

## AUTOMATIC IMAGE DATA ANALYTICS FROM A GLOBAL SENTINEL-2 COMPOSITE FOR THE STUDY OF HUMAN SETTLEMENTS

*Christina Corbane<sup>1</sup>, Panagiotis Politis<sup>2</sup>, Pieter Kempeneers<sup>1</sup>, Martino Pesaresi<sup>1</sup>, Dario Rodriguez<sup>1</sup>, Vasileios Syrris<sup>1</sup>, Pierre Soille<sup>1</sup>*

<sup>1</sup>European Commission, Joint Research Centre

<sup>2</sup>Arhs Developments S.A., Luxembourg

### ABSTRACT

The Copernicus Sentinel-2 mission offers new opportunities for mapping human settlements over large areas and for the update and improvement of the Global Human Settlement Layer. This paper presents the fully automated processing workflows tailored for large scale mapping of built-up areas from Sentinel-2 imagery. The first results provide insights into the capabilities gained either by analyzing separately optimally selected S2 tiles or by the processing a best-available-pixel composite over large areas.

**Index Terms**— Global Human Settlement Layer, built-up areas, Sentinel-2, pixel composite, JRC Big Data Platform

### 1. INTRODUCTION

The successful launch of the Copernicus Sentinel satellites marked a new era in the Big Data landscape and stirred the need for the development of operational image processing workflows that produce actionable, trusted, and robust information for different application areas based on free and open data. Mapping and monitoring of human settlements at a global scale is one particular application area that can greatly benefit from the Earth Observation data revolution brought by the Sentinels:

The two first missions, Sentinel-1 (S1) and Sentinel-2 (S2) operational since October 2014 and December 2015, respectively, provide free time series suitable for monitoring built-up areas changes at global scale. Sentinel-1 is designed as a constellation of two synthetic aperture radar (SAR) satellites, namely Sentinel-1A (launched in April 2014) and Sentinel-1B (launched in April 2016), offering a full systematic coverage of the land surface at a global level in the Interferometric Wide swath mode every six days. With such characteristics, Sentinel-1 gives the possibility to provide up-to-date global information on the status and evolution of human settlements and allows regular updates of built-up areas. Sentinel-2 satellites A and B, with the Multi Spectral Imager (MSI) instrument provide a 5-day revisit, 10 m pixels in visible bands: specifications which cover a number of human settlements mapping requirements. The complementarity of the two sensors can be used to compile a joint cloud-free global image database at a fine spatial resolution for mapping human settlements.

In 2016, the first map of human settlements (GHS\_S1) to be fully derived from a global coverage of Sentinel-1 data was produced in the framework of the Global Human Settlement Layer (GHSL) project of the European Commission [1]. Two main components were key to the success of this Big Data challenge: 1) the advanced machine learning technology used for the automatic information extraction and which builds on the Symbolic Machine Learning (SML) classifier originally designed to deal with big data scenario and 2) the versatility of the Joint Research Centre Earth Observation Data and Processing Platform (JEODPP) [2] that allowed the selection, download, storage and mass processing of 6,721 S1 scenes used in the production of the GHS\_S1 layer and the associated global mosaic [3].

The latest developments presented in [1] in terms of the computationally efficient SML classifier combined with the growing capacity of the JEODPP to deploy consolidated information extraction workflows and the opportunity to leverage on the systematic coverage of Sentinel-2 are of great interest for the purpose of human settlements at a global scale [2]. The potential and added-value of Sentinel-2 data for improving high-resolution human settlement mapping was demonstrated in [4] in a pilot study covering selected areas in Italy. Scaling up the methods to cover large geographical areas involves new challenges related to: 1) the adaptation of the workflows to the characteristics of the large and heterogeneous coverage of Sentinel-2 imagery, 2) the need to optimize the selection the images to address the access, storage and computations requirements, 3) the automation of the information extraction methods while allowing flexibility in the choice of the area to be processed and efficiency in I/O.

The present work proposes two automated workflows tailored for large scale mapping of built-up areas from Sentinel-2 imagery. The workflows take into account the need to reduce the computations requirements while still enabling the coverage of large areas such as full countries, continents or even all landmass. The underlying idea is to provide solutions for automatic information extraction from Sentinel-2 data feeds that can work both in cloud environments or standard clusters.

## 2. OPTIMIZED GLOBAL COVERAGE OF SENTINEL-2 INPUT DATA

### 2.1. Optimized selection of S2 tiles

Since February 2018, S2 mission started fully exploiting the two satellite units of the constellation and delivering over 4 Terabytes of daily data on the Copernicus portals. For the purpose of mapping of human settlements at a global scale, it is needed to select from the millions of available S2 images, the best subset that covers the full landmass and minimizes the cloud coverage and the amount of data to be stored and processed.

The selection was performed at the 100 x 100 km tiles according to the Military Grid Reference System (MGRS) in which the S2 images are provided by the European Space Agency (ESA). The selection process itself was based on a floating forward search of all available quicklooks and cloud masks [5]. The quicklooks represent a spatial and spectral subset of the level 1C products. At each iteration the most significant quicklook image was included in order to obtain the minimum number of images required for a cloud free composite.

The maximum number of overlapping images was constrained to five. As a result, less than 5% of the available S2 images in 2017 were selected. On average, 3.16 overlapping images were needed for a cloud free global land cover composite (see FIGURE 1). Due to a lack of snow mask, an important number of selected images were covered with snow. Therefore, the selection process was repeated for latitudes above 45 degrees North, selecting only images acquired during summer season.

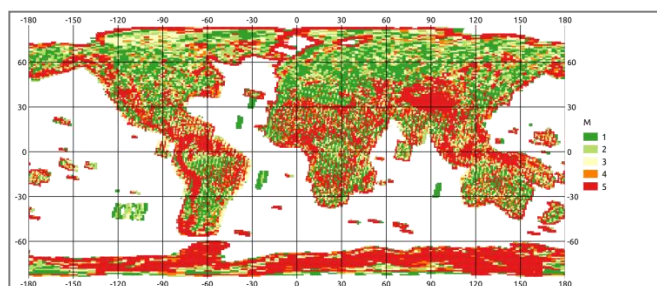


FIGURE 1. Number of overlapping image tiles (M) in the optimal subset obtained from the selection algorithm with  $\max(M)=5$ .

### 2.2. Generation of a global cloud-free image composite

Based on the selection of quicklook images as discussed in section 2.1, the level 1C products at full spatial and spectral resolution were downloaded. A maximum composite was then calculated, based on the maximum normalized difference vegetation index (NDVI) for each pixel (see FIGURE 2).

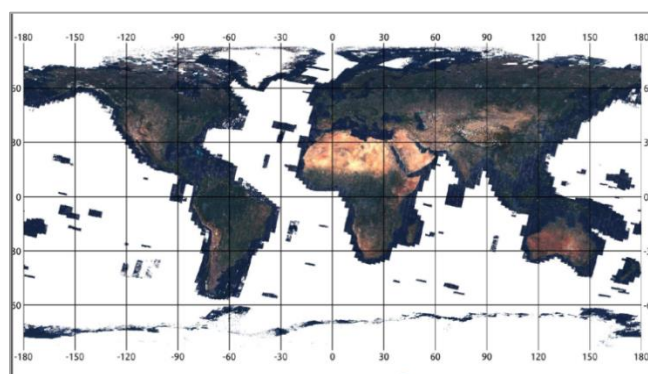


FIGURE 2. World composite based on selected Sentinel-2 quicklooks.

### 2.3. Atmospheric correction of input data

All 92,985 downloaded tiles were stored on the JEODPP and atmospherically corrected using the Sen2Cor L2A processor (version 2.5.5; [European Space Agency. http://step.esa.int/main/third-partyplugins-2/sen2cor/](http://step.esa.int/main/third-partyplugins-2/sen2cor/) (accessed May 2018), which performs topographic correction and transforms top-of-atmosphere reflectance (TOA) to bottom-of-atmosphere reflectance (BOA). Scene classification and cloud masks are produced for each scene in the Sen2Cor process to allow for cloud and shadow masking prior to further analysis. On the basis of the scene classification, the percentage of cloud /shadow coverage over land was calculated.

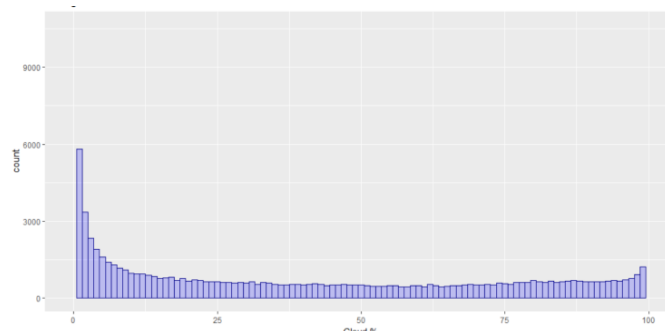


FIGURE 3. Distribution of the number of S2 tiles by percentage of cloud/shadow pixels over land derived from L2A scene classification.

Figure 3 shows the distribution of the number of tiles by percentage of cloud/shadow pixels over land. It shows that despite the quality check and the selection criteria, there are around 1,480 selected tiles with 100% cloud/shadow coverage over land. This is related either to the persistent cloud coverage (e.g. over mountainous areas, tropical zones) and to the non-consideration of the cloud shadows and the land surface in the selection schema that is based on image quicklooks and the poor quality of the cloud masks in vector format delivered with the imagery. A better cloud mask as well as the introduction of a new mask indicating cloud shadows could greatly improve the results.

### 3. PROCESSING FLOWS FOR BUILT-UP AREAS EXTRACTION FROM SENTINEL-2

The GHSL production workflow builds on the Symbolic Machine learning (SML) method that was designed for remote sensing big data analytics. The SML schema is based on two relatively independent steps:

- (1) Reduce the data instances to a symbolic representation (unique discrete data-sequences);
- (2) Evaluate the association between the unique data-sequences subdivided into two parts:  $X$  (input features) and  $Y$  (known class abstraction derived from a learning set).

In the application proposed here the data-abstraction association is evaluated by a confidence measure called ENDI (evidence-based normalized differential index) which is produced in the continuous  $[-1, 1]$  range. Details on the SML algorithm and its eligibility in the framework of big data analytics may be found in [6]. This classification technique has been successfully applied for the processing of Landsat data records of the past 40 years and for generating the first GHSL multi-temporal global product (GHS-Landsat) [1].

#### 3.1. Tile-based processing workflow

A first proof-of-concept demonstrated the added-value of S2 data in improving global high-resolution human settlement mapping [4]. In the current study, the initial algorithm proposed and which builds on the SML classifier is extended to exploit the key features of S2 data: i) the availability of four 10 m spatial resolution bands (B2-Blue, B3- Green, B4- Red and B8- Near Infrared), ii) the availability of six bands at 20 m resolution especially in the Near Infrared and Shortwave Infrared (B5, B6, B7, B8a in Near Infrared and B11, B12 in Shortwave Infrared), iii) the output classification of Sen2cor that can be used for a stratified learning of built-up areas by landcover class.

The following features ( $X$ ) derived from Sentinel-2 are used for the classification of the Sentinel-2 image with the SML approach: i) Spectral features: the three 10 m resolution and the seven 20 m bands, ii) Textural features: a textural feature derived from the brightness (corresponding to the maximum of the visible bands at 10 m) by applying the Pantex methodology [7]. The textural feature is used for refining the output confidence layer by eliminating over-detections, especially roads and open spaces identified as built-up. The learning set ( $Y$ ) is based on the built-up as derived from the GHSL-Landsat. The rough classification output of the Sen2Cor is used during the associative analysis for stratifying the learning set of built-up derived from GHSL-Landsat. This allows tailoring the training set to the image under processing especially in the presence of clouds or cloud shadows and hence allows reducing commission and omission errors. The output confidence is further refined using a global annual composite of maximum Normalized Difference Vegetation Index (NDVI) derived

from Bands B4 and B8 of all S2 images. This global layer was calculated in Google Earth Engine using TOA S2 images acquired in 2017. The diagram in Figure 4 presents a simplified version of the workflow for the classification of S2 image tiles. The processing chain is implemented using a massively parallel workflow that runs at tile level. The output confidences of overlapping and redundant are then tiled and mosaicked hence achieving reduced computation time, allowing easy replacement of image tiles in case of availability of better quality data and ensuring continuity across reference years in the case of updating the product specifications.

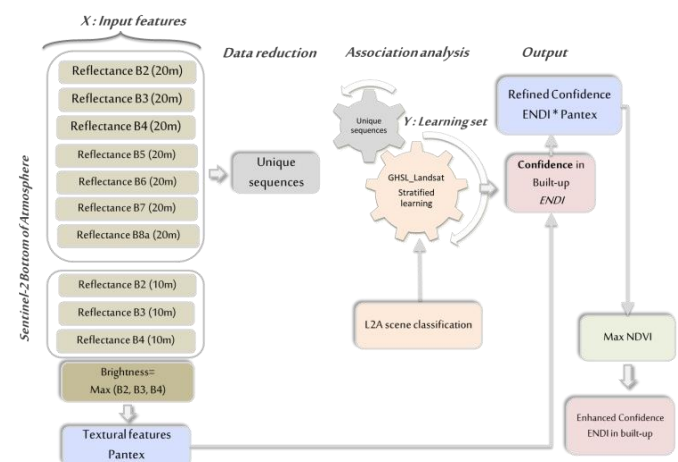


FIGURE 4. Tile-based fully automated workflow for built-up areas extraction from S2 surface reflectance data.

#### 3.2. Composite-based processing workflow

An alternative workflow has been also developed for processing of pixel-based image composites as opposed to scene (tile)-based image processing. The workflow aims at exploiting the best-available-pixel composite by offering a novel opportunity to generate built-up information products that are spatially contiguous over large areas and in a manner that is dynamic, transparent, systematic, repeatable, and spatially exhaustive. The main differences with respect to the tile-based workflow are the following:

- The applicability to both TOA and BOA input data,
- The exclusion of the L2A scene classification from the learning schema because of its irrelevance in the case of the pixel-based composite,
- The use of the 10 meter bands as input, instead of all 10 and 20 m bands, as a compromise between memory efficiency and the need to cover large areas (i.e. equivalent to 5 x 5 S2 images tiles of 100 x 100 km each).

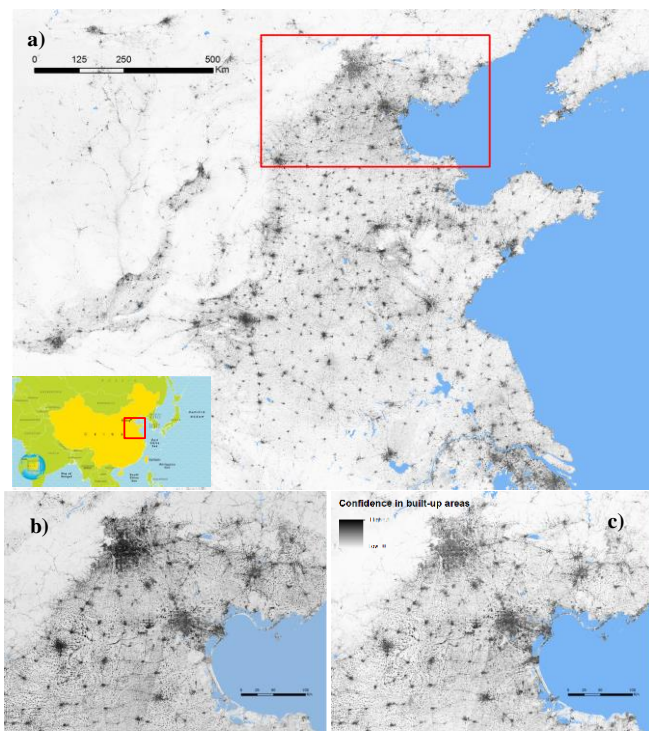
To avoid artifacts in the output confidence layer due to the arbitrary partitioning of the composite, the workflow is executed with a block processing approach including an

overlap of 25% across neighbouring blocks. The outputs confidence layers from overlapping blocks are then merged using the average operator.

## 4. FIRST RESULTS

### 4.1. Visual assessment of the results

The workflows were tested in a pre-operational setting on large areas covering China, Italy, France and selected cities in Asia and Africa. The figures below show the output confidence layer of built-up areas obtained from the processing of 6,068 S2 tiles covering China (Figure 5). A close view over Beijing shows artifacts in the tile-based processing workflow due to the artificial partitioning of S2 images into footprints (Figure 4b). Despite the mosaicking of the results, these artifacts can still be observed especially in the case where two adjacent tiles are acquired in two different seasons with significant differences in the density of the vegetation coverage.



**FIGURE 5.** A) Results of built-up areas extraction in China shown in terms of confidence measure (ENDI rescaled in the range [0,1]). B) Close view of the output of the tile-based workflow over Beijing compared to C) the output from the composite based workflow.

### 4.2. Visual assessment of the results

The two workflows were compared in terms of performance. A large scale test for assessing the performance has been implemented for the same extent in China, covering a total area of 6,210,000 Km<sup>2</sup>. The processing was accomplished using a conventional cluster,

consisting of 16 processing nodes (E5-2650v2@2.60 GHz) with a total amount of 256 GB of RAM. The operating system is CentOS 6.9. Memory usage constraints bounded the number of concurrent jobs to a total of 2 jobs for the composite workflow, resulting in a total processing time of ~12 hours. With 10 concurrent jobs, the tile-based workflow was completed in 15 hours. The results of the test are summarized in Table 1. They show the suitability of both workflows for the processing of large areas and give indications on the scalability potentials of the methods.

**TABLE 1.** Performance assessment of the two workflows

	Tile Based workflow	Composite based workflow
Input	1865 S2 tiles (100 x 100 km tiles)	276 blocks (150x150 km blocks)
Processing time	15 h	12 h
Number of concurrent jobs	10	2
RAM requirements per job	22 GB	120 GB

## 5. CONCLUSION AND OUTLOOK

In this paper two workflows for large scale automatic extraction of built-up areas from Sentinel-2 imagery were presented. Both methods build on the SML classifier, but are tailored to the processing of either single tiles or pixel composites of S2 tiles. The results for both methods are promising, suitable for information retrieval from big volumes of S2 data and offer new prospects for large scale mapping of built-up areas from S2 data. The combination of outputs from both methods is foreseen for the mapping of human settlements at the global level in view of updating the GHSL.

## REFERENCES

- [1] C. Corbane *et al.*, "Mass processing of Sentinel-1 and Landsat data for mapping human settlements at global level," in *Proc. of the BiDS'17*, 2017, pp. 52–55.
- [2] P. Soille *et al.*, "A versatile data-intensive computing platform for information retrieval from big geospatial data," *Future Gener. Comput. Syst.*, vol. 81, pp. 30–40, 2018.
- [3] V. Syrri, C. Corbane, and P. Soille, "A global mosaic from Copernicus Sentinel-1 data," in *Proc. of the BiDS'17*, 2017, pp. 268–271.
- [4] M. Pesaresi, C. Corbane, A. Julea, A. J. Florczyk, and V. Syrri, "Assessment of the added-value of Sentinel-2 for detecting built-up areas in the frame of the Global Human Settlement Layer," *Remote Sens.*, 2015.
- [5] P. Kempeneers and P. Soille, "Optimizing Sentinel-2 image selection in a Big Data context," *Big Earth Data*, vol. 1, no. 1–2, pp. 145–158, 2017.
- [6] M. Pesaresi, V. Syrri, and A. Julea, "A New Method for Earth Observation Data Analytics Based on Symbolic Machine Learning," *Remote Sens.*, vol. 8, no. 5, p. 399, May 2016.
- [7] M. Pesaresi, A. Gerhardinger, and F. Kayitakire, "A Robust Built-Up Area Presence Index by Anisotropic Rotation-Invariant Textural Measure," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 1, no. 3, pp. 180–192, Sep. 2008.