UNECE – HLG-MOS Machine Learning Project
Imagery Theme Report

Organisation: INEGI - Mexico
Author(s): Abel Coronado, Jimena Juárez

Date: 01/10/2020
Version: 1.0

Content

Preliminaries

Currently, large volumes of satellite information are available, such as the one announced in March 2015[1] by NASA, giving public access to the complete collection of LANDSAT 8 satellite images with 30-meter resolution, in the AMAZON cloud[2]. This facilitates access to large volumes of satellite information that cover the entire Earth, the frequency these satellite travels generate images of the whole globe is in periods of 16 days, which means approximately 8 terabytes of information is generated per year. NASA also offers access to images from MODIS satellites with a resolution of 500 meters which generate a complete image of the entire earth on a daily basis. It is also possible to access Sentinel-2 images[3] that have a resolution of 10 meters and cover the Earth every 5 days.

[1] http://landsat.gsfc.nasa.gov/?p=10221

[2] http://aws.amazon.com/es/public-data-sets/landsat/

[3] https://eur-lex.europa.eu/eli/reg_del/2013/1159/oj

The availability of satellite information is growing more and more (Toth & Jóźków, 2016) . Today there are private companies with constellations of nanosatellites that are capable of generating an image at a resolution of 3-5 meters of the entire earth daily (Curzi, Modenini, & Tortora, 2020), (Safyan, 2020). The wide availability of free and commercial satellite images opens opportunities to take advantage of these sources of information through Machine Learning methods. While on the other hand the demand for information on monitoring natural resources and statistical variables that can be observed in images such as the growth of urban areas is growing. This demand for information is evident in international projects such as the one expressed in the United Nations document: "Transforming our world: the 2030 Agenda for Sustainable Development" where an action plan is established with broad scopes in favor of people, the planet and prosperity, in the three dimensions of sustainable development: economic, social and environmental. This wide reach is achieved through 17 sustainable development goals (SDGs) and their corresponding targets. In March 2016, the indicators that will allow to continuously monitor the fulfillment of these established goals were first defined during the meeting of the Inter-agency and Expert Group on SDG Indicators (IAEG-SDGs) of the United Nations Statistical Division by the member countries. Some of the indicators have significant potential to be estimated by processing of large volumes of satellite images through computer vision and Machine Learning techniques (Holloway & Mengersen, 2018). Therefore, in this report, the results of four pilot projects are presented, which correspond to pilot projects carried out by Australia, Netherlands, Switzerland and Mexico.

Generic Pipeline for Production of Official Statistics Using Satellite Data and Machine Learning

After noting the lack of a generalized approach to describe how satellite data can be used by NSOs, as well as, acknowledging that the issue is even more complicated because use of satellite data often requires ML techniques which themselves are being experimented and not yet integrated in the production process in many NSOs, the development of the generic process pipeline is one of the first the deliverables in the Imagery Theme team. A generic process model describes high-level activities that need to be followed to achieve a certain objective or to deliver a specific output. This pipeline focuses on the specific use of satellite data to produce official information.

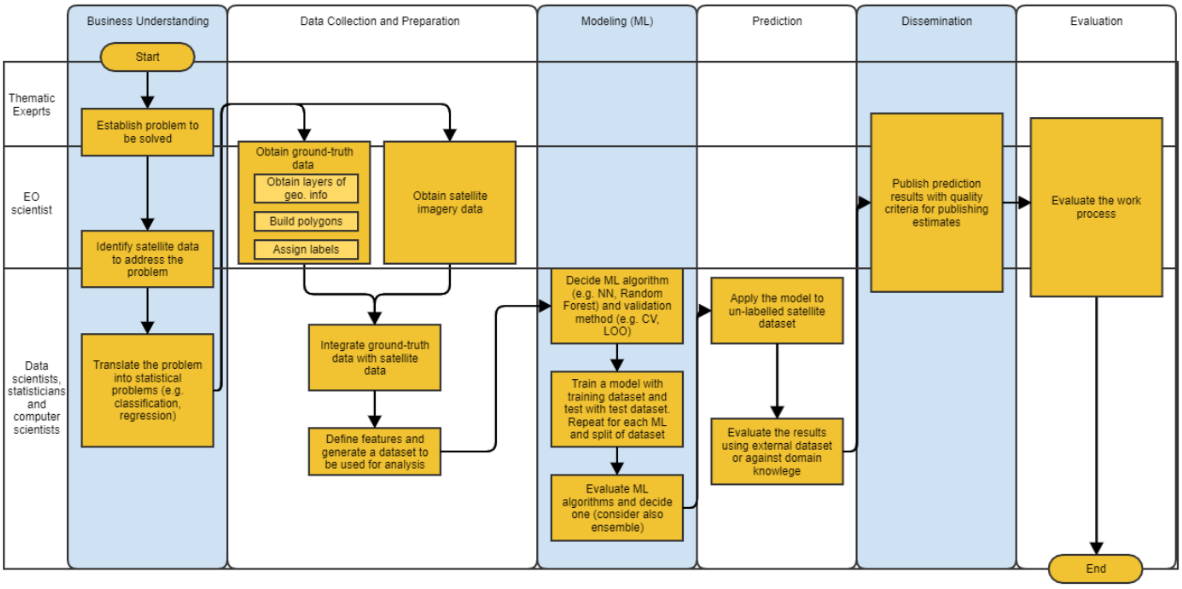
This pipeline aims to address following issues:

- There is lack of understanding about business process needed to use satellite data for statistical production.
- Processing and analysing satellite data require techniques that are not in traditional skill set of statistical organizations.
- Lack of common reference points to consolidate and link them, even though there is increasing body of works related to use of satellite data for production of official statistics.

The pipeline developed as in diagram and outlines the six main stages (business understanding, data collection and preparation, ML modelling, prediction, dissemination, evaluation) and the main specialized roles (thematic expert, E0 scientist, data scientists, statisticians and computer scientists) involved in each of the steps.

The diagram of the pipeline is provided below. More detailed description for this activity can be found in the specific report as well as additional examples related to the pilot projects of the Imagery Theme team.

Diagram of the pipeline


Business Understanding
Data Collection and Preparation
Modeling (ML)
Prediction
Dissemination
Evaluation
Thematic Exeprts
EO scientist
Data scientists, statisticians and computer scientists
Start
Establish problem to be solved
Identify satellite data to address the problem
Translate the problem into statistical problems (e.g. classification, regression)
Obtain ground-truth data
Obtain layers of geo. info
Build polygons
Assign labels
Obtain satellite imagery data
Integrate ground-truth data with satellite data
Define features and generate a dataset to be used for analysis
Decide ML algorithm (e.g. NN, Random Forest) and validation method (e.g. CV, LOO)
Train a model with training dataset and test with test dataset. Repeat for each ML and split of dataset
Evaluate ML algorithms and decide one (consider also ensemble)
Apply the model to un-labelled satellite dataset
Evaluate the results using external dataset or against domain knowlege
Publish prediction results with quality criteria for publishing estimates
Evaluate the work process
End

Motivation

In order to explore the potential of alternative data sources to those already known in the Official Statistics (Censuses, Surveys and Administrative Records) or to enrich existing projects, several projects were carried out aiming to take advantage of satellite images with Machine Learning (ML) techniques.

This document is intended to summarize the pilot projects carried out by Australia, the Netherlands, Switzerland and Mexico.

Machine Learning involves the automatic discovery of patterns in the data using computational algorithms and, from those regularities, proceeding to carry out tasks such as the detection of various categories (Bishop, 2006) in a training set. This is called *Supervised Learning*. The pilot projects reported in this document belong to this category and show the application of various classification algorithms that seek to relate the implicit or explicit patterns found in data carefully labeled by experts, with equivalent patterns in unlabeled data, intending for the algorithms to identify generalization rules that allow assigning categories to objects that have not been manually analyzed. Once the algorithms assign the "predicted" category, it is important to perform the evaluation of the ability of the algorithm to generalize with previously labeled testing sets, but never used in training procedures, and reporting the corresponding performance metrics for each project.

Each country wrote a detailed report of their work and corresponding experiments, we invite the reader to review the specific details of each country, in this document we will present the essential aspects.

Problem to solve

Each Statistical Office established the characteristics of the pilot test to be carried out, in which satellite images were used in the context of Machine Learning applications in order to solve specific office problems. As stated by the NSOs themselves, they try to solve problems related to the reduction of human intervention in the process of updating the Address Register (AR) or the measurement of statistical variables such as poverty or expansion of urban areas, as well as the detection of change in land use and land cover (LULC). Regularly finding a link to satellite images implies having some type of geographically referenced statistical information, as well as field work for validation, which is the basis for training automatic classification algorithms. The participating countries have a georeferenced source for such training.

The countries established the main motivation of their pilot test, identifying a relevant motivation that allows them to explore the validity of the approach, through the execution of the pilot project and a subsequent evaluation with respect to the original motivation once the project is completed. Some countries are still in the preliminary stages so definitive results are not yet available in some cases.

The expectations of the participants involve the need to create a new process that complements the activities of the NSOs or simply to improve existing processes. Either way, progress will be based on the application of Machine Learning techniques to satellite images.

Country	Problem to Solve	Contribution	Value Assessment
Australia	Use a model of ML to reduce the amount of manual intervention required during regular Address Register (AR) maintenance processes.	Reduce costs (time) by improving the current process that is a resource intensive process.	The number of automatically classified addresses.
Netherlands	Explore the potential of ML for detecting poverty and population distribution from aerial or satellite imagery.	Learn how to use machine learning to exploit imagery as a new data source in the production of official statistics and assist other countries who do not have income data in measuring poverty from imagery.	A working computer prototype.
Switzerland	Facilitating land use and cover classification and by improving change detection	Improvement existing process to reduce costs (time). At present, internal resources are almost entirely allocated to visual interpretation, at the expense of other activities.	A working computer prototype that allows to demonstrate the innovative potential of the FSO in the use of artificial intelligence to process images.
Mexico	Detect the extension of urban areas nationwide using ML	Reduce time and money. Generate information products that contribute to the cartographic update. 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	Clear objectives with links to potential impacts on existing and future data products.

Organizational Context

This section identifies the institutional priorities that led to the pilot exercise and identifies the stakeholders that support the execution of the exploratory project.

Country	Relevance to the Institution	Stakeholders involved and relevance for them	To go beyond the demonstration phase
Australia	Freeing manual classification experts from the simple work that can be performed by automatic algorithms, allowing them to focus their efforts in more complex Address that cannot be classified automatically. Results from Automated Image Recognition (AIR) can be used in conjunction with other administrative data sources to strengthen confidence and quality in the AR.	ABS officers. Since the Address Register forms the population frame for survey sampling it is important that it is of the highest quality and truly reflects the Australian population and its housing stock.	Initially this project was created for the Address Register, but downstream effects were always considered throughout. Consultation with Household Surveys and Census continue to ensure that their expectations are met. This project has now moved to production but has taken many steps to gain acceptance within the organization.
Netherlands	The Center for Big Data Statistics was launched in 2016 and has attracted data scientists with strong expertise in machine learning and computer science. This project has greatly stimulated the collaboration between the two groups. The study has also stressed the importance of specialized hardware and IT skills needed to be able to apply deep learning.	MAKSWELL is a project funded by the European Union to harmonize indicators on sustainable development and well-being. Work Package 3 of MAKSWELL focuses on the measurement of regional poverty. Register-rich countries like the Netherlands can compile regional income-based poverty statistics from a complete enumeration of tax income data.	At this time, it is too early for the project to determine if it will go beyond the demonstration phase.

Switzerland	The FSO's land use statistics are an invaluable tool for long-term spatial observation with an acquisition period that has been gradually reduced from 12 years (in 1979) to 6 years today. At present, internal resources are almost entirely allocated to visual interpretation, at the expense of other activities. Therefore, having a tool that simplifies the task of visual interpreting experts will allow them to generate information more quickly and allow them to contribute to additional activities.	A non-exhaustive list for stakeholders in Switzerland is as follows: Federal Administration (ARE. BAFU, swisstopo, BLW), Regional statistics (CORSTAT), Regional geoinformation centres (KKGEO), Regional spatial planning offices (KPK)	Yes, helping us move forward with a concept that integrates visual interpretation (by experts) and automatic classification and detection of changes.
Mexico	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.	The General Directorate Sociodemographic Statistics. Additionally, it is possible to contact the INEGI cartographic update areas. The General Directorate Sociodemographic Statistics is interested in incorporating quarterly predictions that detect the change in growth in cities to incorporate their values in the population estimation models. Additionally, the cartographic update areas could also take advantage of the quarterly estimates.	There is but no roadmap is fixed yet.

Data Context

In a ML project with satellite images it is very important to define the data sources with which to work. Each country determined the study region and proceeded to acquire satellite or aerial information. With this action, one of the key aspects of this type of projects is identified. Additionally, raster information handling capabilities are required such as specialized software like Geographic Information Systems (ArcGIS, QGis) or the developing processing algorithms through specialized libraries in programming languages like Python (Rasterio, RasterFrames) or R (raster).

The images used for the classification processes were mainly aerial images with submetric resolution (~ 25 cm) per pixel, for which the countries have developed infrastructure and invested in specialized flights which generate the appropriate aerial images for visual interpretation processes carried out by experts or, in another cases, they used open sources such as Landsat images with a resolution of 30 meters per pixel. Each type of image has a specific amount of bands or also known as channels that keep the information corresponding to each color; once the images are available, a segmentation and labeling procedure is carried out, based on the manual work of experts in visual interpretation and/or in field work activities. In the labeling process, a specific number of classes, that depend on the specific objectives of each country, are identified; this process is expensive in time and money. Hence, ML processes aim to contribute to the automatic labeling processes in order to ease the workload of manual processes.

The following table shows a summary of the characteristics of the images and the number of classes that were labeled in the pilot tests carried out in each country.

Country	Imagery	Pixel Resolution	Image Resolution (pixels)	Channels/Bands	Number of Labeled Images	Number of Classes Labeled
Australia	Aerial Images	~23cm	150 x 150	3 (red, green, blue)	6,000	6
Netherlands	Aerial Images Landsat 8	25cm 30 m	400 x 400 4 x 4	3 (red, green, blue) 11 (aerosol, blue, green, red, nir, swir1, swir2, pan, cirrus, tirs1, tirs2)	70,000 70,000	5
Switzerland	Aerial Images Landsat 8	25cm 30 m	200 x 200 Unknown	3 (red, green, blue) 11 (aerosol, blue, green , red, nir, swir1, swir2, pan, cirrus, tirs1, tirs2)	420,000	73
Mexico	Landsat 5,7	30 m	33 x 33	6 (blue, green , red, nir, swir1, swir2)	40,000	2

In addition to the images, complementary information from the study area is usually used that can be used to enrich the characterization processes, for example, georeferenced information in vector format like ESRI Shapefiles or the open GeoPackage format, which contain statistical or geographic information that can be the basis for new labels or contribute to the classification processes. It is also possible to incorporate digital elevation models, which correspond to reticular information where the values of the pixels represent the elevation with respect to sea level and from which it is possible to generate additional information such as the calculation of slopes.

Data preparation and Feature Extraction

Once the data is available, the data is typically processed into a format that is compatible with the classification algorithms, it is at this time that data augmentation can be performed, which is very common in Deep Learning workflows (pipelines). This consists in carrying out systematic variations of the original images to expand the number of labeled examples available, for example rotating the images, changing the scale, etc. This is done in order to prevent or reduce the chances for the algorithm to overfit when using very small data sets.

In these exercises, all the countries carried out Data Augmentation processes to increase the amount of information used by the algorithms.

The feature extraction is a procedure that consists of characterizing the images according to analysis processes, for example, texture calculation, characterizing aspects like the shape, or definition of spectral indices. Sometimes, such as the case of the pilot test in Mexico, feature extraction is performed manually, which means the experts determined the characterization strategy. In the case of the rest of the countries, they relied on the capabilities offered by the convolutional algorithms of deep neural networks for the automatic extraction of characteristics.

Machine Learning Solutions

There is a great variety of ML algorithms (Ferreira, Iten, & Silva, 2020), (Youssef, Aniss, & Jamal, 2020) applied to observations of the Earth. In the case of the pilot projects, two types of algorithms were used in the case of Australia, Netherland and Switzerland, who used state-of-the-art methods based on convolutional neural networks (CNN), these operate based on basic building blocks (convolution filters, pooling layers) that are organized in architectures identified by the state of the art or built by data scientists, according to the needs of each project. Due to the complexity in the training of these algorithms, some tools have been developed (Tensorflow, CNTK, PyThorch, Keras) that take advantage of the computational power of specialized hardware (Graphics Processor Unit (GPU) and Tensor Processing Unit (TPU)). Besides using Deep Learning algorithms, Mexico, Netherlands, and Switzerland used more "traditional" ML methods such as Extremely Randomized Trees (ET), Random Forest (RF) and Support Vector Machines (SVM).

Country	Algorithm	Python Library	Achievement
Australia	Custom 12 layers CNN Architecture	Tensorflow (CPU)	Moved to production
Netherlands	CNN Architecture based on VGG16 and ResNet50 RF and SVM	Tensorflow (GPU) Scikit-learn	Still proof of concept
Switzerland	CNN Architecture based on Xception RF	Tensorflow (GPU) Scikit-learn	Still fine-tuning the algorithm
Mexico	CNN Architecture based on LeNet Extra-Trees	Tensorflow (CPU) Scikit-learn	Still proof of concept

ML is iterative and incremental, and hence, results can improve as experience is gathered in the application of the methods as well as in the specific problem. The countries consider that the algorithms used are well known in the literature; however, as knowledge is gained from the application of the methods, it is possible to reach customized adjustments that improve the results achieved so far.

Results

The results achieved by the countries are in a proof-of-concept stage, with the exception of Australia who have moved their project to production, and it is currently operating in the institution.

Country	Best Model	Overall Accuracy
Australia	Custom CNN	96.9 %
Netherland	ResNet CNN	74.0 %
Switzerland	Xception CNN	~ 90 %
Mexico	Extra Trees	93.87 %

Although they are in different stages, all the countries consider going beyond the demonstration phase. As mentioned before, Australia is already in that phase, followed by Switzerland, who are in the validation and integration stage of established methodologies of the process they are improving; Mexico is in talks with key skate-holders to use their results of the pilot test in real models to validate the impact, specifically, by taking advantage of the results of the 2020 Population Census to validate the results with field data. In the case of Netherlands, they are in the stage of transferring the pilot project to a closed environment that protects the confidentiality of the data to test with confidential records in order to validate the model.

The challenges faced to carry out the pilot test were diverse. For Australia, it was crucial to have a solid business case to convince their organization to launch the project, it was also very important to define the problem to make it as simple as possible so that the goals were achievable and with that, generate reasonably fast value to the organization.

In the case of Netherlands, a bottleneck was not having the specialized hardware (GPU) to train the convolutional neural network in its computing center to avoid having confidential data in open environments. Therefore, their first exercise was carried out with an open data variant that allowed to validate the proof of concept while they manage to get a specialized equipment to work in a secure environment.

In Mexico, it is considered that more iterations should be made in the algorithm training process to achieve better results in the validation phases.

Benefits Obtained

The countries were able to generate value in various ways from the pilot exercises conducted. Australia demonstrated that based on their solid business case and well-defined work, they were able to push the automatic address classification pipeline into production, managing to process a large volume of information in a reliable and fast way. Freeing up analysts and allowing efforts to be focused on the most complex issues that previously could not be addressed due to insufficient resources. The Netherlands, Switzerland and Mexico were able to show that it is possible that ML algorithms associate certain statistical variables and learn from aerial or satellite images to later recognize these variables in images that were not in the training set, meaning these algorithms were able to generalize the process for learning these variables.

In their reports, the countries acknowledge that the results achieved will make it possible to create and implement more ML solutions as other application needs are identified, and that collaboration between methodologists and data scientists has been strengthened; which shows the value that ML can bring to the processes established within institutions.

Learned lessons

Have a solid business case for the ML project

Narrow down the problem; do not try to solve a very complex problem, just enough complexity to add value.

DNN training involves the use of specialized hardware and, if it is desired to use it to train models with large amounts of confidential data within the secure environments of the institutions, it is required to incorporate this hardware into the internal computer centers.

To be able to carry out classification exercises based on satellite and aerial images. It is essential to have high quality training sets validated by experts in visual interpretation, field work as well as complementary data sets from administrative records, surveys or censuses.

References

Curzi, G., Modenini, D., & Tortora, P. (2020). Large Constellations of Small Satellites: A Survey of Near Future Challenges and Missions. *Aerospace, 7*, 133. doi:10.3390/aerospace7090133

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning.* USA: Springer.

Ferreira, B., Iten, M., & Silva, R. G. (2020). Monitoring sustainable development by means of earth observation data and machine learning: a review. *Environmental Sciences Europe, 32*, 120. doi:10.1186/s12302-020-00397-4

Holloway, J., & Mengersen, K. (2018). Statistical Machine Learning Methods and Remote Sensing for Sustainable Development Goals: A Review. *Remote Sensing, 10*, 1365. doi:10.3390/rs10091365

Safyan, M. (2020). Handbook of Small Satellites, Technology, Design, Manufacture, Applications, Economics and Regulation. 1057-1073. doi:10.1007/978-3-030-36308-664

Toth, C., & Jóźków, G. (2016). Remote sensing platforms and sensors: A survey. *ISPRS Journal of Photogrammetry and Remote Sensing*, 22-36.

Youssef, R., Aniss, M., & Jamal, C. (2020). Machine Learning and Deep Learning in Remote Sensing and Urban Application: A Systematic Review and Meta-Analysis. *Proceedings of the 4th Edition of International Conference on Geo-IT and Water Resources 2020, Geo-IT and Water Resources 2020.* New York, NY, USA: Association for Computing Machinery. doi:10.1145/3399205.3399224