INEGI

SUSTAINABLE DEVELOPMENT GOALS

IAEG-SDGs
Inter-agency Expert Group on SDG Indicators

Integrating EO with Official Statistics using Machine Learning in Mexico
(Work in progress)

March 10, 2020

INEGI
abel.coronado@inegi.org.mx

UNECE Machine Learning Project

This is one of the pilot projects in the **Machine Learning Project** of the UNECE **High-Level Group on Modernization of Official Statistics**.

Participating NSI's

SUSTAINABLE DEVELOPMENT GOALS IAEG-SDGs

UNECE Machine Learning Project

Objectives

- Investigate and demonstrate the value added of ML in the production of official statistics, where "value added" is increase in relevance, better overall quality or reduction in costs.
- Advance the capability of national statistical organisations to use ML in the production of official statistics.
- Enhance collaboration between statistical organisations in the development and application of ML.

INEGI

SUSTAINABLE DEVELOPMENT GOALS IAEG-SDGs

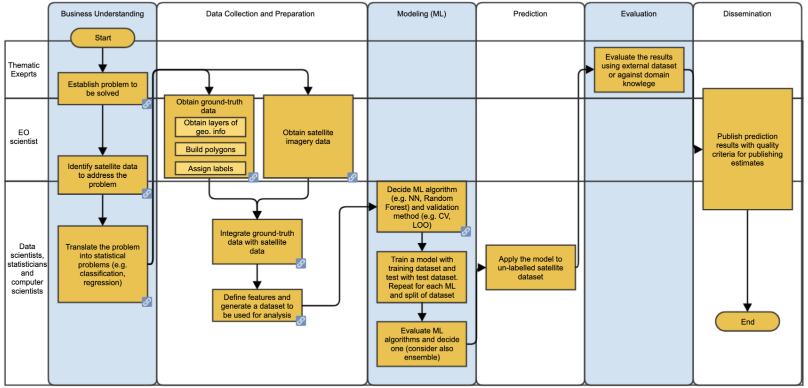
Draft of Imagery Pipeline

Imagery Pipeline

In order to have a road map.
We build an abstract machine learning pipeline for Imagery

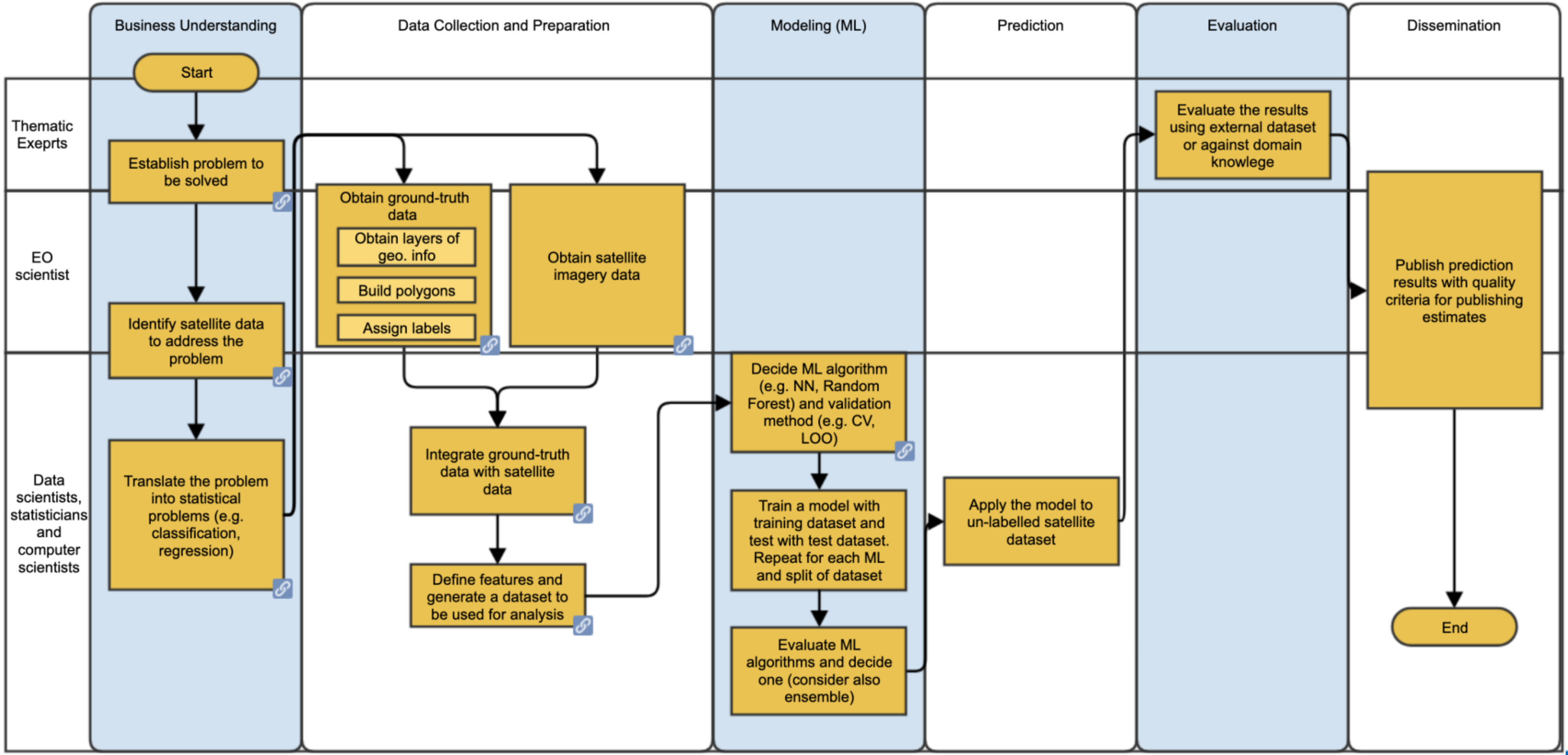
Based on:

IBM CRISP-DM
Cross Industry Standard Process for Data Mining
&
Microsoft Data Science lifecycle



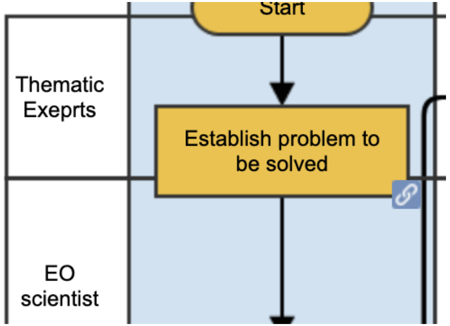
InKyung Choi; Refactoring

Imagery Pipeline



Pilot Project

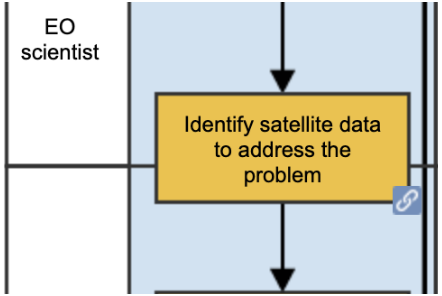
Establish the problem to be solved



Monitor the growth of urban locations of Mexico, which would generate a more timely input for :

- Cartography update
- Estimation of the population in non-census years
- Related with:
- SDG **Indicator 11.3.1**: Ratio of land consumption rate to population growth rate
- SDG **Indicator 15.3.1**: Proportion of land that is degraded over total land area

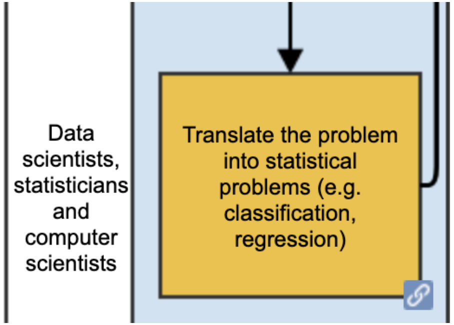
Identify satellite data to address the problem



In our case, the data to monitor the growth of cities can be Landsat images (NASA & USGS):

- They are open data.
- There is a constant record since the 70s, although they are available from 1985 to date.
- The spatial resolution of Landsat images is 30 meters.
- Temporal resolution is 16 days and 8 days with combined satellites.
- Spectral resolution, in this pilot project we use 6 spectral bands
- All the data we use is Analysis Ready Data (ARD)
- We take advantage that we have just built the Mexican Geospatial Data Cube, with all this information.

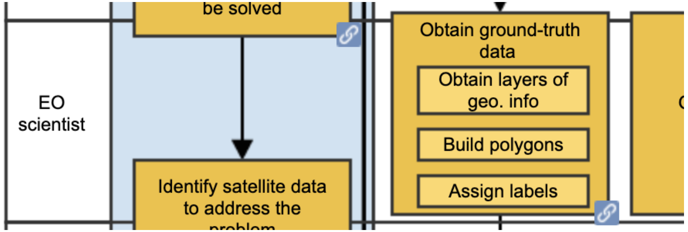
Translate the problem into statistical problem



We define that is a classification problem:

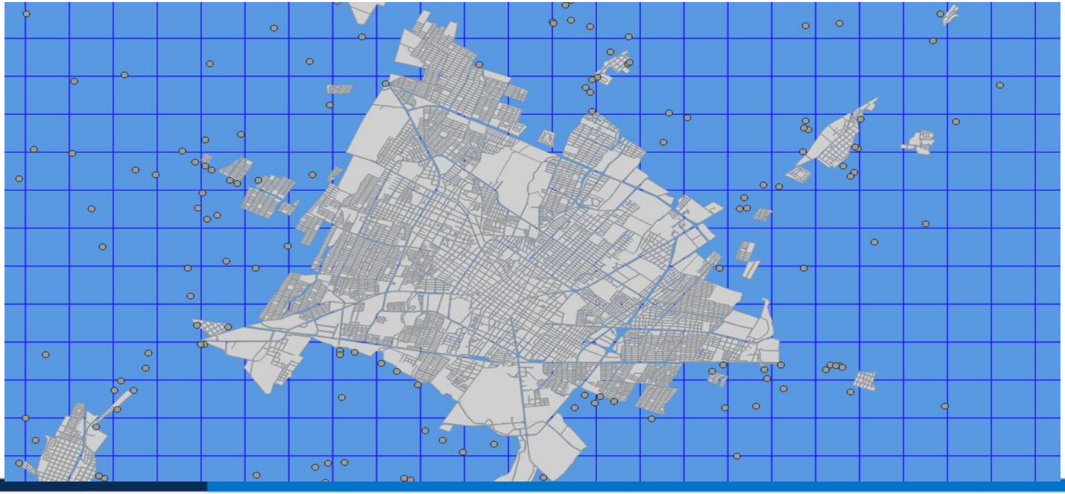
- Unit of analysis: 1km x 1km squares covering the whole country: 1’ 975, 719
- 2 classes were designated:
- Urban
- Non-Urban

Obtain ground-truth data

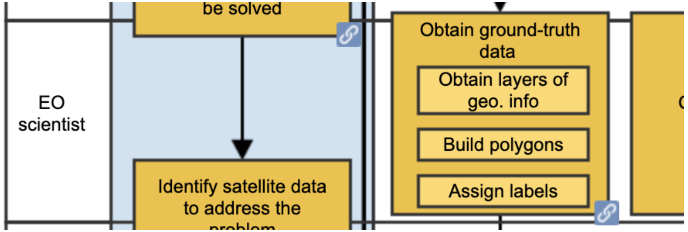


Obtain layers of geographical information:

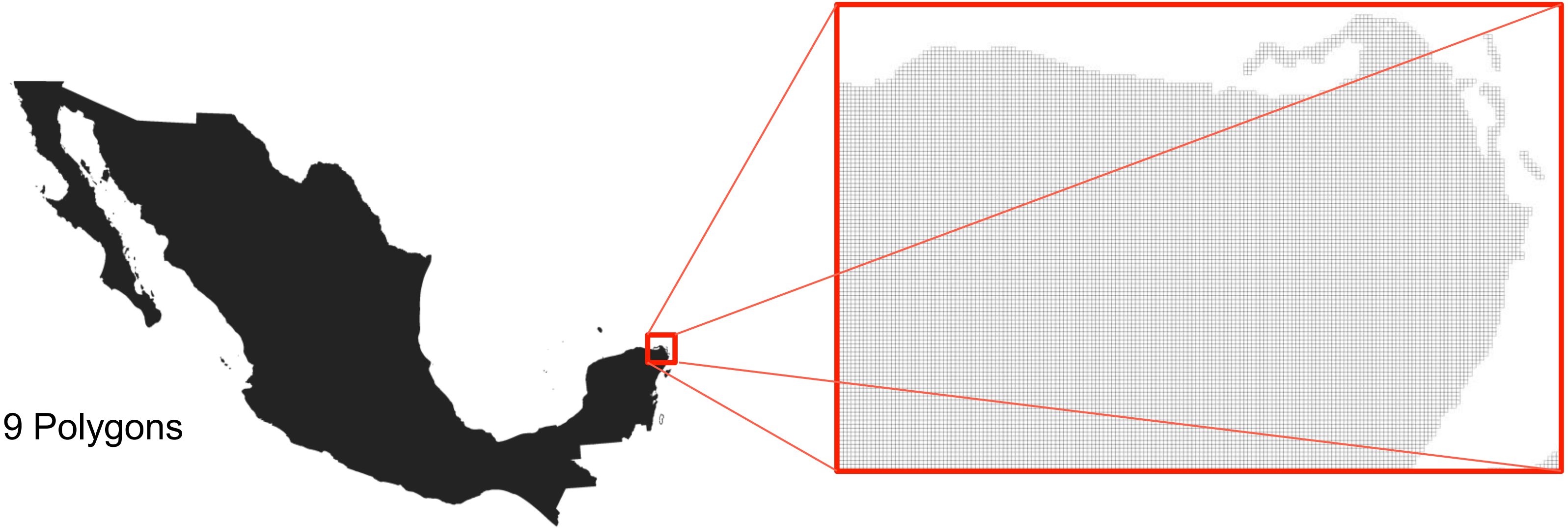
- Georeferenced Population Census 2010 (Block Level Aggregation)
- Georeferenced Economic Census (Economic Unit Level)



Obtain ground-truth data



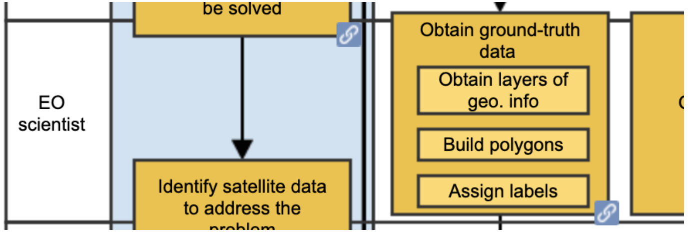
Build the polygons (1 km – 1 km)

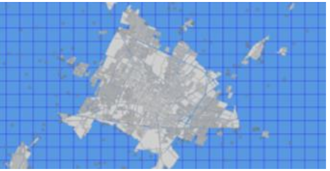


1’ 975, 719 Polygons

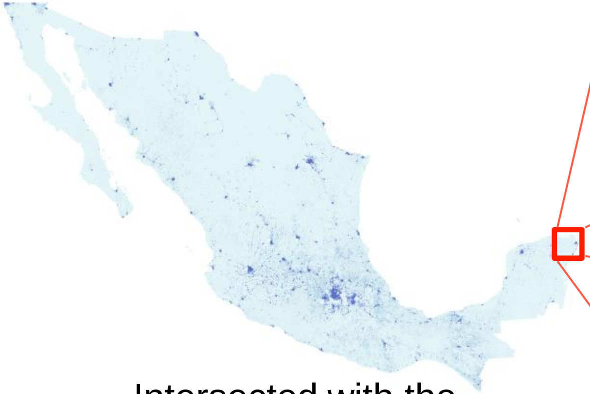
Obtain ground-truth data

Assign Labels

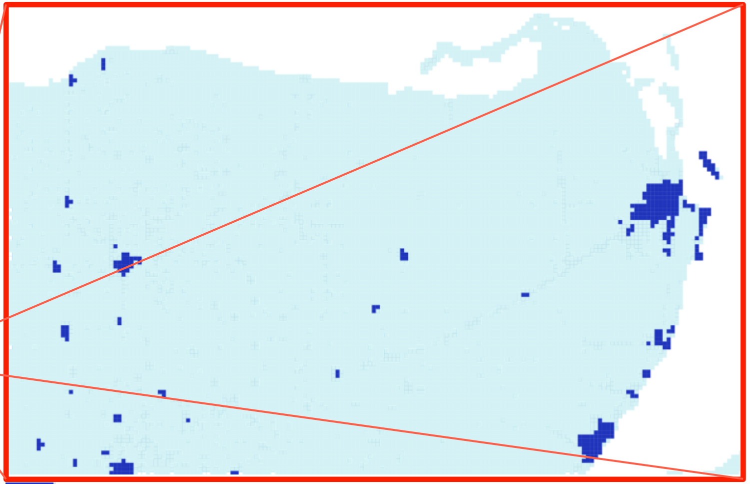




Georeferenced Census Results (Population and Economic)

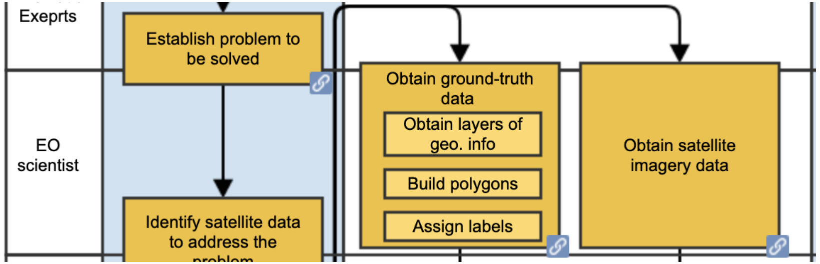


Intersected with the 1km national grid



	Urban	36,759
	Non-Urban	1,938,960

1km by 1km regions

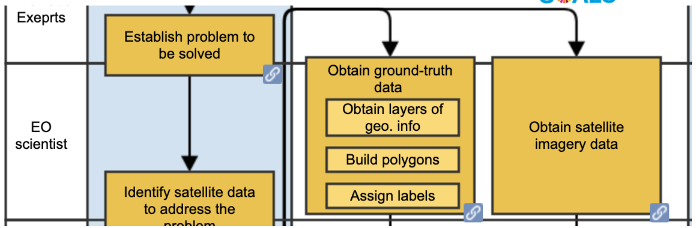


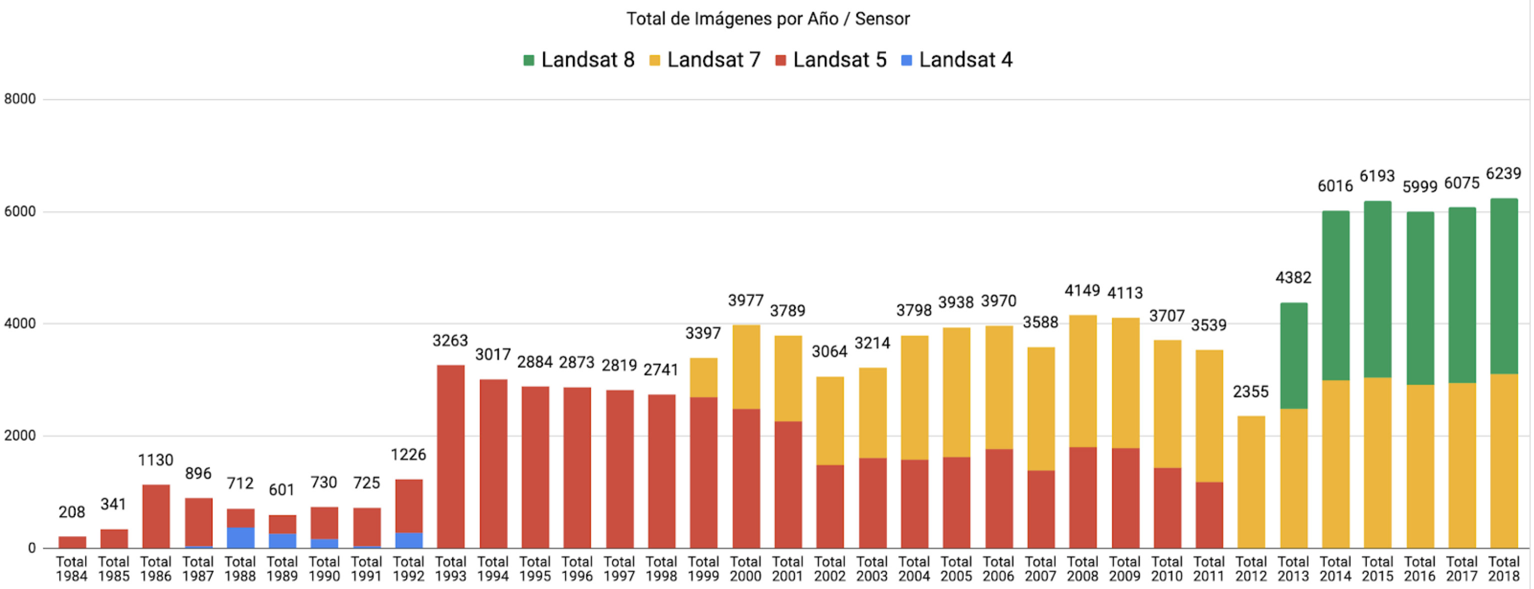
Obtain satellite imagery data

- 32 TB of Images in external discs.
- 90 TB decompressed
- March 4, 2019
- The images are ARD, Analysis Ready Data.
- In essence ARD means that the pixels of the entire time series are aligned and comparable.

Obtain satellite imagery data

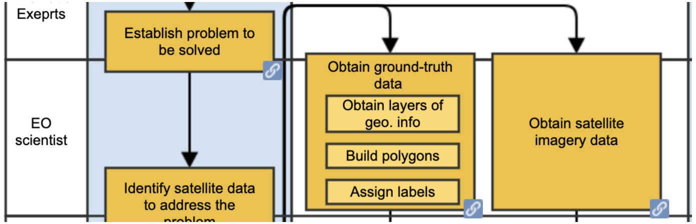
- 36 Years

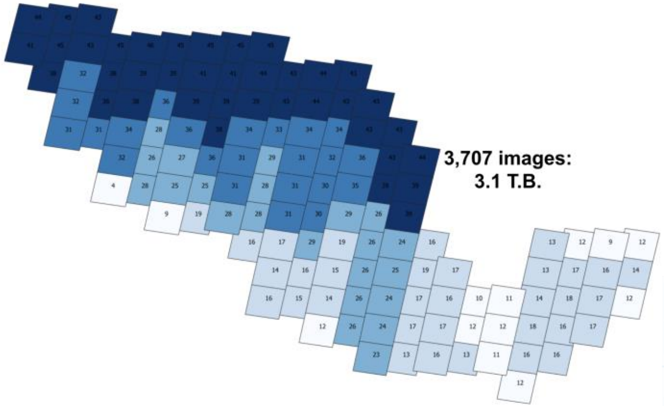




Obtain satellite imagery data

- 2010 same year of the Census

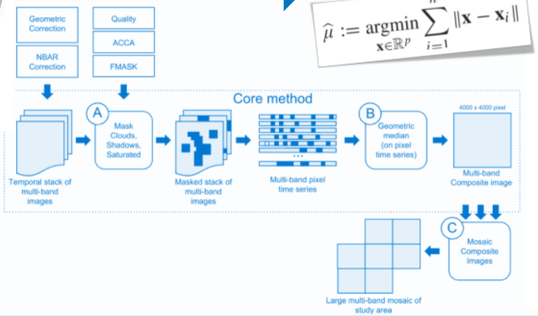




3,707 images: 3.1 T.B.

Australian Government
Geoscience Australia
OPEN DATA CUBE

$$\hat{\mu} := \underset{\mathbf{x}\in\mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^{n} \|\mathbf{x} - \mathbf{x}_i\|$$



2,737,273,075 pixels – 35 Gb

National Cloud-Free Mosaic

Use the ODC to generate annual summaries

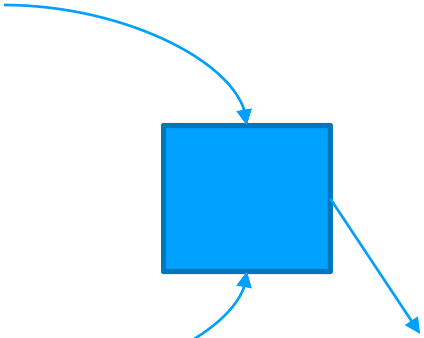
Comparador de Geomedianas

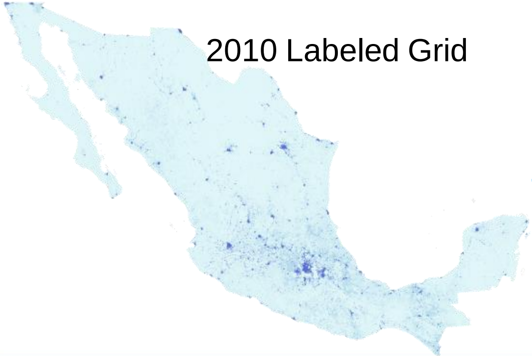
Geomedians

Integrate ground-truth data with satellite data

2010 National Cloud-Free Mosaic

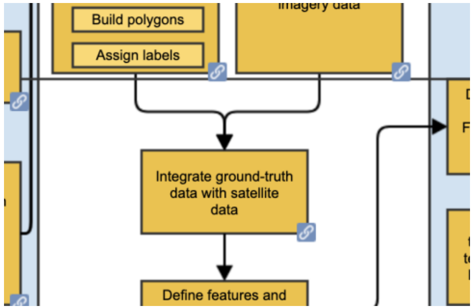
33 x 33 pixels
30 meters of resolution

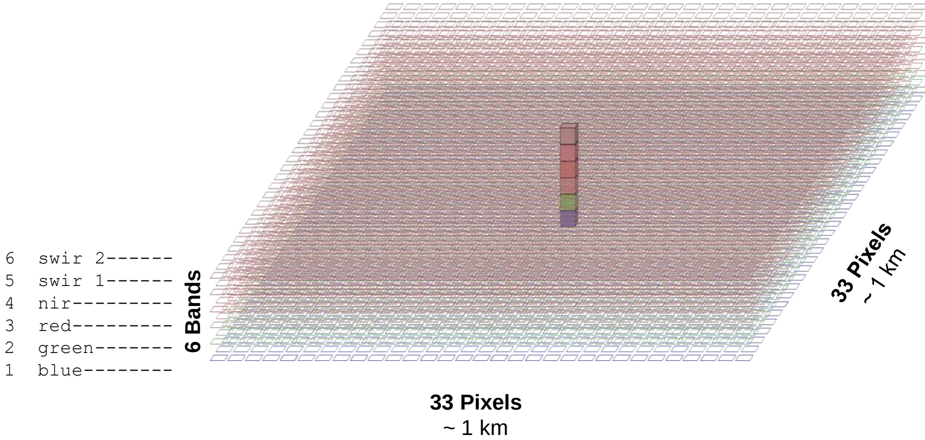




2010 Labeled Grid

1 km x 1 km



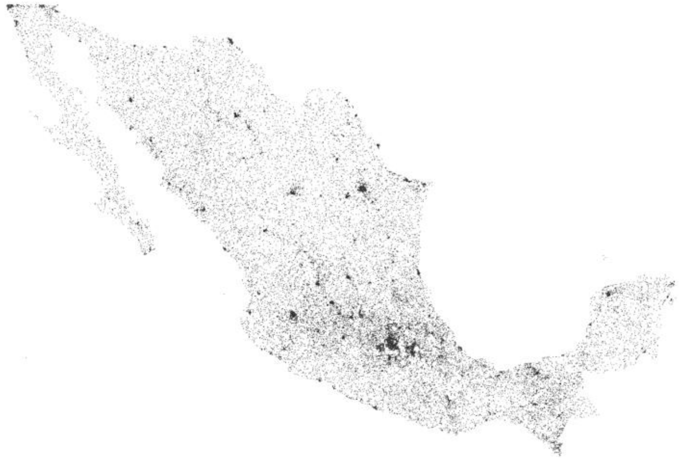


Labeled Multispectral Image Patch

Integrate ground-truth data with satellite data

Take a Random Sample

- From national grid with the urban and non-urban labels, a random sample of 40,000 elements was taken: 20,000 for each class. Then, image patches were extracted from the cloud-free mosaic. Resulting in 40,000 images labeled, each one with 33 pixels x 33 pixels, with 6 spectral bands (or layers).

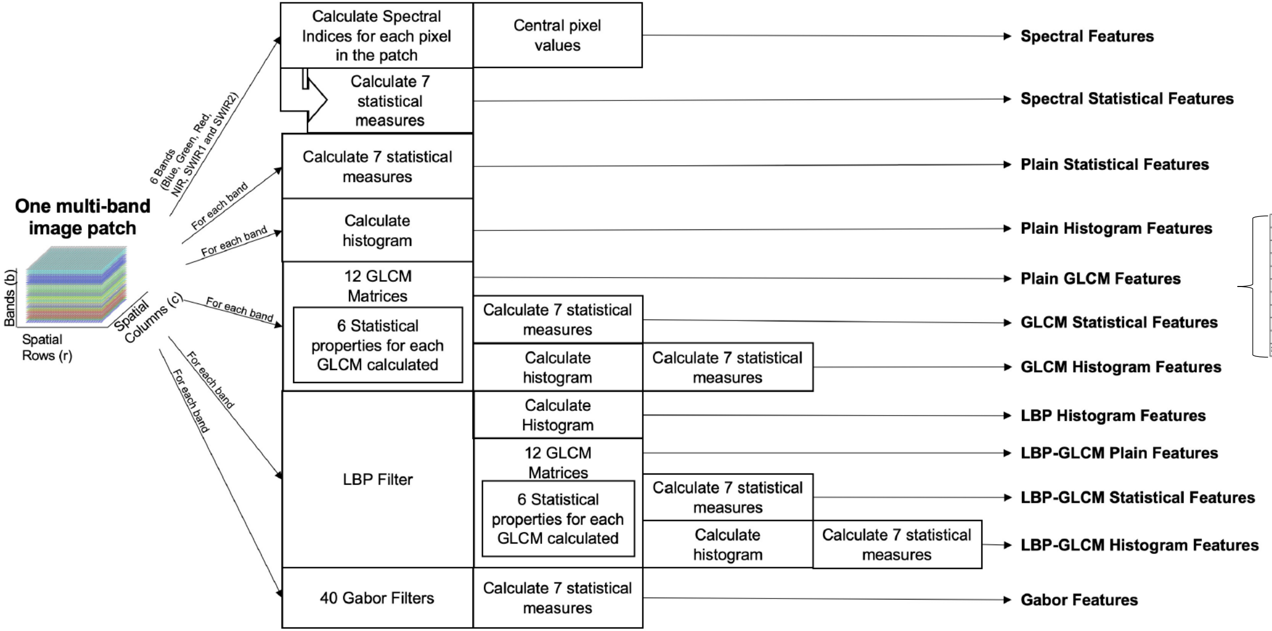


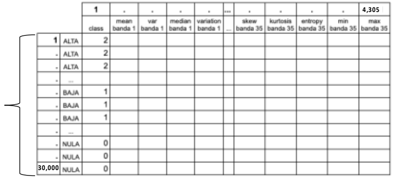
SUSTAINABLE DEVELOPMENT GOALS IAEG-SDGs

Define features and generate a dataset to be used for analysis

- Feature Engineering

40,000 tagged regions
1km x 1km
33 pixels x 33 pixels x 6 Bands

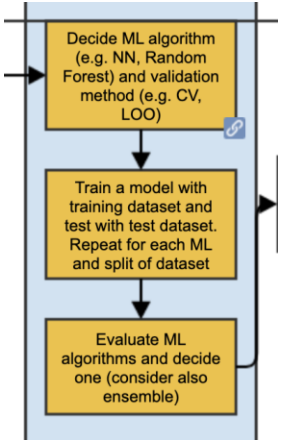


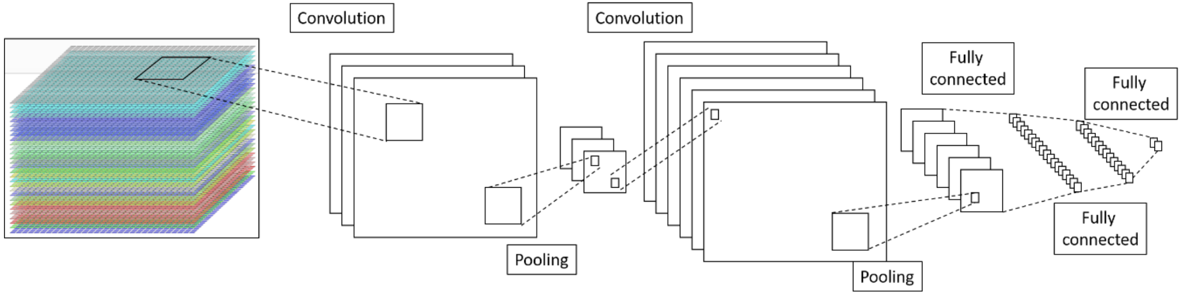


40,000 rows in a data matrix

Models tried

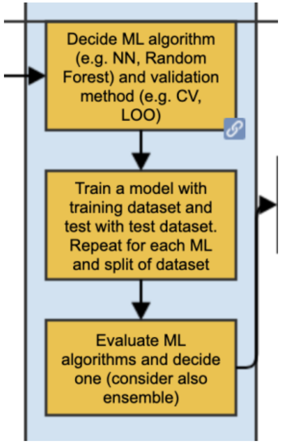
Two different models were tested, an Extra Trees model also known as Extremely Randomized Trees and a LeNet Convolutional Neural Network





Results

The evaluation with training data was performed 10-fold cross-validation, for both methods.



Extra Trees

	precision	recall	f1-score
Non-Urban	0.92312	0.93532	0.92916
Urban	0.93438	0.92218	0.92821
O.A.			**0.92870**
macro avg	0.92873	0.92875	0.92868
weighted avg	0.92882	0.9287	0.92870

LeNET

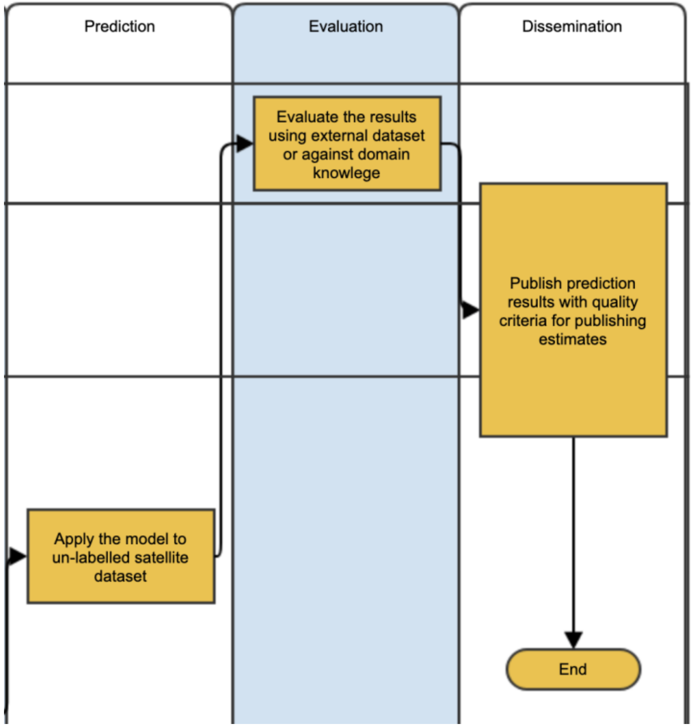
	precision	recall	f1-score
Non-Urban	0.91372	0.90296	0.90808
Urban	0.90465	0.91445	0.90932
O.A.			**0.90873**
macro avg	0.90919	0.90868	0.90869
weighted avg	0.90917	0.90873	0.90872

Detailed results

Next Steps

Next Steps

- Finish other iteration, no later than one 2 months.
- The grid is also an important innovation and is evolving, it is likely that we will have a new version very soon and we should update the classification.
- Apply the model to un-labelled years, for example 2019 first semester.
- Validate a sample of the un-labelled year, requesting support for the area of visual interpretation of the geography division, to have a measure of product quality.
- Hold more meetings with the area of sociodemographic statistics, involve them in the exercise to improve the potential benefit in the population estimate.
- Hold meetings with the cartographic update area, to receive feedback.



INEGI

SUSTAINABLE DEVELOPMENT GOALS

IAEG-SDGs

GRACIAS!

abel.coronado@inegi.org.mx

Conociendo
México

01 800 111 46 34
www.inegi.org.mx
atencion.usuarios@inegi.org.mx

INEGI Informa