Enhancing Vegetation Mapping in Mexico through Artificial Intelligence and Remote Sensing Techniques

Rodolfo Orozco Gálvez [1], Humberto Ramos Ramos [2], Juan Carlos Camacho Pérez [3], José Luís Ornelas de Anda [4], Carlos Manuel López López [5], Alexis Karim Ahedo Díaz [6]

[1] INEGI, Deputy General Directorate of Natural Resources and Environment, Aguascalientes, Mexico – rodolfo.orozco@inegi.org.mx
[2] INEGI, Directorate of Natural Resources, Aguascalientes, Mexico – humberto.ramos@inegi.org.mx
[3] INEGI, Directorate of Analysis of Information on Natural Resources and Environment, Aguascalientes, Mexico – juan.camacho@inegi.org.mx
[4] INEGI, Deputy Directorate General of Research, Aguascalientes, Mexico – jose.ornelas@inegi.org.mx
[5] INEGI, Deputy Directorate of Information Integration on Natural Resources and Environment, Aguascalientes, Mexico – carlos.lopezl@inegi.org.mx
[6] INEGI, Department of Vegetation and Land Use, Aguascalientes, Mexico – alexis.ahedo@inegi.org.mx

Keywords: INEGI, Vegetation, Land Use, AI, Deep Learning.

Abstract

Mexico, recognized for its exceptional biodiversity, is home to over 58 vegetation types and nearly 30,000 documented plant species. This remarkable ecological variety is influenced by the country’s complex topography, diverse climates, and varying soil conditions. Since 1978, the National Institute of Statistics and Geography (INEGI) has been pivotal in understanding the distribution and condition of Mexico's vegetation. INEGI’s methods have progressed from traditional analogue mapping to sophisticated digital formats, utilizing satellite imagery, other ancillary geospatial data layers and advanced photointerpretation techniques.
The data generation process follows rigorous methodologies that are publicly accessible, and dedicated teams across Regional Directorates and State Coordination Offices oversee the mapping of nearly 2 million km² of territory with a lean workforce of just 30 personnel. This information serves as a National Interest Information, mandated for use by government entities.
Recent advancements have underscored the need for innovative modelling and processing capabilities. INEGI are currently exploring artificial intelligence applications, particularly the use of multilayer perceptron neural networks, to enhance vegetation and land-use detection. Robust quality assurance and control measures aligned with ISO-2859 standards are integrated. This article showcases how these initiatives leverage AI to improve data accuracy and processing efficiency, thereby revolutionizing national vegetation mapping and contributing to sustainable land management practices. By highlighting collaborative efforts and outcomes achieved, this work aims to foster a deeper understanding of ecological dynamics and resource management in Mexico.

1. Introduction

Mexico’s vast and diverse ecosystems, spanning nearly 2 million square kilometers of continental territory, present a significant challenge for accurate vegetation mapping. The country’s complex topography and wide range of vegetation types necessitate precise and efficient mapping techniques to support sustainable resource management and ecological conservation. The National Institute of Statistics and Geography (*INEGI*) has been instrumental in advancing vegetation mapping methodologies, evolving from traditional analogue approaches to modern digital processes that harness remote sensing technologies.

Vegetation mapping in Mexico has been conducted for over 45 years and is currently published every five years under the title "Land Use and Vegetation Map Information at a 1:250,000 Scale" (*CUSUEV*), with the most recent edition corresponding to 2018 (Series VII).

The primary objective of this research is to harness deep learning techniques and satellite imagery to enhance vegetation mapping in Mexico. Specifically, this study focuses on optimizing workflows for satellite data processing, refining classification models, and evaluating the resulting improvements in classification accuracy.

2. Methodology

The expertise of fieldwork professionals was combined with the interpretation of remote sensing imagery to identify vegetation types and landscapes, along with the capabilities of artificial intelligence tools—particularly deep learning, in order to develop a supervised classification.

The analysis and development of the predictive model were carried out using 42 individual variables, organized into raster bands stored in GeoTiff format.

The data employed in implementing a new methodology for digital vegetation mapping in Mexico were obtained from various satellite platforms, such as *Landsat-8* and *Sentinel-2*, which provide multispectral images essential for analyzing diverse geographic phenomena.

2.1 Mexican Geospatial Data Cube (*CDGM*)

The *CDGM* is a system that facilitates user access to satellite imagery of interest, based on specific locations and acquisition dates. Since 2018, *INEGI* has promoted this project, which serves as an implementation instance of the *Open Data Cube* tool (https://www.opendatacube.org), developed by *Digital Earth Australia*. The *CDGM* employs an index grid and an *Albers*

Equal Area projection with the *GRS80* ellipsoid, which are well-suited for raster data (see Figure 1).

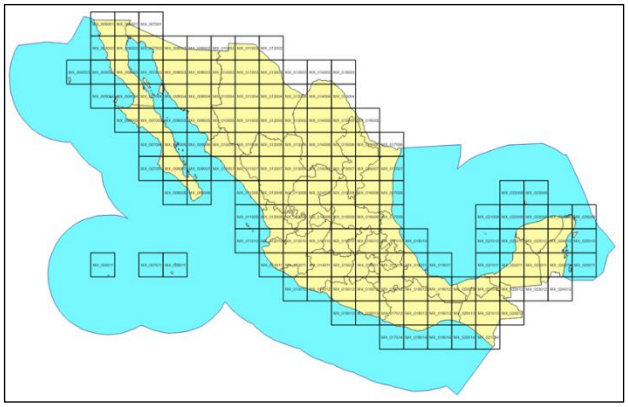

Figure 1. General grid of the *CDGM*, composed of 144 tiles, each measuring 150 km per side.

Among its many advantages, this grid enables compatibility across various raster datasets, such as *Landsat*, *Sentinel*, and *MODIS*, by nesting their diverse resolutions. It also optimizes data processing times through computational parallelization techniques. Regarding this projection, it facilitates homogeneous area measurements at a semi-continental scale, which is ideal for the region of Mexico.

The grid's extreme coordinate specifications, expressed in meters, are (900 000, 300 000) for the minimum and (4 200 000, 2 400 000) for the maximum.

Through the *CDGM*, it is possible to generate spatiotemporal statistics, such as Geometric Median (*Geomedian*), Tasseled Cap Transformations (*TCWBG*), Normalized Difference Vegetation Indices (*NDVI*), and Water Observations from Space (*WOfS*), among others, using the *datacube_stats* module. Additionally, more specific products can be developed by creating custom programming codes within the CDGM to meet analytical needs.

2.2 Main Variables of Analysis

The base year was established as 2018, with the aim of comparing the results against the official publication of the latest version of the *CUSUEV*. Variables that exhibit a strong influence in the vegetation types were identified. Data derived from these variables were registered to the previously described homogeneous coordinate system and spatial resolution, thereby allowing the integration of inputs from diverse sources within a single geospatial grid.

2.2.1 Surface Reflectance Spectral Data

2.2.1.1 Landsat Geomedian

The spectral bands from the *Landsat 8 Collection 2 Level 2* satellite were used as analysis variables, employing the *Geomedian*, which provides representative spectral values for the selected time period by removing "noise" (clouds, terrain or cloud shadows, and cirrus clouds) using the quality band (*pixel_QA*) provided by the *United States Geological Survey (USGS)*, which ensures spatial consistency even at the boundaries between source scenes or index grid cells.

The following raster bands were considered as analysis variables:

- Coastal aerosol
- Blue
- Green
- Red
- Near-infrared (*NIR*)
- Short-wave infrared 1 (*SWIR1*)
- Short-wave infrared 2 (*SWIR2*)

2.2.1.2 Principal Components Analysis (*PCA*)

PCA is an algorithm within a group of techniques aimed at reducing the dimensionality of a dataset caused by the presence of three or more variables. The algorithm exploits the dependencies among the variables to represent them in a simpler form without significant data loss, making *PCA* one of the most widely used and robust dimensionality reduction techniques. The *PCA* algorithm was applied to the blue, green, red, *NIR*, *SWIR1*, and *SWIR2* variables, and the first three principal components were extracted to be used as features in the classifier model:

- PCA1
- PCA2
- PCA3

2.2.2 Spectral Indices

2.2.2.1 Normalized Difference Vegetation Index (*NDVI*)

NDVI is one of the most widely employed vegetation indices among the extensive array of available indices. Its values range from -1 to 1, indicating the density of green vegetation and the level of photosynthetic activity. This index exploits the reflectance characteristics of green vegetation, which exhibits low reflectance in the red portion of the spectrum (*RED*) and high reflectance in the near-infrared (*NIR*) range.

NDVI processing was conducted using the *CDGM*, incorporating all available image observations for the base year. From these data, the following statistical metrics (analysis variables) were computed:

- Mean
- Maximum
- Minimum
- Median
- Standard Deviation

These metrics enable the detection of changes in vegetation as well as in land cover and land use. *NDVI* is calculated using Equation (1), and subsequently, the listed statistical measures are derived for all satellite observations available during the study period.

$$NDVI = \frac{NIR - RED}{NIR + RED}, \quad (1)$$

where $NDVI$ = normalized difference vegetation index
NIR = near-infrared band
RED = red band

2.2.2.2 Modified Normalized Difference Water Index (*MNDWI*)

MNDWI is a metric used for the detection and assessment of water bodies in terrestrial landscapes. This index is based on the normalized difference between the *GREEN* and *SWIR* spectral bands of an image.

$$MNDWI = \frac{GREEN - SWIR}{GREEN + SWIR}, \quad (2)$$

where $MNDWI$ = modified normalized difference water index
$GREEN$ = green band
$SWIR$ = short-wave infrared band

The resulting *MNDWI* value is obtained using Equation (2) and ranges from -1 to 1, where positive values indicate the presence of water, since water absorbs radiation in the *SWIR* bands and reflects radiation in the *GREEN* band. Conversely, negative values typically represent terrestrial areas, which tend to reflect more radiation in *GREEN* than *SWIR*.

2.2.2.3 Normalized Difference Tillage Index (*NDTI*)

The *NDTI* is a metric employed for the detection of tilled soil areas. Its incorporation as an analysis variable has contributed to distinguishing agricultural zones from bare soil. The index is calculated using Equation (3), which employs the *SWIR1* and *SWIR2* bands from *Landsat* mission sensors, demonstrating enhanced performance in identifying silt and clay soils.

$$NDTI = \frac{SWIR1 - SWIR2}{SWIR1 + SWIR2} \quad (3)$$

where $NDTI$ = normalized difference tillage index
$SWIR1$ = short-wave infrared 1 band
$SWIR2$ = short-wave infrared 2 band

2.2.2.4 Bare Soil Index (*BSI*)

BSI is used to assess the extent of bare soil areas within a landscape. This metric is defined by Equation (4).

$$BSI = \frac{(RED + SWIR) - (NIR + BLUE)}{(RED + SWIR) + (NIR + BLUE)} \quad (4)$$

where BSI = bare soil index
RED = red band
$BLUE$ = blue band
$SWIR$ = short-wave infrared band
NIR = near-infrared band

The resulting *BSI* value ranges from -1 to 1, where negative values indicate the presence of bare soil, since surfaces lacking vegetation reflect more radiation in the *NIR* band than in the *RED* band. Conversely, positive values suggest the presence of vegetation cover, as vegetation reflects similarly in both the *NIR* and *RED* bands.

2.2.3 Topography

For the topographic analysis, the *NASA Shuttle Radar Topography Mission (SRTM) Version 3* (V3, 2014) was employed. One of the most notable features of the *SRTM* mission is its capability to acquire global-scale topographic elevation data. From the *SRTM* data, a Digital Elevation Model (*DEM*) of the Earth was generated with an approximate spatial resolution of 30 m for most regions. The *DEM* provides detailed information about the terrain, including mountains, valleys, and land surfaces, and is used in various applications ranging from cartography and natural resource management to urban planning and scientific research.

Vegetation species are distributed across specific elevation intervals, which, to some extent, determine their access to moisture, precipitation, temperature, solar exposure, and groundwater depth. Likewise, the presence of steep slopes can limit water retention from precipitation and result in higher soil erosion rates, leading to shallow and poorly developed soils. For this reason, the variables selected for analysis include slope, topographic position index, roughness, and illumination, with elevation serving as the central datum.

2.2.3.1 Slope

Slope is derived from the *DEM* using surface analysis techniques. At each point in the *DEM*, the slope is determined as the change in elevation along a specific direction. The resulting slope value, expressed in degrees, facilitates the interpretation of terrain inclination. This information is crucial for identifying mountainous regions and areas with limited water retention, among other applications.

2.2.3.2 Topographic Position Index (*TPI*)

TPI evaluates the relative position of a point based on its elevation compared to that of its surroundings. It is calculated by subtracting the elevation of the point from the average elevation of its neighbouring points (see Equation 5), with positive values indicating elevated positions (peaks) and negative values representing depressions. This index is useful in geomorphological studies, terrain classification, and canyon detection, providing key insights into landscape structure.

$$TPI = Z - Z_{avg}, \quad (5)$$

where TPI = topographic position index
Z = elevation of the point in question
Z_{avg} = average elevation of the neighbouring points in a given area

2.2.3.3 Terrain Roughness

This variable, calculated according to Equation (6), measures the variability of the topography in an area by determining the standard deviation of the elevations within a region of the *DEM*. High values indicate rugged terrain with variable elevations, whereas low values reflect smoother, more uniform surfaces.

$$R = \sqrt{\frac{\sum (Z_i - Z_{avg})^2}{N}}, \quad (6)$$

where R = terrain roughness
Z_i = elevation of each pixel within the area of interest
Z_{avg} = average elevation of all the pixels in the area of interest
N = total number of pixels within the area of interest

2.2.3.4 Terrain orientation (Illumination)

To approximate terrain orientation, hillshade derived from the *DEM* was employed using a light source set at 180° (from the south) with a 45° elevation angle, utilizing the *GDAL* tool. Although hillshade is typically used for visual terrain analysis, it was adopted in this study as a quantitative expression of terrain orientation (i.e., aspect or exposure).

Terrain orientation significantly influences local and microclimatic conditions, particularly on slopes. In the northern hemisphere, south-facing slopes receive more solar radiation, resulting in warmer and drier conditions, whereas north-facing slopes tend to be cooler and more humid, thus promoting more robust vegetation development. This variation may determine the

presence of distinct vegetation types, making terrain orientation a relevant variable for the analysis.

2.2.4 Climatology

2.2.4.1 Cloud cover percentage

Using the quality band (*pixel_QA*) from *Landsat* images, the percentage of cloud presence per pixel was quantified according to Equation (7). This value was calculated by dividing the number of cloud detections by the number of valid observations and multiplying by 100. This analysis resulted in the product termed *Cloud Observations from Space (COfS)*, derived through CDGM.

$$COfS = \frac{cloud}{total} \times 100, \quad (7)$$

where $COfS$ = cloud cover percentage
$cloud$ = number of cloud detections
$total$ = number of valid observations

2.2.4.2 Temperature

Environmental temperature is a crucial factor in the biological processes of plants, influencing transpiration, respiration, and growth with both immediate and long-term effects.

In technologically managed agriculture, such as in greenhouses, thermal conditions are controlled because the temperature of the plant does not always match that of the surrounding environment—plants can cool through evaporation or warm via irradiation. Specific temperature intervals can either promote or limit the development of plant species.

To evaluate the thermal conditions of a pixel, four statistical variables were integrated into the model:

- Annual mean temperature
- Maximum temperature of the warmest month
- Minimum temperature of the coldest month
- Coefficient of variation

2.2.4.3 Precipitation

Precipitation is a key climatic factor for vegetation, as it regulates the availability of water for plants. Plant species exhibit diverse water requirements, ranging from those that thrive in high-precipitation regions to those adapted to seasonal or permanent water scarcity, which ultimately defines their distribution across various vegetation types. Precipitation is influenced by atmospheric humidity, condensation temperature, and the transport of moisture by prevailing winds. Considering the water needs of the plant species present in Mexico, vegetation classes were delineated based on precipitation value intervals by integrating the following four variables into the model:

- Annual total precipitation
- Precipitation of the driest month
- Precipitation of the wettest month
- Annual precipitation variability

2.2.4.4 Martonne Aridity Index

Developed in 1926 by Philippe Martonne, the *Martonne Aridity Index* is a key metric for assessing aridity and classifying climatic conditions based on the availability of water in the soil. It is based on the relationship between annual precipitation and the average temperature of the warmest month, as defined by Equation (8).

$$MAI = \frac{P}{T} + 10, \quad (8)$$

where MAI = Martonne Aridity Index
P = annual precipitation in millimetres
T = average monthly temperature of the warmest month in degrees Celsius

A higher index value indicates more humid climatic conditions, while lower values reflect arid conditions. The climatic categories include arid, semi-arid, sub-humid, humid, and super-humid zones, allowing for a detailed analysis of climatic regimes across different regions.

2.2.5 Vegetation structure (Canopy height)

The forest canopy height was estimated using data obtained from *NASA's Global Ecosystem Dynamics Investigation (GEDI)* mission. This system employs *Light Detection and Ranging (LiDAR)* technology mounted on the *International Space Station* to capture the three-dimensional structure of vegetation in forest ecosystems worldwide. Data for the national territory were downloaded and processed for analysis.

2.2.6 Radar Backscatter Data

As a complement to the optical *Landsat* imagery, products derived from the radar scenes of the *Sentinel-1* satellite were utilized. These scenes were pre-processed to the *Radiometric Terrain Correction (RTC)* level using the On-demand platform of the *Alaska Satellite Facility (HyP3)*. Subsequently, a weighted composite of the available scenes from both orbits was generated—known as the *Local Resolution Weighted Composite (LRW)*—using the sensor's dual polarization. To represent the composite in *RGB* format, a new band was created based on the ratio between the gamma polarizations.

The following bands were considered as variables:

- $\gamma^{\theta}_{VV} LRW$ (Vertical-Vertical)
- $\gamma^{\theta}_{VH} LRW$ (Vertical-Horizontal)
- $\gamma^{\theta}_{VV} / \gamma^{\theta}_{VH} LRW$

2.2.7 Geospatial location

Geospatial location is fundamental for the classification of vegetation and land-use, as it determines the climatic, geographic, and topographic conditions that influence their distribution. Factors such as latitude and longitude shape unique patterns that explain the predominance of certain vegetation types and land uses in different regions.

2.3 Sampling sites

2.3.1 Geospatial database

Both the training and validation points used for the generation and evaluation of the model are stored in a relational *PostgreSQL* database. This database incorporates tables, views, triggers, domains, rules, functions, and extensions that optimize data management and manipulation, ensuring the integrity, centralization, uniformity, and availability of records.

The database includes the *PostGIS* extension, which enables the storage, indexing, and querying of geospatial data, facilitating its visualization and analysis through Geographic Information Systems (*GIS*) such as *QGIS*, which is employed in this case study.

2.3.2 Training points

The sampling plan implemented for field data collection followed a stratified random approach, in which strata were delineated based on the vegetation formations classification proposed by INEGI in 2014. The spatial distribution of the strata was determined by considering the area occupied by each formation within the national territory, ensuring the sampling of all ecosystems—including those with restricted distribution and subject to intense disturbances, such as the Mountain Mesophilic Forest (Gual-Díaz and Rendón-Correa, 2014).

Random sampling within each vegetation formation was designed to capture a wide range of conditions, including:

a) Variability in land use and vegetation, ranging from primary areas to disturbed or secondary sites, as well as agricultural and built-up zones.
b) Ecologically significant sites, such as protected natural areas, recharge zones, and priority conservation areas.
c) Human settlements and urbanized zones.

Sampling intensity was defined based on the area of each vegetation formation within the national territory, adjusted using Equation (9) to optimize point distribution. Key environmental variables were considered to minimize territorial heterogeneity and maximize sampling representativeness (Priego-Santander et al., 2013; Díaz et al., 2012).

$$I = \frac{n}{N} \times 100, \quad (9)$$

where I = sampling intensity expressed as a percentage
n = sample size
N = population size

Each sampling site covered 900 m² (30 x 30 m), equivalent to the coverage of a *Landsat* sensor pixel. To date, 231,000 sampling sites have been generated, distributed proportionally to the area occupied by each vegetation type, and used for training through photointerpretation of satellite images (see Table 1).

2.3.3 Training point labels

The assignment of labels to the training points consisted of classifying each sampling site according to the vegetation type, its condition, or the land use present. This classification was performed through a combined visual interpretation of *Landsat* imagery and very high-resolution images, such as those provided by *Maxar* or *Google Earth*. This approach enables a precise and contextualized assignment of labels in the classification model.

Key	Vegetation Type	National Area (%)	Number of Samples
ACUI	Aquaculture	0.1	140
AGR_AN	Annual Agriculture	15.8	36 603
AGR_PER	Permanent Agriculture	0.9	2 148
AH	Human Settlements	1.1	2 565
BC	Cultivated Forest	0.0	88
BCO/P	Primary Coniferous Forest	6.5	15 025
BCO/S	Secondary Coniferous Forest	2.0	4 712
BE/P	Primary Oak Forest	5.4	12 585
BE/S	Secondary Oak Forest	2.6	6 035
BM/P	Primary Mountain Mesophilic Forest	0.7	1 524
BM/S	Secondary Mountain Mesophilic Forest	0.3	588
EOTL/P	Primary Special Woody Other Types	0.1	282
EOTL/S	Secondary Special Woody Other Types	0.1	208
EOTnL/P	Primary Special Non-Woody Other Types	0.1	180
EOTnL/S	Secondary Special Non-Woody Other Types	0.0	1
H2O	Water Bodies	1.3	3 107
MXL/P	Primary Xerophytic Woody Shrubland	9.3	21 487
MXL/S	Secondary Xerophytic Woody Shrubland	1.2	2 699
MXnL/P	Primary Xerophytic Non-Woody Shrubland	17.1	39 490
MXnL/S	Secondary Xerophytic Non-Woody Shrubland	1.6	3 624
OT	Other Lands	0.5	1 224
P	Grassland	15.7	36 273
SC/P	Primary Deciduous Tropical Forest	5.4	12 523
SC/S	Secondary Deciduous Tropical Forest	3.6	8 416
SP/P	Primary Evergreen Tropical Forest	4.1	9 407
SP/S	Secondary Evergreen Tropical Forest	1.0	2 287
SSC/P	Primary Semi-Deciduous Tropical Forest	1.4	3 281
SSC/S	Secondary Semi-Deciduous Tropical Forest	0.6	1 406
VHL/P	Primary Woody Hydrophilic Vegetation	0.6	1 290
VHL/S	Secondary Woody Hydrophilic Vegetation	0.0	99
VHnL/P	Primary Non-Woody Hydrophilic Vegetation	0.7	1 697
VHnL/S	Secondary Non-Woody Hydrophilic Vegetation	0.0	6

Table 1. Samples assigned for each vegetation type.

2.4 Implementation of Deep Learning in image classification

2.4.1 Deep Learning

Deep Learning (DL) is a specialized subfield of *Machine Learning (ML)* focused on the progressive extraction of hierarchical representations from data. It is distinguished by the construction of multiple successive layers, each capturing increasingly significant patterns and features within the dataset. The depth of a model refers to the number of layers involved in its architecture. Unlike shallow learning, which typically works with one or two layers, deep learning models can incorporate dozens or even hundreds of layers, all learned automatically through exposure to training data. This hierarchical approach enables a rich and contextualized representation of information, facilitating advanced applications in complex tasks such as computer vision, natural language processing, and pattern recognition.

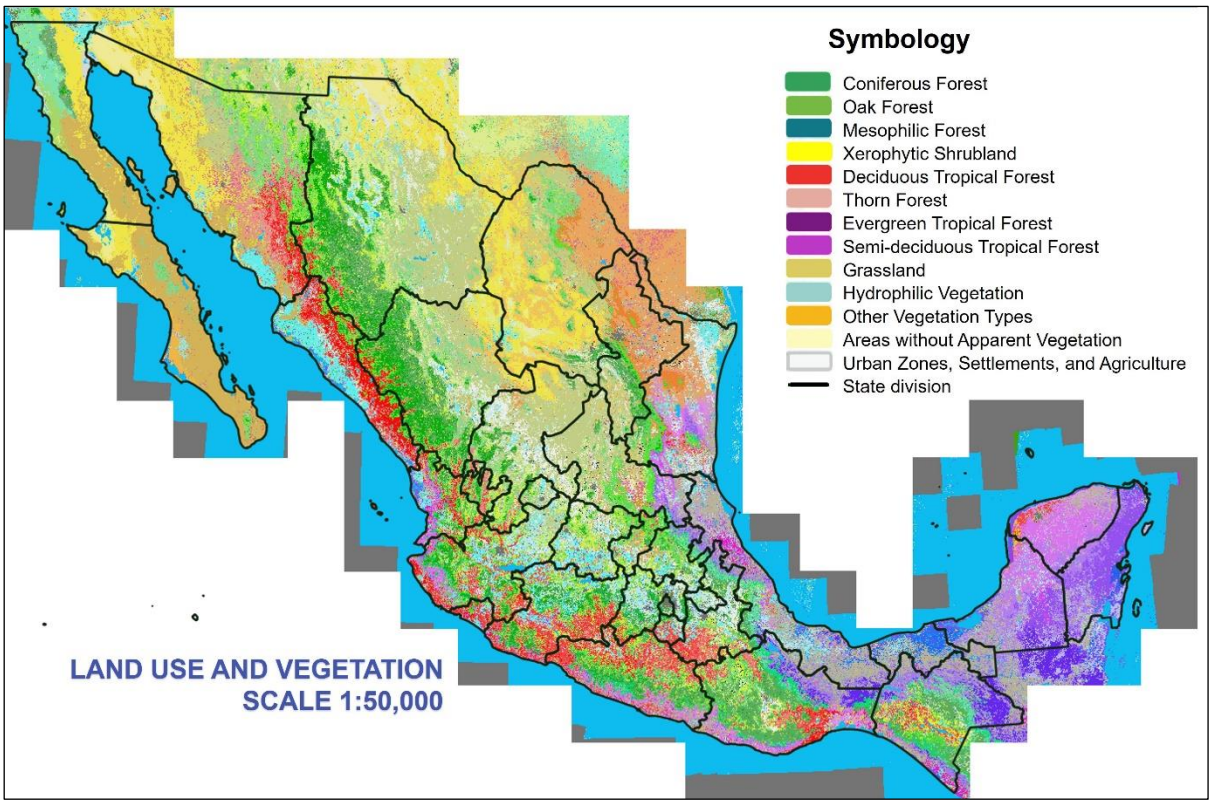

Figure 2. Land Use and Vegetation Mapping in Mexico Using Deep Learning Techniques.

2.4.2 Construction of the Prediction Model

In the context of DL, classifying each pixel in an image to assign it a specific category is known as semantic segmentation. This approach typically employs architectures based on *Convolutional Neural Networks (CNNs)*. For generating geospatial data on vegetation and land-use, a supervised classification approach was adopted. However, due to the limitation of having only point-based training data, the application of *CNNs* was not feasible.

Given that each training point contained relevant features related to vegetation, a fully connected neural network, or multilayer perceptron, was employed. This model consists of five hidden layers with 325 neurons each and an output layer for classifying into 123 categories, resulting in a total of 477,548 parameters.

The deep neural network was implemented using the *Python* programming language and features an input layer composed of a vector containing the 41 normalized variables defined earlier in this study. Throughout the successive layers, these variables are progressively transformed until reaching the output layer, which is configured with 123 elements corresponding to the classification categories. Each output element represents the probability that a pixel belongs to a specific category. The final label is assigned based on the category with the highest probability.

During model training, a loss function is applied that returns a value of "0" in the case of a correct classification; if the label is incorrect, the model adjusts its parameters using gradient descent and regularization until stability is achieved. To prevent overfitting, regularization techniques were implemented, such as the *Adam optimizer* and *dropout*, which randomly deactivates 50% of the neurons at the output of each layer during training. This improves the model's generalization ability when encountering new data. The trained model was then applied to the 144 tiles covering the country, classifying each pixel. The process execution was optimized using *Python* tools such as *GDAL*, *Keras-TensorFlow*, and *NumPy*, along with the use of *GPUs*. This allowed the entire national mosaic to be processed in approximately four hours.

This approach followed an agile development scheme, constantly integrating advanced techniques and methodological adjustments through trial and error and the discovery of new results, ensuring the model's adaptability to national geospatial data.

3. Results

Figure 2 presents a graphic representation of the results obtained from the application of the predictive model for vegetation and land use classification in Mexico.

3.1 Validation of Results

The validation of the classified images constitutes a crucial stage within the production analysis design of *INEGI's Statistical and Geographical Process Model (MPEG)*. This process not only involves evaluating the classification performed but also measuring the performance of the adjusted model in terms of accuracy and precision, using an independent validation set distinct from the training data.

To ensure robust and reliable results, the validation set must adhere to the same classification schema used for the training points, thereby enabling the generation of relevant metrics for assessing model performance. The primary objectives of this phase are:

- **Overall Accuracy:** Evaluating the percentage of the area that has been correctly classified at the national level.
- **Class Accuracy:**
  - *User Accuracy:* For the i-th class, defined as the proportion of the area classified as class i that truly belongs to class i.
  - *Producer Accuracy:* For the j-th class, defined as the proportion of the area that is currently class j and was correctly classified as class j.
- **Spatial Domain Accuracy:** For example, assessing classification accuracy at the state level.

3.1.1 Model Performance Evaluation

The performance of a machine learning model is measured by its ability to generalize correctly on test and validation data while maintaining low error levels. Among the metrics used is the misclassification rate, which varies according to the importance assigned to different types of errors, particularly in critical applications such as defining conservation policies in protected natural areas. The model errors can be categorized as true negatives, false negatives (also known as omission errors), true positives, and false positives (also known as commission errors).

To identify these errors, various metrics and techniques were employed to evaluate the constructed classification model: confusion matrix, overall accuracy, omission and commission errors, the *Kappa Statistic*, and *F1* evaluation. The confusion matrix is structured with rows representing the actual class and columns representing the class assigned by the classifier. From this matrix, it is possible to calculate metrics such as false positive rates, false negative rates, sensitivity (the ability to detect positive cases), and specificity (the ability to correctly identify negatives). The generation of the *Receiver Operating Characteristic (ROC)* curve provides a detailed visualization of the balance between the model’s sensitivity and specificity.

3.1.2 Field Evaluation and Evaluation of Geospatial Data

The geospatial data generated by the model undergo a rigorous accuracy evaluation protocol, which estimates the percentage of correctly classified area at various levels. To achieve this, a team of interpreters verifies the classifications of the resulting dataset following a probabilistic, stratified sampling design, as implemented by INEGI.

This process was complemented by field campaigns, during which the model's results were validated through comparison with direct observations, allowing for the identification of certain weaknesses in the model regarding label assignment in the training points.

3.2 Limitations

3.2.1 *Geomedian* Image Quality: In regions with frequent or persistent cloud cover, the quality of *Geomedian* images is compromised, which hinders the precise extraction of information.

3.2.2 Errors in Training Data: A rate of up to 50% incorrect labels has been detected in the training points, necessitating conceptual and strategic adjustments in data interpretation by specialists.

3.2.3 Spatial Sampling Imbalance: Although it is not feasible to achieve a completely balanced sample for all vegetation categories—due to the vast extent of some formations (spanning tens of thousands of square kilometres)—a local balance can be achieved to improve the representativeness of less extensive categories.

3.2.4 Heterogeneity in Accuracy and Separability: Some categories, such as mangroves and certain types of tropical forests, exhibit high classification reliability (>90%), while others demonstrate greater difficulty in being accurately identified.

3.2.5 Limitations in Class Separability: These difficulties are attributable both to spectral similarity between some categories and to issues in the initial label assignment, which affects the overall precision of the model.

In summary, the integration of deep learning (*DL*) improved the accuracy of digital vegetation mapping in Mexico, achieving an overall accuracy of 77%, representing a notable improvement over traditional cartographic classification method. The thematic accuracy, which measures the correct classification of vegetation types, was evaluated using ISO-2859 standards and reached 53.1%, indicating misclassification challenges in certain vegetation categories.

4. Conclusions

1. Integrating remote sensing data, fully connected neural network algorithms, and GPU processing enabled the development of a methodology with high expectations for efficiency and scalability in geospatial classification across Mexico. This represents a significant advancement that is expected to replace conventional methods in the medium term.

2. Compared to other supervised classification techniques used in previous exercises, DL techniques provide a significant improvement in the accuracy of vegetation and land use mapping in Mexico.

3. There is a need for further refinement in class differentiation and label consistency. The integration of stratified random sampling and key environmental variables enabled the capture of a diverse range of vegetative conditions, thereby enhancing the model’s predictive performance. However, additional refinements are necessary, particularly in the use of *Geomedian*, NDVI, MNDWI, and other spectral indices derived

from *HLS (Harmonized Landsat-Sentinel)* imagery, aiming to improve both input resolution and output accuracy. Furthermore, incorporating vegetation-related variables, such as canopy height and cover—obtainable from the LiDAR sensor aboard the International Space Station—has been identified as a critical improvement. These advancements would lead to enhance classification accuracy and a more representative depiction of vegetation dynamics.

4. An assessment employing confusion matrices, Kappa statistics, and field validation campaigns strengthened the reliability of the model’s outcomes while also revealing potential areas of refinement in its implementation.

5. Issues such as low quality of *Geomedian* in regions with persistent cloud cover and errors in training data labelling underscore the need to optimize data collection and preparation processes, including the validation of labels assigned by the teams of specialists.

6. Integrating variables enabling the monitoring of seasonal vegetation dynamics is recommended, along with the exploration of advanced neural network architectures to improve the identification of categories with low spectral separability.

References

Benavides, J. E., 2021. Estimación del recurso arena blanca para determinar la disponibilidad de la oferta en el crecimiento urbano de la ciudad de Iquitos, Loreto, Perú-2018. Tesis de licenciatura, Universidad Nacional de la Amazonía Peruana. https://hdl.handle.net/20.500.12737/7581.

Dalponte, M., ... Frizzera, L., Gianelle, D., 2023. Spectral separability of bark beetle infestation stages: A single-tree time-series analysis using Planet imagery. Ecological Indicators 153. https://doi.org/10.1016/j.ecolind.2023.110349

Díaz, V., Sosa-Ramírez, J., Pérez-Salicrup, D., 2012: Distribución y abundancia de las especies arbóreas y arbustivas en la Sierra Fría, Aguascalientes, México. Polibotánica 34: 99-126.

Gazzea, M., ... Arghandeh, R., 2022. Tree Species Classification Using High-Resolution Satellite Imagery and Weakly Supervised Learning. IEEE Trans. Geosci. Remote Sensing 60, 1–11. https://doi.org/10.1109/TGRS.2022.3210275

Gual-Díaz, M., Rendón-Correa, A., 2014. Bosques mesófilos de montaña de México: diversidad, ecología y manejo. Comisión Nacional para el Conocimiento y Uso de la Biodiversidad. México. 352. ISBN. 978-607-8328-07-9.

Instituto Nacional de Estadística y Geografía., 2014. Uso del Suelo y Vegetación escala 1:250 000. Serie VI. INEGI-Aguascalientes, México.

Killough, B., 2018. Overview of the Open Data Cube Initiative. *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, Valencia, Spain, 8629-8632. doi.org//10.1109/IGARSS.2018.8517694.

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

Llorente-Bousquets, J., y Ocegueda S., 2008. Estado del conocimiento de la biota, en Capital natural de México, vol. I: Conocimiento actual de la biodiversidad. Conabio, México, 283-322.

Piragnolo, M., ... Grigolato, S., 2021. Responding to Large-Scale Forest Damage in an Alpine Environment with Remote Sensing, Machine Learning, and Web-GIS. *Remote Sensing* 13, 1541. https://doi.org/10.3390/rs13081541

Priego-Santander, A., Campos, M., Bocco, G. y Ramírez-Sánchez, L., 2013. Relationship be-tween landscape heterogeneity and plant species richness on the Mexican Pacific coast. Applied Geography, 40, 171-178. doi.org/10.1016/j.apgeog.2013.02.013.

Roberts, D., Dunn, B., Mueller, N, 2018: Open Data Cube products using high-dimensional statistics of time series. *IGARSS 2018 IEEE International Geoscience and Remote Sensing Symposium*, 8647-8650. doi.org/10.1109/IGARSS.2018.8518312.

Roberts, D., Mueller N., Mcintyre, A, 2017: High-Dimensional Pixel Composites From Earth Observation Time Series. *IEEE Transactions on Geoscience and Remote Sensing* 55(11), 6254-6264. doi.org/10.1109/TGRS.2017.2723896.

Small, D., 2012. SAR backscatter multitemporal compositing via local resolution weighting. *Geoscience and Remote Sensing Symposium*, Munich, Germany, 2012, 4521-4524. doi.org/10.1109/IGARSS.2012.6350465.

Toth, C., Jóźków, G., 2016: Remote sensing platforms and sensors: A survey. *ISPRS Journal of Photogrammetry and Remote Sensing*, 115, 22–36. doi.org/10.1016/j.isprsjprs.2015.10.004