code) for analysis at the state and municipal level.

4.2 Stage 2: Satellite Image Preprocessing

With the municipal geographic structure duly configured in Google Earth Engine, the satellite image collections that constitute the data sources for the extraction of geospatial variables were pre-processed.

4.2.1 Temporal and Quality Filtering

Systematic filters were applied to the satellite imagery collections for the year 2020 analysis, establishing specific quality thresholds: cloud cover less than 10% for Sentinel-2 optical images (CLOUDY_PIXEL_PERCENTAGE < 10), and data availability greater than 80% for derived products. The full-time range was defined from January 1 to December 31, 2020, to ensure consistent and homogeneous annual coverage among all municipal units (Kassa et al., 2023; Zanaga et al., 2022; Pekel et al., 2016).

4.2.2 Generation of Temporary Compounds

Once the quality filters were applied, specific temporal compounds were generated for each variable: median (.median()) for Sentinel-2 NDVI to minimize atmospheric anomalies, annual average (.mean()) for VIIRS night lights to reduce seasonal variability (Gibson et al., 2021; Hu et al., 2015), and unique images for static products (ESA WorldCover 2021 and JRC Global Surface Water with threshold ≥ 90% permanent presence) (Zanaga et al., 2022; Pekel et al., 2016).

4.2.3 Selection and Preparation of Satellite Products

For this deployment, the following satellite products were configured: Sentinel-2 L1C (COPERNICUS/S2) selecting bands B4 (red) and B8 (near-infrared) for NDVI calculation, ESA WorldCover v200 2021 (ESA/WorldCover/v200/2021) for land cover

classification, VIIRS DNB Monthly Composite (NOAA/VIIRS/DNB/MONTHLY_V1/VCMSLCFG) for night light emissions, and JRC Global Surface Water v1.4 (JRC/GSW1_4/GlobalSurfaceWater) for water body identification Permanent.

Stage 3: Extraction and Calculation of Municipal Statistics

4.3.1 Calculation of Municipal Statistics

Once the satellite products were configured, a generic function was developed to calculate statistics at the municipal level using the reduceRegions method of Google Earth Engine (Soares et al., 2023; Kumar & Mutanga, 2018; Amani et al., 2020). This function made it possible to obtain average values or proportions depending on the variable analyzed, considering the native resolution of each dataset: 10 m for NDVI and land cover, 500 m for night lights and 30 m for bodies of water. Municipal-level aggregation was chosen to ensure direct compatibility with official CONEVAL poverty estimates, which represent the finest administrative level at which these measurements are published in Mexico. Nevertheless, this choice may conceal intra-municipal heterogeneity, particularly in large or socioeconomically diverse municipalities, and the heterogeneous native resolutions of the variables introduce scale-dependent uncertainty in the aggregated statistics, a limitation consistent with the Modifiable Areal Unit Problem (MAUP).

4.3.2 Calculation of Indices and Transformations

On this methodological basis, the following transformations were implemented: calculation of NDVI by normalized difference

$$\text{NDVI: } (B8 - B4) / (B8 + B4) \quad (1)$$

applying the .normalizedDifference() method (Kassa et al., 2023); direct extraction of intensity values for night lights (Gibson et al., 2021; Hu et al., 2015); binary classification for permanent water bodies by thresholding (occurrence ≥ 90) (Pekel et al., 2016) and extraction of thematic categories of land cover without additional transformation (Zanaga et al., 2022).

4.3.3 Distributed Processing by Administrative Units

The application of these transformations was executed in a distributed way for each municipality by iterating on the list of unique municipal codes, applying geo-clipping operations (.clip()) and simultaneous statistical aggregation (Soares et al., 2023; Cardille et al., 2023). A unique identification system was implemented based on the official INEGI codes (CVE_ENT for states, CVE_MUN for municipalities) (INEGI, 2023b).

4.4 Stage 4: Export and Structuring of the Final Dataset

4.4.1 Exporting Results

Once the processing by municipal unit was completed, a dual export system was implemented that generates both tabular statistics and raster images by municipality (Soares et al., 2023). The zonal statistics were exported in CSV format using Export.table.toDrive(), while the georeferenced images were exported individually by municipality using Export.image.toDrive() with different resolutions depending on the source dataset.

4.4.2 Hierarchical Structuring of the Dataset

To ensure the organization of the products, the final dataset was structured with hierarchical organization by municipality, creating individual folders for each administrative unit with standardized nomenclature (state code + municipal code). Each folder contains four georeferenced images corresponding to the variables analyzed: NDVI, Land Cover, NightLights and Water, all with consistent temporal metadata (year 2020). Figure 3 shows an example of the hierarchical structure of the dataset and the nomenclature of the files, using the states of Guerrero and Morelos only as a reference. The complete dataset includes information from all the states of the country (INEGI, 2023b).

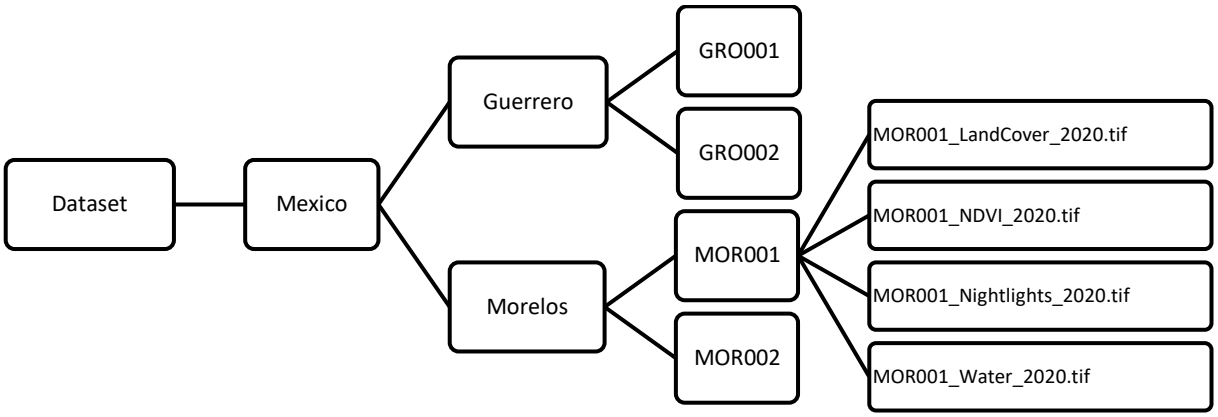


Fig. 3. Dataset hierarchical structure and file naming

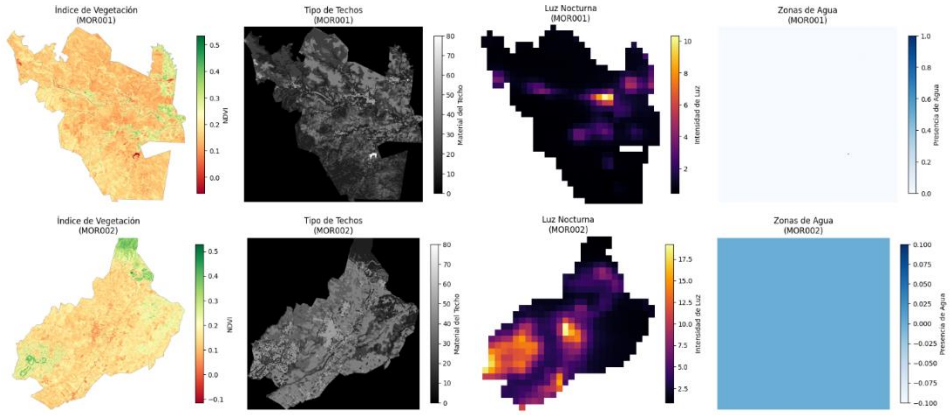
4.4.3 Quality Control and Export Settings

Finally, optimized export parameters were established: maximum pixel limit of 1×10^{13} (maxPixels: 1e13) to handle large municipalities, appropriate data format according to the nature of each variable (Float32 for NDVI and night lights; UInt8 for land and water cover), and automatic clipping to the geometric boundary of each municipality using .bounds() to optimize storage (Cardille et al., 2023; Amani et al., 2020).

5 Contribution

The developed dataset integrates geospatial information processed for the municipalities of Mexico. Its structure includes two main components: aggregated tabular information at the municipal level in CSV format and high-resolution georeferenced images in GeoTIFF format, designed for detailed spatial analysis.

To illustrate the content of the dataset, Figure 4 includes examples of satellite images corresponding to the four main indicators for three representative municipalities of the state of Morelos: NDVI (Sentinel-2, 10 m), land cover (ESA WorldCover, 10 m), night lights (VIIRS, 500 m) and surface water resources (JRC Global Surface Water, 30 m).



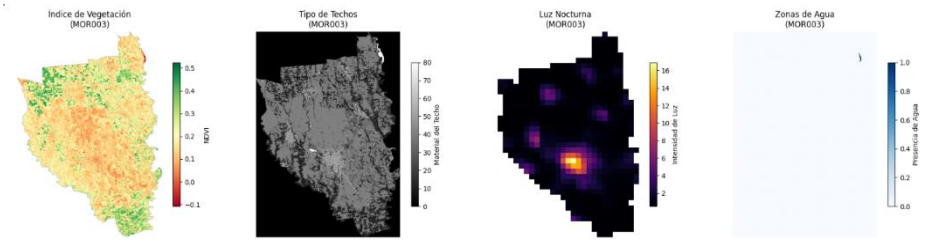


Fig. 4. Satellite variables extracted for three municipalities in the state of Morelos

The visualizations illustrate the ability of satellite indicators to capture relevant socio-economic characteristics.

- The NDVI reveals patterns of agricultural activity (high values in green) versus urban or degraded areas (low values in red), both associated with different levels of economic development.
- Land cover identifies patterns of urbanization and territorial use, where the proportion between urban, agricultural and natural areas indicates the degree of municipal infrastructural development.
- Nightlights reflect the intensity of economic activity and access to electricity. Concentrated patterns indicate active economic centers, while dark areas may indicate less development or limited access to infrastructure.
- Water zones identify the availability of surface water resources, a critical factor for productive activities and quality of life, especially relevant in contexts of rural poverty.

Each municipal record in the dataset includes these four variables, allowing a multidimensional characterization of socioeconomic conditions. The integration of satellite imagery together with tabular data facilitates the training of machine learning models for poverty prediction and territorial development analysis, as well as comparative studies and longitudinal analyses.

The dataset is available in a public GitHub repository (https://github.com/D24CE117-ArmidaCastillo/mexico-poverty-satellite-dataset), under MIT license and in compliance with the FAIR principles to maximize its accessibility and reuse in future research (Jacobsen et al., 2020). The repository includes technical documentation and the processing scripts implemented in Google Earth Engine, ensuring transparency and reproducibility. Although municipal-level tabular statistics are provided in aggregated form, the GeoTIFF images are exported at their original native resolution and clipped to the exact boundary of each municipality, fully preserving intra-municipal spatial variability. Researchers with access to finer-grained socioeconomic ground-truth data — such as AGEB or locality-level indicators — can use the GeoTIFF images to derive satellite variables at those same spatial scales, enabling sub-municipal predictions of such socioeconomic indicators.

The repository contains the GEE extraction script (exportacion_municipiosmx.py) and the technical documentation (INSTRUCTIONS.txt). The dataset is organized hierarchically: M-Mexico/ contains one subfolder per state, each holding a summary CSV file named [StateCode]_Estadisticas_2020.csv and one subfolder per municipality (M-[StateMunCode]/). Each CSV file contains the following fields: state code (CVE_ENT), municipal code (CVE_MUN), municipality name (NOMGEO), variable name (variable), and four zonal statistics — mean, standard deviation, minimum, and maximum — for each of the four satellite-derived variables. Each municipal subfolder contains four GeoTIFF files, one per variable, following the naming convention [StateMunCode]_[Variable]_2020.tif. For example, the folder M-AGS001 corresponds to the municipality of Aguascalientes, whose name and statistics can be found in AGS_Estadisticas_2020.csv, and contains files such as AGS001_NDVI_2020.tif and AGS001_NightLights_2020.tif. Regarding GEE permissions, only a free account registered at earthengine.google.com is required; all satellite products used are publicly available within the GEE data catalogue and require no additional licenses or commercial access.

The design of the dataset is aimed at training machine learning models for the prediction of municipal poverty levels. Its structure favors the implementation of both traditional algorithms (random forest, SVM, gradient boosting) and deep learning architectures capable of integrating tabular data with georeferenced images using convolutional neural networks. The methodology is reproducible, scalable and adaptable, which enables its extension to other states and multiple time periods, enabling longitudinal studies and regional comparative analyses.

The dataset covers 2,478 municipalities and has a total size of 90.0 GB. NightLights and Water achieve complete municipal coverage (100%), while LandCover covers 99.9% and NDVI 99.7% of municipalities. The minimal missing coverage in NDVI and LandCover is attributable to persistent cloud cover in specific municipalities during the 2020 reference period.

5.1 Correlation Analysis with Official Poverty Estimates

To assess the empirical robustness of the geospatial proxies included in the dataset, Pearson and Spearman correlation analyses were conducted between the four satellite-derived variables and official municipal-level poverty indicators published by CONEVAL for 2020.[1] Two poverty measures were examined: multidimensional poverty incidence — defined as the percentage of the municipal population with both insufficient income and at least one social deprivation — and extreme poverty incidence — defined as the percentage of the population with insufficient income and three or more simultaneous social deprivations. The analysis covered 2,461 municipalities for which complete satellite and poverty data were available.

Results are summarized in Table 3. All four variables showed statistically significant correlations with both poverty indicators ($p < 0.001$). NDVI showed a moderate positive correlation with both poverty dimensions, consistent with the association between high vegetation cover and less economically developed rural areas. Land cover presented a moderate negative correlation with both dimensions, indicating that municipalities with greater proportions of urban and agricultural land tend to exhibit lower poverty rates. The presence of permanent surface water bodies — treated as a binary indicator — also showed a negative correlation with both poverty dimensions, suggesting that access to permanent water resources is associated with lower poverty levels. Nighttime light intensity showed a moderate negative correlation with both dimensions, reflecting the well-established link between access to electricity and economic development. Taken together, these results provide preliminary empirical support for the predictive capacity of the dataset and encourage its use in machine learning models for poverty estimation at the municipal level in Mexico.

Table 3. Pearson and Spearman correlations between satellite-derived variables and official municipal poverty indicators (N = 2,461 municipalities).

Satellite variable	Poverty indicator	Pearson r	Pearson p-value	Spearman r	Spearman p-value
NDVI	Poverty (%)	0.4411	<0.001	0.4302	<0.001
NDVI	Extreme poverty (%)	0.4529	<0.001	0.474	<0.001
Night Lights	Poverty (%)	-0.2311	<0.001	-0.2738	<0.001
Night Lights	Extreme poverty (%)	-0.178	<0.001	-0.273	<0.001
Land Cover	Poverty (%)	-0.3526	<0.001	-0.4203	<0.001
Land Cover	Extreme poverty (%)	-0.3979	<0.001	-0.4465	<0.001
Water (binary)	Poverty (%)	-0.3511	<0.001	-0.3635	<0.001
Water (binary)	Extreme poverty (%)	-0.2839	<0.001	-0.3063	<0.001

6 Conclusions

This paper presents a comprehensive methodology for the generation of a multispectral dataset aimed at detecting poverty at the municipal level in Mexico. The integration of multiple geospatial information sources through Google Earth Engine allowed deriving satellite indicators with potential application in machine learning models, which demonstrates the technical feasibility of remote sensing techniques as a complementary tool in the identification of vulnerable areas.

The resulting dataset is a high-value computational resource for researchers in machine learning and applied computer science. Its availability facilitates the development and validation of automated spatial analysis algorithms, based on updated and systematically processed data. Likewise, the proposed methodology is scalable and transferable to other geographical contexts, contributing to the emerging field of geospatial data processing applications for socioeconomic analyses based on artificial intelligence.

Additionally, this paper offers a comparative evaluation of satellite data access platforms, highlighting the suitability of Google Earth Engine for large-scale geospatial big data analysis projects. The implementation of optimized zonal aggregation functions and the dual structure of the dataset (tabular and georeferenced) establish a replicable methodological framework, which can be reused for the construction of similar datasets oriented to machine learning applications in the geospatial domain.

[1] https://www.coneval.org.mx/Medicion/Paginas/Pobreza-municipio-2010-2020.aspx

6 Acknowledgments

This work was supported by the "Instituto de Investigación Aplicada y Tecnología" and the "Universidad Iberoamericana, Ciudad de México".

References

Alto Nivel. (2025, August 13). *México reduce en 8.3 millones la pobreza de 2022 a 2024: INEGI*. https://www.altonivel.com.mx/mexico-reduce-en-8-3-millones-la-pobreza-de-2022-a-2024-inegi/

Amani, M., Ghorbanian, A., Ahmadi, S. A., Kakooei, M., Moghimi, A., Mirmazloumi, S. M., Moghaddam, S. H. A., Mahdavi, S., Ghahremanloo, M., Parsian, S., Wu, Q., & Brisco, B. (2020). Google Earth Engine cloud computing platform for remote sensing big data applications: A comprehensive review. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 13*, 5326–5350. https://doi.org/10.1109/JSTARS.2020.3021052

Cardille, J. A., Crowley, M. A., Saah, D., & Clinton, N. E. (Eds.). (2024). *Cloud-based remote sensing with Google Earth Engine: Fundamentals and applications*. Springer Cham. https://doi.org/10.1007/978-3-031-26588-4

Chen, C., He, X., Liu, Z., Sun, W., Dong, H., & Chu, Y. (2020). Analysis of regional economic development based on land use and land cover change information derived from Landsat imagery. *Scientific Reports, 10*, Article 12721. https://doi.org/10.1038/s41598-020-69716-2

Chitturi, V., & Nabulsi, Z. (2021). *Predicting poverty level from satellite imagery using deep neural networks*. arXiv. https://doi.org/10.48550/arXiv.2112.00011

Comisión Nacional para el Conocimiento y Uso de la Biodiversidad. (2021, October 23). *Imágenes de satélite en línea*. Biodiversidad Mexicana. https://www.biodiversidad.gob.mx/region/imagenes-satelite

Confederación Patronal de la República Mexicana. (2025, June 25). *COPARMEX advierte que la desaparición del CONEVAL afecta la evaluación y la rendición de cuentas en el combate a la pobreza*. https://coparmex.org.mx/coparmex-advierte-que-la-desaparicion-del-coneval-afecta-la-evaluacion-y-la-rendicion-de-cuentas-en-el-combate-a-la-pobreza/

Copernicus Data Space Ecosystem. (n.d.). *Copernicus Data Space Ecosystem*. https://dataspace.copernicus.eu/

Flores, Z. (2025, July 29). Medición de la pobreza en México enfrenta dudas por cifras creíbles y agenda política. *Bloomberg Línea*. https://www.bloomberglinea.com/latinoamerica/mexico/medicion-de-la-pobreza-en-mexico-enfrenta-dudas-por-cifras-creibles-y-agenda-politica/

Gibson, J., Olivia, S., Boe-Gibson, G., & Li, C. (2021). Which night lights data should we use in economics, and where? *Journal of Development Economics, 149*, Article 102602. https://doi.org/10.1016/j.jdeveco.2020.102602

Goldblatt, R., Stuhlmacher, M. F., Tellman, B., Clinton, N., Hanson, G., Georgescu, M., Wang, C., Serrano-Candela, F., Khandelwal, A. K., Cheng, W.-H., & Balling, R. C., Jr. (2018). Using Landsat and nighttime lights for supervised pixel-based image classification of urban land cover. *Remote Sensing of Environment, 205*, 253–275. https://doi.org/10.1016/j.rse.2017.11.026

Hall, O., Dompae, F., Wahab, I., & Dzanku, F. M. (2023). A review of machine learning and satellite imagery for poverty prediction: Implications for development research and applications. *Journal of International Development, 35*(7), 1753–1768. https://doi.org/10.1002/jid.3751

Heitmann, S., & Buri, S. (2019). *Poverty estimation with satellite imagery at neighborhood levels: Results and lessons for financial inclusion from Ghana and Uganda*. International Finance Corporation. https://www.ifc.org/content/dam/ifc/doc/mgrt/ifc-2019-poverty-estimation-with-satellite-imagery-at-neighborhood-levels.pdf

Huang, L. Y., Hsiang, S. M., & Gonzalez-Navarro, M. (2021). *Using satellite imagery and deep learning to evaluate the impact of anti-poverty programs* (NBER Working Paper No. 29105). National Bureau of Economic Research. https://doi.org/10.3386/w29105

Instituto Mexicano de Tecnología del Agua. (2023, September 13). *Información satelital del agua subterránea: Herramienta clave para la sustentabilidad ambiental*. Gobierno de México. https://www.gob.mx/imta/articulos/informacion-satelital-del-agua-subterranea-herramienta-clave-para-la-sustentabilidad-ambiental

Instituto Nacional de Estadística y Geografía. (2025, August 13). *Comunicado de prensa 118/25: Pobreza multidimensional*. INEGI. https://www.inegi.org.mx/contenidos/saladeprensa/boletines/2025/pm/pm2025_08.pdf

Instituto Nacional de Estadística y Geografía. (2025, August 13). *Reporte de resultados 27/25: Análisis de los resultados de la medición de la pobreza multidimensional, 2024*. INEGI. https://www.inegi.org.mx/contenidos/saladeprensa/boletines/2025/pm/pm2025_RR_08.pdf

Instituto Nacional de Estadística y Geografía. (n.d.). *Imágenes de alta resolución*. INEGI. https://www.inegi.org.mx/temas/imagenes/imgar/

Instituto Nacional de Estadística y Geografía. (n.d.). *Mapa Digital de México V6.1*. INEGI. https://gaia.inegi.org.mx/mdm6/

Instituto Nacional de Estadística y Geografía. (n.d.). *Marco Geoestadístico*. INEGI. https://www.inegi.org.mx/temas/mg/

Jacobsen, A., de Miranda Azevedo, R., Juty, N., Batista, D., Coles, S., Cornet, R., Courtot, M., Crosas, M., Dumontier, M., Evelo, C. T., Goble, C., Guizzardi, G., Hansen, K. K., Hasnain, A., Hettne, K., Heringa, J., Hooft, R. W. W., Imming, M., Jeffery, K. G., ... Schultes, E. (2020). FAIR principles: Interpretations and implementation considerations. *Data Intelligence, 2*(1–2), 10–29. https://doi.org/10.1162/dint_r_00024

Kumar, L., & Mutanga, O. (2018). Google Earth Engine applications since inception: Usage, trends, and potential. *Remote Sensing, 10*(10), Article 1509. https://doi.org/10.3390/rs10101509

Land Cover CCI. (2016). *Land Cover CCI product user guide: Version 2.5*. European Space Agency. https://maps.elie.ucl.ac.be/CCI/viewer/download/ESACCI-LC-PUG-v2.5.pdf

Lim, S. L., Sreevalsan-Nair, J., & Daya Sagar, B. S. (2024). Multispectral data mining: A focus on remote sensing satellite images. *WIREs Data Mining and Knowledge Discovery, 14*(2), Article e1522. https://doi.org/10.1002/widm.1522

Martínez Prentice, R., Villoslada, M., Ward, R. D., Bergamo, T. F., Joyce, C. B., & Sepp, K. (2024). Synergistic use of Sentinel-2 and UAV-derived data for plant fractional cover distribution mapping of coastal meadows with digital elevation models. *Biogeosciences, 21*, 1411–1431. https://doi.org/10.5194/bg-21-1411-2024

Okaidat, A., Melhem, S. Z., Alenezi, H., & Duwairi, R. (2021). Using convolutional neural networks on satellite images to predict poverty. In *2021 12th International Conference on Information and Communication Systems (ICICS)* (pp. 164–170). IEEE. https://doi.org/10.1109/ICICS52457.2021.9464598

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

Tamiminia, H., Salehi, B., Mahdianpari, M., Quackenbush, L., Adeli, S., & Brisco, B. (2020). Google Earth Engine for geo-big data applications: A meta-analysis and systematic review. *ISPRS Journal of Photogrammetry and Remote Sensing, 164*, 152–170. https://doi.org/10.1016/j.isprsjprs.2020.04.001

Tiruneh, G. A., Meshesha, D. T., Adgo, E., Tsunekawa, A., Haregeweyn, N., Fenta, A. A., Alemayehu, T. Y., Mulualem, T., Fekadu, G., Demissie, S., & Reichert, J. M. (2023). Mapping crop yield spatial variability using Sentinel-2 vegetation indices in Ethiopia. *Arabian Journal of Geosciences, 16*, Article 631. https://doi.org/10.1007/s12517-023-11754-x

Torres Cruz, I. (2025, August 19). Medición de la pobreza por INEGI, ¿se redujo o se midió convenientemente? *La Crónica de Hoy*. https://www.cronica.com.mx/academia/2025/08/19/medicion-de-la-pobreza-por-inegi-se-redujo-o-se-midio-convenientemente/

United Nations. (2023). *The Sustainable Development Goals report 2023: Special edition*. United Nations. https://unstats.un.org/sdgs/report/2023/

Velastegui-Montoya, A., Montalván-Burbano, N., Carrión-Mero, P., Rivera-Torres, H., Sadeck, L., & Adami, M. (2023). Google Earth Engine: A global analysis and future trends. *Remote Sensing, 15*(14), Article 3675. https://doi.org/10.3390/rs15143675

World Bank. (2020). *Poverty and shared prosperity 2020: Reversals of fortune*. World Bank. https://doi.org/10.1596/978-1-4648-1602-4

Yu, B., Shi, K., Hu, Y., Huang, C., Chen, Z., & Wu, J. (2015). Poverty evaluation using NPP-VIIRS nighttime light composite data at the county level in China. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 8*(3), 1217–1229. https://doi.org/10.1109/JSTARS.2015.2399416

Zanaga, D., Van De Kerchove, R., Daems, D., De Keersmaecker, W., Brockmann, C., Kirches, G., Wevers, J., Cartus, O., Santoro, M., Fritz, S., Lesiv, M., Herold, M., Tsendbazar, N.-E., Xu, P., Ramoino, F., & Arino, O. (2022). *ESA WorldCover 10 m 2021 v200* [Data set]. European Space Agency. https://doi.org/10.5281/zenodo.7254221

Zhao, N., Liu, Y., Cao, G., Samson, E. L., & Zhang, J. (2017). Forecasting China's GDP at the pixel level using nighttime lights time series and population images. *GIScience & Remote Sensing, 54*(3), 407–425. https://doi.org/10.1080/15481603.2016.1276705