

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

## 1. Introduction

### Background

Satellites have revolutionised the way humans monitor the Earth system. The coverage and comprehensiveness of the satellite data is unparalleled and information generated from them is invaluable to solve complex global issues such as climate change<sup>1</sup>. Scientific and academic community have long been using satellite data for monitoring the environment and ecosystem. With national statistics offices (NSOs) actively looking for new sources of data to meet new data demand, satellite data is gaining more and more attention within the official statistics community.

The increasing interest in using the satellite data for official statistics can be attributed to several factors, namely: emerging need for more frequent and more disaggregated data, greater supply of quality satellite data, and advanced capability to make the use of the satellite data within statistical organisations.

With ever fast-changing world, many of issues that a current society faces often requires more frequent monitoring at more disaggregated level. Surveys that statistical organisations use to monitor changes in societies such as urbanisation and land cover, are costly and time-consuming. Satellite data-based statistics can be produced at a much faster pace, allowing a timely decision making. For example, Landsat, one of the biggest Earth Observation (EO) satellites, generates data of global coverage at 30m<sup>2</sup> resolution, every 16 days. The advance of EO technologies has allowed satellite sensors to capture phenomenon on Earth at a much higher spatial and temporal resolution, increasing the accuracy of satellite data reflecting the actual phenomenon. Also, satellite data can support traditional survey-based production (e.g. assisting enumerator visits during survey) which can allow statistical organisations to produce statistics more quickly<sup>2</sup>.

Secondly, it has become easier for users without specialised EO knowledge to obtain the high-quality satellite data. Since its very first satellite half century ago, the number of satellites orbiting around Earth has increased to more than 2,000<sup>3</sup>. While data from many satellites remain non-public (e.g. military satellites, private satellites), several agencies that operate EO satellites have been providing their data as public good and they are now expanding this practice even more. Considerable efforts are being made by the satellite data providers to make these data “analysis-ready”, meaning that raw satellite data are pre-processed in a way that users without EO expertise can start their analysis directly. Moreover, these data have been made freely available on cloud platform which allows users to explore the vast amount of the data without worrying about storage and computing power of their local machines (e.g. [Google Earth Engine](#)). This implies that the entry barriers for non EO-experts to use the satellite data have been lowered significantly.

---

<sup>1</sup> Committee on Earth Observation Satellites (2015) [Satellite Earth Observations in Support of Climate Information Challenges](#)

<sup>2</sup> United Nations Statistics Division (2019) [Guidelines on the use of electronic data collection technologies in population and housing censuses](#)

<sup>3</sup> [UCS Satellite Database](#) (accessed Feb. 2020)

Lastly, capability to process and analyse the satellite data has increased within NSOs. Satellite data can be considered as “big data”. They are often highly structured (with fixed three dimensions: space, time and spectral bandwidth), but one data file can be of gigabytes or even terabytes<sup>4</sup>. For example, the size of satellite data covering Australian continent alone is estimated to surpass 1.5 PB (petabyte) in 2020. Machine learning (ML) techniques is essential to make use of such big data and more and more statistical organisations are investing in developing ML capability in-house. There are already few NSOs that have successfully replaced traditional survey-based statistics with satellite data-based statistics. For example, Statistics Canada replaced its September crop survey with estimates based on satellite data (Section 3.2 for more details).

The potential of satellite data draws great attention and there have been various works at the regional and international levels to introduce this new source of data to official statistical community. Most notably, Satellite Imagery and Geospatial Data Task Team Report from the United Nations Global Work Group (UN GWG) on Big Data<sup>5</sup> provides essential information on a range of topics for statistical organisations to understand the characteristics of satellite data and how they can use the data to produce official statistics. The report includes the overview of available satellite data sources and analysis methods that can be used for the satellite data, and many concrete and practical case studies from NSOs.

More recently, the Conference of European Statisticians (CES) conducted an In-Depth Review on Satellite Imagery and Earth Observation Technology in Official Statistics<sup>6</sup>. The report provides the overview of activities at international and national levels, identifies opportunities and makes recommendations on how to advance the use of satellite imagery for official statistics (e.g. NSOs should collaborate on a generalized approach to EO-data use, NSOs should collaborate with EO organisations to consolidate input requirements).

## **Objective of the Paper**

To make the full use of satellite data for official statistics, there are still challenging issues to address. Although the entry barriers are lower than ever before, the satellite data are relatively new to NSOs and capability needs to be developed to understand pros and cons of the data. The quality implication of using the new data source as well as new analysis methods (e.g. ML techniques) is one of the most prominent issues that has to be tackled. Due to the massive expense and level of technologies involved to operate long-term satellite programmes, the number of satellite data providers is relatively few compared to traditional data providers. This allows international statistical body to establish relationship with the data providers, and discuss and negotiate with them as official statistics community. Such institutional mechanism, however, is not in place yet.

Among the challenges, this paper focuses on the lack of a generalised approach to describe how satellite data can be used to produce official statistics. The development of such general pipeline aims to address following issues:

- There is lack of understanding about business process needed to use satellite data for statistical production. Workflow to use satellite data is often conceived as vastly different from usual business process in statistical organisations. This inhibits statistical organisations to actively seek satellite data as new source of data hence potential of satellite data remains untapped;

---

<sup>4</sup> Lewis, A. et al. (2017) [Remote Sensing of Environment](#)

<sup>5</sup> United Nations Global Working Group on Big Data (2017) [Satellite Imagery and Geospatial Data Task Team Report](#)

<sup>6</sup> Conference of European Statisticians (2019) [In-depth Review on Satellite Imagery and Earth Observation Technology in Official Statistics](#)

- Processing and analysing satellite data require techniques that are not in traditional skill set of statistical organisations. The scope and boundary of works that can be done by statisticians, domain experts and EO experts are not clear which often leads to misunderstanding about tasks involved and their complexity;
- Although there is increasing body of works related to use of satellite data for the production of official statistics, there is no common reference points to consolidate and link them. Useful knowledge about tools, software, techniques and skills exists in isolated manner.

Note that the issue is even more complicated because, as described earlier, 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 remainder of this paper is structured as follow. Section 2 presents the pipeline with description of main steps to follow and related activities for each step. In Section 3, the pipeline is applied to real world examples from NSOs to illustrate the applicability of the pipeline in different contexts. The paper concludes with summary and remarks in Section 4.

## 2. Pipeline

A generic process model describes high-level activities that need to be followed to achieve a certain objective or to deliver a specific output. It provides a common language that can facilitate collaboration among experts in different domains and can be a powerful tool for sharing of knowledge and experience among different organisations. The Generic Statistical Business Process Model (GSBPM)<sup>7</sup>, for example, provides a tool to harmonise the activities undertaken for different survey programmes in statistical organisation which then can be used to identify duplicates and facilitate sharing of services.

Satellite data can serve many purposes in statistical organisations. They can be used for assisting data validation and supporting the field operations in the survey. The use case that this paper aims to describe is for the prediction of “unobserved” values which can then be used to produce existing statistical product more quickly or create a new statistical product. In agricultural survey, for example, surveyed area covers only a part of territory of interest and the rest of territory remains “unobserved”. As satellite data often have global coverage with high temporal frequency, they can be used to predict the unobserved values using relationship learned from comparing “observed” survey data and satellite data with the common coverage area. This can allow agricultural statistics to be produced for much larger scale as well as for non-survey periods.

Based on industry models<sup>8</sup>, a generic pipeline to produce official statistics (through prediction of unobserved values) using satellite data and machine learning is proposed as in diagram (Figure 1) which can be broadly divided into six stages as below:

1. Business understanding: establishing the problem to be resolved; identifying satellite and other data to address the problem; and translating the problem into statistical problem;
2. Data collection and preparation: obtaining satellite data and other data; integrating them; and generating datasets to be used for analysis;
3. ML modelling: deciding ML algorithms and validation method; training ML models; and evaluating the models;
4. Prediction: applying the model to satellite data for prediction; evaluation the results using external dataset;
5. Dissemination: publishing the prediction results with quality criteria;
6. Evaluation

Each stage is further broken down into several steps which are represented as yellow boxes in the columns corresponding to the stage in Figure 1. For each step, Table 1 provides more information regarding:

- What activities should be carried out?
- Who should play the leading role?
- Which GSBPM phase/sub-process can be related to?
- What resources are available?

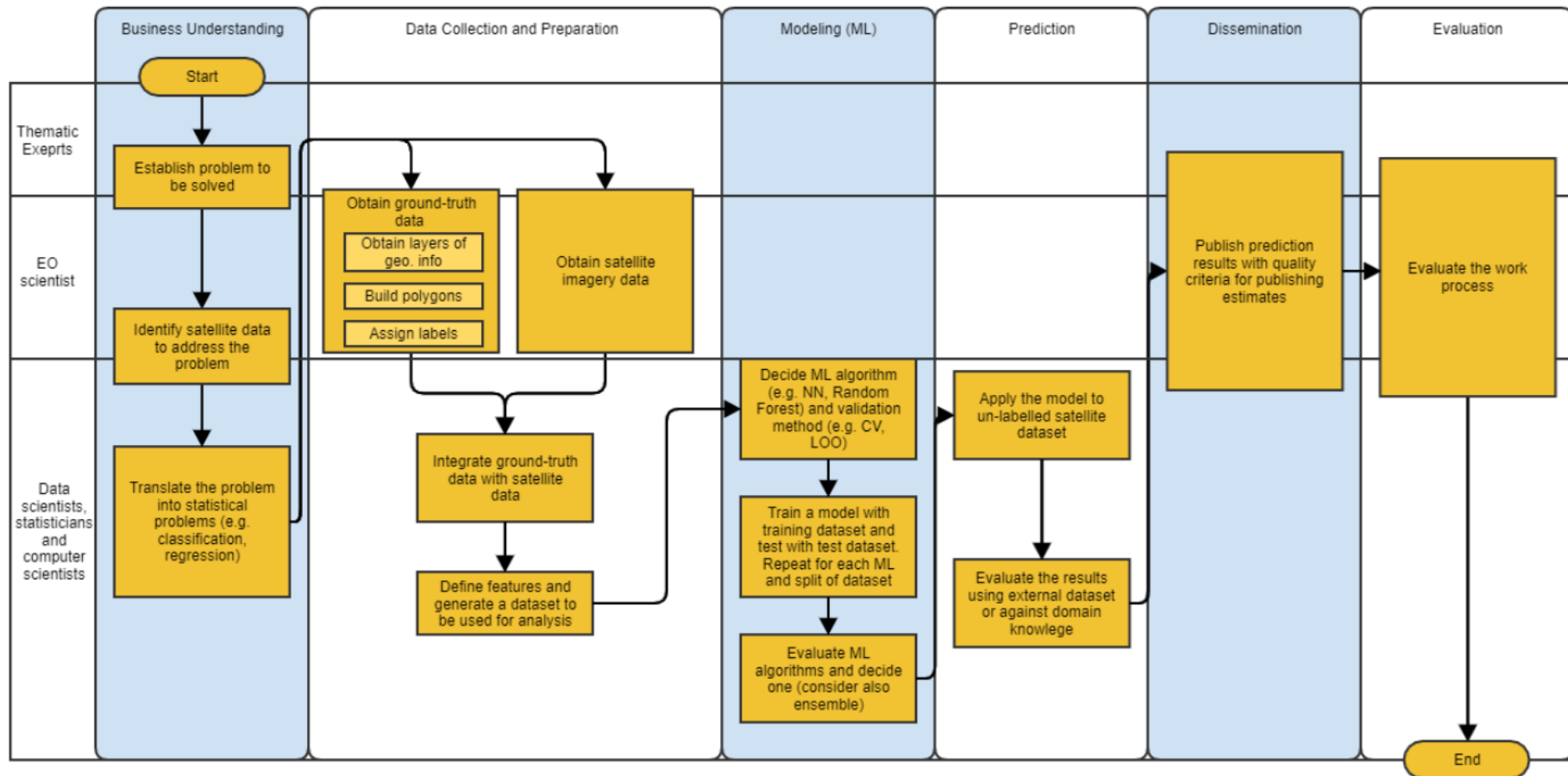
It is hoped that the Table can be further expanded and elaborated with more resources and examples in the future.

---

<sup>7</sup> United Nations Economic Commission for Europe (2019) Generic Statistical Business Process Model (version 5.1)

<sup>8</sup> IBM Cross Industry Standard Process for Data Mining (CISP-DM); Microsoft Team Data Science Process (TDSP); Generic Statistical Business Process Model (GSBPM)

**Figure 1. Diagram of the pipeline**



**Table 1. Description of the pipeline**

	Step	Description	Lead
<b>Stage 1: Business Understanding</b>			
1	Establish the problem to be resolved	<p>In this very first step, domain experts, together with statisticians and EO specialists establish the problem to be resolved with satellite data. This includes initial investigation and identification of what kind of product is needed, how it should be delivered and in what format. The time period and geographical area that the problem is targeting are specified in this step. Although acquiring and preparing satellite data for statistical purpose is resource-intensive activity, once they are ready, the satellite data can serve various purposes. Therefore, it is recommended to engage potential users in the organisation as widely as possible so that their need can be incorporated at the early stage to maximise the return on investment.</p> <p>Related GSBPM sub-process: sub-process 1.1 (Identify needs), sub-process 1.2 (Consult and confirm needs), sub-process 1.3 (Establish output objectives) and sub-process 1.4 (Identify concepts)</p>	Domain experts
2	Identify satellite data and ground-truth data to address the problem	<p>There are several factors to consider when choosing satellite data to address the problem established in the previous step. The problem may require specific temporal resolution (e.g. week, month) and spatial resolution (e.g. 30m, 100m), hence satellite data should meet these requirements. Accuracy of satellite data and completeness of accompanying metadata can impact accuracy (or estimated accuracy) of the resulting statistical product. Continuity plan and open data policy of satellite data provider (as well as political relation with the country operating satellite programme) are also critical factor to consider when choosing which satellite data to use, in particular, if NSO wants to use satellite data for the regular production. Satellite data, in their raw form, requires complex pre-processing which should be conducted by experts with specialised knowledge and skills. Increasingly, satellite data providers provide their data in Analysis-Ready-Data (ARD)<sup>9</sup> form which minimises additional works needed from the user side. Therefore, availability of the satellite data in ADR form is also a critical factor to consider when deciding which source of data to use.</p> <p>Ground-truth data contain direct measure of the actual phenomenon that statistical organisations are interested. For example, agriculture survey can directly capture the yield of crops from farms while what satellite imagery data capture is surface reflectance from the farms (i.e. indirect measure that can be modelled to predict the yield). Ground-truth data are needed to train a prediction model so that the relationship between the actual</p>	EO experts

<sup>9</sup> Sensors on satellite capture energy reflected from objects on Earth. As this reflected energy travels from Earth surface, through atmosphere and above-atmosphere, to the satellite, it is influenced by various atmospheric and exoatmospheric factors. Pre-processing is needed to account for these factors. Although definition of Analysis Ready Data (ARD) is not yet established, it generally refers to satellite imagery data after geometric and radiometric correction.

		<p>phenomenon (represented by ground-truth data) and satellite data is established based on the ground-truth data and this to be used to make a prediction where ground-truth data are not available (i.e. “unobserved”).</p> <p>Depending on the problem identified in the previous step, ground-truth data can be obtained via existing survey, administrative records, field visits, etc. Statistical organisations often have access to micro data which is a great advantage compared to other government agencies or private/academic institutes.</p> <p>Useful resource: UN GWG Handbook Chapter 2. Data Source</p> <p>Related GSBPM sub-process: sub-process 1.5 (Check data availability)</p>	
3	Translate the problem into statistical problem	<p>In this step, the problem identified in previous step is formulated as a statistical problem. This includes specification of unit of analysis and potential (explanatory) variables. The problem can largely fall into two categories: classification (when prediction is for categorical value, e.g. for land cover type) and regression (when prediction is for continuous value, e.g. for crop yield).</p> <p>Useful resource: UN GWG Handbook Chapter 3.5 Analysis</p> <p>Related GSBPM sub-process: sub-process 2.2 (Design variable description)</p>	Statisticians/Data scientists
<b>Stage 2: Data Collection and Preparation</b>			
4	Obtain ground-truth data	<p>In this step, ground-truth data identified in the previous step are acquired. The critical part in this step is to have high quality geo-referenced data. Data are called geo-referenced when the geospatial location of where the data are observed can be identified. For example, farm property or parcel used as sample unit in agricultural survey may have unique ID for administrative purpose. If geospatial information (physical location, area, etc.) linked to the administrative ID exists, the sample unit can be geo-referenced. Having accurate geospatial boundary information of sample units in ground-truth data is crucial as it affects quality of integration with satellite data.</p> <p>Related GSBPM sub-process: sub-processes under phase 4 Collect</p>	EO experts and Statisticians/Data scientists
5	Obtain satellite imagery data	<p>Due to its volume and various pre-processing required, statistical organisation usually do not store bulk satellite data in-house, instead they obtain the data as and when needed. Several satellite operators provide satellite data for free. Most notably, NASA/USGS has a long history of making their data publicly available on open platform such as <a href="#">Earth Explorer</a>. IT companies like Google and Amazon are providing Landsat and Sentinel satellite data on their cloud computing platform. The most ideal situation would be when a country has operational Open</p>	EO experts

		<p>Data Cube (ODC)<sup>10</sup>. ODC provides high-performance computing/data infrastructure that allows users to retrieve ARD data. Agency providing ODC can also provide a product based on ARD which statistical organisations can directly use without obtaining satellite data themselves.</p> <p>Satellite data are often provided with various metadata. In particular, due to pre-processing applied, ADR can have a range of quality measures that are tagged to each pixel. Therefore, it is also important to obtain relevant metadata when obtaining the satellite data.</p> <p>Related GSBPM sub-process: sub-processes under phase 4 Collect</p>	
6	Integrate ground-truth data with satellite imagery data	<p>Once ground-truth data and satellite data are obtained, they should be integrated. Compared to survey data or administrative records that often have matching key variable (e.g. identification key for individual, household, property), satellite data do not have such variable. Instead, satellite data are matched with ground-truth data using geospatial information. While satellite data are provided in pixels, ground-truth data may not exist in such regular shape. Due to this difference in spatial unit and resolution, integrating satellite data with ground-truth data (e.g. irregular shape of property boundary) poses challenges of spatial-matching. For example, pixels could fall completely inside the property boundary but some pixels could fall on the boundary. In such case, a rule should be established to determine when a pixel is considered inside or outside the boundary.</p> <p>After the integration, satellite data can be categorised into two groups: labelled data (pixels in the satellite data that have matching ground-truth data, thus “true” value is known); and unlabelled data (pixels in the satellite data that do not have matching ground-truth data, thus “true” value is unknown). The satellite data whose true values are known are used for developing the prediction model in the later step. In most cases, only small portion of the satellite data has known true values as spatial and temporal coverage of satellite data are often much larger than those of ground-truth data. Notable exception is when census data are used as ground-truth data, in which case all (or almost all) pixels in the satellite data have matching ground-truth data.</p> <p>Useful resources: HLG-MOS Data Integration Project</p> <p>Related GSBPM sub-process: sub-process 5.1 (Integrate data)</p>	EO experts

<sup>10</sup> The objectives of ODC go beyond provision of ARD for specific time or location. They aim to provide spatially/temporally consistent "stacks" of satellite imagery data with quality measures in high performance computing/data infrastructure that allows users to retrieve data for any given time, for any given location, for any given pixel size. According [CEOS Open Data Cube](#), 3 countries (Switzerland, Australia and Colombia) have operational ODC and 7 countries have in-development ODC.



7	Define variables and generate a dataset to be used for analysis	<p>Sensors on satellite captures reflections from Earth at multiple spectral bands. For example, Landsat 8 provides data at 11 bands (e.g. blue band, green band, red band, near infrared band, short-wave band). In this sense, satellite data can be considered as multivariate (spatially/temporally) dataset where each pixel contains values from multiple spectral bands. Satellite data product can also contain composite index derived from bands such as Normalized Difference Vegetation Index (NDVI) that can be used to detect vegetation on the ground. In this step, depending on the problem to be solved, statisticians choose which bands/index to use or define a new variable using bands. After this step, satellite data can be represented as N-by-P matrix where N is the number of pixels in the dataset to be used in ML modelling and P is the number of variables (i.e. bands, index, composite of two).</p> <p>Related GSBPM sub-process: sub-process 5.5 (Derive new variables and units)</p>	EO experts and Statisticians/Data scientists
<b>Stage 3: Machine Learning Modelling</b>			
8	Decide ML method, measure of performance and validation method	<p>Various machine learning methods are applicable for analysing satellite data. Examples include:</p> <p>For classification problem:</p> <ul style="list-style-type: none"> <li>• Logistic and multinomial regression</li> <li>• Support vector machine</li> <li>• Neural networks</li> <li>• Classification trees (CART)</li> <li>• K-Nearest neighbour (K-nn)</li> </ul> <p>For regression problem:</p> <ul style="list-style-type: none"> <li>• Extensions of linear regression (ridge regression, lasso regression, etc.)</li> <li>• Regression tree</li> <li>• Random forest</li> <li>• Functional analysis</li> </ul> <p>Often, multiple methods are considered as candidates to find the most suitable one for given problem and dataset and measure of performance is decided at this step to select a method to make prediction in later step. For classification, measures are constructed based on confusion matrix (error matrix) such as accuracy, precision and recall. For regression, mean squared prediction error or its variations are often used.</p>	Statisticians/Data scientists

		<p>Assessing performance with the same dataset that was used to build the ML model leads to overestimation of performance. Therefore, the labelled dataset is split into training dataset to build a ML model and testing dataset to assess how well the model would work for different sample. Testing results can vary depending on how the dataset is split. To reduce such variability, cross validation (repeating training and testing for different splits) is employed. Cross validation can be exhaustive (e.g. leave-one-out) and non-exhaustive (e.g. k-fold). When splitting the dataset, statistician may need to consider a mechanism to ensure that distribution of response variable in each split is similar to that of the entire dataset.</p> <p>Useful resources: Methodological Approaches for Utilising Satellite Imagery to Estimate Official Crop Area Statistics (2014; ABS)</p> <p>Related GSBPM sub-process: sub-process 2.5 (Design processing and analysis) and sub-process 5.2 Classify and code</p>	
9	Train a model with training dataset and test with test dataset. Repeat for each ML and split of dataset	<p>For each split of the dataset, a ML model for prediction is built using a training dataset. This includes estimation of parameters and feature selection. Training set can be further divided to estimate hyper-parameter that often required for ML. After this, the model is used to make prediction for the testing set. Measure of performance for each method is calculated by averaging measures of performance obtained from multiple testing sets.</p> <p>Related GSBPM sub-process: sub-process 5.2 (Classify and code)</p>	Statisticians/Data scientists
10	Evaluate ML methods and decide one (consider also ensemble)	<p>In this step, a method (or ensemble of multiple methods) is selected based on how well it performed in cross validation. While measure of performance provides important basis to evaluate methods, other factors also play a crucial role in selecting method. For example, if a method requires significantly long computing time than other methods while resulting in only slightly better performance, the method may not the most suitable choice for the regular production.</p> <p>Related GSBPM sub-process: sub-process 5.2 (Classify and code)</p>	Statisticians/Data scientists
<b>Stage 4: Prediction</b>			

11	Apply the model to unlabelled satellite imagery dataset	<p>Unlabelled data refers to data that do not have matching ground-truth data. Obtaining ground-truth data and integrating them with satellite data (step 4 and 6) need significant resources, often requiring manual inspection. Therefore, most parts of satellite data can remain unlabelled and prediction has to be made for this “out-of-sample” dataset using the model that was built based on labelled dataset. The risk of extrapolation has to be carefully examined. In general, the risk is higher when the characteristics of unlabelled dataset is vastly different from labelled dataset (e.g. when time point of unlabelled data is different from that of labelled data).</p> <p>Related GSBPM sub-process: sub-process 5.2 (Classify and code)</p>	Statisticians/Data scientists
12	Evaluate the prediction results using external dataset or against domain knowledge	<p>Once the prediction is made for unlabelled data, it has to be evaluated to ensure the quality of prediction. For the evaluation, statistical organisations can conduct a separate field-visit, consult domain experts or use a Very-High-Resolution (VHR) imagery data. For example, if prediction is made for urban green areas, VHR can be used for validating the results by selecting few predicted areas and manually comparing them with areas in VHR.</p> <p>Related GSBPM sub-process: sub-processes under phase 6. Analyse</p>	Domain experts
<b>Stage 5: Dissemination</b>			
13	Publish the output	<p>In this step, prediction results are published. Depending on quality criteria set by domain experts and statisticians, some parts of the results can be published with flag or entirely suppressed. Given that use of satellite data for official statistics is still at early stage, it is important to make metadata regarding methods and satellite data as accessible as possible when releasing the final output.</p> <p>Useful resource: GSBPM Quality Indicators; HLG-MOS Big Data Project (Framework for the Quality of Big Data)</p> <p>Related GSBPM sub-process: sub-processes under phase 7. Disseminate</p>	Domain experts and Statisticians/Data scientists
<b>Stage 6: Evaluation</b>			
14	Evaluate the work process	<p>In the final step, the evaluation of specific instance of the work process is conducted.</p> <p>Related GSBPM sub-process: sub-processes under phase 8. Evaluate</p>	Domain experts, EO experts and Statisticians/Data scientists

### 3. Application

#### 3.1 Monitoring urbanisation (INEGI, Mexico)

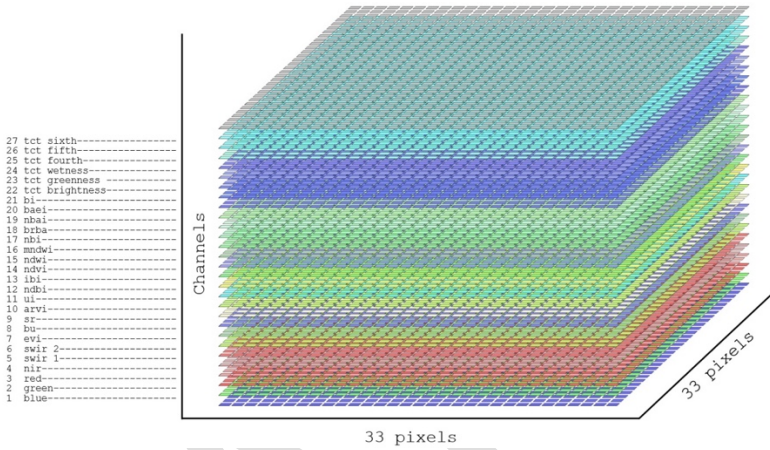
INEGI conducted a pilot study under HLG-MOS Machine Learning Project (2019-2020) on the use of satellite data for the mapping of urban areas in non-census years<sup>11</sup>. The pilot study is mapped to the pipeline in table below. More details regarding method and analysis results can be found in the report of INEGI pilot study.

**Table 2. Activities undertaken in INEGI pilot study mapped to the pipeline**

	Pipeline step	Activities undertaken in INEGI pilot study
1	Establish the problem to be resolved	The use of satellite data was considered to monitor the growth of urban locations in Mexico for non-census years. This would enable the adjustments of the models and samples of the household and business surveys, as well as better planning for the census. When the quality of monitoring system improves, it could also be used to produce new sets of official statistics on its own.
2	Identify satellite and ground-truth data to address the problem	For the satellite data, Landsat 5 and Landsat 7 from NASA/USGS were used. The data have 30m spatial resolution and 16-day temporal resolution (8-day when Landsat 5 and Landsat 7 are combined), and provides 6 spectral bands (blue, green, red, near infrared (NIR), short-wave infrared (SWIR) 1 and short-wave infrared (SWIR) 2). The data are open and provided as ARD (Analysis-Ready-Data).  For the ground-truth data, 2010 Population Census and Economic Census within INEGI were used. These datasets were geo-referenced and available at the census block and economic unit level respectively.
3	Translate the problem into statistical problem	The problem can be formulated as a classification with binary response: urban and non-urban. Given that 2010 census data were available, the focus of the pilot study was on predicting urban areas and non-urban areas for the year 2010.
4	Obtain ground-truth data	(1) Layers of geographical information were obtained from: <ul style="list-style-type: none"> <li>• Georeferenced Population Census (block level aggregation)</li> <li>• Georeferenced Economic Census (economic unit level)</li> </ul> (2) Polygon (rectangular grid) of size 1km x 1km were created covering the territory of Mexico. This results in 1,975,719 grids. (3) Intersecting (1) and (2), each grid was labelled as either urban or non-urban. This results in 36,759 urban grids and 1,938,960 non-urban grids.
5	Obtain satellite data	The satellite imagery data were received from NASA & USGS (32 tera-bytes images in the form of external discs) in March 2019. The data contains Landsat 4, 5, 7 and 8 which, when combined, covers from 1984 to 2018. Using Landsat 6 and Landsat 7 images for the year 2010, a cloud-free national mosaic was created using the geomedian algorithm <sup>12</sup> . This results in 2.7 billion pixels of size 30m x 30m covering the whole territory of Mexico.
6	Integrate ground-truth data with	Although both the satellite data and the ground-truth data were in rectangular forms, the size were different (1km x 1km for the ground-truth data and 30m x 30m for the satellite data). Therefore, (ground-truth) 1km x 1km grid was matched with 33x33 of (satellite data) 30m x 30m grids. Note that as the

<sup>11</sup> INEGI HLG-MOS Machine Learning Pilot Study (accessed March 2020)

<sup>12</sup> Roberts, D., Dunn, B. and Mueller, N. (2018) [Open Data Cube Products Using High-Dimensional Statistics of Time Series](#)

	satellite imagery data	ground-truth data were from census, ground-truth values were available for the whole territory of interest (i.e. Mexico), hence all grids in the satellite data could be “labelled”. Out of 1,975,719 grids, 40,000 grids were selected to be used for ML modelling (20,000 grids for urban and 20,000 grids for non-urban).
7	Define features and generate a dataset to be used for analysis	<p>Starting from 6 spectral bands, 15 index were derived which then used to derive additional 6 index using tasselled cap transformation (27 “channels” in total). In total, for each 1km x 1km grid, 33x33x27 values were obtained as below.</p>  <p>To be used for the ML modelling, features were derived from raw variables as below: for each channel, a patch (consisting of 33x33 grids) was characterized by 7 distributional statistics (mean, variance, median, variation, bias, kurtosis and entropy), frequencies in 10 bin-histogram and value of the central grid. In total, there are 486 features for each 1km x 1km grid (as opposed to 33x33x27 values), which results in 40,000 rows x 486 columns matrix.</p>
8	Decide ML algorithm, measure of performance and validation method	<p>Two ML algorithms were considered: i) Extra Trees (also known as Extremely Randomized Trees); ii) LeNet Convolutional Neural Network</p> <p>For the performance, following measures were used: macro average and weighted average of precision, recall, f1-score and overall accuracy.</p> <p>10-fold cross validation was used.</p>
9	Train a model with training dataset and test with test dataset. Repeat for each ML and split of dataset	The dataset was split into 10 parts. For each split, 90% of the data was used to train and the rest was used to test to produce measures of performances. This was repeated for 10 times (10-fold cross validation).
10	Evaluate ML algorithms and decide one (consider	Extra Trees and LeNet CNN were compared based on the performance measures. Extra Trees performed better than LeNet CNN for all measures

	also ensemble)	while taking less time (fitting Extra Trees took about 1 hour while fitting LeNet took 3 hours).
11	Apply the model to un-labelled satellite dataset	For the ML modelling, 40,000 grids were used. The two models were applied to the rest of the data to produce predication (urban vs. non-urban) for the entire territory in the year 2010.
12	Evaluate the prediction results using external dataset or against domain knowledge	As the whole data were labelled, the same set of performance measures could be obtained as in step 10. Prediction results from the entire territory showed that the Extra Trees model again outperformed LeNet model. However, the precision for urban category decreased compared ML modelling step.
13	Establish quality criteria for publishing estimates and publish data	Not available (the study is for the proof of concept)
14	Evaluate the work process	Not available (the study is for the proof of concept)

### 3.2 Modelling crop yield (Statistics Canada)<sup>13</sup>

The crop survey in Statistics Canada has long history dating back to 1908. The main purpose of the survey is to obtain information on crop areas, yields, production and stocks. The survey results provide valuable information to develop and administer agricultural policies for federal and provincial departments, and to conduct production and price analysis and economic research for relevant academia and private institutions. Until 2015, the survey was conducted at six time points throughout the year: March, June, July, September, November and December. For each time point, farmers provide different information following the seeding, growing and harvesting cycle (e.g. preliminary area for seeding in March, final area seeded in June).

From 2012-13, Statistics Canada started collaborating with Agriculture and Agri-Food Canada (AAFC) and Environment Canada (EC) on a model which could derive crop yield estimates for principal crops grown in Canada. They developed a model-based method to produce a preliminary estimate using satellite data, agroclimatic data and survey data. In 2016, September estimate were replaced by this model-based estimates which has reduced the response burden on the crop producers.

#### **Table 3. Reconstruction of workflow based on the pipeline**

<sup>13</sup> This subsection including the mapping to the pipeline in Table 3 is constructed based on Statistics Canada, [Field Crop Reporting Series](#) (accessed March 2020), Statistics Canada, [Model-based Principal Field Crop Estimates](#) (accessed March 2020) and Statistics Canada, [Integrated Crop Yield Modelling Using Remote Sensing, Agroclimatic Data and Survey Data](#) (accessed March 2020) by author



	<b>Step</b>	<b>Description</b>
1	Establish the problem to be resolved	The purpose is to estimate 19 principal crop yield and production at: i) Provincial level (for Quebec, Ontario, Manitoba, Saskatchewan and Alberta, which account for 98% of agricultural area in Canada) ii) National level
2	Identify satellite and ground-truth data to address the problem	Three types of data were used: i) Satellite data: Normalized Difference Vegetation Index (NDVI) from the Advanced Very High Resolution Radiometer (AVHRR) sensor aboard the National Oceanic and Atmospheric Administration (NOAA) series of satellites (available at a spatial resolution of 1 km <sup>2</sup> ). This spectral vegetation index is often used as a surrogate for photosynthetic potential (values close to +1 indicating high vegetation content while values close to 0 indicating no vegetation). ii) Ground-truth survey data: historical data from Field Crop Reporting Series iii) Ground-truth agroclimatic data: daily temperature and precipitation data at weather stations across Canada
3	Translate the problem into statistical problem	The crop yield takes non-negative continuous value. The problem can be formulated as a regression where yield (November) is the dependent variable and other data (i.e. NDVI, July data, temperature and precipitation) are independent variables.  Although the unit of publication is province and national, the modelling is done at the Census Agricultural Region (CAR) <sup>14</sup> , hence the unit of integration in the later stage is also at CAR level.
4	Obtain ground-truth data	i) Ground-truth survey data: Crop Reporting Series data exist within Statistics Canada. If a crop is abundant in a province, the yield data are available at a lower geographic level (usually, CAR). If not, survey data are available at the province level only. Data for July, September and November from 1987 are used for the model. ii) Ground-truth agroclimatic data: The agroclimatic data are provided by Environment Canada and other partner institutions which are then incorporated into a Versatile Soil Moisture Budget (VSMB) model by AAFC to produce agroclimatic indices used in the analysis. Temperature and precipitation are measured at weather stations located throughout Canada. To obtain a representative value for each CAR, the data from weather stations within the cropland extent of the CAR are averaged. If a CAR lacked climate data, stations from neighbouring CARs were used. Daily agroclimatic indexes were aggregated into monthly sums and means for the months of May to August. Standard deviation (Std) for the month was also calculated and included in the analysis to represent how variable the weather was during the month.
5	Obtain satellite imagery data	Since 1987, Statistics Canada has monitored crop conditions using the AVHRR. The NDVI data were processed on a continuous basis throughout the agricultural growing season (April to October) for the entire land mass of Canada. Statistics Canada has a time series of NDVI data from 1987 to present. Only NDVI pixels that geographically coincide with an agriculture land were extracted to generate the mean NDVI value for cropland within each of the CAR. Three-week moving average of mean weekly NDVI composites were used from week 18 to 36 (May to August).

<sup>14</sup> CARs are composed of groups of adjacent census divisions and each province has multiple CARs

6	Integrate ground-truth data with satellite data	From previous steps, the data are prepared at CAR level, hence integration of three data sets (i.e. satellite NDVI data, ground-truth survey data and ground-truth agroclimatic data) can be done by matching the values using the CAR id. CAR covers much larger geographical area than NDVI pixel (1km <sup>2</sup> ) and both survey and agroclimatic data are available for CARs in the entire territory of interest.
7	Define features and generate a dataset to be used for analysis	From previous steps, the features (e.g. mean, sum, standard deviation of agroclimatic index) are already defined and calculated. Each CAR has 28 years of data (from 1987 to 2014) and 80 explanatory variables.
8	Decide ML algorithm, measure of performance and validation method	<p>Three models were considered:</p> <ul style="list-style-type: none"> <li>i) Stepwise multiple linear regression model (done in SAS)</li> <li>ii) Least Absolute Shrinkage and Selection Operator (LASSO) robust model: to take outliers into account, LASSO method was used for variable selection and MM method for estimation (done in SAS)</li> <li>iii) Least Angle robust model: additionally, another robust method (done in R)</li> </ul> <p>Estimating the November yield is the purpose of the model, hence, to measure performance of each model, relative difference between yield estimates from the model and the November yield estimate was used as below:  Relative difference = 100* (model estimate – November survey estimate)/November survey estimate. Given that September survey estimate used to provide the preliminary estimate for November survey results, the estimates from the model was compared with September survey estimate to check how the model-based approach performs compared to survey-based approach.</p>
9	Train a model with training dataset and test with test dataset. Repeat for each ML and split of dataset	Not available
10	Evaluate ML algorithms and decide one (consider also ensemble)	After evaluating the model estimates using median, 75 <sup>th</sup> and 90 <sup>th</sup> percentiles of the absolute relative difference (see Table 2 and Table 3 in <a href="#">Integrated Crop Yield Modelling Using Remote Sensing, Agroclimatic Data and Survey Data</a> ), LASSO robust model was selected as the final model.
11	Apply the model to unlabelled satellite dataset	Not applicable (all data are labelled and used in step 8)



12	Evaluate the prediction results using external dataset or against domain knowledge	<p>The model estimates are compared with September survey estimates for provincial and national level from 1987 to 2014 in Figure 1 and Table 4-8 of <a href="#">Integrated Crop Yield Modelling Using Remote Sensing, Agroclimatic Data and Survey Data</a>.</p> <p>(After September survey was replaced by model-based approach, July estimates are used to evaluate the prediction (the target of the model is to estimate November estimates, hence difference between the two sets of estimates are to be expected). Subject-matter experts also review the results to identify any questionable estimates. External sources of field crop yield estimates are also used to identify possible errors)</p>
13	Establish quality criteria for publishing estimates and publish data	<p>Modelled yield estimates are produced for crops at the provincial and national levels. A set of rules were established to determine which modelled yields are of an acceptable level of quality to publish. For example, estimates are published if, for each crop,</p> <ul style="list-style-type: none"> <li>i) There exists a minimum of 12 years of historical survey data for both November and July must be available as well as June survey area estimates and July survey yield estimates for the current year; and</li> <li>ii) Coefficient of variation (CV) is smaller than 10%</li> </ul>
14	Evaluate the work process	Not available

## 4. Conclusion

Satellite data hold a great potential for official statistics. They can be used for traditional survey programmes by assisting field operations and data validation, and for producing statistics at a faster pace by allowing prediction on non-surveyed area and period. This paper proposed a generic pipeline covering activities from business understanding to dissemination/evaluation that are needed to make prediction based on satellite data and machine learning. The pipeline was applied to two real world examples to test its applicability. It is hoped that this pipeline can be further extended with more examples and research in the future and contribute to the development of generalised approach on how to use EO data for official statistics.