Service Contract 3506/R0-Copernicus/EEA.59142

Task 6: Use of AI/ML EO-based analytics products for filling gaps in CLC+ instances (CLC Legacy and/or LULUCF)

Final Report

Version 1.2

14.12.2023

EIONET Action Group
EAGLE
Land Monitoring in Europe

Title:

Task 6: Use of AI/ML EO-based analytics products for filling gaps in CLC+ instances (CLC Legacy and/or LULUCF)

Place and date:

Wageningen, 14. 12. 2023

Authors:

Marian Vittek, Gerard Hazeu, Geoff Smith

Contributors:

Barbara Kosztra, Emanuele Mancosu, Filipe Marcelino, Julian Delgado Hernandez

Contact information:

Wageningen Environmental Research
Wageningen University and Research
Droevendaalsesteeg 3
6708 PB, Wageningen
Tel.: +31 317 489 559
Email: marian.vittek@wur.nl

Table of Contents

1 SCOPE AND OBJECTIVES

The objective of this negotiated procedure was to continue the support to EEA with respect to the further development of the EAGLE concept and its application in the context of the CLC+ product suite.

Task 1 addresses the update of the relevant EAGLE documentation and making them available to service providers as well as users to support the implementation of CLC+ Core, respectively the ingestion of data into the database.

Task 2 concerns the update of the EAGLE web presentation and the integration of the web pages into the new Copernicus web portal.

Task 3 reviews and updates the bar-coding concept in light of the lessons learned during the ingestions of data into CLC+ Core. On the other hand, the task shall help to simplify and streamline the bar-coding of CLMS products and other frequently used feature classes by providing a proposal for a standard bar-coding of these data.

Task 4 finally addresses the development of an EAGLE concept for the characterisation and handling of change data in the CLC+ Core database as well as a critical review on the extraction of change data from the database.

Task 5 is about providing support to the organisation of Copernicus related meetings.

Task 6 addresses the scope of using AI and ML technologies (and other commercial EO-based analytics) for the gap filling in CLC+ instances.

Task 7 is about providing and “EAGLE view” and support to the ISO standardisation group.

This task 6 summary report presents the results of a scoping study of using AI and ML technologies (and other commercial EO-based analytics) for gap filling of especially CLC+ Legacy.

2 BACKGROUND TASK 6

2.1 Background

In 2019, a LU inventory focusing on potential input data for a future CLC+ Legacy instance was provided (in form of an excel table) by the Member States. The results of these returns have been analysed by ETC/ULS and EAGLE.

In 2021, MS were asked to refine this first inventory with a focus on a future LULUCF Instance. Additionally, MS were asked to provide LU layers created from their national LU data (gap-filled by CLMS data in case of data gaps).

In tasks 15, 16 and 22 of the Specific Contract EEA.59032, the ETC DI analysed the (land use) information provided by the Member States in response to the request above. The analysis will assess crucial information gaps that need to be addressed to develop the required Instances.

2.2 Objective

The objective of this task is to assess the use of modern technologies (like artificial intelligence/machine learning) and/or recent EO-based products that are being provided by the service industry to support the closing of current data and information gaps to derive the CLC+ LULUCF and Legacy instances.

To reach this objective the following main activities were undertaken:

- Inventory of data gaps for CLC+ LULUCF (also Legacy (if existing)): what type of data is missing (data gap typology)
- Inventory of datasets to be used to determine LU in gaps: EO-based products, CLMS data sets, other spatial data
- Evaluations of suitability AI/ML methods for gap filling – which approach to use to fill the gaps
- Test specific methodologies in case study areas for certain gaps
- Review the usefulness of commercially available analytics layers to support gap filling.

2.3 Inputs

Inputs for this task were the following:

- Reports and outputs of tasks 15, 16, 17 and 22 of the Specific Contract EEA.59032
- Spatial data layers with spatial gaps (e.g., NODATA layers from MS countries regarding LULUCF instance (and/or gap layers regarding CLC+ Legacy (if existing))
- Geospatial layers to be used for calculation of missing data
- Analytics layers from external sources

2.4 Partners

The task is led by WENR (Marian Vittek/Gerard Hazeu). Main contributing partner was Specto Natura (Geoff Smith). Lechner, UBA-V, UMA and DGT were other partners with minimum involvement.

3 DATA GAPS INVENTORY AND SELECTION CRITERIA

3.1 Data gap inventory CLC+ instances

3.1.1 Gaps discovered in CLC+ LULUCF (based on CLMS data/national data)

In task 22 "Support to LULUCF instance development, including prototype testing" (SC59032) a critical review and comparison of available LULUCF prototype instance based on only CLMS data and created by a service provider consortium is done. Also, the development of an own LULUCF instance solely based on national land use information is discussed for some specific MS.

The comparison of the surface areas of the prototype LULUCF instances with the statistical data reported by the countries show deviations. Large deviations exist for the IPCC category Wetlands and Other Lands indicating that data harmonisation (class definitions) and data availability are limiting a correct classification of those categories.

3.1.2 Gaps discovered in CLC+ Legacy (based on CLMS data/national data)

Task 16's report "Support for the creation of a pilot CLC+ Legacy instance for 2018 and comparison with CLC2018" (SC59032) gives an overview of the data situation for CLC+ Legacy. The data situation in the MS regarding land use data (EAGLE Land Use Attributes (LUA) and Land Characteristics (LCH) is diverse.

The availability and the quality (e.g. temporal extent, update frequency) of LUAs and LCHs needed to derive CLC+ Legacy are highly variable between countries making it difficult to produce a harmonised European product. Furthermore, CLMS or CLC data is often used in MS to derive LUA/LCH elements which means without CLMS or CLC EAGLE elements could not be derived.

As reported from the gap analysis in "Support for the creation of a pilot CLC+ Legacy instance for 2018 and comparison with CLC2018" (ETC-DI Report 2022/SC59032 Task 16; D 16.1) Table 1 indicates for each CLC class the gap (expressed in %) as the share of MS where data does not exist. The gap analysis was carried out based on the inventory of EAGLE LUA/LCH present in MS national data. For each LUA/LCH the MS were asked to provide the availability and other characteristics of national datasets. The mapping between LUA/LCH and CLC classes allowed for an assessment of the gap in each CLC class. In the case of CLC classes that consist of several LUA/LCH elements only the summary figure is provided. In Table 1, CLC classes are identified in descending order by the percent gap of relevant layers. The higher the gap, the higher the need to find alternative data sources for this class.

Table 1: CLC classes ordered by their respective gap. The gap is measured as the share of MS where data is not available. Original source: EAGLE SC57755 Task 1: CLC+ Core.

CLC code	CLC Description	Gap % of relevant layers
331	Beaches - dunes - sands	Not analysed
332	Bare rocks	Not analysed
511	Water courses	Not analysed
512	Water bodies	Not analysed
523	Sea and ocean	Not analysed
142	Sport and leisure facilities	89

334	Burnt areas	89
121	Industrial or commercial units	81
133	Construction sites	69
212	Permanently irrigated land	69
311	Broad-leaved forest	69
312	Coniferous forest	69
313	Mixed forest	69
211	Non-irrigated arable land	67
141	Green urban areas	64
323	Sclerophyllous vegetation	62
222	Fruit trees and berry plantations	60
241	Annual crops associated with permanent crops	56
213	Rice fields	50
244	Agro-forestry areas	50
421	Salt marshes	50
423	Intertidal flats	50
521	Coastal lagoons	50
221	Vineyards	49
231	Pastures	47
321	Natural grasslands	47
522	Estuaries	47
322	Moors and heathland	46
412	Peat bogs	45
111	Continuous urban fabric	44
112	Discontinuous urban fabric	44
123	Port areas	44
333	Sparsely vegetated areas	44
242	Complex cultivation patterns	42
243	Land principally occupied by agriculture with significant areas of natural vegetation	42
124	Airports	36
132	Dump sites	36
223	Olive groves	36
324	Transitional woodland-shrub	36

411	Inland marshes	36
122	Road and rail networks and associated land	26
131	Mineral extraction sites	25
335	Glaciers and perpetual snow	0
422	Salines	0

The production of CLC+ Legacy instance for some selected countries also revealed that some CLC classes were more difficult to derive on basis of national data. For The Netherlands the CLC classes 123, 423, 522, mixed classes 242, 243 and 313 and separation of classes 511 and 512 were not possible to derive when producing a national CLC+ Legacy (in a combination of the two national databases Landelijk Grondgebruik Nederland (LGN) and Bestand BodemGebruik (BBG). In addition, some CLC+ Legacy classes had large deviations from the original CLC2018 (e.g. CLC classes 122, 133, 222, 321 and 324).

3.1.3 Gaps filled with OSM data

In task 17 of the SC59032 with EEA the usefulness of OSM for gap filling for CLC+ Legacy was assessed. Although OSM has its limitations in terms of determining its accuracy, timeliness, and homogeneity, OSM is for some themes the only European dataset available (e.g., Sport and leisure facilities).

Conclusions of this study were that OSM can be used for gap-filling the following CLC classes:

- 142 Sport and leisure facilities,
- 121 Industry, commerce, and public facilities.

For these classes a revised tag set is provided in Table 2 of the task 17 report “Testing of OSM and other data for CLC+ Legacy” (SC59032)).

Other promising classes mentioned in the task 17 report, which need however further detailed analysis are:

- 421 Salt marshes
- 412 Peat bogs (possibly in combination with Peatland map)
- 123 Port areas (possibly in combination with EuroRegionalMap)
- 124 Airports (possibly in combination with EuroRegionalMap)
- 132 Dump sites
- 122 Road and rail networks and associated land (possibly in combination with EuroRegionalMap)
- 131 Mineral extraction sites

Class 141 was excluded after more detailed analysis because there is not a good tag set in OSM that represents the CLC description of 141. Other classes were excluded because the OSM data itself was partially (albeit for a small part) derived from CLC in the past[1].

Other data

Task 17 of SC59032 suggested that next to OSM other data sources can be used to fill some gaps. The CLC classes that show the highest gap (above 65%) are *Sport and leisure facilities (142),*

[1] In Portugal, the use of OSM is tested to adapt our National LCLU product (COS) from 1ha to 0,5ha and to add some new thematic detail. OSM have proved to be very useful in land use classes like 142, 121, 123, 132, 122, and 131. For Portugal, OSM is not helpful in land cover classes like 421 and 412.

Burnt areas (334), Industrial or commercial units (121), Construction sites (133), Permanently irrigated land (212), Non-irrigated arable land (211) and *the three Forest classes (Broad-leaved forest, Coniferous forest, Mixed forest) (311, 312 and 313),* see Table 1.

Of those, OSM can be used for gap-filling Sport and leisure facilities and Industrial or commercial units.
For the class Construction sites, given its inherently temporary nature, it is hard to find reliable data sources, rather even not national data in many cases.

The HRL Forests Copernicus products can be used for gap-filling the Forest classes (3**).

The EFFIS, GWIS and Burnt area yearly composite datasets can be used, in combination with HRL Forests, to identify burnt areas and therefore help gap-filling class the class Burnt areas.

Finally, the explored global maps for the classes Permanently irrigated land, Non-irrigated arable land do not meet the general criteria for gap-filling.

3.2 Selection of data gaps

3.2.1 Background

After several studies looking at the feasibility to produce CLC+ LULUCF and CLC+ Legacy instance it became clear that there is especially a lack of harmonised land use data at European level. Land use is particularly important for the production of CLC+ Legacy instance as CLC classes are namely described as a mix of land cover and land use (LC/LU). Also for CLC+ LULUCF instance land use (LU), i.e. management information is needed, however all the information is grouped into 6 reporting categories that apparently have fewer gaps compared with the required 44 classes of CLC. For this reason, the focus in this study is on CLC+ Legacy instance.

Experiences with the ingestion and/or extraction of national and European spatial LC/LU data into CLC+ Core on basis of EAGLE elements it became clear that for the moment we should focus on the gap filling of LU/LC classes and not on EAGLE elements. The final CLC+ instances are spatial datasets mapping LC/LU classes and not elements. Furthermore, training data on EAGLE elements (LUA or LCH) that are needed to develop AI/ML based LC/LU mapping, are more difficult to obtain.

For these above mentioned reasons it was decided to see how AI/ML can potentially contribute to filling gaps in the production of CLC+ Legacy, i.e. can AI/ML help to map CLC+ Legacy classes for which no harmonised European data exist.

The selection of CLC classes to focus on is partly based on the gap analysis performed in tasks performed in previous Copernicus Service Contracts (classes having high shares of MS where data is not available, no thematic correspondence in OSM etc). Next to the gaps discovered during these previous studies, it became clear that some CLC classes were difficult to map when creating CLC+ Legacy at national level for the Netherlands. Furthermore, to get a better insight in the potential of AI/ML to map CLC classes it was also decided to select some classes in both the urban as the semi-natural domain. Also, during the selection of CLC classes, we took the availability of sufficient training data into consideration.

3.2.2 Selection criteria

The general criteria that datasets need to fulfil to be used for gap-filling are the following (as mentioned in task 17 report “Testing of OSM and other data for CLC+ Legacy” (SC59032)):

- Reference date: the reference date of the dataset must match or be close enough to CLC+ Legacy. It must be possible to determine the reference date.
- Regular updates: the dataset must be regularly updated.
- Spatial coverage: the dataset must cover Europe (EEA-38 + UK).
- Spatial resolution: the dataset must have the same or higher spatial resolution of CLC+ Legacy.
- Thematic homogeneity: the dataset thematic definitions must be homogeneous across Europe, and homogeneous/comparable to CLC thematic definitions.
- Accuracy: The accuracy of the dataset must be known.
- Availability: the dataset must be free to use (ideally with open data licence).

In this study on the assessment of the use of modern technologies (like artificial intelligence/machine learning) and/or recent EO-based products these criteria also apply. For the AI/ML derived products especially the reference date and spatial resolution of the data used for training and classifying gaps and the accuracy of the result are important in the consideration if data can be used for gap filling. And in addition to the list above for AI/ML derived products high quality (accurate, extent/area) training data is of utmost importance.

3.2.3 Selected gaps

On basis of the data gaps inventory discussed in section 3.1 and the section 3.2.1 and 3.2.2 we came to the selection of the following data gaps for CLC+ Legacy classes:

- 121 - Industrial sites
- 123 - Port areas
- 141 - Green urban areas
- 324 - Transitional woodland shrub
- 423 - Intertidal flats

3.2.4 Training data

One of the selection criteria is the availability of training data. For the training of the AI/ML models we are using Urban Atlas and Coastal Zones hotspot datasets 2018. The more detailed classes of these datasets can be relatively easy aggregated as they have a hierarchical relation to the above mentioned CLC classes.

4 SELECTION AND SUITABILITY OF AI/ML METHODS FOR GAP FILLING

4.1 AI/ML definition

Artificial intelligence (AI) is multidisciplinary field including variety of technologies such as machine learning (ML) and deep learning (DL). AI is the theory and development of computer systems able to perform tasks normally requiring human intelligence, such as visual perception, decision making, speech recognition and image patterns recognition. As shown in Figure 1, DL is a part of ML as well as a part of the broad AI field. AI incorporates human behaviour to machines or systems, while ML is the method to learn from data or experience, which automates analytical model building. DL also represents learning methods from data where the computation is done through multi-layer neural networks and processing. Deep learning methodology uses term “deep” to refer to the concept of multiple layers of state through which data is processed for building a data-driven model.

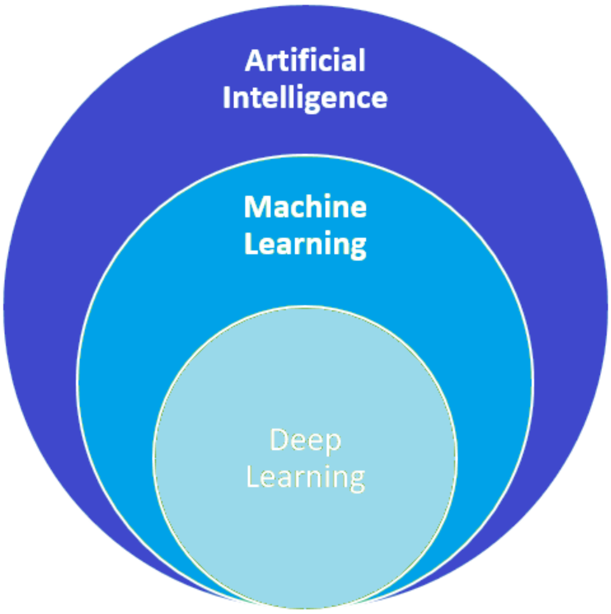


Figure 1. Overview position of artificial intelligence, machine earning and deep Learning.

4.2 Overview of AI/ML methodologies

AI in general offers wide range of methodologies which are in principle divided into two main categories: supervised and unsupervised. Supervised ML techniques are applied when there is a sample of data which needs to be predicted or explained. It could be done by using previous data of inputs and outputs to predict an output based on a new input. Unsupervised ML is focused on ways to relate and group data points without the use of a target variable to predict. In other words, it evaluates data in terms of traits and uses the traits to form clusters of items that are similar to one another. In the field of DL there are different groups of algorithms which differs on type of information to be extracted from image dataset such as object detection and image by image classification. In the field of remote sensing is most relevant instance segmentation performed by pixel-wise segmentation.

4.3 Potential of AI/ML technologies for land cover/use mapping

ML methods are often applied in land cover mapping thanks to its ability to learn automatically complex patterns and relationships in large and high-dimensional datasets.

DL techniques, particularly convolutional neural networks (CNNs), are increasingly being applied in land cover classification tasks due to their ability to automatically learn hierarchical features from spatial data. In the process of selecting the right algorithm, it is important to take into consideration several factors such as the size of dataset, the complexity of the land cover classes, and the computational resources available, among others.

5 Test specific AI/ML methodologies in case study areas

5.1 Study area

We selected two study areas with the size of 100 x 100 km which corresponds to EEA grid E37N32 in Netherlands and E30N16 in Spain (Figure 2). The main reasons for selection of these areas were:

- Test for different climate, environment, landscape
- Presence of selected classes / data gaps
- Different availability (coverage) of training data
- Availability of satellite imagery (Sentinel-2) for several time steps within the reference year 2018

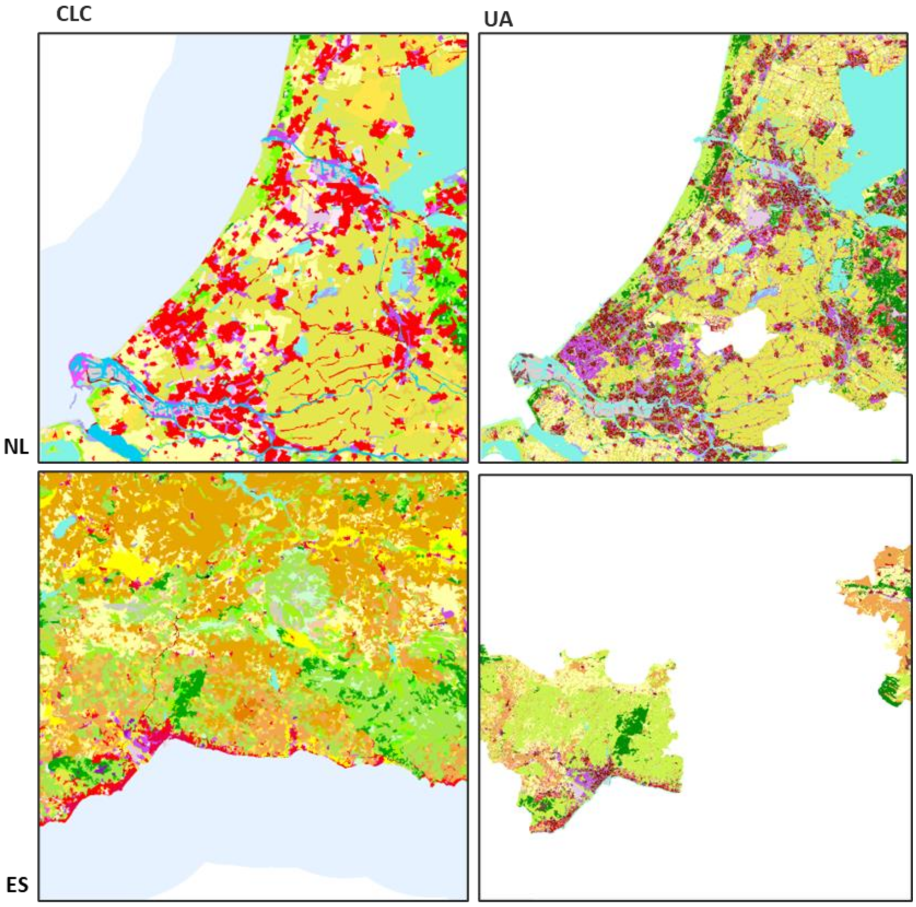


Figure 2. Selected study sites: Above – Netherlands E37N32, below - Spain E30N16 with coverage of CLC (left) and UA (right).

The list below shows an overview of the used datasets and their translation into CLC classes.

- **Urban Atlas**
  - 12100 Industrial, commercial, public, military and private units -> 121
  - 12300 Port areas -> 123
  - 14100 Green urban areas -> 141
  - 32000 Herbaceous vegetation associations -> 324
- **Coastal zones** (level 2/3)
  - 112 Industrial, commercial, public and military units -> 121
  - 123 Port areas and associated land -> 123

    - 14 Green urban, sports and leisure facilities -> 141
    - 34 Transitional woodland and scrub -> 324
    - 723 Intertidal flats -> 423
- **Sentinel 2** satellite images (16)
    - 4 bands (10m)
    - 4 seasons in the year 2018

The UA and CZ maps have been merged and translated into CLC classes. The merged dataset was created as a Union and UA has priority over CZ. So, the focus in the study was on the CLC classes 121 (Industrial or commercial units), 123 (Port areas), 141 (Green urban areas), 324 (Transitional woodland/shrub) and 423 (Intertidal flats).

5.2 Methods

This task is focused on the evaluation of methodologies for mapping of CLC classes by means of artificial intelligence methods and classification of remote sensing imagery. Target areas for mapping are those areas where high-resolution datasets for direct translation into CLC classes are not available. We used in this task AI based approaches which are largely data oriented and computationally intensive. Advantage of using such a methodology is that workflows could be highly automated and could be repeated as soon as new data are added. From several methods available, we selected Random Forest and Neural network as a representation of machine learning and deep learning approach, respectively.

5.2.1 Random Forest

As a first method Random forest (RF) was selected, which is a supervised ML method working on the principle of constructing a multitude of decision trees at training stage and the majority vote (mode) across them in the classification stage. RF builds multiple decision trees by training each tree on a random subset of the training data within process is known as bagging. This helps to reduce overfitting and increases the model's stability. Since RF uses random selection of individual pixels and shape of entities being detected does not play important role, this method could be more suitable for mapping classes with irregular shapes.

In the process of setting up the classification model, the hyperparameter tuning was performed by random search tool to determine optimal values for the following parameters: number of estimators and maximal depth of decision trees. Both the random search tool and the RF model use scikit learn v. 1.3.2[2] an open-source machine learning library in Python.

For the RF implementation we selected the following methods of organising and processing data: spatial split (section 5.2.1.1) and pixel based random split (multiclass and per class classification) (section 5.2.1.2).

5.2.1.1 Spatial split

Datasets in both study sites were spatially divided for training (70%) and validation (30%). For the Dutch study site a division was made in direction north – training, south – validation and for the Spanish study site in direction east – training, west - validation in order to capture the variability of LC classes being mapped.

[2] https://scikit-learn.org/

5.2.1.2 Pixel based random split

In the pre-processing phase 2000 samples were selected for each class what gives in total 12000 random samples over whole dataset (5 CLC classes and a ‘unknown’ or unclassified class). RF model was used to predict CLC classes in the whole dataset.

5.2.1.2.1 Multiclass classification

In the pre-processing phase 2000 random samples within the training subset of the study area were selected for each class in LC dataset (labels) and corresponding images layer (covariates) in order to create a balanced training dataset. The pixel values of these samples were used as an input for the training of the classification model. Following, the established model was used to predict pixel values representing classes for the validation subset of the study area.

5.2.1.2.2 Per class classification

Each of 5 classes were classified separately following the same steps as for the multiclass approach.

5.2.2 Neural Networks

The second method applied in this task is neural networks (NN). NN works on the principle of using layers of interconnected nodes (artificial neurons), combined together to process and learn from data. In particular, Convolutional Neural Networks (CNN) are relevant for use in semantic segmentation of images due to their ability to capture spatial hierarchies of features detected on satellite images. By employing hierarchical layers, the CNN model architecture is capturing broader spatial relationships and contextual information that can aid in making more accurate segmentation predictions. Contextual information helps the model understand the global structure of the scene and improves its ability to differentiate between different objects or regions within an image and could be more accurate in the urban environment featuring regular shape of objects.

As for the input, there were 12000 random points selected over the study area. Input satellite image stack and label layer were split into tiles of size 32 x 32 pixels that were used in the sequential deep neural network model which we applied[3].

5.3 Results

5.3.1 The Netherlands

5.3.1.1 Random forest

5.3.1.1.1 Spatial split

Figure 3 shows the mapping result after applying the RF algorithm for the Dutch test area. Multiclass classification resulted in a low overall accuracy (47%) (see Figure 4). The accuracy of individual classes when evaluating f1 score (harmonic mean of user accuracy (UA) and producer accuracy (PA) i.e. (2*((PA*UA)/(PA+UA)))) ranges from highest for Industrial areas (34%) through Port areas (30%), Green urban areas (18%), Transitional woodlands-shrub (11%) and Intertidal flats (3%) with the lowest f1 score.

Class Green urban areas were best classified with a producer accuracy (PA) of 72% followed by Port areas (64%) and Industrial areas (53%). Industrial areas were often classified as Port areas or Green urban areas, but also many Port areas were classified as Industrial areas. This is to be

[3] https://keras.io/guides/sequential_model/

expected as Industrial areas and Port areas are spectrally very similar. Lowest producer accuracies were for Transitional woodland/shrub (41%) and Intertidal flats (8%). Large parts of Transitional woodland-shrub were misclassified and assigned to Green urban areas. Also this could be expected as both classes have similarities in their spectral reflectance. Similarly, the majority of Intertidal flats class were classified as Port areas. From the areas not mapped into any of the target classes the largest areas were assigned to Industrial and Green urban areas.

The user accuracies (UA) of all classes are relatively low. UA from 25% (Industrial areas) till 2% (Intertidal flats) are low which means large overestimations of the classes (high commission errors). The unclassified class has by far the highest UA.

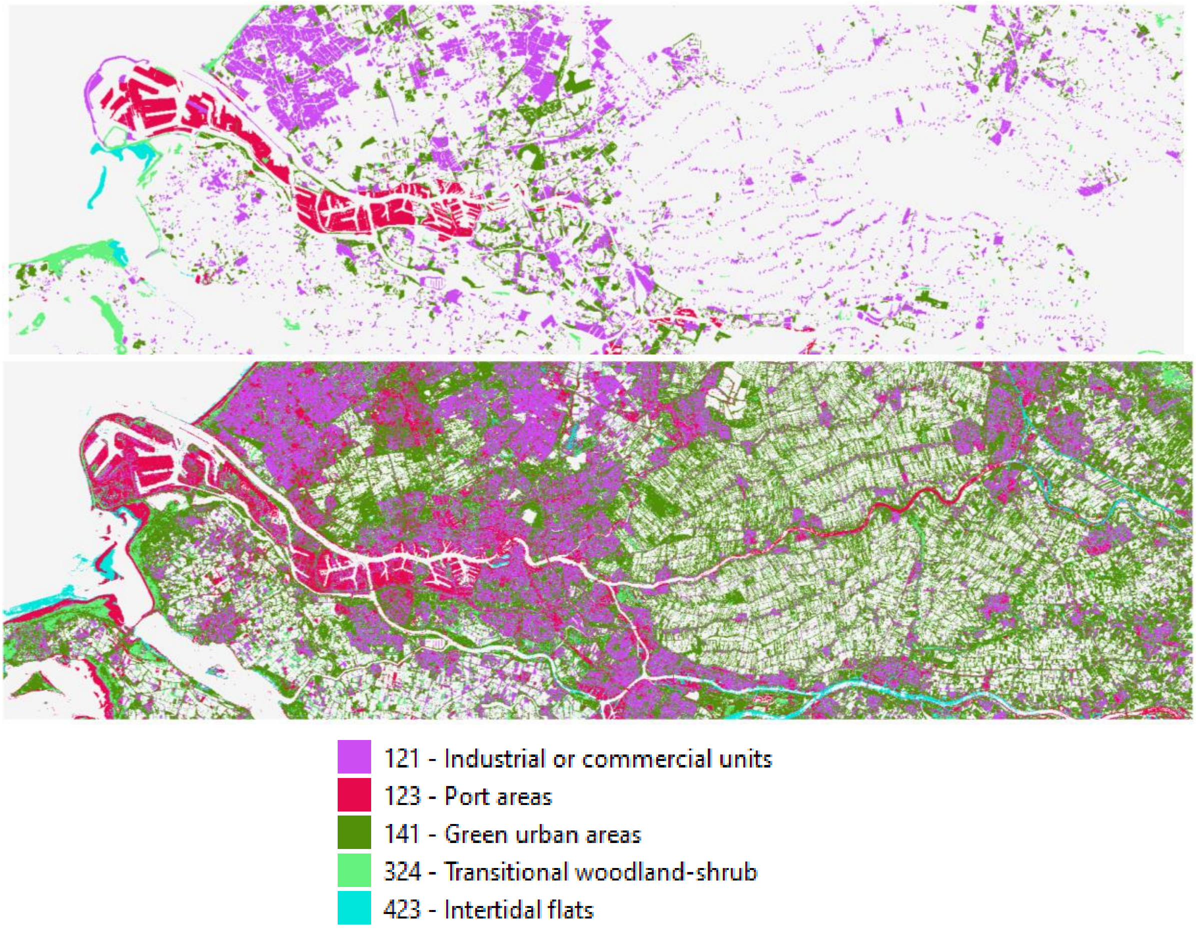

121 - Industrial or commercial units
123 - Port areas
141 - Green urban areas
324 - Transitional woodland-shrub
423 - Intertidal flats

Figure 3. Original 5 CLC classes derived from UA and CZ (above) and predicted CLC classes using RF model at the study site in Netherlands (spatial split).

		PREDICTED							
		Industrial	**Ports**	**GreenUrban**	**TransWoodlands**	**IntertidalFlats**	**Unclassified**	**SUM**	**PA**
TRUE	**Industrial**	1358586 4.53%	539069 1.80%	470114 1.57%	143284 0.48%	4682 0.02%	35410 0.12%	2551145	53.25%
	Ports	163513 0.55%	437475 1.46%	46310 0.15%	22766 0.08%	3373 0.01%	6713 0.02%	680150	64.32%
	GreenUrban	172677 0.58%	36123 0.12%	935377 3.12%	106577 0.36%	9419 0.03%	33484 0.11%	1293657	72.30%
	TransWoodlands	14847 0.05%	11744 0.04%	136277 0.45%	113880 0.38%	378 0.00%	3772 0.01%	280898	40.54%
	IntertidalFlats	1820 0.01%	43572 0.15%	306 0.00%	2344 0.01%	5989 0.02%	18183 0.06%	72214	8.29%
	Unclassified	3656894 12.19%	1184876 3.95%	7456345 24.85%	1313425 4.38%	338246 1.13%	11172150 37.24%	25121936	44.47%
SUM		5368337	2252859	9044729	1702276	362087	11269712	30000000	
UA		25.31%	19.42%	10.34%	6.69%	1.65%	99.13%		**46.74%**

Figure 4. Confusion matrix showing true label of original and predicted CLC classes using RF model at the study site in Netherlands (spatial split).

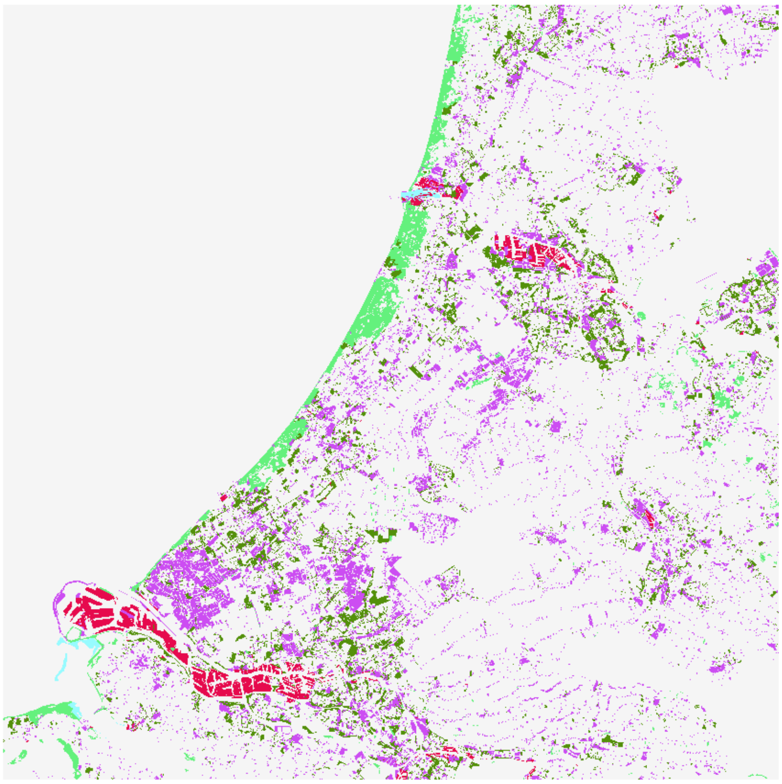
5.3.1.1.2 Pixel based random split

Multiclass classification

Figure 5 shows the mapping result after applying the RF algorithm for the Dutch test area (pixel based random split – multiclass classification). The multiclass classification with pixel based random split resulted in an overall accuracy of 64% (see Figure 6). Most remarkable is the overestimation of target classes when a lot of unclassified areas were assigned to different classes. Accuracy of individual classes when evaluating f1 score (harmonic mean of user and producer accuracy i.e. (2*((PA*UA)/(PA+UA)))) ranges from highest for Transitional woodland/shrub and Industrial areas (31%) through Port areas (22%), Green urban areas (20%) and Intertidal flats (11%).

Producer accuracy (PA) is particularly high for Intertidal flats (94%) and Transitional woodland/shrubs (81%) meaning that a low number of pixels were missed in the classification (low omission error). Industrial areas and Green urban areas have the lowest producer accuracy 46% and 66%, respectively.

User accuracy (UA) is expressing the rate of a class not being overestimated relative to the total area of the class is highest for Industrial areas (23%) and Transitional woodland/shrubs (19%) and lowest for Green urban areas (12%) and Intertidal flats (6%). These relatively low user accuracies (UA) indicate that these classes are highly overestimated (high commission errors).



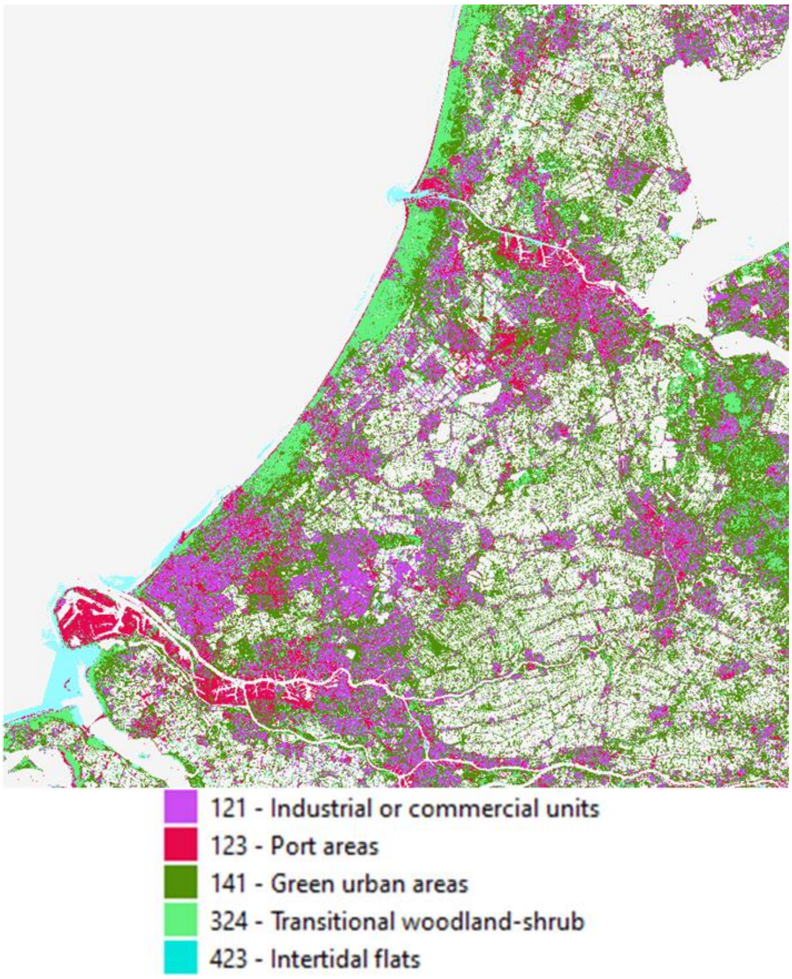
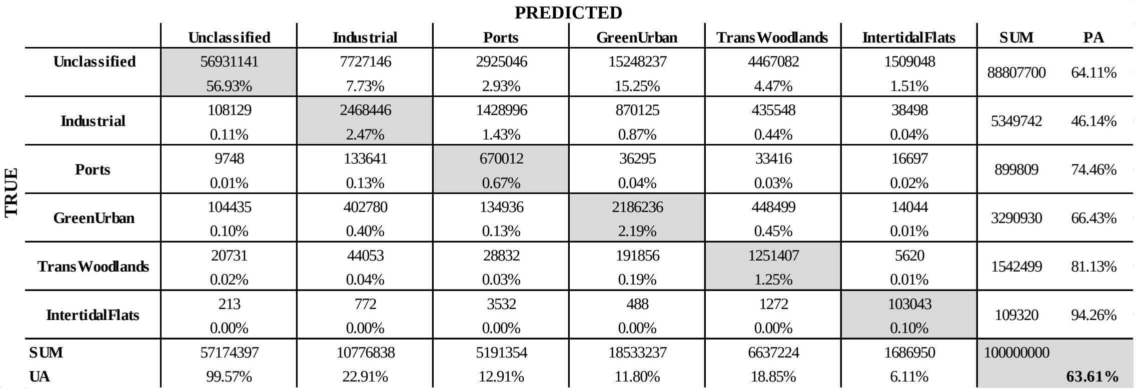

🟪 121 - Industrial or commercial units
🟥 123 - Port areas
🟩 141 - Green urban areas
🟩 324 - Transitional woodland-shrub
🟦 423 - Intertidal flats

Figure 5. Original 5 CLC classes derived from UA and CZ (above) and predicted CLC classes using RF model at the study site in Netherlands (pixel based random split – multiclass classification).



		PREDICTED							
		Unclassified	**Industrial**	**Ports**	**GreenUrban**	**TransWoodlands**	**IntertidalFlats**	**SUM**	**PA**
TRUE	**Unclassified**	56931141 56.93%	7727146 7.73%	2925046 2.93%	15248237 15.25%	4467082 4.47%	1509048 1.51%	88807700	64.11%
	Industrial	108129 0.11%	2468446 2.47%	1428996 1.43%	870125 0.87%	435548 0.44%	38498 0.04%	5349742	46.14%
	Ports	9748 0.01%	133641 0.13%	670012 0.67%	36295 0.04%	33416 0.03%	16697 0.02%	899809	74.46%
	GreenUrban	104435 0.10%	402780 0.40%	134936 0.13%	2186236 2.19%	448499 0.45%	14044 0.01%	3290930	66.43%
	TransWoodlands	20731 0.02%	44053 0.04%	28832 0.03%	191856 0.19%	1251407 1.25%	5620 0.01%	1542499	81.13%
	IntertidalFlats	213 0.00%	772 0.00%	3532 0.00%	488 0.00%	1272 0.00%	103043 0.10%	109320	94.26%
	SUM	57174397	10776838	5191354	18533237	6637224	1686950	100000000	
	UA	99.57%	22.91%	12.91%	11.80%	18.85%	6.11%		**63.61%**

Figure 6. Confusion matrix showing true label of original and predicted CLC classes using RF model at the study site in Netherlands (pixel based random split – multiclass classification).

Per class classification

The classification of Industrial areas as individual class resulted in an overall accuracy of 80% when class area itself and the rest of study area were evaluated together. When considering only the class area, f1 score (harmonic mean of user and producer accuracy) reaches 32%. Producer accuracy is proportionally higher (87%) than user accuracy (19%) meaning that many pixels of the background were classified as industrial areas while a lower number of pixels for this class were missing in the classification. In other words, a high commission error respectively low omission error (see Figure 7).

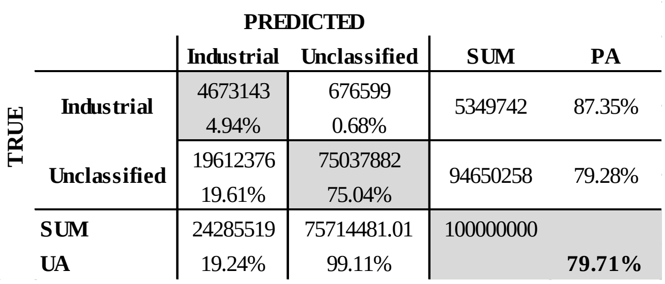
The classification of Port areas as individual class has an high overall accuracy (88%) when class area itself and the rest of study area were evaluated together. When considering only the class area, f1 score reaches 12%. Producer accuracy is much higher (92%) than the user accuracy (7%) meaning that a very large number of background pixels were classified as Port areas while a low number of pixels for this class were missing in the classification.

The classification of Green urban areas as individual class reaches overall accuracy 72% when class area itself and the rest of study area were evaluated together. When considering only the class area, f1 score reaches 17%. Producer accuracy is much higher (90%) than user accuracy (10%) meaning that a very large number of background pixels were classified as Green urban areas while a low number of pixels for this class were missing in the classification.

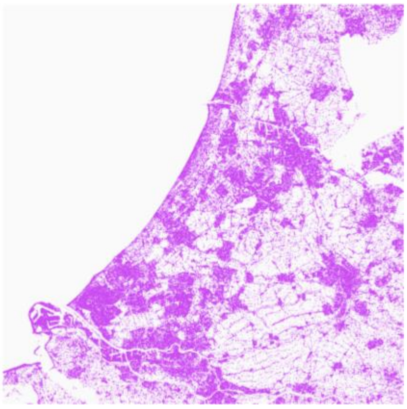
The classification of Transitional woodland/shrub areas as individual class has high overall accuracy (87%) when class area itself and the rest of study area were evaluated together. When considering only the class area, f1 score gets 18%. Producer accuracy is much higher (93%) than user accuracy (10%) meaning that a very large number of background pixels (relative to total number of pixels in class) were classified as Transitional woodland/shrub areas while a low number of pixels for this class were missing in the classification.

The classification of Intertidal flats as individual class has very high overall accuracy (97%) when class area itself and the rest of study area were evaluated together. When considering only the class area, f1 score reaches only 7%. Producer accuracy is much higher (98%) than user accuracy (3%) meaning that a very large number of background pixels were classified as Intertidal flats while a low number of pixels for this class were missing in the classification. In other words, a high commission error respectively low omission error (see Figure 7).

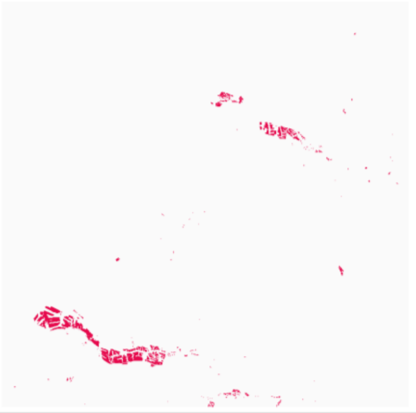
The general pattern is that individual classes are being overestimated when classified individually. The most probable reason is that background areas are very heterogenous including classes very similar to the target class.



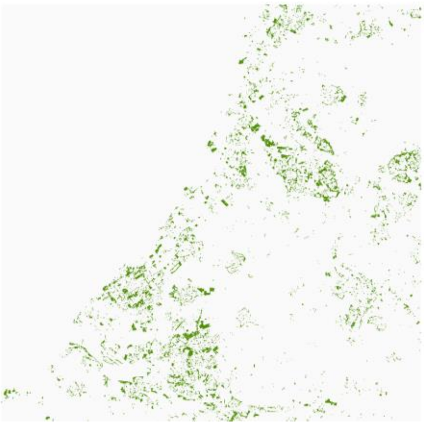
		PREDICTED			
		Industrial	Unclassified	SUM	PA
TRUE	Industrial	4673143 4.94%	676599 0.68%	5349742	87.35%
	Unclassified	19612376 19.61%	75037882 75.04%	94650258	79.28%
	SUM	24285519	75714481.01	100000000	
	UA	19.24%	99.11%		**79.71%**



		PREDICTED			
		Ports	Unclassified	SUM	PA
TRUE	Ports	829557 0.84%	70252 0.07%	899809	92.19%
	Unclassified	11699136 11.70%	87401055 87.40%	99100191	88.19%
	SUM	12528693	87471307	100000000	
	UA	6.62%	99.92%		**88.23%**



		PREDICTED			
		Urb.Green	Unclassified	SUM	PA
TRUE	Urb.Green	2970186 3.07%	320744 0.32%	3290930	90.25%
	Unclassified	27667811 27.67%	69041259 69.04%	96709070	71.39%
	SUM	30637997	69362003	100000000	
	UA	9.69%	99.54%		**72.01%**



		PREDICTED			
		Shrubs	Unclassified	SUM	PA
TRUE	Shrubs	1428970 1.45%	113529 0.11%	1542499	92.64%
	Unclassified	13344860 13.34%	85112641 85.11%	98457501	86.45%
	SUM	14773830	85226170	100000000	
	UA	9.67%	99.87%		**86.54%**

		PREDICTED			
		Int.Flats	Unclassified	SUM	PA
TRUE	Int.flats	106841 0.11%	2479 0.00%	109320	97.73%
	Unclassified	3056235 3.06%	96834445 96.83%	99890680	96.94%
	SUM	3163076	96836924	100000000	
	UA	3.38%	100.00%		**96.94%**

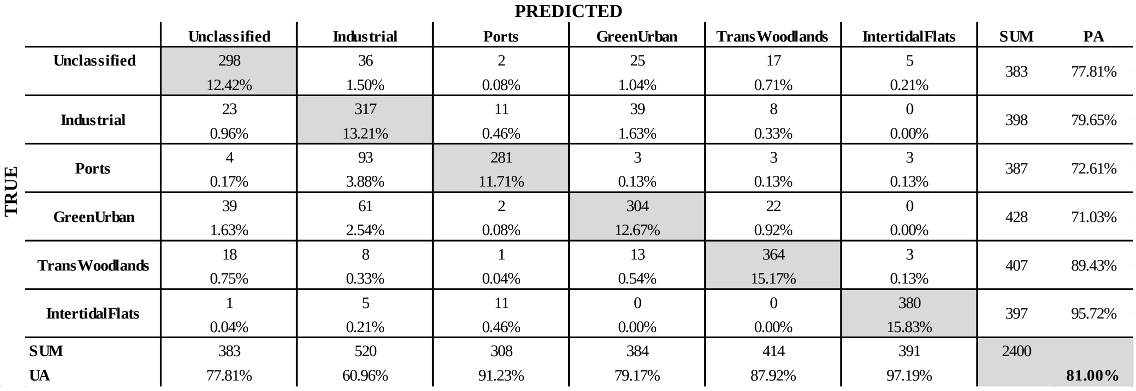
Figure 7. Confusion matrices evaluating true and predicted labels (left) and map coverage (right) of individual CLC classes and using RF model at the study site in Netherlands (Industrial areas, Port areas, Green urban areas, Transitional woodland/shrub and Intertidal flats from top to bottom).

5.3.1.2 Neural Networks

Figure 8 shows the confusion matrix for the Dutch test area after applying the neural network methodology. The multiclass classification with pixel based random split is validated with a sample dataset containing 2400 data points resulted in a relatively high overall accuracy (81%). The accuracy of individual classes when evaluating f1 scores ranges from highest for Intertidal flats (96%), through Transitional woodlands/Shrubs (89%), Port areas (81%), Green Urban areas (75%) and Industrial Areas (69%).

The highest producer accuracy is for Intertidal flats (96%) and for Transitional woodland/shrubs 89% while for other classes it ranges between 70-80%.

The user accuracy reaches highest values for Intertidal flats (97%) and Port areas (91%) while its lowest value is for Industrial areas (61%). The high user accuracies mean that compared to other methodologies applied the commission error (or overestimation) is relatively low.



		PREDICTED							
		Unclassified	**Industrial**	**Ports**	**GreenUrban**	**TransWoodlands**	**IntertidalFlats**	**SUM**	**PA**
TRUE	**Unclassified**	298 12.42%	36 1.50%	2 0.08%	25 1.04%	17 0.71%	5 0.21%	383	77.81%
	Industrial	23 0.96%	317 13.21%	11 0.46%	39 1.63%	8 0.33%	0 0.00%	398	79.65%
	Ports	4 0.17%	93 3.88%	281 11.71%	3 0.13%	3 0.13%	3 0.13%	387	72.61%
	GreenUrban	39 1.63%	61 2.54%	2 0.08%	304 12.67%	22 0.92%	0 0.00%	428	71.03%
	TransWoodlands	18 0.75%	8 0.33%	1 0.04%	13 0.54%	364 15.17%	3 0.13%	407	89.43%
	IntertidalFlats	1 0.04%	5 0.21%	11 0.46%	0 0.00%	0 0.00%	380 15.83%	397	95.72%
	SUM	383	520	308	384	414	391	2400	
	UA	77.81%	60.96%	91.23%	79.17%	87.92%	97.19%		**81.00%**

Figure 8. Confusion matrix showing true label of original and predicted CLC classes using neural networks at the study site in Netherlands.

5.3.2 Spain

5.3.2.1 Random forest

Figure 9 shows the mapping result after applying the RF algorithm for the Spanish test area. Multiclass classification resulted in an overall accuracy of 63% (Figure 10). The accuracy of individual classes when evaluating f1 scores ranges from highest for Transitional woodland/shrub (51%) through Industrial areas (9%), Green urban areas (9%), and Port areas (3%) with the lowest.

Producer accuracy is particularly high for Transitional woodland/shrubs (82%) and Port areas (77%) meaning that a low number of pixels were missed in the classification (low omission error). Industrial areas has the lowest producer accuracy with 47%.

User accuracy is also highest for Transitional woodland/shrubs (37%) and lowest for Port areas (2%). User accuracies are far lower than the producer accuracies. The low user accuracies mean that the classes are overestimated (high commission errors) as it was also registered in the Dutch test case.

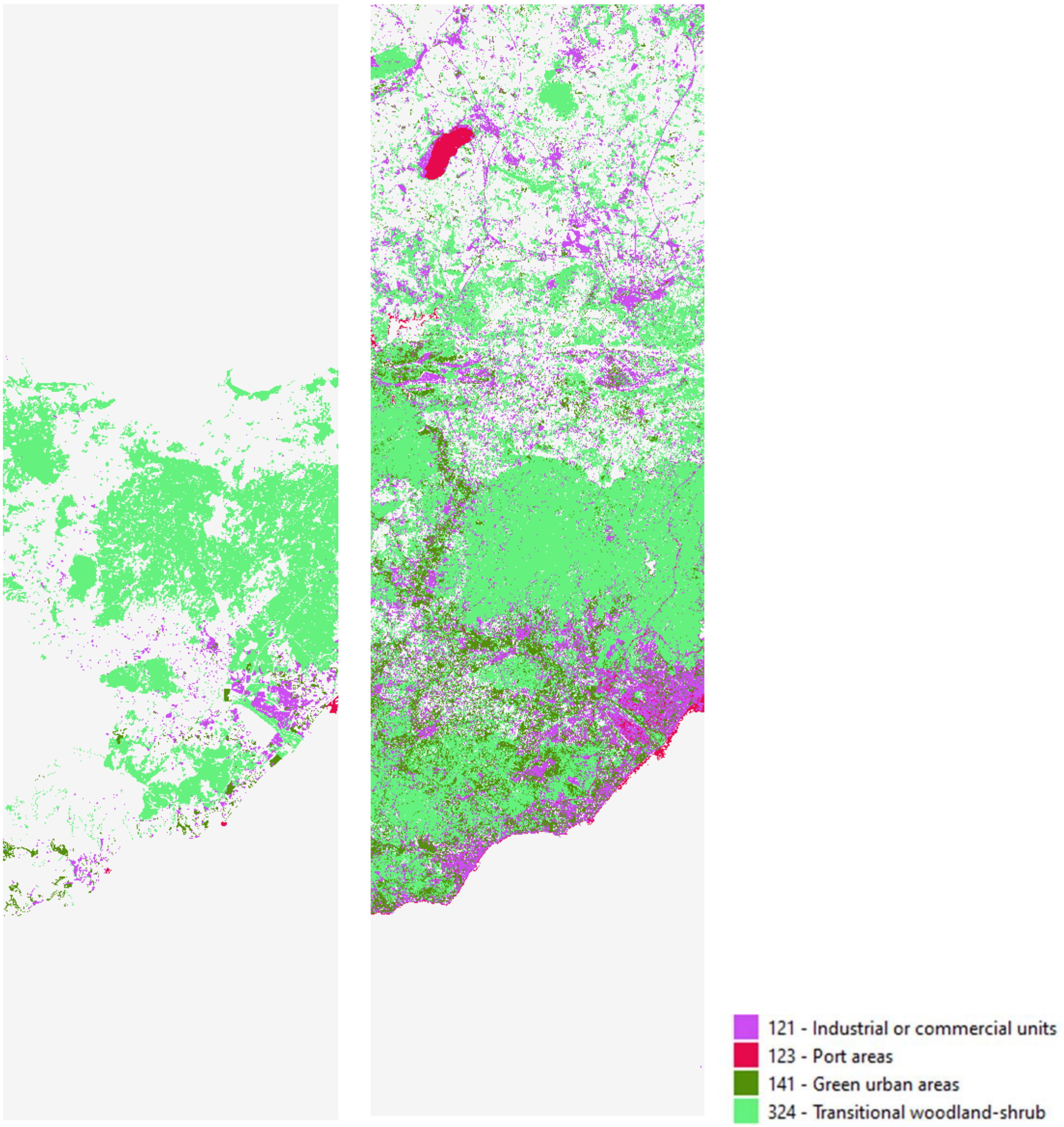
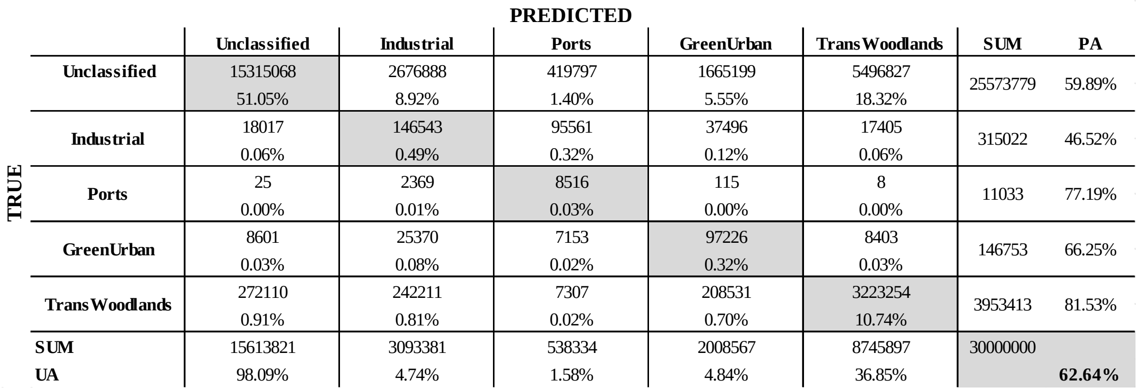

121 - Industrial or commercial units
123 - Port areas
141 - Green urban areas
324 - Transitional woodland-shrub

Figure 9. Original 4 CLC classes derived from UA and CZ (left) and predicted CLC classes (right) using RF model at the study site in Spain.



		PREDICTED						
		Unclassified	**Industrial**	**Ports**	**GreenUrban**	**TransWoodlands**	**SUM**	**PA**
TRUE	**Unclassified**	15315068 51.05%	2676888 8.92%	419797 1.40%	1665199 5.55%	5496827 18.32%	25573779	59.89%
	Industrial	18017 0.06%	146543 0.49%	95561 0.32%	37496 0.12%	17405 0.06%	315022	46.52%
	Ports	25 0.00%	2369 0.01%	8516 0.03%	115 0.00%	8 0.00%	11033	77.19%
	GreenUrban	8601 0.03%	25370 0.08%	7153 0.02%	97226 0.32%	8403 0.03%	146753	66.25%
	TransWoodlands	272110 0.91%	242211 0.81%	7307 0.02%	208531 0.70%	3223254 10.74%	3953413	81.53%
	SUM **UA**	15613821 98.09%	3093381 4.74%	538334 1.58%	2008567 4.84%	8745897 36.85%	30000000	 **62.64%**

Figure 10. Confusion matrix showing true label of original and predicted CLC classes using RF model at the study site in Spain.

5.4 Conclusions and Discussion

In overall, the deep learning approach using the neural networks classification model shows higher overall accuracy (81%) than the machine learning using the random forest classification model (64%). This comparison concerns the results of pixel based random split since it was used for both modelling approaches. Classification results differ in the order of individual class f1 scores. The neural networks model resulted with highest f1 score for the classes Intertidal flats (96%), Transitional woodland/shrub (89%) and Industrial areas (80%), while random forest model has the highest f1 score for the classes Transitional woodland/shrub, Industrial areas (both 31%) and Port areas (22%). The high accuracy rate for Intertidal flats might not be representative since the number of sampled pixels is relatively high compared to the overall low number of pixels belonging to the class in the study area. Therefore, the pixel based random approach might use the same pixels for training and validation resulting in high accuracies.

In general, the results of per class classification of individual classes has lower accuracy than the ones coming from the multiclass classification. The one exception is the Industrial areas class with f1 score 32% for the per class classification (individual class) compared to 31% for the multiclass classification. Classes with f1 scores of the same magnitude are Green urban areas and Intertidal flats with 17% and 7% (per class classification) compared to 20% and 11% for the multiclass classification, respectively. Relatively high differences in f1 scores exist between the per class and the multiclass classification for the Port areas and Transitional woodland/shrub classes, i.e. 12% and 18% versus 22% and 31%, respectively. The reason for lower f1 scores for the individual classifications could be that the unclassified or background areas have a larger extent and they include more area spectrally similar to the target classes to be classified.

The random forest approach applied on data selected, for training and validation, with the pixel based random split approach resulted in an higher overall accuracy (64%) compared to the spatial split approach (47%). The reason could be that the selection of samples from a broader area could have a positive impact on capturing the characteristics of the class representative for the entire study area and not only for the training area as in the case of the spatial split approach. However, the accuracy in the pixel based random split approach could be a little overestimated since when calculating accuracy for the entire area, training pixels are also included (2000 per each class). This is not the case for neural networks classification where random sample pixels are selected as different than the ones used for the classification.

The results of random forest classification using spatial split shows higher overall accuracy at the study site in Spain (63%) compared to The Netherlands (47%). In the Spanish study site f1 scores are much higher for class Transitional woodland/shrubs, i.e. 51% versus 11% for The Netherlands. For the classes Industrial, Port and Green urban areas the f1 scores in The Netherlands (34%, 30% and 18%, respectively) are higher than in Spain (9%, 3% and 9%, respectively). Intertidal flats class is occurring only in the Netherlands (f1 score = 2%). The reason for these differences between the study sites might be that some classes are more difficult to separate from their surroundings due to the fact that their spectral properties are more similar in Spain. Except for the Transitional woodland/shrub which is more accurately mapped in Spain than in The Netherlands.

The user accuracies are in general lower than the producer accuracies (except the ones for the neural networks). The user accuracies for the classes classified with the neural network method are much higher than the ones from the random forest method which means the overestimation or commission errors are much lower for the classification based on the neural network method compared to the other methodologies applied.

5.5 Recommendations

AI/ML methods showed the possibility and potential to be applied in filling gaps in CLC data. The methods have the benefit that they can cover large area by using a consistent approach with minimal human intervention. The AI/ML methods are using large amount of source data in the form of satellite images (and other covariates) and they are computational intensive. However, a drawback might be the relation/dependency between the class definition or data label used for the model training and the covariate layers – predictors. The class definition is not always reflected by the properties of the remote sensing data used as predictors, i.e. the spectral properties and spatial organisation of elements defining the LC/LU class are not always reflected by the satellite image properties. Therefore, the improvement and increasing the amount of training data would be beneficial.

Also, the classification with AI/ML could be improved by using the knowledge of LU/LC information typically surrounding a target class to be mapped. This could be partially captured by using NN approach where surrounding area is considered, additional information can be added in form of underlaying data layers, masks or buffer zones. For example, Port areas can occur only within certain distance from sea and rivers, Urban green areas can only be present within built up areas and Transitional woodlands/shrubs occur only outside of urban areas.

In order to get better classification results it is an option to include more covariate layers such as elevation, Copernicus high resolution layers etc.

Accuracy of certain class detection depends also on level of aggregation. Mapping accurately CLC classes such as Industrial and Port Areas is relatively difficult if based only on remote sensing data. However, after aggregation to higher level (Industrial, commercial and transport units / Artificial surfaces) it might be easier to separate them from other higher level aggregated CLC classes. On other hand some classes are better distinguishable at more detailed level or non-aggregated. E.g. individual buildings are better distinguishable then urban fabric.

The large non-mapped or background areas, i.e. the areas not mapped by the target classes, in the original dataset of unclassified are very heterogeneous and sometimes spectrally looks similar to the target classes. For this reason the non-mapped areas might being wrongly mapped as one of the target classes. For avoiding or minimising the mixing-up of target classes with the background areas we think a two-step approach could help. A first step is the division of the background, i.e. the area not mapped as one of the target classes, into urban and non-urban background areas. The second step is to use the urban background in the mapping of the urban CLC classes (classes 121, 123 and 141) and to use the non-urban background for mapping the non-urban CLC classes (324, 423). We suggest to use for those two different groups of classes different AI/ML methods, possibly a neural network for the urban classes and a RF method for the non-urban classes.

In overall, deep learning approach showed higher precision on mapping of classes selected in this task. In order to improve the results there are still different possibilities to adapt neural network model architecture being more customized to particular target classes.

While RF and NN represents existing reliable methods for LU/LC mapping, additional methods such using generative AI could also show a good potential for the filling gaps in the LU/LC data. Generative AI is being typically used to generate synthetic data layers, augment existing datasets or simulate various environmental conditions. Generative models such as Generative Adversarial Networks involves generating realistic synthetic images that can be added to the original dataset, thereby increasing the diversity of the training data.

6 EXPLORATION POTENTIAL FOR DERIVED ANALYTICS FROM COMMERCIAL EO SYSTEMS

6.1 Introduction

This sub task aimed at exploring the potential for analytic layers derived from commercial EO systems to populate some of the EAGLE elements which are beyond the capabilities of Sentinel data and existing land cover land use sources at a pan-European level. These analytics layers would be ingested into CLC+ Core to provide additional capabilities for those developing extraction rulesets and for the production of CLC+ Instances.

The sub task was in part built on the 'Earth Observation methods for gap filling' review undertaken in the scope of the CLC+ Core Data Need report (Task 1 of service contract 3436/R0-Copernicus/EEA.57755) and focused on potentially usable EO methods for filling data gaps when deriving CLC+ LULUCF and CLC+ Legacy instances. It concluded that some EO-based intermediate products could be used to address around a quarter of the known data gaps at the time. However, not all of these were part of operational services and thus their delivery in a sustainable fashion could not be guaranteed.

In this work, analytic layers / intermediate products from the commercial sector (e.g., via commercial satellite operators and aggregators such as Planet, Capella, SkyFi, etc.) were assessed to see if they could potentially support the more challenging elements of the EAGLE data model.

For example, transient features and detailed spatial patterns / textures may be difficult to detect with the spatial resolution and temporal repeat frequency of the Sentinels. Some commercial systems have enhanced spatial and temporal performance even if they are more limited in terms of quality.

The first part of the sub task reviewed **the available analytics layers and then attempted to bar code** them with the latest version of the EAGLE data model.

The second half of the task **tested potential options for, and identify issues with, the ingestion of the analytics layers into the CLC+ Core**, including the likely licensing and access arrangements with the various data suppliers. It also demonstrated a small number of use cases to illustrate potential operational deployment.

Some of the political context for this sub task was set out in early 2022 by the European Commission. At the 14th EU Space Conference in January, Thierry Breton[4] noted that public procurement lowers commercial risk and provides long-term prospect to stabilise the business of a small companies, in particular start-ups. It also has a positive effect on private investors. Then at the Copernicus Horizon 2035 event[5] in February, which brought together the brightest minds in Earth Observation, Mauro Facchini suggested that the Commission would be open to purchasing analytics products from the commercial sector to support Copernicus activities.

This also coincided with increased awareness of land characterisation from the commercial sector with the President of Product and Business at Planet expressing an interest in the EAGLE data model and how these types of developments could be supported (pers. comm.).

[4] https://ec.europa.eu/commission/presscorner/detail/es/speech_22_561

[5] https://www.youtube.com/watch?v=iWtSfxIzEG0

6.2 Analytics layers

The **concept of analytics layers is analogous to intermediate products,** which are commonplace within many manufacturing and production-oriented systems where a process can be broken down into a number of common components or sub-assemblies, which can be produced independently and often in an automated way[6].

6.2.1 Background

A good example of an intermediate product is 'flour' within the food industry. The raw material, wheat, is milled into flour by an established process on a large scale with consequential increases in efficiency, consistency, quality and reductions in cost. The flour is then a common ingredient used in a wide range of food products and it passes on the cost savings and quality benefits provided by its efficient central production.

The intermediate product concept grew out of the increasing complexity of end products and the need for production to become more efficient in terms of cost and resources. As businesses became more responsive to unique consumer requirements and the range of products grew to meet these unique configurations, the production of all components from their raw materials became unmanageable. Specialist businesses and production centres developed to provide the individual components or groups of components. It then became critical to understand the breakdown of the final product into its intermediate products and the structure in which they can be assembled most efficiently.

An intermediate product is therefore one that is more likely to require further processing or combination with other products before it is useable or saleable to the ultimate end user. The additional processing might be done by the original producer, another processor or the end user themselves. Thus, an intermediate product might be a final product for one company and an input for another company that will process it further. To allow these approaches to work it is vital that clear specifications and standards are adhered to by all parties along the production workflow.

So, in the context of EO, the schematic in Figure 11 attempts to show how intermediate products or analytics layers fit into the overall EO-based product workflow. Once the EO data has been processed to analysis ready data (ARD), then at each step afterwards there are potentially multiple onward uses. The first step should be a set of standard analytics layers, which can be easily explained to the end user and may already be found as part of their business systems. These layers can be used by multiple applications and each application can use one or more analytics layer along with non-EO ancillary data and domain knowledge. The final applications are therefore being driven by information familiar to the end users, which is consistent across applications and thus builds traceability and trust in the end user domain. In the case of this work the application is the population of the CLC+ Core with information which will support later ingestion rulesets.

[6] https://medium.com/@baggiesgeoff/bespoke-or-off-the-peg-eo-needs-to-adapt-its-offering-afdb6e151823

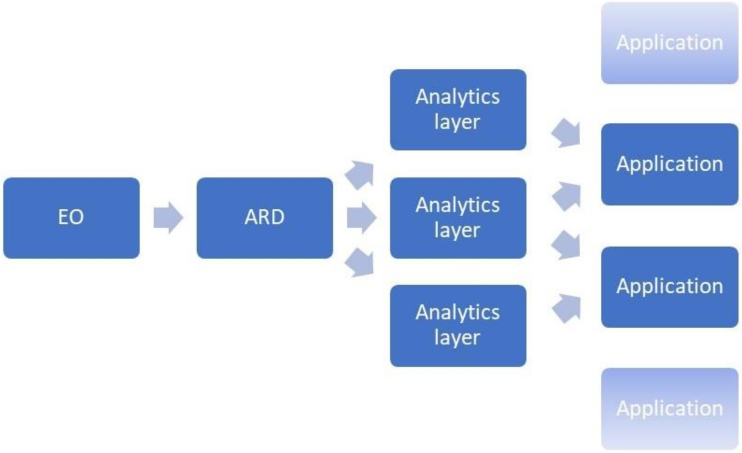


Figure 11. A possible schematic for breaking down the EO workflow into a series of self-contained intermediate products.

So, if we consider the EAGLE data model then there may be an equivalence between the elements of the model and some intermediate products. The origins of each of the intermediate products must be considered in case they are highly correlated, but potentially the deployment / population of the EAGLE data model can be supported by the use of analytics layers. Also, given the relatively simple nature of most of the analytics layers they are likely to be closer to a one-to-one relationship with the EAGLE elements than when ingesting existing land cover / land use datasets.

6.2.2 Potential analytics layers

The following sections examine current examples of intermediate product implementations and production system.

Public systems

Although this sub task was aimed at commercial analytics layers it is also worth considering analytics layer developments in the public sector as they provide background and context.

The most established processes for the development and exploitation of what could be described as analytics layers are at the global level. The NASA Moderate Resolution Imaging Spectrometer (MODIS) programme generates a range of land, atmosphere, ocean and ice products built on the heritage of processing Advanced Very High Resolution Radiometer (AVHRR) data to produce multi-date image and NDVI composites. Similar products were also produced by the European MEdium Resolution Imaging Spectrometer (MERIS) which was on board the Envisat platform before it failed. Over the last decades MODIS (and MERIS) have introduced daily to bi-weekly, super-spectral repeat global coverage of the Earth. The super-spectral nature of these instruments supports the derivation of biophysical measures such as vegetation indices, Leaf Area Index (LAI) and fraction of Absorbed Photosynthetically Active Radiation (fAPAR). The MODIS products are grouped under a number of subheadings such as radiation budget variables, ecosystem variables and land cover characteristics. The ecosystem variables include vegetation indices, LAI, fAPAR, gross and net primary productivity, whilst the land cover characteristics include thermal anomalies, land cover and vegetation continuous fields.

In a similar fashion the global component of the Copernicus Land Monitoring Service (CLMS) derived comparable products from a range of sensors including Sentinel-3. Unfortunately, the medium spatial resolution (300 – 500 m) of the data is too coarse to resolve many of the landscape features of interest when populating a grid database with a cell size of 1 ha. The pan-

European component of CLMS includes a set of intermediate products referred to as the High spatial Resolution Layers (HRLs) which provide information about imperviousness, forests, natural grasslands, wetlands, and permanent water bodies. There are also dynamic monitoring products related to snow and ice and vegetation phenology and productivity. The HRLs are produced from 10 / 20 m spatial resolution optical satellite imagery from Sentinel-2 through a combination of automatic processing and interactive rule-based classification. Pan-European wall-to-wall products will cover the 39 European Economic Area (EEA) countries and are produced from a short time window (+/- 1 year) of image data.

The USGS is now also beginning to generate level-3 products as part of the Landsat Collection 2 processing. These analytics layers will include Dynamic Surface Water Extent, Fractional Snow-Covered Area and Burned Area products. They can be delivered for every Landsat scene and provide information with a 30 m spatial resolution.

Commercial systems

Over the last decade there has been a rapid growth in commercial EO data providers as part of the new space trend. Some of these are pure data suppliers and others began with a potential market which they aimed to support with analytic information. As both these types of companies have evolved, they have started to add analytics layers to their product portfolios. Many have the potential to produce bespoke analytics layers for particular customers, but few are routinely producing them.

The most prominent in relation to operational analytics layer development is the US company **Planet**. It operates a fleet of satellites offering very high spatial resolution optical imagery. The main image product is derived from their Super Dove satellite constellation and is marketed as PlanetScope. It provides daily acquisitions globally with a spatial resolution of 3 m for 8 spectral bands in the visible and NIR region. Internally and through the acquisition of other companies Planet have begun to develop a series of analytics layers produced from their own image data and external sources that are referred to as **Planetary Variables and Analytics Feeds**. The Planetary Variables are related to biophysical properties of the surface and are provided at a range of spatial and temporal resolutions. They include soil water content (100 m), biomass proxy (10 m), land surface temperature (100 m), above ground forest carbon (< 5 m), woody vegetation canopy height and canopy cover (< 5 m). The Analytics Feeds are based on Planetscope data so are highly detailed. They include monthly or weekly buildings and infrastructure detection, monthly or weekly overview of new and existing roads, and automated change detection for early identification of the construction of roads and buildings. Planet is now also looking at the development of a land cover / land use analytics product by collaboration with **Impact Observatory** (see below).

The commercial SAR operators are also developing analytics products but they are more focused around change detection and feature identification given the capabilities of SAR systems. **Iceye** and **Capella** have change detection products based on high temporal frequency repeat acquisitions. Capella is also offering vessel detection which is appropriate given the contrast in microwave response between water surfaces and chips etc.

Other options

Although not strictly analytics layers, there are a number of regularly produced land cover / land use products at the global level which could potentially have some value compared to CLC+ Backbone.

Inspired by the 2017 WorldCover conference the European Space Agency (ESA) initiated the **WorldCover** project. In October 2021 a freely accessible global land cover product at 10 m spatial resolution for 2020 was released based on both Sentinel-1 and Sentinel-2 data. It contains 11 land cover classes ("Tree cover", "Shrubland", "Grassland", "Cropland", "Built-up", "Bare / sparse vegetation", "Snow and Ice", "Permanent water bodies", "Herbaceous Wetland", "Mangrove" and "Moss and lichen") and was independently validated with a global overall accuracy of about 75%. In 2022 a new version of the product with even higher quality, WorldCover 2021, was released with a global overall accuracy of 76.7%.

In June 2021 Esri released a new high spatial resolution global land cover map as part of the company's **Living Atlas**. The map was built on the Copernicus Sentinel-2 satellite image archive and produced using a machine learning workflow from Impact Observatory (IO) supported by Microsoft. The product is described as a 10 m spatial resolution raster product which records 10 classes for the 2020 reference year. Since then, IO have continued to develop innovative AI-powered methods for automated Land Use Land Cover mapping and monitoring in near-real-time. IO Monitor uses a deep learning approach to classify 14 land use and land cover categories (including clouds) globally using Copernicus Sentinel-2 imagery. Custom land use and land cover change maps are available for any area of interest, over user-specified time periods, from 2018 to the present (refreshed daily).

To help turn satellite imagery into more useful information for quantifying change, Google worked with the World Resources Institute (WRI) to create Dynamic World. Powered by Google Earth Engine and an AI Platform, Dynamic World provides global, near real-time land cover data at a 10 m spatial resolution for 9 classes.

Although, the ESRI and Google products appear to have very impressive specifications and are visually appealing at small scales, their actual representation of the landscape is not what would be expected and are not comparable with the more conventional CLC+ BackBone and WorldCover datasets (Figure 12). Given the availability of CLC+ BackBone to CLC+ Core, these global datasets will be less useful although should be kept under consideration as they continue to develop their capabilities and in particularly are trying to deliver dynamic information on landscape processes.

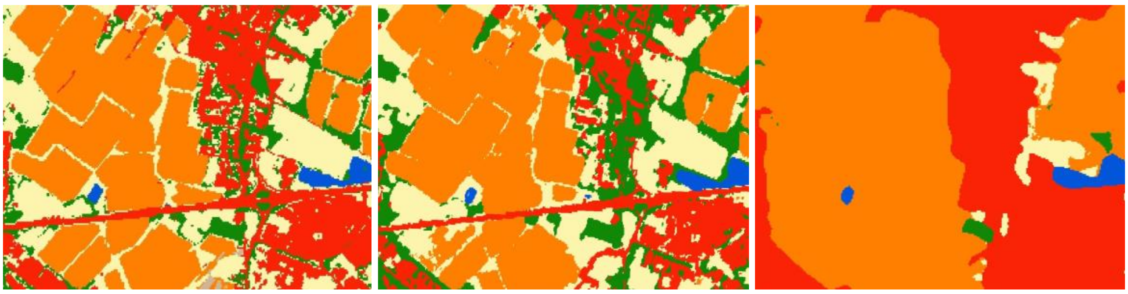


Figure 12. A comparison of CLC+ BackBone (left), ESA WorldCover (middle) and ESRI Living Atlas (right) using a common legend.

6.2.3 Selecting viable options

In principle any analytics layer could be integrated into CLC+ Core, but in practice, there are a number of criteria which self-select a smaller subset of options:

- Spatial resolution – When working with a 1 ha grid in CLC+ Core it is important to allow each cell to be characterised by a representative number of pixels thus a maximum spatial resolution of around 30 m would be acceptable. This is at the top end of the spatial resolutions used in the CLMS pan-European products (early version of CLC) and what might be considered appropriate for representing the European landscape.
- Temporal resolution – Surface properties can change gradually (e.g., tree growth) or episodically (e.g., clear felling, fires or landslides) whereas others changes are related to dynamic and / or cyclical processes (e.g., hay cutting and grazing). To

reliably capture these changes and to improve the confidence in the reported properties it will be necessary to have high-cadence acquisitions. In the case of habitat monitoring in Europe the frequency of acquisitions should ensure that representative seasonal profiles can be produced, thus the revisit frequency needs to be weekly or better.

- Time series extent – Also, for detecting seasonal changes a multi-year time series is required to disentangle variations in behaviour due to the impact of climatic variations and management practices.
- Spatial extent – To support the CLC+ Core and the CLC+ Instances it is most appropriate for the analytics layers to be available for the same spatial extent as the CLC+ product suite, i.e., EEA38 + UK.

Given the above factors the most obvious sources for analytic layers currently available are some of the individual Planet Planetary Variables and Analytics Feeds. The Iceye and Capella offerings and the Planet land cover products are less relevant to the EAGLE data model and they are currently produced only on demand.

6.3 Cross reference with EAGLE

The aim of this work was to identify which EAGLE elements within CLC+ Core can be addressed by the commercial analytics layers. As by default they are basic surface properties there is no need for full *bar coding of each layer, but only to identifying the one of more elements that they could populate within CLC+ Core and their relevant barcode values*. Also, there are only quite limited definitions and lists of inclusions and exclusions.

Analytics Feed – Building and infrastructure footprints (10 m)

• LCC-1_1_1_1 Buildings	3
• LCC-1_1_1_2 Specific Structures and Facilities	3
• LCH-1_2 Built-Up Pattern	1

Analytics Feed – Roads (10 m)

• LCC-1_1_1_3 Open Sealed Surfaces	5
• LUA-4_1_1 Road Network	5
• LCH-1_8_1 Road Network Type	5

Planetary Variable - Woody vegetation canopy cover (5 m)

• LCC-2_1 Woody Vegetation	5
• LCH-3_13 Crown Cover Density	5

Planetary Variable - Woody vegetation canopy height (5 m)

• LCC-2_1 Woody Vegetation	5
• LCH-9_1_4 Object Height	5

Planetary Variable - Biomass proxy (10 m)

• LCC-2_2 Herbaceous Vegetation	5
• LUA-1_1 Agriculture	5
• LCH-5_1_1_1 Cropland	5

6.4 CLC+ Core Ingestion / Integration

The above lists of EAGLE elements show that these analytics layers could be relatively easily ingested into the CLC+ Core given the current version of the EAGLE data model and the platform setup. In most cases only a single element from each of three element groups were required.

The analytics layers related to change were not considered in this sub task as there are currently issues around ingesting change layers into the CLC+ Core and various approaches are suggested with in Task 4 of this contract.

6.5 Use cases

The inclusion of the analytics layers listed above in CLC+ Core will offer a number of opportunities. In a conventional sense the elements they populate can be used in extraction rules, but their native fine spatial and temporal resolutions offer other possibilities for spatial patterns and temporal behaviour to be reported.

6.5.1 Green roof

A common example used by the EAGLE group when explaining the need for land characterisation rather than classification is the situation and challenges posed by green roofs (Figure 13). A green roof or living roof is a roof of a building that is partially or completely covered with vegetation and a growing medium, planted over a waterproofing membrane. So, when mapped by remote sensing these artificial structures have the appearance of vegetation and are often miss classified as urban green space or even fields due to them commonly having a geometric shape.

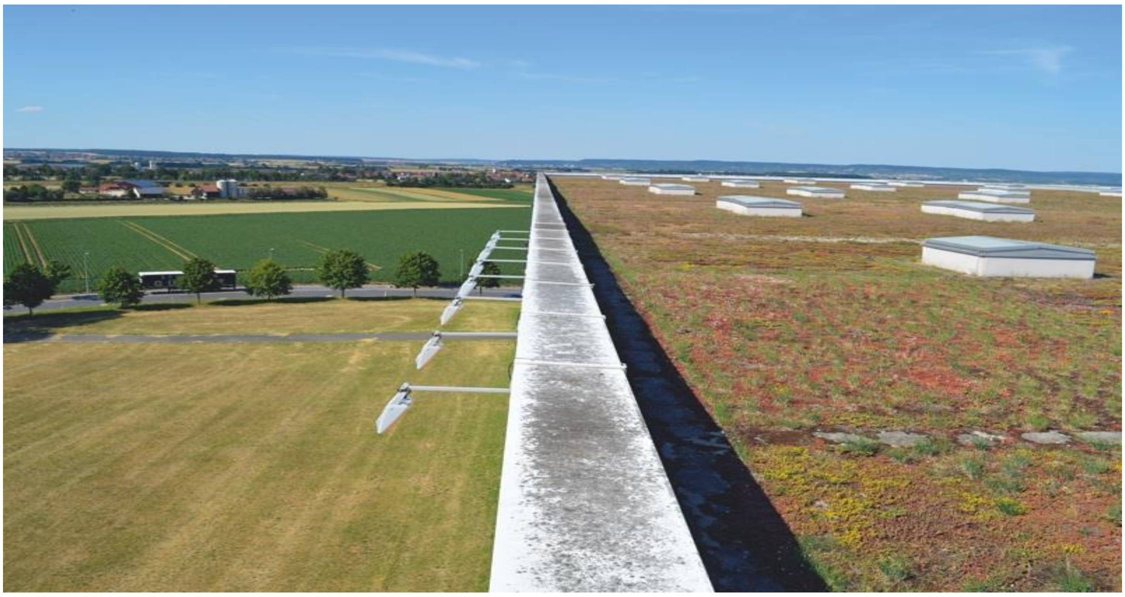


Figure 13. An example of green roof on a large industrial building.

In the above example, this industrial building would be difficult to distinguish from the adjacent grassland, especially when using satellite imagery with a spatial resolution of 10 m or greater. However, by combing a layer for herbaceous vegetation (LCC-2_2) from CLC+ BackBone with the building footprints (LCC-1_1_1_1 and LCC-1_1_1_2) from the Planet Planetary Variable it should be possible to identify the locations of building with green roofs. Of course, in CLC+ Core the buildings and / or herbaceous vegetation would need to cover a significant proportion of the grid cell to get a reliable result.

6.5.2 Elaborating on built up pattern

Within the heading LCH-1_2 Built-Up Pattern, the EAGLE data model contains five separate categories:

• LCH-1_2_1	Scattered Single Houses, Discontinuous
• LCH-1_2_2	Single Blocks, Discontinuous
• LCH-1_2_3	Suburban Row Houses, Terraced, Semi-Detached Houses
• LCH-1_2_4	City Street Blocks, Closed Front
• LCH-1_2_5	Large Complex Buildings, Big Halls

Although the building footprint layer does not contain this level 3information directly, it could be derived from the 10 m spatial resolution analytics layers summarised within the 1 ha cells of CLC+ Core. In Figure 14 there are a number of built-up patterns that can be identified given the appropriate geospatial analysis. Also, in this example the roads analytic layer can be used to support a regionalisation process.

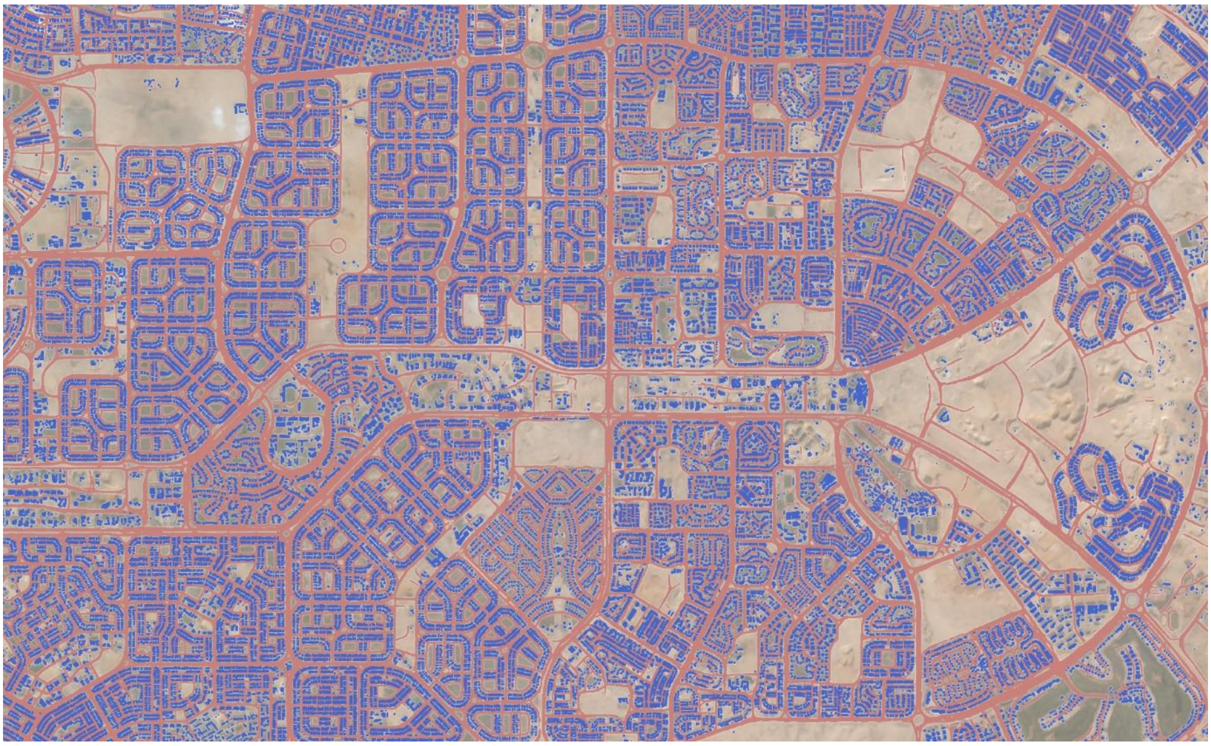


Figure 14. Examples of different building patterns in the Building Footprint Analytics Feed from Planet.

6.5.3 Data gaps in CLC+ LULUCF and Legacy

A clear driver for considering the commercial analytics layers as input to the CLC+ Core was the need to plug gaps in the production of CLC+ Instances.

Given the availability of other datasets the roles in supporting the CLC+ Instances may be quite limited. However, the increased spatial and temporal resolutions of their native specifications may allow to more clearly identify the changes subtle within a 1 ha grid cell. For instance, changes in configuration in the development of brownfield sites and the detection of management practices in forestry and agricultural areas.

6.6 Conclusions

It is clear that some of the commercial analytics layers described above could provide input to the useful information for extractions if ingested in CLC+ Core.

Those described in detail in section 6.3 should be considered the most viable at the moment, but analytics layers are an area of considerable a rapid development at the moment, so it is likely that the list will grow significantly in the coming years. Those in charge of the maintenance of CLC+ Core and the reviewers of user feedback should maintain a watching brief to see when new layers come available and what they might offer.

At the present time, support that they can offer to the CLC+ Instances might appear limited, but as they continue to evolve it is likely that they may be able to address a number of the more specific elements of the EAGLE data model, particularly in the Land Characteristics (LCH) section. The building footprint example shows how some additional processing or more advanced ingestion rules of an analytics layer could potentially address elements which have so far been difficult to populate at the European scale or without complex harmonisation of diverse MS data.

As a more general move in EO, more general users are starting to balk at the prospect of having to select and process images from the ever-increasing data stream, or drink from a fire hydrant as it has been described. In such a world intermediate products / analytics layers are only going to become more popular / practical as can be seen by the considerable interest surrounding the High Resolution Vegetation Phenology and Productivity (HR-VPP) product. The EEA should keep this in mind when developing the CLMS product suite and suggesting layers for inclusion in CLC+ Core.