ML APPLICATION TO USE SATELLITE DATA IN COMBINATION WITH CENSUS DATA TO PRODUCE NEW INFORMATION

Organisation: INEGI

Author(s): Abel Coronado; Jimena Juárez; Ricardo Bucio

Date: 03 / 03 / 2020

Version: 1

1. Background and why and how this study was initiated

This pilot project proposes the use of Landsat satellite data for the mapping of urban areas in non-census years, using as ground truth an urban density grid of 1km x 1km generated from the field data of the 2010 Population Census at block level and the updated Georeferenced Business Register to date 2010. This Machine Learning application will generate national classifications that identify the expansion of cities year by year. With which it will be possible to generate information products that contribute to the cartographic update, since the classifications of recent years will identify the growth of cities nationwide. Alerting urban map specialists that there is a new area to map. It will also be possible to incorporate urban growth data into the population estimation models. Finally, it will be possible to generate new types of statistics that allow observing the evolution of the extension of the cities of Mexico throughout the period of time for which the images are available, from the years 1984 to 2020. This type of project It can be carried out with the participation of the INEGI Data Science Laboratory and the areas of IT and Geography.

2. Data

2.1 Input Data (short description)

National cloud-free mosaic with resolution at 30 meters, generated from the analysis of the Landsat 5 and 7 image time series. With color depth of 16 Bits and 6 Multispectral bands (Blue, Green, Red, NIR, SWIR 1 and SWIR 2).

1 km x 1 km grid with 1,975,719 cells covering the national territory (Mexico).

1,741,553 were labeled as Urban (36,759) and Non-Urban (1,704,794) according to census data on population and housing 2010, economic data and road infrastructure. The remaining 234,166 correspond to an undefined class, where no data was available.

2.2 Data Preparation

INEGI built a Geospatial Data Cube (https://www.opendatacube.org/) with all the available Landsat images for Mexico with the highest quality level, also referred to Analysis Ready Data

(ARD). Using the 3707 Landsat 5 and 7 images for 2010, a cloud-free national mosaic was calculated using the Geomedian algorithm (more details: https://ieeexplore.ieee.org/document/8518312).

From the 1 km x 1 km grid labeled with the urban and non-urban classes, a random sample of 40,000 elements was taken: 20,000 for each class. Then, image patches were extracted from the cloud-free mosaic, corresponding to each region of approximately 1 square km selected in the random sample. Resulting in 40,000 images labeled, each one was 33 pixels x 33 pixels (only the pixels that fall within the 1km cell were used), with 6 spectral bands (or layers).

Next, we expand each of these 40,000 images by generating for each pixel 21 additional indexes, each one based on different pair combinations from the the 6 layers of the initial cloud-free mosaic. These 21 additional indexes are divided into 15 spectral indexes: evi, bu, sr, arvi, ui, ndbi, ibi, ndvi, ndwi, mndwi, nbi, brba, nbai, baei, bi; and 6 more that come from tasseled cap transformations. The 21 added to the initial 6 variables, resulting in 27 “layer” values per pixel. Finally, matrices of 33 pixels x 33 pixels x 27 channels were obtained: figure 1.

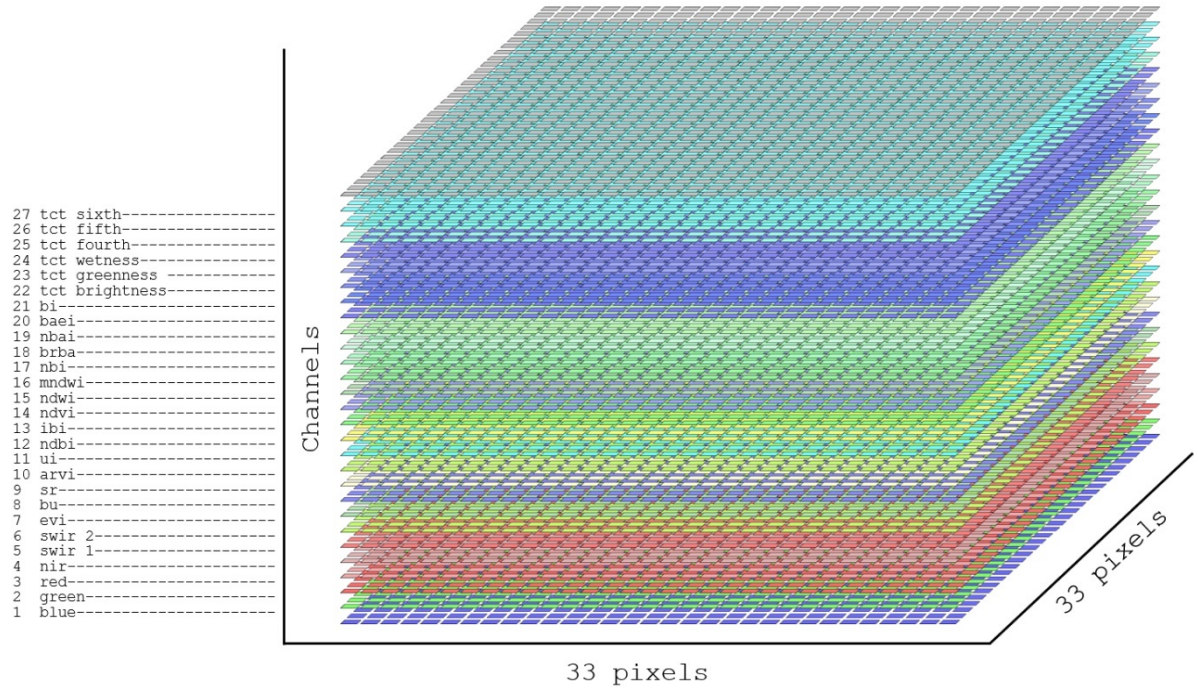

Image 1.- Visual representation of one of the images used in this study.

For the experiment with a pipeline based on extra-trees, each patch was characterized by generating 7 statistics of this set of pixels in each channel: mean, variance, median, variation,

bias, kurtosis and entropy; In addition, for each channel, the frequency of the data in 10 bins was calculated by calculating the histogram and the values of the central pixel, the descriptive data obtained were organized in a matrix of 40,000 rows and 486 columns.

For the CNN experiment, the 33 x 33 x 27 patches were normalized.

2.3 Feature Selection

For the experiment with a pipeline based on extra-trees, with the matrix of pre-calculated characteristics, with dimensions 40,000 rows by 486 columns; A combination of feature selectors was used: first the family-wise error rate using the probability of making one or more false discoveries, in python is implemented in the method: *sklearn.feature_selection.SelectFwe* and subsequently we use the method *sklearn.feature_selection.VarianceThreshold* that removes all low-variance features.

In the case of the neural network, we let the 27 layers of information be used, no manual feature selection is needed.

2.4 Output data

National grid of regions of 1 square km classified according to urban or non-urban classes.

3 Machine Learning Solution

3.1 Models tried

Extra Trees is a classifier very similar to Random Forest, as it consists of many decision trees. The prediction of each tree is considered. The final prediction is reached by majority vote. Two important differences are that the Random Forest uses bootstrap replicas, that is, it subsamples the input data with replacement, while the extra trees use the entire original sample. And selecting breakpoints to divide the nodes. Random Forest chooses the optimal division, while Extra Trees chooses it at random, making it more efficient by omitting optimization at the cut point. Finally, Extra-Trees are more computationally efficient and deal better with noisy features.

LeNet is a simple, convolutional neural network based on multiple layers of pre-processing and feature extraction. It uses convolutional layers and subsampling of images connected to a fully connected neural network to generate its predictions.

Those two different models were tested, an Extra Trees model also known as Extremely Randomized Trees and a LeNet Convolutional Neural Network, figure 2.

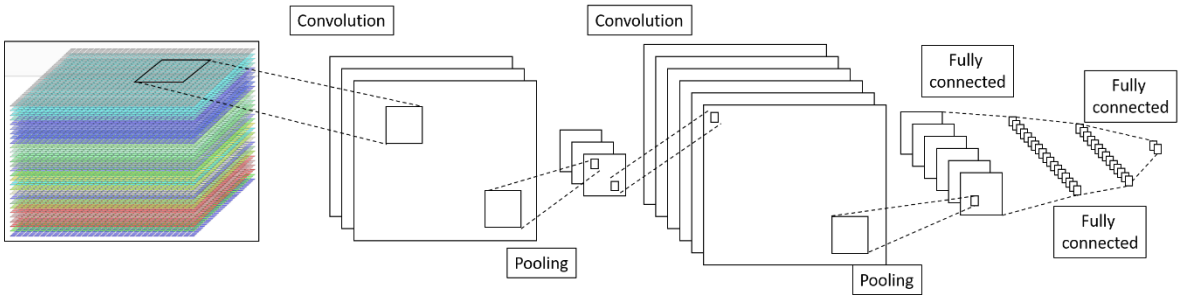

Figure 2.- LeNet (Deep Neural Network) used in the experiments.

3.2 Model(s) finally selected and the criterion

(i.e.: which model was why seen being the best?)

Extra Trees, due to its better performance than the Deep Neural Network in validation tests and acceptable speed in training and classification.

3.3 Hardware used

- Dell Precision 7820 Tower
- Processor Inter(R) Xeon(R) Gold 6230, 2.1GHz
- 256 GB RAM DDR4 2934MHz
- 512 Gb Solid state hard drive de
- 4 TB conventional hard drive
- Windows 10 Pro for Workstations

3.4 Runtime to train the model

(e.g.: 2 hours for 500,000 training samples and 25 features)

For Extra-Trees, 1:07 hours for a sample of 40,000 items with 486 features

For the LeNET Neural Network, 3:21 hours for 40,000 images with 27 channels of information.

4 Results

(e. g. in terms of RMSE, MAE, distributional accuracy [*], F1 (micro or macro), recall, accuracy, (threshold,) ..., perhaps as a table for different situations (if available))

[*]: If used: How did you measure distributional accuracy? By proportions, moments, quantiles, correlations, ...?

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	0.92875	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	0.90870	0.90869
weighted avg	0.90917	0.90873	0.90872

Subsequently, using the training data, that is, the information from the 40,000 images. A classification of the whole country was made in 2010. It was taken advantage of the fact that there is a whole country labeled for that census year:

Extra Trees Confusion matrix:

	Predicted					
Actual		NON-URBAN	URBAN	Total	Recall	Macro Recall
	NON-URBAN	1,599,258	105,536	1,704,794	93.81%	95.17%
	URBAN	1,277	35,482	36,759	96.53%	
	Total	1,600,535	141,018	1,741,553		
	Precision	99.92%	25.16%		**93.87%**	**O.A.**
	Macro Precision	62.54%				
	F1	96.77%	39.92%		**68.34%**	**Macro F1**

LeNET Confusion matrix:

	Predicted					
Actual		NON-URBAN	URBAN	Total	Recall	Macro Recall
	NON-URBAN	1,472,381	232,413	1,704,794	86.37%	91.51%
	URBAN	1,227	35,532	36,759	96.66%	
	Total	1,473,608	267,945	1,741,553		
	Precision	99.92%	13.26%		**86.58%**	**O.A.**

	Macro Precision	56.59%				
	F1	92.65%	23.32%		**57.99%**	**Macro F1**

5 Code/programming language

(e.g. the Python code is stored in GitHub)

The Python programming language, version 3.6 was used in an anaconda 4.7 virtual environment.

6 Evolution of this study inside the organisation

(e. g.: Collaboration within the organisation? Has this study advanced ML within the organisation?)

The study has been presented as a proposal in different areas of the institute as an auxiliary method for planning and generating indicators, and has been well received, generating interest in its potential future. We will begin to work together with the Data Science Research Area, created at the end of last year,to improve the classification algorithms. As well as having conversations with the area responsible for the updates of the urban map. To generate valuable products for the task they perform. Derived from the Machine Learning results of the experiments shown.

7 Is it a proof of concept or is it already used in production?

(If it is a proof of concept: Was it successful? How will its results prospectively be used in the future?)

Based on the models obtained, the classification at the national level must be evaluated by experts in the area of visual interpretation of images to obtain feedback.

7.1 What is now doable which was not doable before?

The main added value will be the generation of national classifications with anual update times for the cartographic update. Tests could even be done to verify if a national classification is possible every 3 months. With the purpose of contributing to the quarterly estimates of the population totals that INEGI intends to make in the area of Sociodemographic Statistics.

7.2 Is there already a roadmap/service journey available how to implement this?

Because it is still in the testing and optimization phase, a complete production strategy is not yet available.

7.3 Who are the stakeholders?

Who are the people or units in your organization that will or could be affected by this change? How have they been consulted or involved in the change?

The General Directorate Sociodemographic Statistics is very interested in incorporating quarterly predictions that detect the change in growth in cities to incorporate their values in the population estimation models. After the Population Census, the predictive capacity of the algorithm can be evaluated with better field data, so that a process of incorporating the data in production can be initiated. Additionally, it is possible to contact the INEGI cartographic update areas so that you can take advantage of the quarterly estimates.

7.4 Robustness

What fail checks are in place or planned to ensure that the ML solution is consistently meeting or exceeding the set gold standard (e.g. in terms of quality, speed, costs, maintaining the model (continuous learning), etc.)?

Design a random scheme of manual validation by specialists in visual interpretation of satellite images. The manual validation of the results will allow us to measure the quality of the ML product. This will give us a better understanding of how the algorithm works compared to manual analysis. Looking for a collaborative work between the Data Science team and the current Visual Interpretation teams.

7.5 Fall Back

Is a fall back plan in place or planned to mitigate the risk of the ML solution failing in production? Will there be resource left in place to go back to e.g. manual imputation or the use of rule-based scripts? (as many sentences as necessary, as few as possible)

In the event that the Machine Learning solution fails, we can rely on manual validation and field work. But with the consequence of having much longer delivery times.

8 Conclusions and lessons learned

(e.g.: ML can be used for editing but one has to have the following points in mind ...)

What has your ML study and the ML project brought to you, your work unit and your organisation so far? Has it advanced the knowledge and use of ML?

This Machine Learning pilot is possible thanks to the existence of field work and geographic data that can be incorporated into the classification processes. Additionally, INEGI has a Geospatial Data Cube that provides the satellite information used in Machine Learning processes.

We identify that the algorithm has a lot of confusion at the limits of cities, so we must work to improve that situation. However, the results are promising and in the coming months we will start tests with the areas that have expressed interest.

Finally, we consider that we must develop internal processes to have a continuous manual validation of the results. To monitor the quality and adjust the training sets in case we detect any bias.

9 Potential organisation risk if ML solution not implemented

(as many sentences as necessary, as few as possible)

That is an important way of looking at the challenge. We must not look at ML being the only risk, but also the risk of not using ML. That brings us back to the first sentences of the blue-sky report on the threats that NSOs are facing. If, for an organisation, ML is still just a “nice to have” or still mostly a buzzword, then the challenge to advance it to the production process is multiplied ten-fold.

Massive sources of information such as satellite images require too much manual labor for years, to generate value from the analysis of the almost 2 million square kilometers that Mexico covers. Machine Learning can be a key differentiator especially in the recognition of easily separable categories, in the most complex cases human intervention is required to generate knowledge that eventually can properly instruct the algorithm. Performing a continuous and incremental update of training sets. Machine Learning does not replace field work, nor manual validation, but it can complement and cover those aspects that have reached enough maturity to be automated.

10 Has there been collaboration with other NSIs, universities, etc?

(yes/no, if yes: which ones?)

Yes, USGS / NASA provided us with Landsat satellite images from Mexico. Geoscience Australia accompanied us in the process of building the Data Cube. The rest of Machine Learning's work is INEGI's internal work.

11 Next Steps

(as many sentences as necessary, as few as possible)

We are not yet ready to go to production; however, it is time to start the collaboration with the production areas. Since the results are promising, it is time to expand the working group within INEGI. To evaluate with them, the contribution to the current processes and define a possible route of incorporation to the productive processes.

Identify alternative sources of information to improve validation processes, in addition to incorporating manual validation of samples by experts in visual interpretation and fieldwork.