UNIVERSITAT POLITÈCNICA
DE CATALUNYA
BARCELONATECH

UNIVERSITAT DE
BARCELONA

UNIVERSITAT
ROVIRA i VIRGILI

Master in Artificial Intelligence

Characterization of wastewater methane emission sources with computer vision and remote sensing

Master's thesis by:
Miguel Castrillo Melguizo
Barcelona Supercomputing Center

Thesis advisor:
Carlos Gómez González
Barcelona Supercomputing Center

Internal Examiner:
Sergio Escalera
Dept. of Mathematics and Informatics, UB

BARCELONA SCHOOL OF INFORMATICS (FIB) - FACULTY OF MATHEMATICS AND INFORMATICS (UB) - SCHOOL OF ENGINEERING (URV)

UNIVERSITAT POLITÈCNICA DE CATALUNYA (UPC) - BARCELONATECH - UNIVERSITAT DE BARCELONA (UB) - UNIVERSITAT ROVIRA I VIRGILI (URV)

Barcelona, 14 October 2022

Summary

Global warming is one of the most critical challenges that modern society faces and will face in the years ahead and the contribution of human activities to this climate crisis has become increasingly evident. Methane emissions are responsible for at least one-third of the total anthropogenic climate forcing and current estimations expect a significant increase in these emissions in the next decade. Consequently, methane offers a unique opportunity to mitigate climate change while addressing energy supply problems. From the five primary methane sources, residual water treatment provided 7% of the emissions in 2010. This ratio will undoubtedly increase with global population growth.

For this reason, locating sources of methane emissions is a crucial step in characterizing the current distribution of GHG better. Nevertheless, there is a lack of comprehensive global and uniform databases to bind those emissions to concrete sources in the ground and there is no automatic method to accurately locate sparse human infrastructures such as wastewater treatment plants (WWTPs). WWTP detection is a state-of-the-art open problem posing many obstacles due to the lack of freely accessible high-resolution imagery, the variety of morphologies and sizes and their usual proximity to other human infrastructures. Additionally, most plants have small infrastructures whose salient features fade out in medium and small-resolution images.

In this work, we tackle this state-of-the-art complex problem and go one step forward by trying to infer capacity using one end-to-end Deep Learning architecture and multi-modal remote sensing data. This goal has a groundbreaking potential impact, as the obtained data could help estimate mapped methane emissions with application in improving current emission inventories, predicting future scenarios, and designing mitigation and adaptation policies. Moreover, it has obvious applicability in estimating water supply conditions.

We will address the problem as a combination of two parallel inference exercises by proposing a novel network to combine multimodal data based on the hypothesis that the location and the capacity can be inferred based on characteristics such as the plant situation, size, morphology, and proximity to water bodies or population centers. We explore technical documentation and literature to develop these hypotheses and validate their soundness with data analysis.

The architecture that we propose is composed of two well-differentiated branches. On the one hand, there is a fully convolutional network (FCN) in charge of identifying and delimiting WWTPs via pixel-wise semantic segmentation, outputting the wastewater infrastructures' location, shape, and area. We use land reflectance information (multispectral imagery) and other input sources that may help the network make decisions. Following examples in the literature, we introduce topographic and land-use (water bodies in

particular) information. On the other hand, a second branch made of fully connected (FC) layers performs the second part of the characterization exercise by regressing the WWTPs' capacities. We validate the capability of the model to learn to predict this capacity from the visual and contextual features of the WWTPs. Visual features include aspects such as the inferred plants' size and morphology, while context features can be the closeness to human settlements or the magnitude of this population.

To validate the architecture and the hypotheses, we develop a model and a dataset in parallel with a series of ablation tests. The process is facilitated by an automatic pipeline, also developed in this work, to create datasets using various sources and train and validate our models leveraging those datasets. This allows for evaluating many possibilities, including data, architecture, and hyperparameter variations. However, although we take or discard some of these hypotheses as we develop the neural architecture and dataset, the final validation involves transferring the model to a similar domain. We test the best-obtained model at scale on a mosaic composed of satellite imagery covering the region of Catalonia. The goal is to find plants not previously labeled but present in wastewater treatment plant (WWTP) databases and to compare the distribution and magnitude of the inferred capacity with the ground truth.

Results show that we can achieve state-of-the-art results by locating more than half of the labeled plants with the same precision ratio and by only using orthophotos from multispectral imagery. Moreover, we demonstrate that additional data sources related to water basins and population are valuable resources that the model can exploit to infer WWTP capacity, although their contribution to improving the segmentation of the plants is more discrete. To be more specific, the total predicted capacity of models trained with water bodies or population data after the domain transfer exercise is in the same order of magnitude that the ground truth. During the process, we also demonstrate the benefit of using negative instances to train our model and the impact of using an appropriate loss function such as Dice’s loss.

These promising results and conclusions pave the way to different research lines to improve the model skill. Additionally, two independent works published during the development of this study, not only prove that WWTP detection is an open issue but also provide additional data and ideas that we can exploit to progress in this investigation.

1. Content table

2. Introduction

Global warming is one of the most critical challenges that modern society faces and will face in the years ahead. The associated climate change is expected to cause between 8 and 80 million human excess deaths this century (Bressler, 2021), depending on the policies adopted and their impact. Moreover, global warming, one of the major drivers of massive vertebrate extinction, is expected to cause 11 to 58% of the projected extinctions of vertebrate species by 2,100 (Pereira et al., 2010). On the other hand, the contribution of human activities to climate change has become increasingly evident. There is scientific consensus that human influence has, without doubt, warmed the global climate system since pre-industrial times (Eyring et al., 2021), with greenhouse gasses (GHG) being the main driver for hot and cold extremes and intensifying precipitation.

Of the human-emitted GHGs, methane is the second in importance (14%) after carbon dioxide. However, it is responsible for at least one-third of the total anthropogenic climate forcing, even if it is emitted in smaller quantities and subsists less in the atmosphere than carbon dioxide. Its significant contribution to global warming is due to its capacity to retain thermal energy, which is 21 times greater than that of carbon dioxide. Current estimates indicate that without significant efforts to reduce methane emissions, these will likely ramp up by 20 percent between 2010 and 2030 (U.S. Environmental Protection Agency, 2012). Consequently, methane offers a unique opportunity to mitigate climate change while addressing energy supply problems, given that we can trap it and use its energy.

In fact, quick action to mitigate methane emissions can produce near-term returns and it is the single most effective strategy to achieve the compromise of limiting global warming to 1.5º (Climate & Clean Air Coalition Secretariat, 2021). With that purpose, the Global Methane Pledge launched at COP 26 achieved the commitment of more than 100 countries to contribute to reducing global methane emissions by at least 30 percent from 2020 levels by 2030. Those emissions originate from five primary sources: agriculture, carbon mines, municipal solid waste, local wastewater, and hydrocarbon systems. Among them, residual water treatment provided 7% of methane emissions in 2010. This ratio will undoubtedly increase with the global population growth, while other activities such as coal or hydrocarbons should become less important in the overall sources of methane emissions.

In residual water treatment, methane appears mainly during the anaerobic decomposition of organic matter. Such processes occur in different moments and ways, usually depending on the governing waste policy and the technology of the plants. In developed countries, treatment is usually carried out in centralized treatment plants. In these plants, aerobic treatment prevails. However, they produce biosolids

that are potential methane emission sources. On the other hand, in developing countries, there is a more extensive presence of anaerobic systems, entailing higher methane emissions.

Nevertheless, independently of the procedure, locating these sources of methane emissions, whether they are primary or secondary, is a crucial step in characterizing the current distribution of GHG better. Potential applications range from improving current emission inventories, predicting future scenarios, and designing mitigation and adaptation policies. Furthermore, even though there are top-down approaches to attribute methane emissions from the surface, there is a lack of comprehensive global and uniform databases to bind those emissions to concrete sources in the ground. In this context, and given the increasing availability of satellite products, artificial intelligence methods have the potential to fill this gap (Zhu et al., 2022) by identifying human infrastructures which are potentially emission sources.

This work aims to detect and subsequently characterize wastewater methane emissions sources by leveraging computer vision models fed with multi-modal remote sensing data. We will address the problem as a combination of a segmentation task and a regression one. With that aim in mind, we will create a new neural network architecture based on two branches: a fully convolutional one for segmentation and a dense one for regression. The branches will be connected but will also have separate inputs. Consequently, we will combine various data provided in alternative formats based on the hypothesis that the location or the capacity can be inferred based on characteristics such as the plant situation, size, morphology, and proximity to water bodies or population centers.

Therefore, the novelty of this study lies in developing a new Deep Learning architecture and combining various remote sensing information according to an initial study in which we will establish a series of hypotheses. We will pre-validate some of these hypotheses while developing the neural architecture and the dataset. The final validation will apply our model at scale over a mosaic composed of satellite imagery covering the Catalonia region. We will use it to find plants not previously labeled but present in wastewater treatment plant (WWTP) databases.

Provided the hypotheses are validated, this model will be a stepping stone toward improving current emission inventories and its potential contribution is twofold. On the one hand, it will allow discovering WWTPs in regions lacking detailed databases. On the other hand, it will inform emission models to help with their task of spatially mapping coarse emission budgets.

3. State-of-the-art

Remote sensing images have been demonstrated to be valuable for different tasks such as classification and monitoring. Neural network applications have a long history in this field. However, it was not until the advent of Deep Learning architectures that the community shifted from other machine learning methods, such as support vector machines or random forests (Ley-Ma et al., 2019). Semantic segmentation is the most widely used method for assigning and delimiting land cover areas (including human structures). Some of the most employed architectures in these studies are the Convolutional Neural Networks (CNNs) (LeCun, 1989), particularly those composed of encoder-decoder subnetworks (Ma et al., 2019), such as the U-Net (Ronneberger et al., 2015).

CNNs are a family of Artificial Neural Networks (ANNs) with considerable success in image detection, classification and segmentation tasks. They excel at these tasks due to particular characteristics that make them appropriate for grid-like topologies (Goodfellow et al., 2016). In the first place, CNN architecture contains inductive bias or priors inspired by the cells in animals' visual cortex. These priors work as small-size kernels applying nonlinear filters to detect local patterns. In turn, their pyramidal stack architecture allows them to achieve increasingly global patterns from this spatial locality. Secondly, these filters apply the convolution operation sliding all over the visual field during learning, sharing weights and biases. This operation grants translation equivariance, and at the same time, it converts them into regularized versions of Multilayer Perceptron (MLP). The weight-sharing combined with a downsampling procedure (using *Max Pooling* layers) gives them a degree of translation invariance to small translations of the input (Goodfellow et al., 2016).

Pixel-wise semantic segmentation, or just segmentation, is a structured output task in which CNNs predict a per-pixel classification (a vector of per-class probabilities) rather than, for example, class predictions for the entire input. Therefore, they can categorize all the input grid points, drawing objects and object boundaries to highlight and delimitate the elements represented in an image. The U-net (Ronneberger et al., 2015), which evolved from the idea of Fully Convolutional Neural Networks (FCN) for semantic segmentation (Long, 2014), stands out in this task. The novelty of the FCN was an upscaling subnetwork following the classical contracting area of the network to recover the original resolution, aided by links crossing the subnetworks combining coarser and finer information to improve the detail. The U-Net authors increased the number of filters (feature channels) on the upscaling side to propagate context information, providing an entirely symmetrical. Since its initial application in biomedical image segmentation, its usage has been extended to other domains, such as remote sensing.

Although delayed compared to classification tasks, semantic segmentation applied to remote sensing has recently received increasing attention, experiencing exponential growth in the number of works and applications during the 2010s (Ma et al., 2019). During this time, the most popular applications involved classification tasks, especially for land-use and land-cover (LULC), and mostly using CNN models. However, only a fraction of them addressed the practical use of these models, and many relied on image data with spatial resolutions below two meters. In particular, most frameworks were composed of encoder-decoder networks and one of the main challenges was, and still is, addressing resolution constraints that make classifying small objects difficult. As a result, different strategies have been suggested and adopted, such as using skip-connections (Liu et al., 2017; Chen et al., 2018), atrous convolutions (Sherrah, 2016), multi-scale feature extraction (Marmanis et al., 2016, Wang et al., 2017) or doing some post-analysis to improve the results (Liu et al., 2017).

In the last few years, due to the success of vision transformers (ViTs) in general computer vision tasks, a number of works based on this new paradigm have emerged in the remote sensing domain. Transformers are Deep Learning models based on the concept of self-attention and multi-head attention, which gives them the capability of learning different and complex long-range dependencies or relations for the same data. In remote sensing, most of the approaches usually propose hybrid architectures combining CNNs and transformers (Aleissaee et al., 2022).

The progress in Deep Learning architectures and satellite imagery availability have favored the advent of studies based on artificial intelligence to map human infrastructures, including some focusing on potential methane emission sources. Interestingly, while developing this work, some groups published promising works on Deep Learning from remote sensing data to locate WWTP infrastructures. On the one hand, Li et al. (2022) use two complementary networks to detect candidate boxes (through Single Shot Multibox Detection) containing WWTPs and perform dual-task classification of LULC and WWTP. On the other hand, Zhu et al. (2022) offer a new dataset for methane source infrastructures in North America and provide results for multi-label segmentation tasks with a DenseNet.

Notably, one usual collateral effect of classifying small objects is the imbalance of the class distribution along the image pixels. Some examples have addressed this problem through changes in the loss function, like using a medium frequency balancing (Liu et al., 2017) or novel loss functions. Milletari et al. (2016) addressed the same problem in the medical context by introducing a loss function based on Dice's score. Finally, there are also methodologies established on the idea of taking advantage of supplementary information, like those using multi-spectral bands and multi-modal data. Some have done it in a direct way, such as Lagrange et al. (2015), which combine RGB images with data from a Digital Surface Mode (DSM) and also LiDAR information, or Volpi et al. (2016), which stacked color infrared (CIR) and

normalized digital elevation models. On the other hand, some studies defined complex architectures with different subnetworks to learn features from the different modalities. For example, Sherrah (2016) used this approach to process CIR and DSM images, whose convolutional features were concatenated, while Marcu and Lordeanu (2016) processed images of different scales in different pathways. Finally, Hu et al. (2016) proposed a feature-fusing strategy that merges two independent convolutional streams at the point that they have a similar dimensionality to minimize the imbalance.

Concerning the satellite products used in these studies, the American Landsat and the European Sentinel-2 are two of the most employed because of their temporal and spatial availability and multispectral nature. We will employ Sentinel-2 orthophotos as the main resource too. However, we think that there is a gap in using data from additional modalities. Some works already tackle the issue of using multi-modal information, but to our knowledge, all of them focus on the task of locating human infrastructures, such as WWTPs. As we explain in the following chapters, one novelty that we propose is performing segmentation and inference in a single shot with a new proposed architecture using data from different modalities. Furthermore, we create an automated pipeline to generate datasets from different information using WTTP labels and evaluate them automatically.

4. Methodology

This work aims to detect and characterize wastewater treatment plants using remote sensing data. With characterization, we refer to the location and delimitation of the WWTPs, together with the determination of their capacities. Some very recent works in the literature (Li et al., 2022; Zhu et al., 2022) address the issue of locating the WWTPs. We tackle this novel and complex problem and go one step forward by inferring capacity in an end-to-end learning fashion. The goal has a great potential impact, as the obtained data could help estimate methane emissions mapped on the territory.

Our characterization problem is composed of two parallel inference exercises. On the one hand, we detect and locate WWTP structures over the territory. We address this problem as semantic segmentation, outputting the wastewater infrastructures' situation, shape, and area. We use land reflectance information (multispectral imagery) and other input sources that may help the network make decisions. Following examples in the literature, we introduce topographic and land-use (and water bodies in particular) information. In turn, we also infer the capacity of the identified WWTPs. We expect the model to learn to predict this capacity from the visual and contextual features of the WWTPs. Visual features include aspects such as the inferred plants' size and morphology, while contextual features can be the closeness to human settlements or the magnitude of this population. Moreover, our proposed architecture will perform these two procedures automatically and concurrently. The network will read data from different modalities and output pixel-level semantic classification categories and the associated numerical capacity. We hypothesize that the area and shape of the WWTP class resulting from the semantic classification will inform the model to perform the capacity regression, aided with additional input data.

In the next two subsections, we explain our foundations for the hypothesis upon which we build the architecture. For that purpose, we get information from technical documentation and literature and validate it with data analysis. We usually restrict the analysis to the WWTP subset that we will use to develop the dataset and subsequently as ground truth to train and evaluate the model. This subset corresponds to 1,000 of the 2,344 plants located in Spain according to the European Urban Waste Water Treatment Directive[1] (UWWTD) dataset (Directorate-General for Environment & European Environment Agency, 2020). This subset underwent a labeling and validation process, including the correction of the georeferentiation of many of the plants.

[1] Contains information about WWTPs reported by the European Member States, including capacity, that is given in population equivalent (p.e.) units.

4.1. First hypothesis: WWTP detection from visual and contextual information

Deep Learning-based pixel-wise segmentation from medium-resolution satellite images is challenging due to the lack of fine structures in the data (Ma et al., 2019). Moreover, the problem of visually identifying WWTPs is especially complex due to the diversity of scales and morphology available in these infrastructures. Different water treatment processes involve different methods, implying changes in morphology and visual appearance. However, there are some commonalities, and wastewater treatment usually involves a sequence of steps, some of which are almost ubiquitous.

In these plants, the process usually starts with separating the most voluminous solids by employing physical barriers. Afterwards, suspended solids are separated into primary settlers. This step is called primary treatment and usually opens the way to the subsequent action, which is the degradation of the organic matter: the secondary treatment. This step is one of a plant's key signatures and can vary broadly among different facilities[2]. Active sludge, stabilization ponds, and trickling filters are some of the most common. However, concerning our objective of identifying potential methane sources, anaerobic treatment of water and sludge is the main source of methane, and sludge appears in all the mentioned stages. The treatment of this sludge will differ between WWTPs and ultimately impact the exact point of the emissions. While small facilities can directly use this sludge for a secondary activity or transfer it to a larger plant, bigger plants can manage it internally if they have the appropriate equipment or transfer it to a waste deposit. In any case, identifying at least these bigger plants is vital to track possible sources of primary or secondary emissions.

As mentioned, the WWTPs' morphology varies with their associated equipment, which is related to the kind of treatments they perform. However, most of them stand out due to their circular-shaped primary settlers. The secondary treatment determines if the WWTP will also feature tanks with a rectangular base, as is the case of the active sludge, or other varied typologies associated with the different biofilm-based treatments, such as biodiscs or tricking filters. From those, trickling filters are also circular infrastructures that can be domed or not. Open ones can be mistaken for primary settlers if not for the multiple pipes that cross their surface. Most of the WWTPs also features secondary settlers, either circular or rectangular.

Thus, circular settlers and trickling filters are hallmarks for most WWTPs, especially if they are accompanied by rectangular (often with rounded corners) tanks. Hence, we hypothesize that we can train a model with remote sensing images, even if it will probably be biased towards these circular and rectangular shapes. A review of the one thousand labeled plants reveals that only 14% do not present open circular structures. Of the rest, many have domed circular structures, most probably corresponding to

[2] Some plants have a tertiary treatment to remove any persistent pollutants from water.

trickling filters. Only a tiny fraction is almost indistinguishable because they do not have primary settlers, are indoor facilities, or are under construction.

Figure 1. WWTPs present broad differences in their visual appearance. However, in orthophotography, they usually stand out by their circular and rectangular-shaped infrastructures. These and Figure A1 images correspond to very high-resolution crops from the Spanish National Geography Institute (IGN).

As a second prior, we also will introduce topographic information obtained from a DEM (Figure B2). Reasoning and observation tell us that WWTPs are usually located in valleys following the river courses, and the network could learn this information. The third assumption we establish that can help locate WWTPs is their proximity to water flows, lakes, or seas. An analysis of the labeled WWTPS shows (Figure 2) that while one-third of them are 900m close to water bodies, the ratio jumps to 0.45 when we increase the radius to 1.5 km. Only a small fraction of these plants are close to the sea.

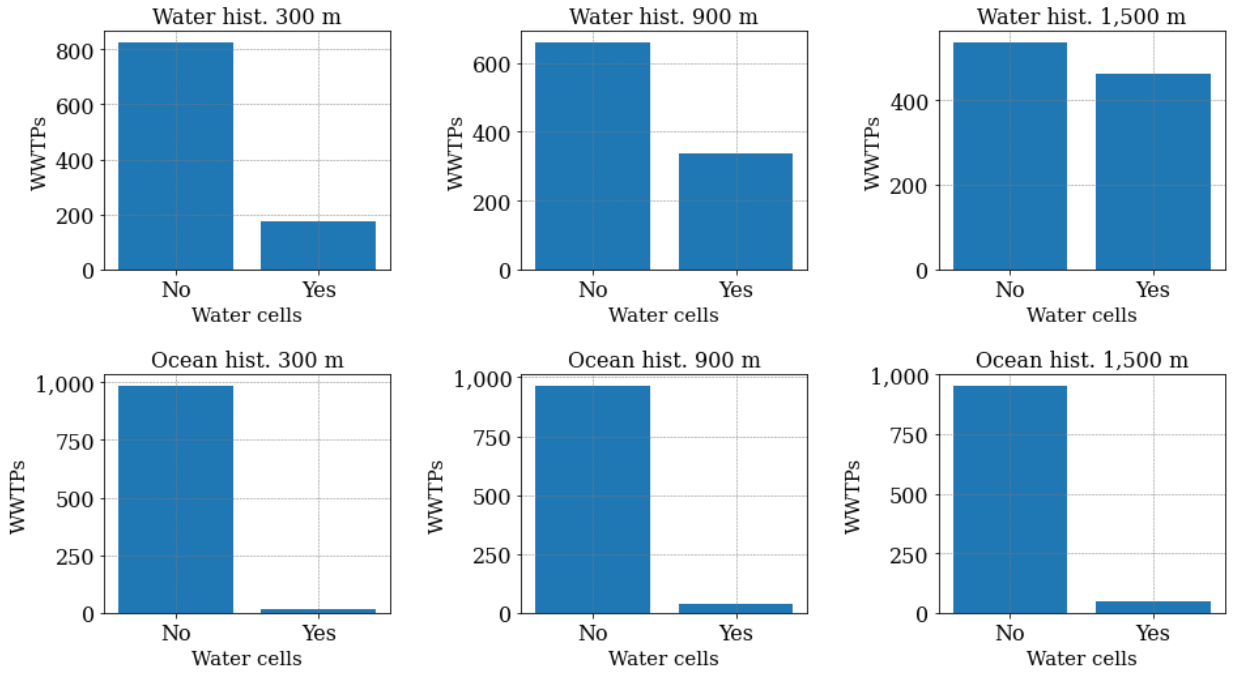

Figure 2. Histograms of the presence of water bodies in the surroundings of the labeled plants for different cutoff distances. Top: inland water. Bottom: ocean water.

4.2. Second hypothesis: WWTP characterization from visual and contextual information

The task of regressing the WWTP capacity can be as hard as their detection. Just as heterogeneous as size or morphology, the associated capacity can vary widely between plants. For this reason, we ensured that our set of labeled plants was representative of the one at the regional level. We verified this assumption by comparing the capacity distribution of the labeled plants with those of the set of plants in the Iberian Peninsula and Europe. Despite the broad variations between plants, Figure 3 shows similar distributions, with most capacities ranging between a few thousand and low tens of thousands.

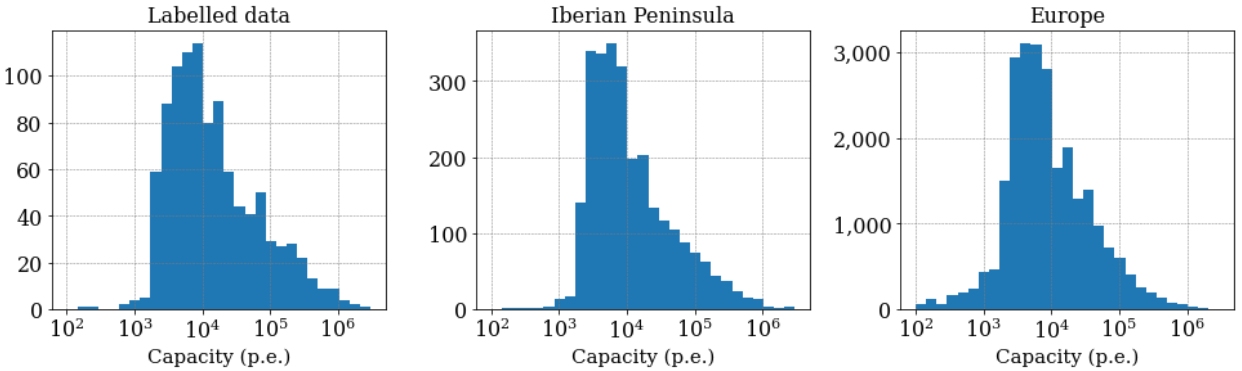

Figure 3. Distribution of WWTP capacities. Left: labeled data. Center: Iberian Peninsula. Right: Europe.

When trying to infer the capacity associated with WWTPs, it is reasonable to assume that higher capacities need more extensive facilities to deploy all the required infrastructure. In turn, more populated areas require a higher water treatment capacity. To illustrate these assumptions, Figure 4 shows the correlation between the size and capacity of the plants in the labeled set, which is higher for larger magnitudes of both variables. We infer that smaller and medium-sized facilities can be more heterogeneous and thus use their area with unequal efficiencies. Nevertheless, in our case, larger sizes and hence the reliability of the prediction aligns with our interest in detecting, with higher priority, those plants contributing more to the global capacity.

Meanwhile, Figure 5 shows the relation between capacity and surrounding population for the labeled WWTPs. In the same sense, correlation is generally high but decreases in higher distances for smaller capacities and less inhabited areas. There can be multiple reasons for this: smaller plants can serve small populations connected to larger cities, and great differences between permanent residents and floating populations (smaller towns and villages receiving occasional residents) can bias the evaluation. However, we can conclude that higher-capacity plants are usually close to cities with higher population densities.

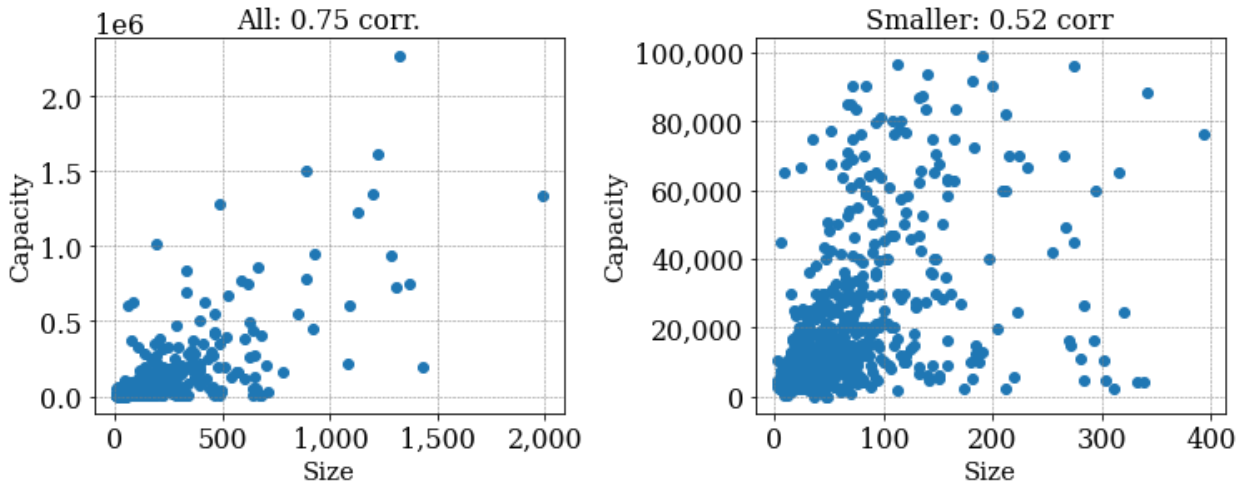

Figure 4. Relation between WWTP area (in label pixels; 1 pixel = 100m²) and capacity. Left: All labeled plants. Right: labeled plants with labels smaller than 400 pixels and capacity lower than 100,000 p.e.

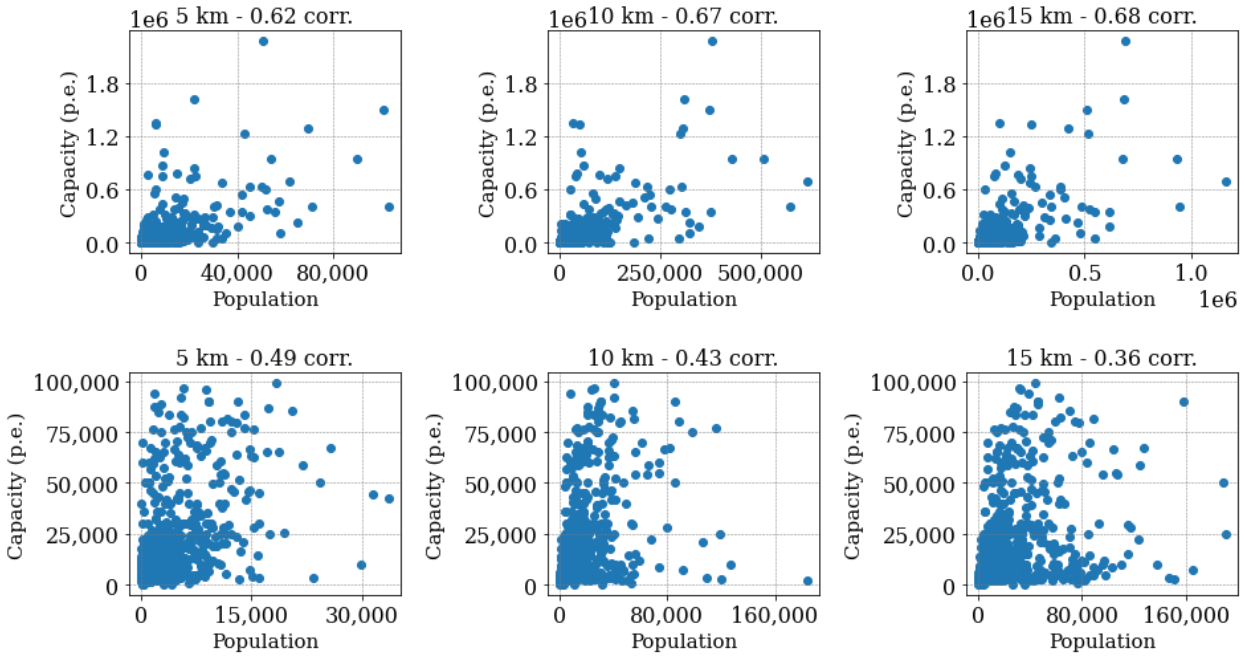

Figure 5. Relation between surrounding population and WWTP capacity for different cutoff distances. Top: all labeled plants and populations. Bottom: labeled WWTPs with less than 100,000 p.e. capacity and close to less than 200,000 inhabitants

Additionally, we think our suggestion to use water bodies data to locate plants could also help to improve capacity prediction. Broader inland water flows may need a capacity paired with the volume of liquid. In the same way, the treatment capacity of a WWTP should not overload the water flow dilution rate (related to the flow volume), especially in the presence of persistent contaminants (Ehalt Macedo et al., 2021).

In light of these facts, we will evaluate the impact of the described factors to locate and characterize WWTPs using different Deep Learning models. In particular, we will try different approaches based on providing the supplementary data in different modalities and adapting the network architecture to introduce this information in the best way possible.

4.3. WWTP capacity as a proxy of the estimated methane emissions magnitude

To test our assumption that capacity can be a good proxy for inferring the associated emissions, we looked up the emissions reported in the European Pollutant Release and Transfer Register (E-PRTR). In Spain, the report contains yearly emissions for plants situated in areas exceeding 100,000 inhabitants. Since WWTP codes differ from the ones in the UWWTD database, we related both databases by crossing the plants' coordinates. We calculated correlations between capacity and reported yearly emissions for the matching plants. Figure 6 shows that emissions correlate with the associated capacity for plants reporting a positive number of emissions. We understand that plants reporting no emissions do not treat their sludge in place or have advanced procedures for processing the methane they generate.

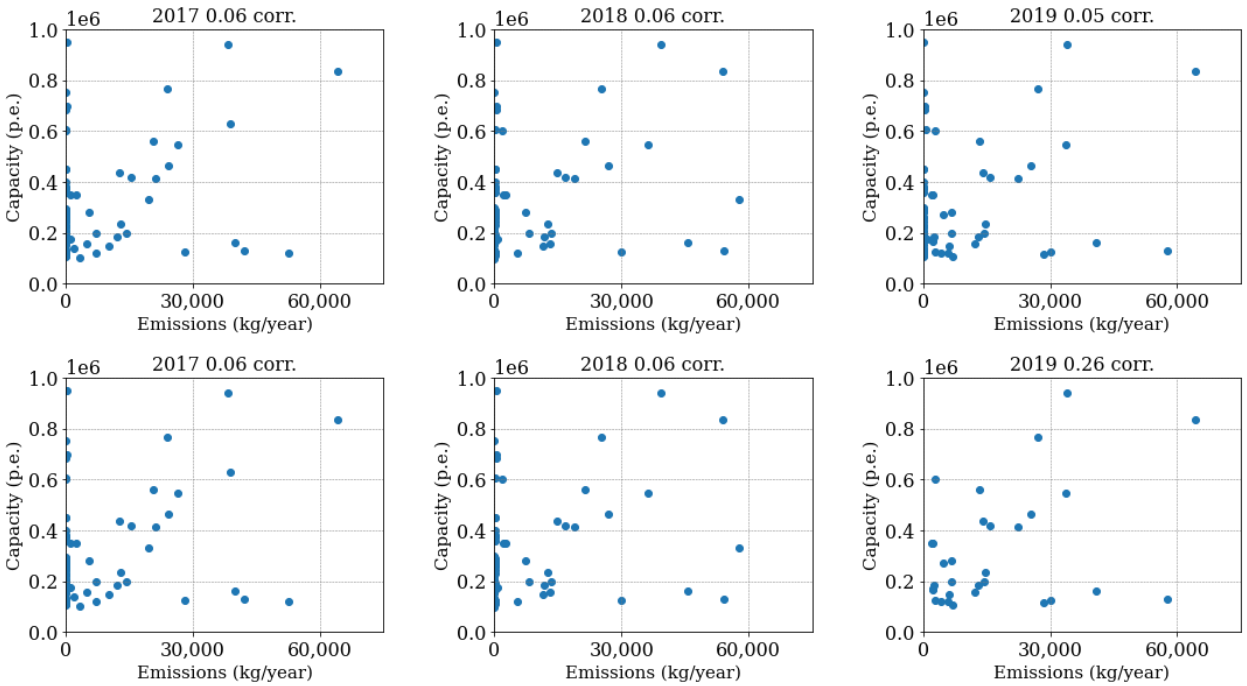

Figure 6: Correlation between WWTP capacity and yearly reported emissions. Top: all labeled plants in the E-PRTR. Bottom: only those labeled plants emitting more than 2,000 kg and less than 100,000 kg yearly.

4.4. Proposed architecture

In our study, we build upon the well-known U-Net architecture, a state-of-the-art neural network for semantic pixel segmentation. In this case, this FCN reads satellite raster data and outputs pixel-level classification labels. Subsequently, we extend (Figure 7) the fully-convolutional architecture by adding a dense regression branch which, in turn, has two input sub-branches. One input reads from the U-Net encoder (bottleneck) output, while another processes vector data from different input sources. These two sub-branches are composed of two dense layers of 64 and 32 neurons. The final outputs of the sub-branches are predictions of the WWTP capacities as a combination of their individual outputs.

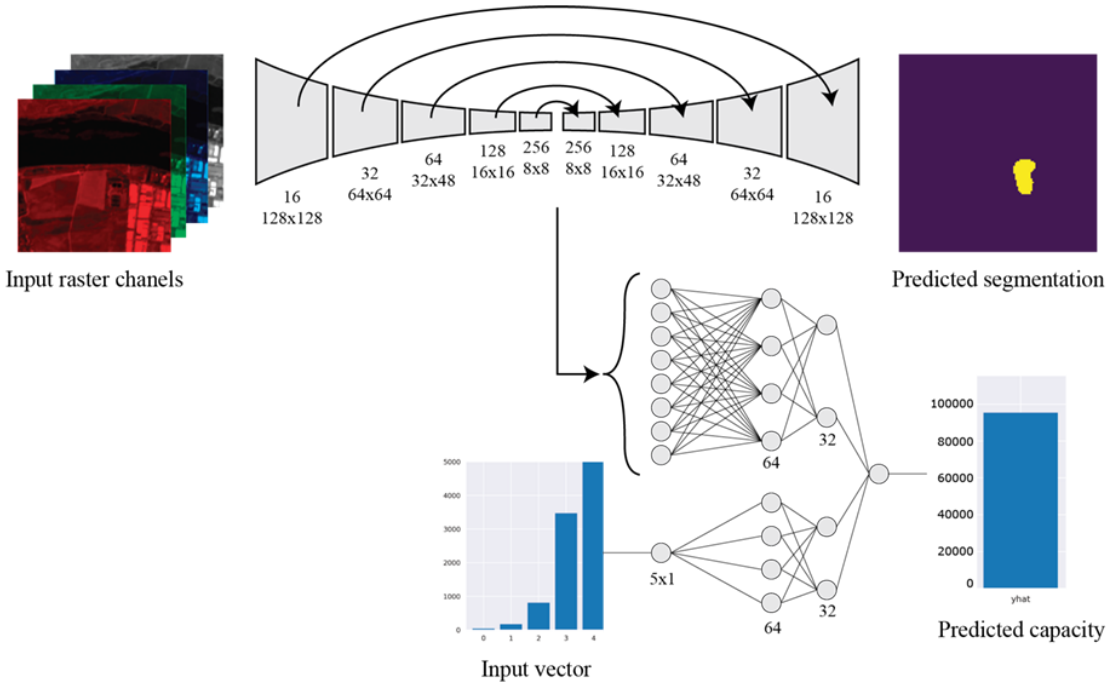

Figure 7: Proposed segmentation and regression architecture. The FCN branch (top) reads input raster channels to infer a pixel-wise segmentation. The dense branch (bottom) takes the output of the FCN encoder and vector data with separate sub-branches. Their output is concatenated to regress the capacity.

4.5. Dataset and workflow

In this work, we create a dataset from freely accessible remote-sensing products to train a model that validates our hypotheses. The same data will be extracted and post-processed in alternate ways to produce raster channels or numeric vectors that will feed one or both model branches. As a result, the extraction process is also an important step of this work as we optimize the dataset and not only the architecture.

This process will be intimately related to the nature of the data used to create the dataset. For this reason, we will detail the different sources employed in the study.

Sentinel-2 is a constellation of Earth-orbiting satellites operated by the European Space Agency (ESA) in the framework of the Copernicus Program. Its primary purpose is to monitor agricultural areas, vegetation, lakes and coastal zones. Sentinel-2 comprises two satellites moving in the same orbit and phased by 180º. Together, they revisit the same point with a 5-day frequency. They equip multispectral cameras providing imagery in 13 different bands covering 290km with up to 10m spatial resolution. Blue, Green, Red and Near-Infrared (NIR) are the bands with greater detail. Sentinel-2 imagery is served by different products distributed in the Copernicus Open Access Hub (European Space Agency - ESA, 2022). Level-2A is the product corresponding to data at the bottom of the atmosphere reflectance and the one used for the project. The images we extract are stacked crops of much larger Earth surface reflectance data granules (tiles) served in different bands in 16-bit depth GeoTIFF format. We only will use the bands with greater detail.

Contrariously, water bodies' information is not extracted from one but from alternative satellite products whose contribution will be compared in the study. The Terra Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) Global Water Bodies Database (ASTWBD) (NASA/METI/AIST/ Japan Spacesystems and U.S. / Japan ASTER Science Team, 2019) and the Global Forest Cover Change (GFCC) Water Cover 2000 Global dataset (Towshend, 2016) provide surface-water information with a similar resolution: 1-arc-second for ASTERWD (ASTER from now onwards) and 30-meter resolution for GFCC (GFCC30 from now onwards). However, ASTER is given in World Geodetic System 1984 (WGS84[3]) grid, while GFCC30 is in UTM/WGS84 projection. Cell values for ASTER are 8-bit integers, with the zero value representing land and the first three natural numbers used to classify water areas in oceans, rivers and lakes. Conversely, GFCC30 has only one value for water, reserving the rest of the values for land, snow/ice, cloud and no data. Although they have a similar resolution, a quick visual inspection is enough to observe that GFCC30 provides greater detail, especially for narrow rivers and streams that do not appear in ASTER.

Finally, we also leverage a novel synthetic near real-time (NRT) land use land cover (LULC) referred to as Google Dynamic World (Brown et al., 2022). Dynamic World is also the name of the Deep Learning model generating this dataset from all existing and new Sentinel-2 10-meter imagery. Accordingly, data is available in the same grid (UTM/WGS84) and provides the same detail as the Sentinel-2 data tiles (Figure B1). In other words, theoretically, no interpolation or reprojection is needed to couple both data sources.

[3] WGS84 is a geographic coordinate system based on a spheroid and measured in angular units (degrees), while UTM is a projection in a 2D plane whose units are meters.

Finally, in opposition to ASTER and GFCC30, Dynamic World is a Land Cover dataset containing nine categories corresponding to different land uses, one representing water cells.

Population data is also coupled and tested at different sources and resolutions. The Global Human Settlement Layer (GHSL) (Schiavina et al., 2019) from the European Commission provides spatial world population data in raster format at medium (250m) and low (1km) resolution in World Mollweide (EPSG:54009) grid. Data derive from satellite observations, census data and crowdsources. The second source of information comes from WorldPop's (Bondarenko et al., 2020) constrained dataset, which utilizes Random Forest models to disaggregate and map population counts to built areas. It is freely available at 100m resolution in WGS84. The three datasets contain integer population counts as cell values.

Finally, the last input data source combined in this work is geographic altitude, extracted from the ASTER Global Digital Elevation Model (GDEM) Version 3 (ASTGTM) (NASA/METI/AIST/Japan Spacesystems and U.S./Japan ASTER Science Team, 2019), available in WGS84 at the 1-arc-second resolution, as the water-bodies dataset from the same initiative. In this dataset, cell values are absolute elevations measured in meters.

As aforementioned, the raster data input to the fully convolutional branch can come from Sentinel-2 land reflectance bands (orthophotos), elevation (DEM), or even water bodies or population databases. We will take advantage of the spatial nature of these sources to enrich the multi-spectral Sentinel-2 images with mapped numerical and categorical information. We do it by stacking the different Sentinel-2 bands (three or four bands) with additional channels to create multidimensional array inputs. In order to complement the Sentinel-2 raster tiles with the additional data sources, they must be reprojected to the same projection. We extract rasters from water bodies, population, or elevation databases on the same coordinates, reprojecting and stacking them on top of the RGB and NIR channels, having at the end four, five or even six dimensions. However, we test if spatiality is crucial by inputting data in other modalities and comparing the results. Since we employ multi-modal data and combine it in different and novel ways to find an optimal dataset and model architecture, we implement an automated pipeline (Figure C1) to create and test remote sensing datasets. The pipeline is flexible and efficient enough to extract this information from different local databases according to the specifications and a set of coordinates indicating the extraction points in a few minutes.

Moreover, we extend our dataset with different augmentations, including flips, rotations, and zooms. All augmentations have a random component: the directions to flip (four possible values), the rotation's magnitude, and the zoom's scale factor. We apply them consistently by considering projection, resolution

or data types aspects to employ the appropriate interpolations. We elaborated the zoom to adapt it better to the problem of study, in the sense that the scaling factor is limited by the minimum and maximum label areas in the distribution.

When using the complementary data to feed the additional regression branch, we follow similar steps, except that the additional rasters do not stack on top of the Sentinel-2 band images. We extract population or water area counts in increasing diameter circles and generate a five-dimension vector that the network reads.

For a more robust hypothesis validation, we test our model in a different domain from the splits of images generated from Sentinel-2 tiles. We scan (Figure C2) a mosaic of the Catalonia area by dynamically generating images on runtime and providing these images to our model to perform inference. At the same time, we collect the population or water-bodies data for the model to read it together with the images. As we are using the entire domain, we are certain that the model will read images with previously unseen WWTPs. For that reason, we also match each predicted WWTP with respect to the complete UWWTD database, provided that there might be errors in the coordinates of some of these facilities, as we detected during the validation process. We consider a prediction valid if any segmented pixel is less than ten pixels away from the ground truth coordinates, knowing that this will lead to an excess of false positives when looking at non-validated plants'. To account for this problem, we will also take note of the false positives in the surroundings of the ground truth coordinates for further evaluation.

5. Development

In this study, we address the suitability of using CNNs to locate and characterize WWTPs from remote sensing images. We tackle the detection issue as a semantic segmentation exercise, using an architecture suitable for this problem, and adapt it to read vector fields and perform regression in parallel to the pixel-wise classification. Our final objective is to train a model with known, labeled plants to subsequently detect unknown plants in different scenarios.

We will search and verify the best dataset and architecture to obtain a trained model. Then, we will validate the hypothesis by performing inference on a different imagery source than the one used in the training process, including known and unknown plants. Finally, we will compare the distribution and magnitude of the inferred capacity with the ground truth.

This section details the execution of the steps prior to the testing phase, including the data obtention and the development of the different components (extraction, segmentation and scan software).

5.1. The input data

In two rounds, we downloaded Sentinel-2 data from the Copernicus Open Access Hub. First, we obtained all tiles with less than 8% cloud coverage available for the Iberian Peninsula and the Balearic Islands for the last two weeks of July 2020. Secondly, we extended this 1.2 terabyte dataset with another download containing tiles for the same area over the whole 2021 year. This time the cloud coverage threshold was set at 15%, resulting in 2.4 additional terabytes. The reason for the dataset extension was to increase its variability by including data captured in different seasons.

The part of the dataset corresponding to the year 2021 includes more than 10,000 different representations of the labeled WWTPs, averaging ten images per plant. Importantly, the largest part of the dataset tiles does not include any plant representation due to their sparsity. On the other hand, the July 2020 data comprises around 4,700 images of labeled plants, averaging four shots per plant. After a few tests, we observed that the impact of using the 2021 extension in the results did not pay the increase in computing time. Therefore, we decided to only work with 2020 data, which presents less cloud coverage and more uniform light conditions. We will refer to this one for the rest of this report when we mention the dataset.

ASTER (water bodies and elevation datasets) and GFCC30 products were obtained from the NASA Earthdata download platform. As the Sentinel-2, these are offered as independent GeoTIFF tile files,

occupying 1.9 and 5 GB each. ASTER has no no-data cells and no overlap between tiles, while GFCC30 presents no-data and cloud-covered areas, with some colliding areas.

Concerning Digital World, the only possibility to acquire this data is through the Google Earth Engine. We generated a composite image of the Iberian Peninsula for July 2020 with the available Javascript API. To generate the composite, we got all the available tiles in that month and averaged the probabilities for each pixel and category to get a pixel-level class. The composite comes as a single GeoTIFF file in the WGS84 coordinate system, given that the original tiles belong to different UTM projection zones.

We directly downloaded population data from the European Commission and Worldpop sites. Population data is much coarser and fits in a single GeoTIFF file. The UWWTD database was also downloaded from the European Commission portal.

To evaluate the performance of the trained models in the real world, we ultimately test if they can detect and characterize every plant in a particular region using a different dataset. We use a mosaic of Catalonia made up of continuous Sentinel-2 tiles, downloaded from the Institut Cartogràfic i Geològic de Catalunya (Institut Cartogràfic i Geològic de Catalunya, n.d.) website. Concretely, the mosaic is divided into two raster files, one comprising the red, blue and green bands, and another devoted to the infra-red.

5.2. Image extraction

The image extraction software is an entire package created in this project to automatically and rapidly generate datasets from multispectral images and, optionally, additional data. The software reads a collection of raster tiles in the provided input directory, obtaining crop images in the coordinates listed in an input database. We move image centers according to a set of random shift ranges during extraction to produce variations in the dataset and reduce the location bias of the model trained with these images. Increasing the range of the random shifts, we can also increment the chance of having negative (without any foreground-class pixel) examples, which will help to make training more robust and generalizable. The provided image labels and the extra data layers or vectors will be shifted with the images.

There are several options that the user can define with command-line arguments. We specify the number of extracted shift instances[4] per location and tile, which can provide different shots of the plant at the same date and time or negative examples (not containing any plant) if the shift is high. Images are composed by stacking the RGB bands by default. However, we can optionally extend the stack with the

[4] In this work, we refer with the name “instance” to every repetition of a WWTP-tile image (a shot of a WWTP in particular time), which usually are only differentiated by having a different shifts or other kind of augmentation.

NIR band and additional channels containing elevation, water bodies, or population information. Water bodies and populations can also be extracted as vectors, together with the WWTP capacity. All the additional channels are aligned using the images' central coordinates, reprojecting them if necessary (if they have different original projections).

Other options, such as the output image size, their format and depth (8-bit or 16-bit), or the rescaling ranges, are parametrized too. For the Sentinel-2 crops, a maximum input value allows discarding input reflectances above a threshold value. For water, elevation, and population data, maximum and minimum values act as low-pass and high-pass filters to create raster masks. The software also allows for extending the output dataset with additional shifts, flips, rotations, zooms, or a combination of them. The extracted area is extended for rotations to ensure that the final image does not contain missing data values in the margins due to the transformation. In the case of zoom, it considers the original resolution of the input sources to adapt the rescaling to each channel.

The software traverses all the input Sentinel-2 granules at runtime, iterating simultaneously and for each output image, over elevation, water-bodies or population tiles. Due to the number of input tiles from different sources, the extraction can be a slow process. To speed up the extraction, we adopted two strategies. On the one hand, the software executes the loop over the Sentinel-2 tiles in parallel. On the other hand, it indexes the additional sources' input tiles and keeps their associations with the plants. This indexation severely limits the number of searches. As a result, performance is bound by the disk latency and bandwidth in computers with a minimum number of processing units (cores). Datasets containing thousands of images are generated in a few minutes (500 to 3,000 images per minute).

5.3. Segmentation pipeline

The segmentation package is in charge of performing the training and validation of the models. It divides the data into reproducible splits, repeating the process a number of times and averaging the results. Traceability is ensured by saving the split links, model, dataset statistics, and plots in different locations. Therefore, we can quickly recover and test any trained model to perform inference or generate new plots.

The model is coded in Keras and Tensorflow. A custom data loader is in charge of providing input batches, dynamically rescaled and normalized. We implement z-score normalization as an option at the instance and the feature level. This option needs the data metrics extracted from the dataset or provided during inference. The data pipeline can run in parallel with a number of workers to asynchronously fetch and normalize the data, which speeds up training.

The architecture of the evaluated models is based on a classical U-Net architecture using five or six convolutional layers with an increasing number of filters in the encoding and decoding subnetworks. Specifically, we use 3x3 kernels with Rectifying Linear Unit (ReLU) activations. The alternative architecture adds a secondary branch with two different inputs followed by two dense layers with ReLU activations for regression training and inference. One of the inputs is connected to the convolutional encoder network's output (the bottleneck). The regression output is the concatenation of the dense layers' outputs. While the U-Net uses a sigmoid output function to calculate the class probabilities for every one of the pixels, the regression has a linear output. We use Adam solver for all the tests and different starting learning rates with a linear decaying schedule after four epochs. The training length is variable and is stopped ten epochs after the validation Intersection Over Union (IoU) stops improving.

In this regard, we use three metrics to evaluate the performance of the convolutional branch: IoU, recall and precision. During the tests, we evaluate two main loss functions with some variations: binary cross-entropy and Dice's score. Conversely, we always use the Mean Absolute Error (MAE) for the regression as an evaluation metric and loss function.

5.4. Test in the real world (domain transfer)

Lastly, the scan package is the one we use to evaluate the performance of the models in a new scenario. It uses a trained model to sweep the entire area in a GeoTIFF raster, optionally using additional input sources (such as water bodies or population databases) to segment WWTPs and regress their capacity. The package reports the sensitivity and the specificity concerning one or more ground truth databases. Moreover, the package also reports borderline situations and the total predicted capacity. The feature is essential because non-labeled plant coordinates are subject to errors. Nevertheless, the software also includes an option to plot the ground truth and predicted capacity over the domain. Visual validation is essential to verify the spatial capacity distribution and the errors' importance.

Sweeping an entire domain covering thousands of square kilometers with 10-meter precision entails reading and analyzing dozens of thousands of images. This task may take several hours if performed sequentially. More if water bodies and population data, generated dynamically, is used too. For that purpose, we parallelized the reading of the images, passing them in batches to the Graphical Processing Unit (GPU) for evaluation. Additionally, we benefit from the parallel segmentation data loader.

6. Results

This section follows the sequence of steps and results to search for an optimal dataset and architecture to detect and characterize WWTPs, and test the resulting model on a different scenario. When improving the initial dataset and model, the number of options and available decisions is overwhelming, considering the amount of data, different modalities and possible hyperparameters. Hence, we followed an incremental approach with a series of ablation tests. Starting from simplified versions of the datasets and architecture, we develop them in parallel, taking decisions from the results obtained in the process.

Every round of tests comprises three independent runs in which the dataset is split into three disjoint splits using reproducible seeds. We do the partition at the WWTPs level, keeping 80% for training, 10% for validation, and 10% for testing. As the ratio of images per location is not constant, the number of images per split will differ slightly between tests. The final score is the average of the test metrics obtained in the three runs. When specified, the model will be trained and validated with splits of one dataset and tested with another dataset's (test) split. An example of crossing datasets is when we are interested in having a higher ratio of negative instances at inference to be closer to real-world conditions.

Until we state the opposite, the U-Net architecture will contain five layers of an increasing number of filters. We use a batch size of eight images and set the learning rate as 2e-4. Meanwhile, the initial dataset contains the red, blue, and green bands from the July 2020 dataset, extracted at the coordinates of the labeled WWTPs with shifts of 50 pixels, ensuring that the plants never exceed the image boundaries. One image is collected per WWTP and tile, producing 4.7 thousand pictures of 128x128 pixels.

We performed all the model training and testing in the CTE IBM Power9 cluster at the Barcelona Supercomputing Center (BSC-CNS), part of the MareNostrum 4 Supercomputer. Learning is only parallelized at the data pipeline level. Therefore, each test runs over one Graphics Processing Unit (GPU), representing one-fourth of one node's computing capacity. Specifically, we run every test on ten cores (40 maximum threads) and one NVIDIA V100 (Volta) GPU.

Similarly, we perform the final testing (scan) over the same resources, with only one difference. Since the image data is read online from the mosaic and does not use the asynchronous data pipeline, besides being processed in parallel with multiple threads, we use the node's Non-Volatile Memory express (NVMe) disks for all the input operations. Using this private partition of the node implies copying the input data at the beginning of each run, but the disk bandwidth and latency gained largely compensates for the copy.

Finally, the dataset generation is usually done in the Nord III cluster of the BSC, more appropriate for data operations that do not need GPUs or the increased power of MareNostrumIV. As aforementioned, the extraction is a mostly I/O-bound task. Hence, we usually allocate at least half a node to have enough memory and I/O bandwidth but only use a fraction of the computing cores for computation.

6.1. Input image format

We start the dataset's creation by deciding the format of the images included, namely the 16-bit channels extracted from Sentinel-2 bands and their final precision. These bands contain land reflectance values peaking values in the 23 thousand units. However, most values group on the first quarter of the range and we consider removing outliers. To illustrate this concentration, Figure 8 shows the mean and standard deviation of the dataset for the RGB channels after using different low-pass filter thresholds. Paying attention to this information, it seems safe to filter reflectance values over 4,000 units.

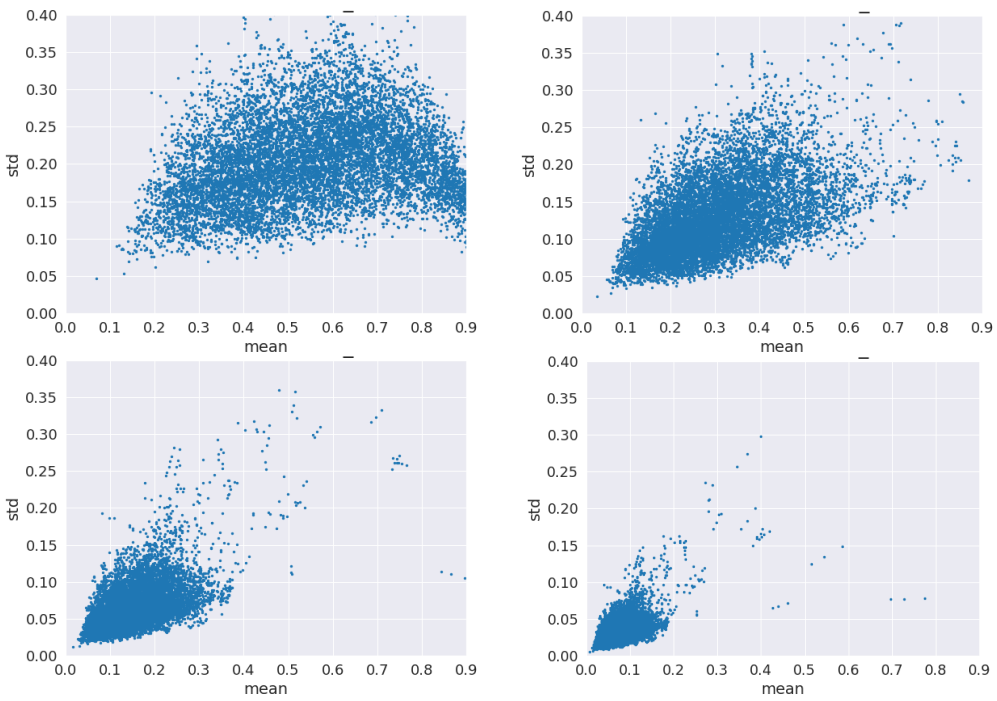

Figure 8: RGB dataset average reflectance and standard deviation. Filtering values over 2,000 (top-left), 4,000 (top-right), 8,000 (bottom-left) and 16,000 (bottom-right).

We test the performance of models trained over 8-bit RGB images extracted after applying the different evaluated thresholds. We repeat the test after stacking the NIR band on top of the RGB channels, producing 4-channel RGB+NIR images. Results represented in Figure 9 show that RGB+NIR images from values below 4,000 land reflectance units give the best overall performance. Furthermore, filtering all values above 2,000 would have been enough if only considering RGB. However, NIR band values are generally higher, and the range between 2,000 and 4,000 is essential.

On the contrary, Figure 11 shows that when evaluating the preferred image depth, the results do not tip the scales toward either option (8-bit and 16-bit). However, we preferred to proceed with 16-bit to avoid the risk of losing information when we work with additional channels not evaluated at this stage.

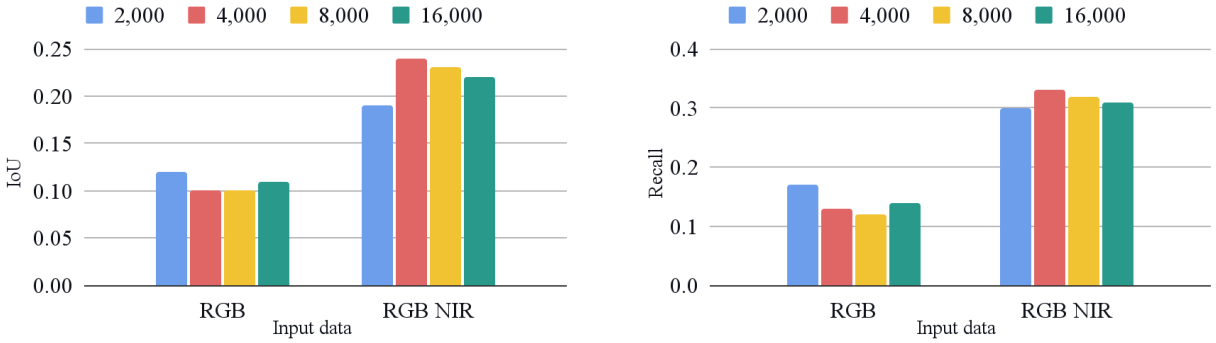

Figure 9: Test IoU (left) and recall (right) of FCN models trained with RGB and NIR orthophotos after filtering reflectance values over 2k, 4k, 8k and 16k.

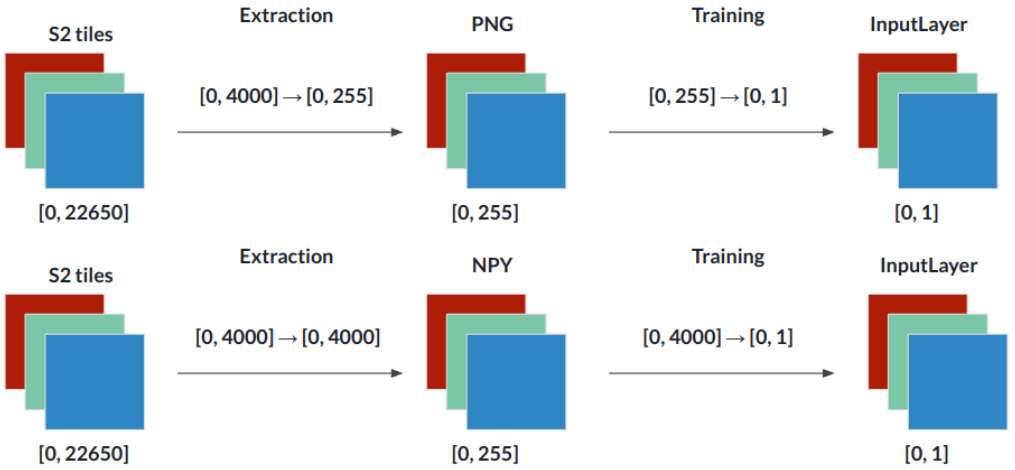

Figure 10: Rescaling workflow during extraction and model training for 8-bit (top) and 16-bit (bottom) images.

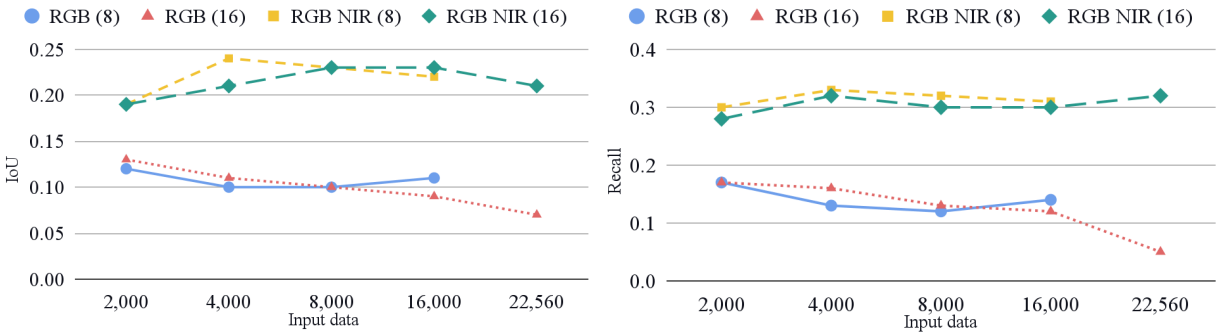

Figure 11: Impact on the test IoU (left) and recall (right) of the image depth (8-bit or 16-bit) for models trained with RGB and NIR images.

6.2. Class imbalance and supplementary raster channels

6.2.1. Moderate class imbalance

We follow by evaluating how the subsequent models learn and perform when we increase the inter-class imbalance by adding negative (without any plant) images. Even if the imbalance is already high at the image level, it is lower than in the real world due to the sparsity of the WWTP infrastructures. Moreover, we intend to increase the variance of infrastructures seen by the model. We progressively incorporate additional shifts, increasing the total number of extracted images and obtaining more than one image instance per WWTP and input tile. At the same time, we test the impact of stacking supplementary channels (resulting in 5-channel images) with water bodies or population data, hypothesizing that this information could help the model locate WWTPs and then characterize their capacity.

Evaluating the information in Figure 12, we can observe that increasing the class imbalance reduces IoU and recall. However, expanding the dataset helps mitigate this loss. When looking at the contribution from including additional channels, the impact of adding water bodies data is more noticeable when the dataset is smaller. On the other hand, adding raster population data is not helpful.

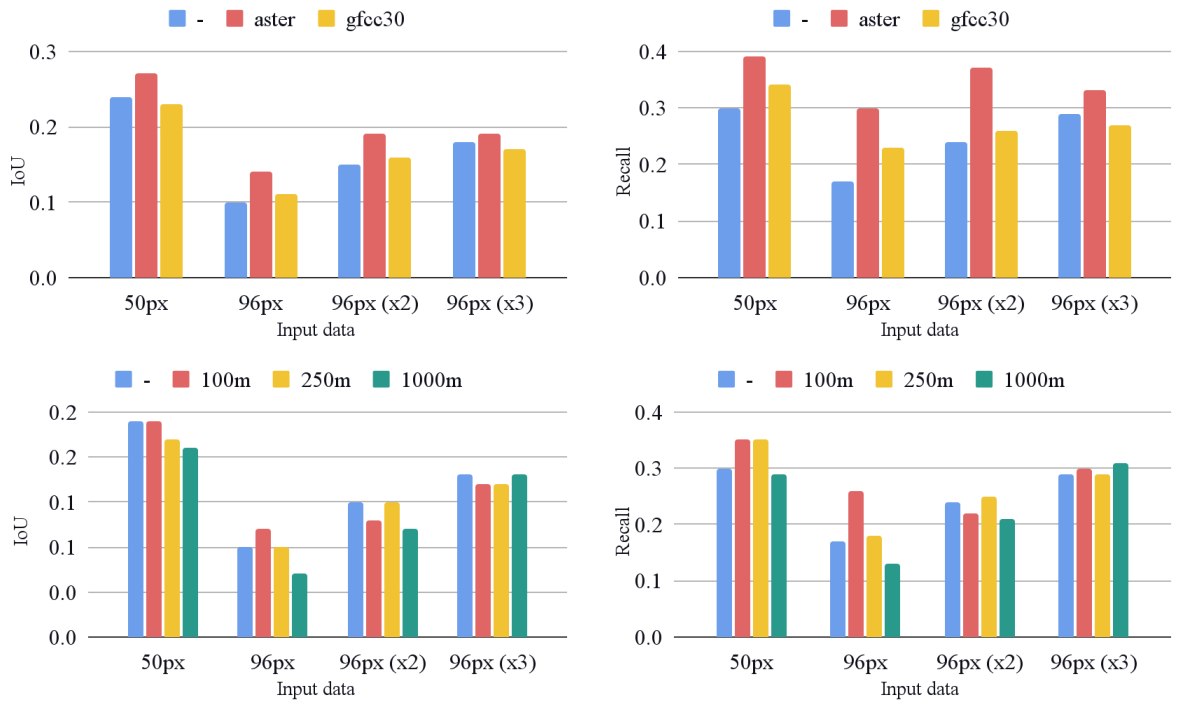

Figure 12: Test IoU (left) and recall (right) of models trained with 4-d RGB+NIR, 5-d RGB+NIR+water-bodies (top), and 5-d RGB+NIR+population (bottom) images, using ASTER and GFCC30 data (water-bodies), and 1000 and 250m GHG and 100m WorldPop data (population). In the horizontal axis, the maximum shift (and the number of shifts), resulting in 4.700, 4.700, 9.400, and 14.100 images, respectively.

6.2.2. Increasing class imbalance and introducing Dynamic World as raster data

In this step, we increase the dataset's negative instances ratio further to reduce the gap between our tests and real-world, sparser scenarios. We train models with two different datasets having a fair amount of negative examples (0.45 and 0.7 negative ratios), and we test them on a third dataset with a superior ratio (0.87) of images not containing any foreground class pixel. Ratios are raised by preserving the number of positive instances (using lower shifts) and adding negative instances (using higher shifts), meaning that these datasets have an increasing number of images. As aforementioned, we do the splits over the set of WWTPs. Therefore, different splits will never contain images of the same plant, even if we cross datasets.

Figure 13 shows that the model performance decreases concerning the previous dataset. Note that we increase the imbalance (at the image level) of an already imbalanced dataset (at the pixel level). Nevertheless, recall stays steady and precision increases in some cases. In turn, adding the ASTER water-bodies channel improves recall and precision. Surprisingly, Dynamic World does not contribute to improving the metrics despite its finer resolution. Note that here we show results from stacking the water-bodies channel as a mask made from water class values. Detailed results from tests using the rest of the land-use classes can be found in Table D5.

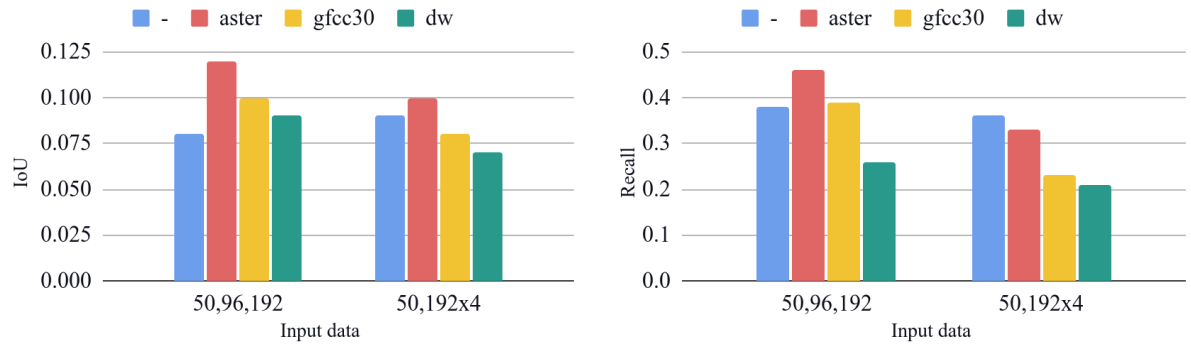

Figure 13: Test IoU (left) and recall (right) of models trained with 4-d RGB+NIR, and 5-d RGB+NIR+water-bodies images, using ASTER, GFCC30 and Dynamic World data. In the horizontal axis, the maximum shift (and the number of shifts), resulting in 14,100 and 23,500 images, respectively.

6.3. Introducing a loss function for datasets with pixel-level class imbalance

Given the destructive impact of increasing the negative examples ratio on the result, we evaluate loss functions that are more appropriate for a problem with a class imbalance at the pixel level. In particular, we test a loss function based on the Sørensen–Dice coefficient (Dice's coefficient from now on). We expect IoU to improve by using a loss function based on the same principle: the only difference between

Dice's coefficient and IoU is the double count of true positives in the former. To convert the coefficient to a loss function (Eq. 1), we subtract it from the unity.

$$DSCL = 1 - \frac{2TP}{2TP + FP + FN} \quad \text{(Eq. 1)}$$

We compare Dice's and cross-entropy loss with slight variations in the architecture and hyperparameters (feature-based z-score normalization, increasing the network depth, and adding batch normalization). This time, we also add a series in which we train the model with the sparser dataset (the same used for testing). To illustrate the importance of using an appropriate loss function, Figure 14 shows that Dice's loss improves test IoU by five points on average, and Figure 15 indicates that it reduces overfitting. The recall also benefits from the change but comes at the cost of decreased precision. It seems clear that while cross-entropy loss is more focused on keeping high precision, Dice's loss is more greedy and exploratory.

A side effect of this loss function is its tendency to fall into local minima. However, this is not the case when using batch normalization, which also provides the best results. Here batch normalization is added in each convolutional layer (excluding the last and first layer of the encoding and decoding subnetworks, respectively) before the ReLU activation. We evaluate this collocation in the following steps.

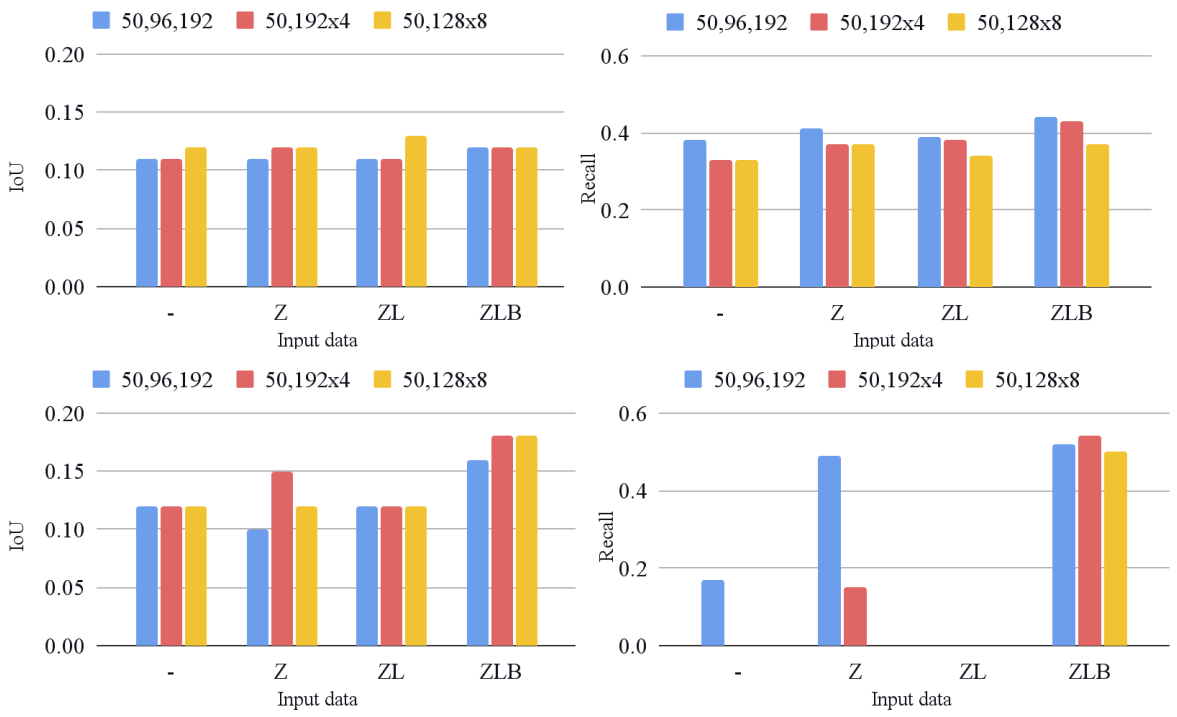

Figure 14: Impact of the loss function on test IoU (left) and recall (right) of models trained with 4-d RGB+NIR images, with different datasets (14,000, 23,500, and 42,300 images). Top: cross-entropy. Bottom: Dice's score. In the horizontal axis: no normalization (-), z-score normalization (Z), z-score + additional conv. layer (ZL), z-score + conv. layer + batch-normalization (ZLB).

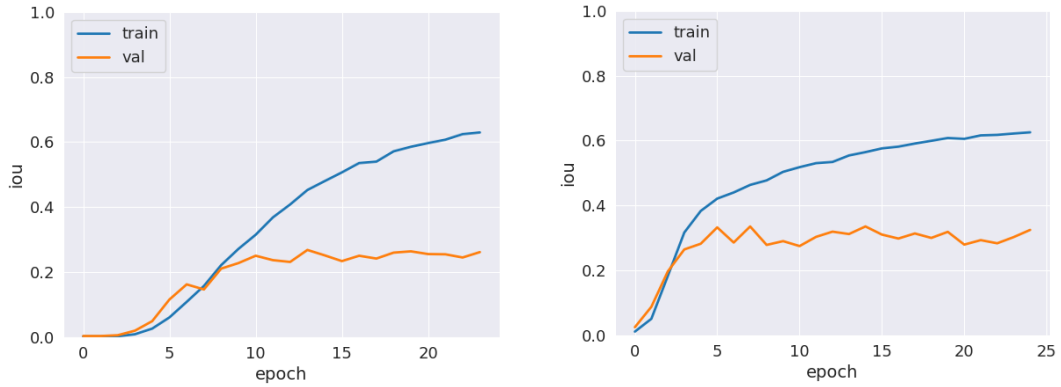

Figure 15: Training and validation IoU curve over 25 epochs for one of the three iterations when using cross-entropy (left) and Dice's (right) loss.

To address the frequent stalling during learning, we evaluated a variant of Dice's coefficient in which we square the denominator terms. This variant avoids having a zero derivative when the prediction perfectly matches the ground truth (Milletari, 2016). The result of this new test (Figure 16) reveals a gain in general training stability but a degradation of the results obtained with batch normalization, which was the best solution until now. Consequently, we choose to proceed with the original Dice's loss while using batch normalization in the convolutional layers.

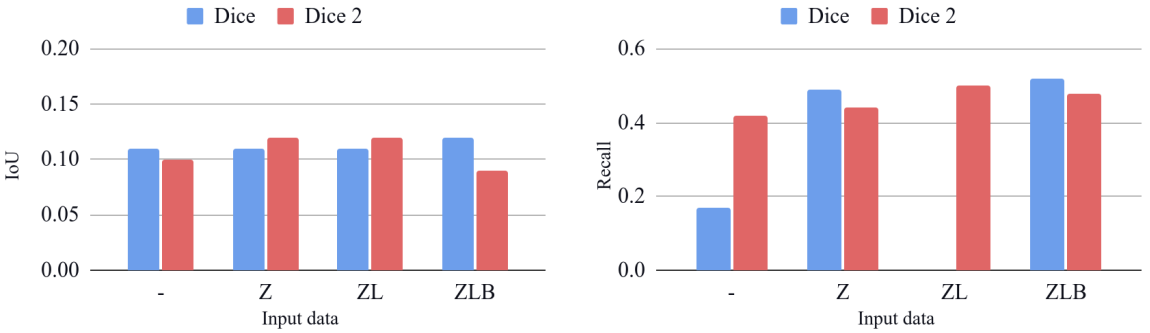

Figure 16: Impact on the test IoU (left) and recall (right) of using Dice's (left) and Dice's with squared terms (right) losses with different datasets of 4-d RGB+NIR (14,000, 23,500, and 42,300) images. In the horizontal axis, no normalization (-), z-score norm. (Z), z-score+additional conv. layer (ZL), z-score+layer+batch-normalization (ZLB).

6.4. Optimizing the network architecture and hyperparameters

After the tests performed in the previous section, we conclude that Dice's coefficient is a better loss function for this problem, and batch normalization improves stability and performance. However, we have yet to demonstrate if batch normalization is the only one responsible for this improvement or if, otherwise, the increased layer depth contributed too. Moreover, we are not sure if the positioning of the normalization is the optimal one. For this reason, we evaluate these aspects separately, including a

different placing of the normalization, after the ReLU activations (and before the next convolutional layer's input). Finally, we verify if we can increase the batch size to speed up learning.

The outcome of this evaluation, as Figure 17 illustrates, is that the increased network depth contributes to better performance. In addition, increasing the batch size does not help to generalize better or decrease the learning time since we already use most of the GPU's capacity. Interestingly, the repositioning of the batch normalization reduces overfitting and makes learning more steady (Figure 18).

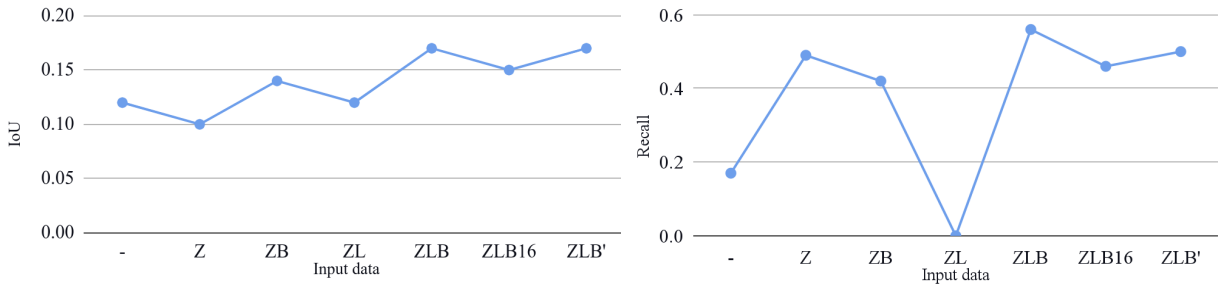

Figure 17: Impact on the IoU (left) and recall (right) of subsequent architecture and hyperparameter modifications. In the horizontal axis: no normalization (-), z-score normalization (Z), z-score+batch-norm. (ZB), z-score+additional conv. layer (ZL), z-score+additional layer+batch-norm (ZLB), same with batch size 16 (ZLB16), z-score+additional layer+modified batch-norm (ZLB').

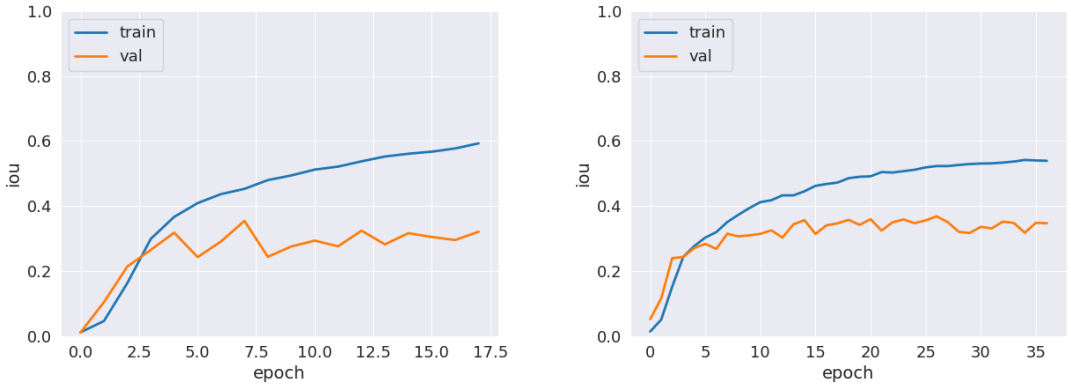

Figure 18: Training and validation IoU curve using early-stopping for one of the three iterations with batch normalization before (left) and after (right) the layer activation.

6.5. Re-introducing supplementary raster channels

After updating our architecture in an aspect as critical as the loss function, we repeat the assessment of the impact of the water-bodies raster data on the performance. Not only was its contribution not completely evident with the cross-validation loss, but the results obtained with Dice's loss and the architecture

changes beat the previous results when using five-channel images. Furthermore, we also evaluate the impact of adding a channel with topographic information obtained from a DEM.

Figures 19 and 20 indicate that the network does not benefit from using the additional stacked channels. Even in the case of ASTER, which previously showed good results, we see no obvious benefit from its use. Elevation slightly improves recall, but usually at the cost of a decreased precision (Tables D9 and D10).

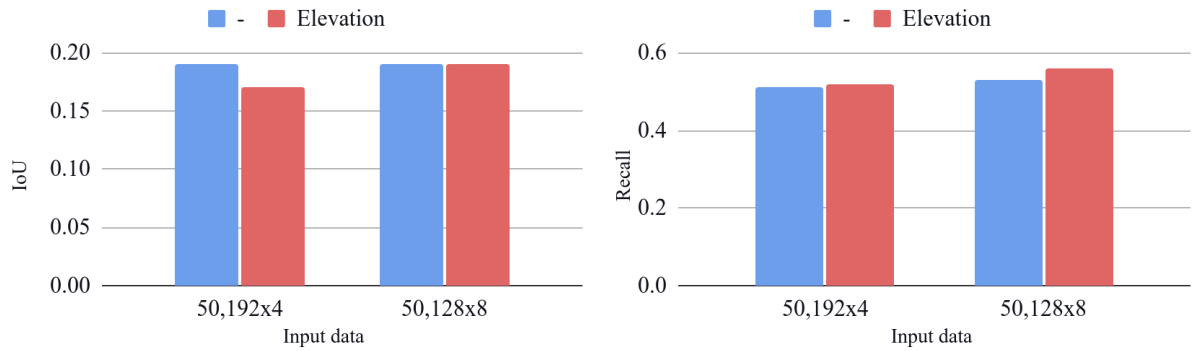

Figure 19: Test IoU (left) and recall (right) of models trained with 4-d RGB+NIR, and 5-d RGB+NIR+elevation images. In the horizontal axis, the maximum shift (and the number of shifts), resulting in 23,500 and 42,300 images, respectively.

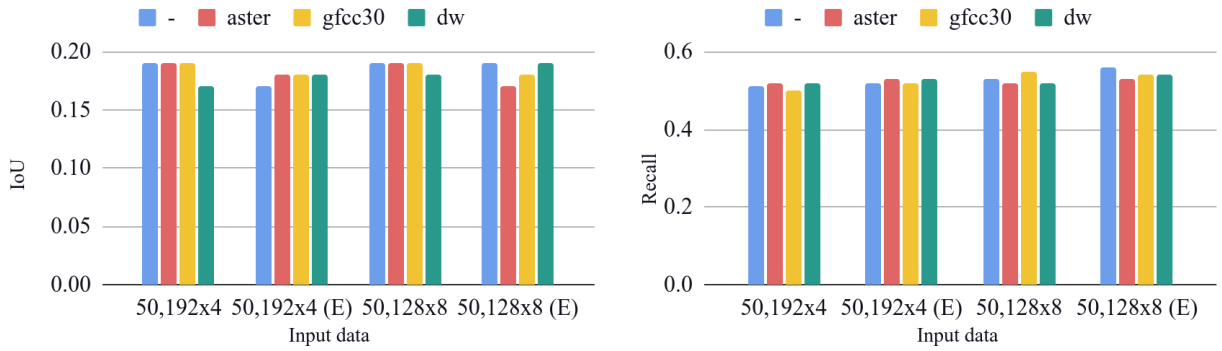

Figure 20: Test IoU (left) and recall (right) of models trained with 4-d RGB+NIR, 5-d RGB+NIR+water-bodies, and 6-d RGB+NIR+water-bodies+elevation images, using ASTER, GFCC30 and Dynamic World data. In the horizontal axis, the maximum shift (and the number of shifts), resulting in 23,500 and 42,300 images, respectively, without and with (E) elevation.

6.6. Optimizing the dataset size and negative instance ratio

The tests in 6.2 give us an idea that IoU and precision can improve by increasing the number of instances and the ratio of negative images, respectively. This section finds the optimal quantities for the dataset size

and proportion of negative instances by doing an exhaustive search based on the independent scaling of these values.

From our tests, depicted in Figure 21, we conclude that the test IoU increases with the size of the dataset, but the precision does not. Moreover, the IoU levels off from 24,000 images onwards. On the other hand, precision increases with the ratio of negative images. We optimize these numbers in new tests summarized in Figure 22, breaking down the upper range of the negative ratio. Focusing on the IoU, it looks like it is not worth employing more than 24,000 images for training, considering the increased training time and energy. On the other hand, precision is better when keeping a ratio close to 0.7.

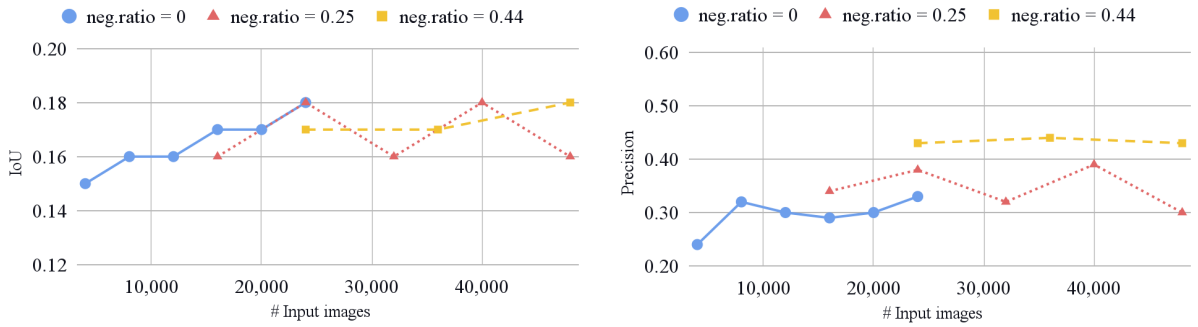

Figure 21: Impact on the test IoU (left) and precision (right) of increasing dataset size and negative instance balance during training with 4-d RGB+NIR datasets.

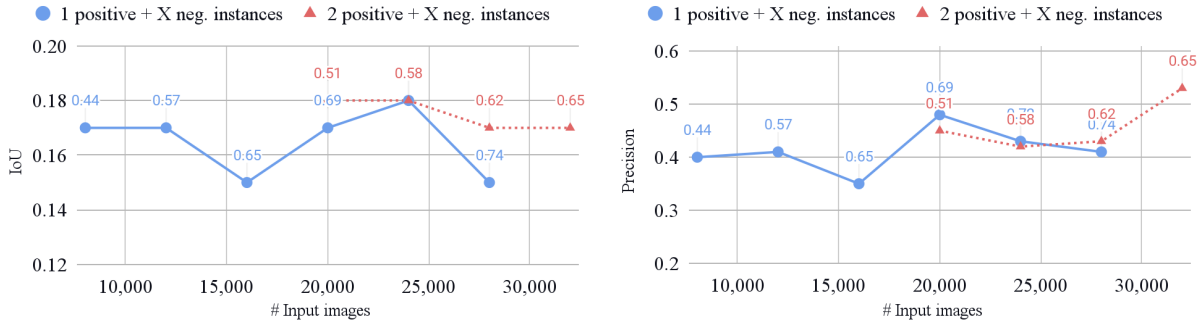

Figure 22: Impact on the test IoU (left) and precision (right) of increasing training dataset size and negative instance balance (represented with label values) by using 1 or 2 positive images per WWTP and tile and a varying number of negative instances, with 4-d RGB+NIR datasets.

As a result, we go on with a dataset of 20,000 images, ensuring at least two positive instances out of every five instances of the same image and a total 0.5 negative instance ratio, considering that we will further extend it with augmentations, which will impact the dataset size.

6.7. Data augmentation

As a final step in the process of optimizing the raster dataset, we extend it with data augmentation techniques. Though the model overfitting has progressively decreased by changing the loss function, increasing the ratio of negative instances, or introducing batch normalization layers in the optimal locations, we aim to decrease it further, as data augmentation is typically beneficial when dealing with pixel-wise segmentation.

Therefore, we assess three different and common transformations: horizontal and vertical flips, rotations, and zooms. However, after rapidly noticing that flips are always helpful, we evaluate all the possible combinations of transformations including flips. We extend the dataset composed of five instances of every plant and tile (close to 20,000 images) with additional augmented copies. Augmentations are applied progressively, from adding one transformed instance to doubling the total number of instances and, consequently, total images.

In the first place, Figure 23 shows that IoU and precision increase with the addition of augmented images. Secondly, the contribution of contribution the rotations and the zooms is unclear. According to the plots, combining all the augmentations is the most secure and stable option. However, following our hypothesis that the plant's size and capacity are related, we prefer to not perturb the dataset with changes in scale (zooms) that may hinder the subsequent regression performance. Therefore, from now on we always augment the dataset using flips and rotations.

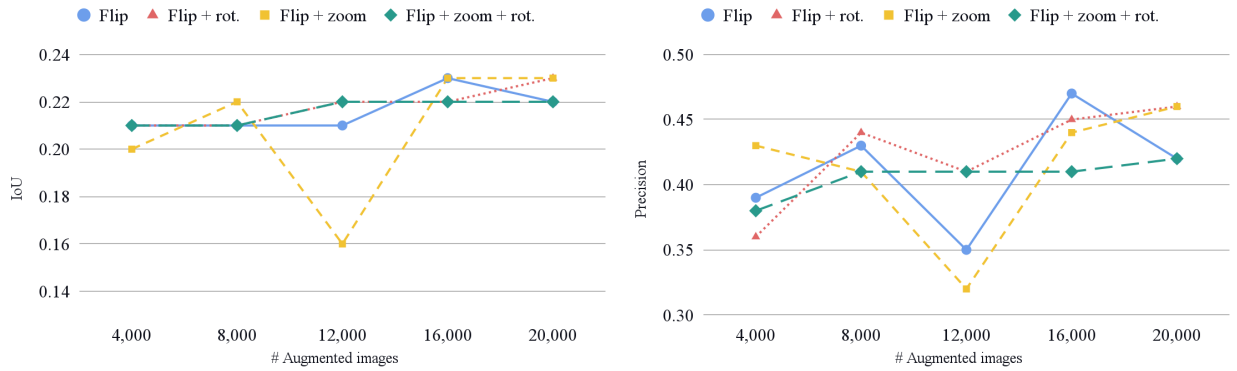

Figure 23: Impact on the test IoU (left) and precision (right) of adding different kind and quantity of augmented instances to 4-d RGB+NIR datasets.

6.8. Regression of the plant capacity

The goal of this subsection is to verify the hypothesis that WWTP's capacity can be inferred from its size, shape and complementary information as the close population. Additionally, we investigate if the magnitude of close water masses can be another useful resource to infer capacity. In order to do so, we enrich the best pixel-level segmentation model obtained so far by adding a branch that processes the information encoded by the U-Net, together with this additional information.

Considering the results in Figure 24, data from the population (at 100-meter resolution) and water bodies (especially at ten-meter resolution) help the model estimate (lower mean absolute error) the global capacity, and it does it better when there are more images to learn.

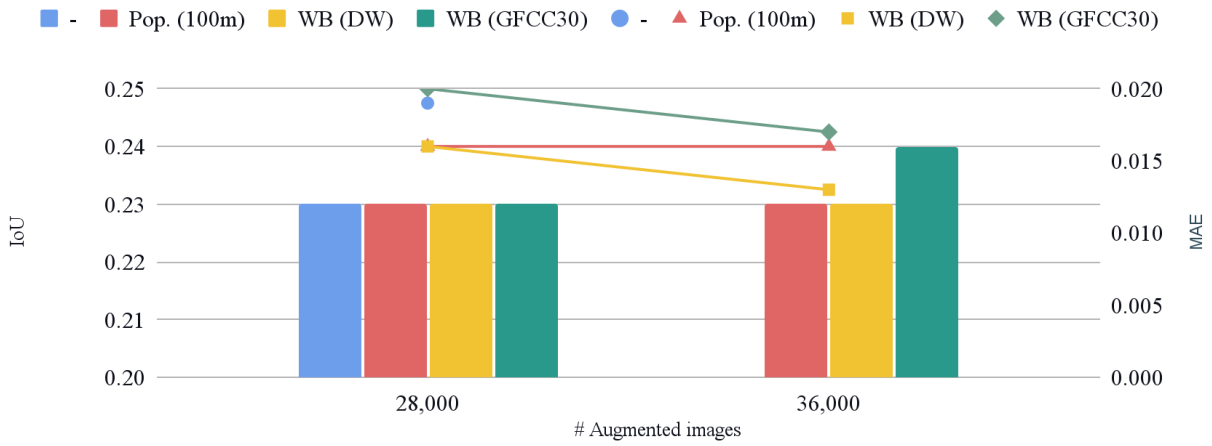

Figure 24: Impact on regression (MAE) from using different vector input modalities (no data, population or water bodies) and raster dataset sizes. Segmentation performance (IoU) for the different 4-d RGB+NIR datasets.

6.9. Test by sweeping an entire domain

For a more robust hypothesis validation, we test our model in a different domain than the images extracted from Sentinel-2 tiles, as we have done in the previous steps. We start from the best models obtained, which achieve an IoU close to 0.25 and find 50% of the total WWTPs with 50% precision. We scan a mosaic of the Catalonia area by generating images on running time and providing these images to our model to perform inference. We sweep the domain with a stride of 64 pixels (half of the image extent) to ensure that every WWTP will fit entirely in the visual field at least once. At the same time, we collect population and water-bodies data for the model to read together with the images.

Starting with the WWTP location and segmentation skill (Figure 25), no model is clearly superior. In general, there is a true positive rate decrease compared to the previous tests using Sentinel-2 tiles. For example, the models using Dynamic World or population at 100m data can find more WWTPs, but they also produce more false positives (note the order of magnitude difference in the scale). Moreover, there is also no clear benefit from using a complete dataset (using only training and validation splits) for training: models return more true positives, but also more false positives. More interesting is the total inferred capacity, that in most cases is in the same order of magnitude[5] that the real one (Tables D18 and D19).

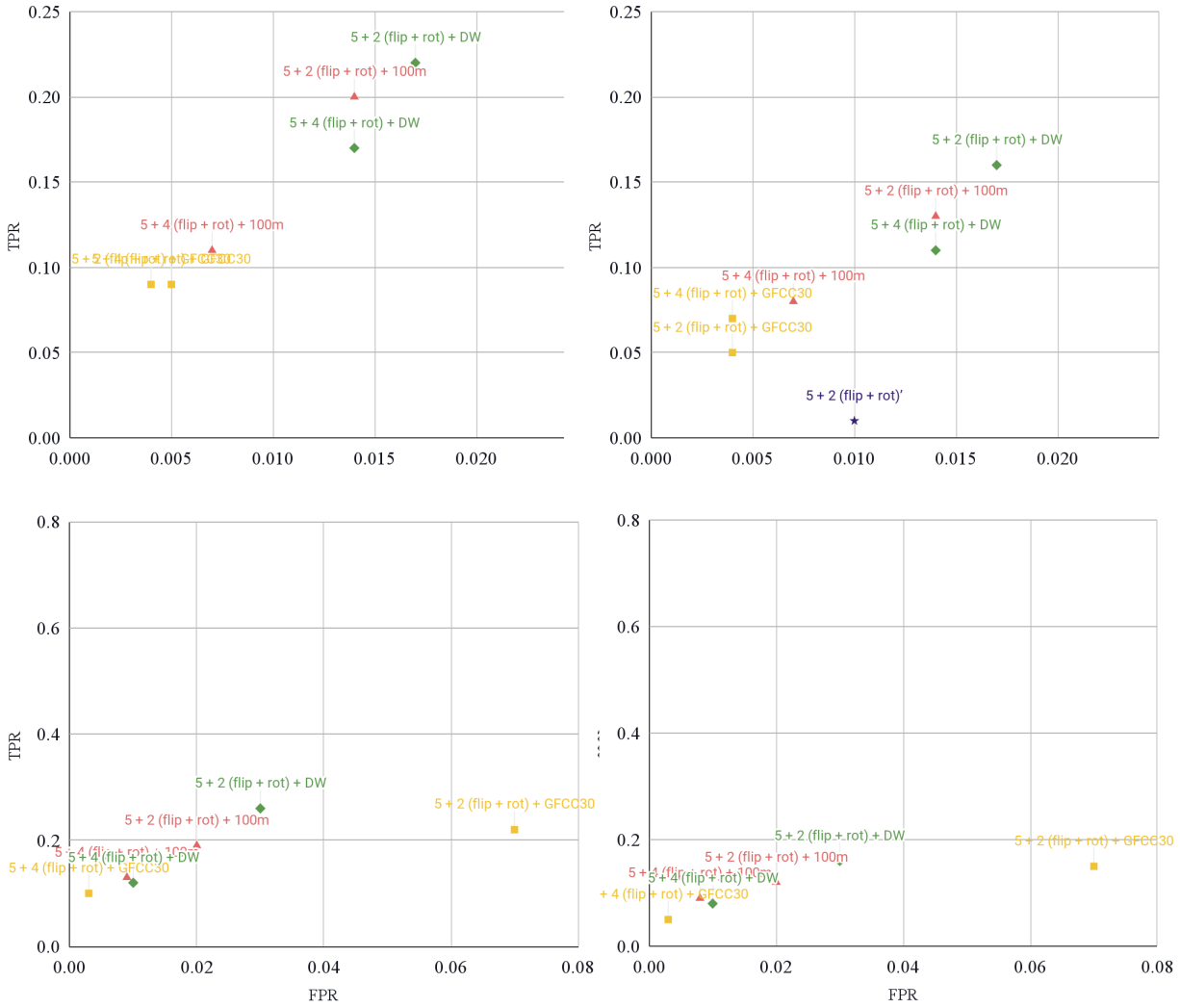

Figure 25: True (TPR) against false (FPR) positive rate for models trained with vector data coming from no data (violet), population(red), GFCC30 (yellow) and Dynamic World (green) water-bodies data. Top: using 80% of the dataset as training split (20,000 images extended with 4,000 or 8,000 flips and rotations). Bottom: using 90% of the dataset as training split (24,000 images extended with 4,700 or 9,400 flips and rotations). Left: Labeled WWTPs as ground truth. Right: UWWD dataset as ground truth.

[5] Each image quadrant is read four times during the sweep; then, we expect the inferred capacity to be four times the actual one.

7. Discussion

Along subsequent steps, we have used the dataset extraction and segmentation pipeline to create a dataset and a model for WWTP characterization. We started focusing on the segmentation and essential decisions to assemble the data, like deciding a threshold value for the rescaling. Then, we proceeded with a series of ablation tests to decide the number and kind of channels that work better in our pixel-wise segmentation exercise. Next, we finished defining the model architecture and refined our dataset with a series of new ablation tests. After the optimization, we added a regression branch and assessed our model for WWTP characterization, finally performing new tests on a related domain.

From these results, we conclude that the 4,000 lower land reflectance values of the original Sentinel-2 tiles make better use of the range used by the network for training. The reason is that higher values are usually outliers that often correspond to reflections from bright objects or clouds introduced in the image.

Results demonstrate the superior performance of the loss based on Dice's coefficient with respect to the cross-entropy when classes have a high imbalance at the pixel level. Such is the case, with the smaller WWTPs occupying only two pixels of the Sentinel-2 tiles. As we are interested in improving IoU, it seems reasonable to use a loss function based on a related concept. However, we saw that Dice's loss could easily fall into local minima. To address this issue, we tested a variant with squared denominator terms that is more stable but less performant. Consequently, we decided to use our first Dice's loss implementation together with batch normalization, which also improves results.

Concerning the dataset size and the kind of information it holds, we saw that increasing the number of instances has a positive impact. That is reasonable since we are providing more examples for the network to learn from. However, the more images, the more expensive the training is, and we choose a size that seems to be at the scalability limit. On the other hand, training with only positive images is prejudicial when the test has a significant imbalance at the image level. Not only are we letting the model infer that every image will contain a foreground class, but also we are reducing the complexity by removing other kinds of structures from the training process. After running different tests, our approach was to use slightly more negative than positive images during training.

Once the dataset and model were defined, we proceeded to verify the hypotheses on evaluation. We inferred that the proximity of water-bodies sources could guide to localizing WWTPs. To validate it, we extended the dataset with additional channels obtained from different sources of water-bodies databases. We also evaluated other information, such as topographic or population data. It seems reasonable that WWTPs are close to human settlements and prefer low altitude gradients since water flows accumulate on depressions driven by the gravitational force.

Looking at the test's results, if the model seems to benefit from using water-bodies information (especially from ASTER dataset) when following the gradient of a cross-entropy loss, the effect is much less clear after moving to Dice's score loss. Furthermore, elevation and population data do not provide any clear impact. A possible explanation is that the model can already deduce the water basins location through one or multiple Sentinel-2 bands such as NIR. Also, the multidimensional information (we input arrays of five or six dimensions) can be difficult to process, and the surrounding context provided by the images may be too small. Moreover, even though this information could help estimate the probabilities of finding a WWTP on a particular location, segmenting the plant's contours is a big leap. Sometimes, the information is too coarse (like in the case of the population) or incomplete (ASTER shows the main water flows and lakes), and in other cases, too noisy. To illustrate this point, note that the Dynamic World dataset classifies water bodies with a ten-meter resolution. The amount of water pixels is enormous; sometimes, these pixels are located even within the WWTP contour (inside their settlers and tanks), while others are not.

The second main hypothesis is validated through our regular tests involving data splits and finally by testing our model at scale in the Catalonia area. In the first case, results show that models using water bodies or population data are better at inferring the correct capacity of the test images. In the second, the impact is evident at many levels, as we will explain.

First, the reason why models (Tables D18 and D19) lose close to 50% of the performance when passing from inferring over Sentinel-2 tiles to working on the Catalonia mosaic would need a deeper study. However, we can observe that models using water bodies or population information infer a global capacity in the same order of magnitude (Tables D18 and D19) as the ground truth, while models not using this information tend to overestimate. Then, Figure 26 shows that models trained with population data have the greatest variability. Interestingly, they tend to attribute an inferior capacity to false positives, most of them probably found in less inhabited areas with smaller stations.

Moreover, in the map plots (Figure E2-E5), we can observe the mapping of this capacity and verify that models trained with the Dynamic World dataset better predict the capacities of plants near rivers and other water sources. Conversely, models trained with population data better estimate capacities in bigger areas like Barcelona. Finally, models trained without supplementary data show no variability in the predicted capacity, implying that the model cannot infer capacity from the data size and morphology. That is the most probable reason why the model attributes positive capacities to images where it has not segmented any pixel. This fact would invalidate one of our hypotheses, but it can be most probably due to the dense network's inability to interpret the encoded data passed by the convolutional branch.

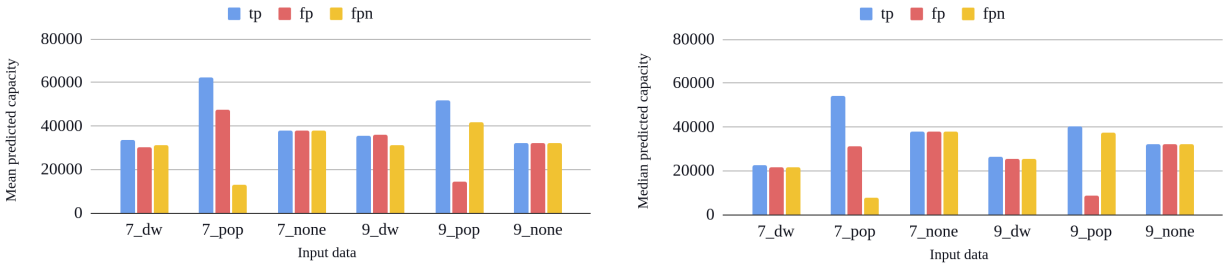

Figure 26: Mean of the averages (left) and mean of the medians (right) of the inferred capacities for true positives (tp), false positives (fp), and false positives in the same image as the ground truth (fpn). In the horizontal axes, the different data using to test the model. The number indicates the instances per WWTP and tile, and the suffix points to the additional data.

8. Conclusions

The deployment of an automatic tool that successfully detects and characterizes wastewater treatment plants could be a groundbreaking achievement due to the potential impact on the skill to map methane emission sources and develop effective decarbonization policies. Moreover, it also has obvious applications for estimating water supply conditions and capacity. However, as shown in the literature (Zhu et al., 2022) it is a state-of-the-art task posing many obstacles due to the variety of morphologies and sizes and their usual proximity to other human infrastructures. Moreover, most plants have small infrastructures whose salient features fade out in medium and small-resolution images.

In this work, we proposed a novel architecture to segment WWTPs and regress their capacity in one shot with a dual-branch deep neural network. We connected these network branches under the hypothesis that visual features help to infer capacity. We also hypothesize that additional land-cover classification data such as water bodies or topographic information can help locate these infrastructures. In the same manner, we hypothesize that water bodies and population data can help estimate the capacity of the plants.

For the validation of the architecture and the hypotheses, we developed a model and a dataset in parallel, with a series of ablation tests. The process was facilitated by an automatic pipeline, developed also in this work, to create datasets using various sources and train and validate our models using those datasets. That allowed for evaluating a significant number of possibilities, including variations in the dataset, architecture choices, and hyperparameters.

Results show that we can locate more than 50% of the labeled plants with slightly more than 50% precision only using orthophotos from multispectral imagery. To be more precise, using the NIR band showed to be of great value. However, the model's performance decreases when transferring our model to a slightly different scenario, with seen and unseen infrastructures. Not only is there a clear loss of skill when transferring the model to this domain, but also the spatial resolution and the ubiquity of human infrastructures trigger many false positives. Finally, the model cannot exploit additional raster information to improve results (maybe because it does not contribute much to the Sentinel-2 bands, specially NIR) and the regression branch cannot take advantage of the receptive fields passed by the convolutional encoder.

However, we demonstrate that additional data sources, such as the magnitude and proximity to water basins and population are valuable resources that the model can exploit to infer capacity in parallel with the segmentation of the WWTP contour. The segmentation result can be leveraged as a mask to improve the prediction of the total domain capacity, which in the case of the models trained with water bodies or

population data is in the same order of magnitude that the ground truth. Moreover, we also demonstrated the benefit of using negative instances to train our model and the impact of using an appropriate loss function such as Dice's loss.

These promising results pave the way to different research lines to improve the model skill. For example, we can improve the passing of information between the two branches or add nonlinearities in the regression branches to combine the information in a more effective way. Secondly, we can jointly combine water bodies and population information to regress capacity. In the third place, we could use the elevation (Figure B2) at the macro-level and train a model to post-process data and remove WWTPs in less likely locations.

Following the recent literature, there are also other potential ways to explore, like using novel benchmarks (Zhu et al., 2022) to pretrain our model before tuning it for our own purpose, using images of different resolutions or sizes (Marcu & Lordeanu, 2016; Chen et al., 2018; Li et al., 2022), leveraging OpenStreetMap labels for wastewater basins (Li et al., 2022), learn spacio-temporal priors (Mac Aodha et al., 2019) or, finally, test architectural changes such as atrous convolutions (Sherrah, 2016) or attention modules (Li et al., 2022).

9. Bibliography

Aleissaee, A. A., Kumar, A., Anwer, R. M., Khan, S., Cholakkal, H., Xia, G.-S., & Khan, F. S. (2022). Transformers in Remote Sensing: A Survey. *arXiv*. https://doi.org/10.48550/arxiv.2209.01206

Audebert, N., Le Saux, B., & Lefèvre, S. (2017). Beyond RGB: Very High Resolution Urban Remote Sensing With Multimodal Deep Networks. *arXiv*. https://doi.org/10.48550/arxiv.1711.08681,

Bondarenko, M., Kerr, D., Sorichetta, A., & Tatem, A. J. (2020). Census/projection-disaggregated gridded population datasets for 189 countries in 2020 using Built-Settlement Growth Model (BSGM) outputs. *WorldPop, University of Southampton, UK*. 10.5258/SOTON/WP00684

Bressler, R. D. (2021). The mortality cost of carbon. *Nature Communications*, *12*(4467 (2021)). https://doi.org/10.1038/s41467-021-24487-w

Brown, C. F., Brumby, S. P., & Guzder-Williams, B. (2022). Dynamic World, Near real-time global 10 m land use land cover mapping. *Sci Data*, *9*, 251. https://doi.org/10.1038/s41597-022-01307-4

Chen, G., Zhang, X., Wang, Q., Dai, F., Gong, Y., & Zhu, K. (2018). Symmetrical Dense-Shortcut Deep Fully Convolutional Networks for Semantic Segmentation of Very-High-Resolution Remote Sensing Images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *11*(5), 1633-1644. https://doi.org/10.1109/JSTARS.2018.2810320

Climate & Clean Air Coalition Secretariat. (2021). Global Methane Pledge. Retrieved October, 2022, from https://www.globalmethanepledge.org/

Directorate-General for Environment (DG ENV), & European Environment Agency (EEA). (2020). *Waterbase - UWWTD: Urban Waste Water Treatment Directive*. Datasets stored in reportnet by member state reporters. Version 8. Retrieved March, 2022, from https://www.eea.europa.eu/data-and-maps/data/waterbase-uwwtd-urban-waste-water-treatment-directive-7

Ehalt Macedo, H., Lehner, B., Nicell, J., Jim and Grill, G., Li, J., Limtong, A., & Shakya, R. (2021). Global distribution of wastewater treatment plants and their released effluents into rivers and streams. http://dx.doi.org/10.5194/essd-2021-214

European Space Agency - ESA. (2022). *Sentinel-2A*. Copernicus Open Access Hub. Retrieved 03, 2022, from https://scihub.copernicus.eu/

Eyring, V., Gillett, N. P., Achutarao, K. M., & Barimalala, R. (2021). Chapter 3: Human influence on the climate system. *IPCC AR6 WG1 2021*.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press. http://www.deeplearningbook.org

Hazirbas, C., Ma, L., Domokos, C., & Cremers, D. (2016). FuseNet: Incorporating Depth into Semantic Segmentation via Fusion-Based CNN Architecture. *ACCV*.

Hu, J., Mou, L., Schmitt, A., & Zhu, X. X. (2017). FusioNet: A two-stream convolutional neural network for urban scene classification using PolSAR and hyperspectral data. *2017 Joint Urban Remote Sensing Event (JURSE)*, 1-4. https://doi.org/10.1109/JURSE.2017.7924565

Institut Cartogràfic i Geològic de Catalunya. (n.d.). *Ortoimágenes Sentinel-2 mensuales. Institut Cartogràfic i Geològic de Catalunya*. ICGC. Retrieved August, 2022, from https://www.icgc.cat/es/Descargas/Imagenes-aereas-y-de-satelite/Ortoimagenes-Sentinel-2-mensuales

Instituto Geográfico Nacional. (2022). *Iberpix*. Retrieved February, 2022, from https://www.ign.es/iberpix2/visor/

Lagrange, A., Le Saux, B., Beaupère, A., Boulch, A., Chan-Hon-Tong, A., Herbin, S., Randrianarivo, H., & Ferecatu, M. (2015). Benchmarking classification of earth-observation data: From learning explicit features to convolutional networks. *2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, 4173-4176. https://doi.org/10.1109/IGARSS.2015.7326745

LeCun, Y. (1989). Generalization and network design strategies. *CRG-TR-89-4, University of Toronto*, 330-352.

Li, H., Zech, J., Hong, D., Ghamisi, P., Schultz, M., & Zipf, A. (2022). Leveraging OpenStreetMap and Multimodal Remote Sensing Data with Joint Deep Learning for Wastewater Treatment Plants Detection. *International Journal of Applied Earth Observation and Geoinformation*, *110*, 102804. https://doi.org/10.1016/j.jag.2022.102804

Liu, Y., Minh Nguyen, D., Deligiannis, N., Ding, W., & Munteanu, A. (2017). Hourglass-shapenetwork based semantic segmentation for high resolution aerial imagery. *Remote Sensing*, *9*(6), 522. https://doi.org/10.3390/rs9060522

Long, J., Shelhamer, E., & Darrell, T. (2014). Fully Convolutional Networks for Semantic Segmentation. *arXiv*. https://doi.org/10.48550/arxiv.1411.4038

Ma, L., Liu, Y., Zhang, X., Ye, Y., Yin, G., & Johnson, B. A. (2019). Deep learning in remote sensing applications: A meta-analysis and review. *ISPRS Journal of Photogrammetry and Remote Sensing*, *152*, 166-167. https://doi.org/10.1016/j.isprsjprs.2019.04.015

Mac Aodha, O., Cole, E., & Perona, P. (2019). Presence-Only Geographical Priors for Fine-Grained Image Classification. *CoRR*, *abs/1906.05272*. https://doi.org/10.48550/arXiv.1906.05272

Marcu, A., & Leordeanu, M. (2016). Dual Local-Global Contextual Pathways for Recognition in Aerial Imagery. *arXiv*. https://doi.org/10.48550/arxiv.1605.05462

Marmanis, D., Datcu, M., Esch, T., & Stilla, U. (2016). Deep Learning Earth Observation Classification Using ImageNet Pretrained Networks. *IEEE Geoscience and Remote Sensing Letters*, *13*(1), 105-109. 10.1109/LGRS.2015.2499239

Milletari, F., Nassir, N., & Seyed-Ahmad, A. (2016). V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. *Fourth International Conference on 3D Vision (3DV)*, *2016*, 565.571. 10.1109/3DV.2016.79

NASA/METI/AIST/Japan Spacesystems and U.S./Japan ASTER Science Team. (2019). *ASTER Global Digital Elevation Model V003 [Data set]*. NASA EOSDIS Land Processes DAAC. Retrieved 09 20, 2022, from https://doi.org/10.5067/ASTER/ASTGTM.003

NASA/METI/AIST/Japan Spacesystems and U.S./Japan ASTER Science Team. (2019). *ASTER Global Water Bodies Database V001 [Data set]*. NASA EOSDIS Land Processes DAAC. Retrieved 09 20, 2022, from https://doi.org/10.5067/ASTER/ASTWBD.001

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. *arXiv*. https://doi.org/10.48550/arxiv.1505.04597

Schiavina, M., Freire, S., & MacManus, K. (2019). GHS-POP R2019A - GHS population grid multitemporal (1975-1990-2000-2015). *European Commission, Joint Research Centre (JRC)*. 10.2905/0C6B9751-A71F-4062-830B-43C9F432370F

Sherrah, J. (2016). Fully Convolutional Networks for Dense Semantic Labelling of High-Resolution Aerial Imagery. *arXiv*. https://doi.org/10.48550/arxiv.1606.02585

Townshend, J. (2016). *Global Forest Cover Change (GFCC) Tree Cover Multi-Year Global 30 m V003 [Data set]*. NASA EOSDIS Land Processes DAAC. Retrieved 09 20, 2022, from https://doi.org/10.5067/MEaSUREs/GFCC/GFCC30TC.003

U.S. Environmental Protection Agency. (2012). *Global Anthropogenic Non-CO2 Greenhouse Gas Emissions: 1990-2030*. EPA 430-R-12-006. https://www.epa.gov/global-mitigation-non-co2-greenhouse-gases/global-non-co2-ghg-emissions-1990-2030

Volpi, M., & Tuia, D. (2017). Dense Semantic Labeling of Subdecimeter Resolution Images With Convolutional Neural Networks. *{IEEE} Transactions on Geoscience and Remote Sensing*, *55*(2), 881-893. https://doi.org/10.48550/arXiv.1608.00775

Wang, H., Wang, Y., Zhang, Q., Xiang, S., & Pan, C. (2017). Gated Convolutional Neural Network for Semantic Segmentation in High-Resolution Images. *Remote Sensing*, *9*(5), 446. https://doi.org/10.3390/rs9050446

Zhang, H., & Deng, Q. (2019). Deep Learning Based Fossil-Fuel Power Plant Monitoring in High Resolution Remote Sensing Images: A Comparative Study. *Remote Sensing*, *11*(9), 1117. http://dx.doi.org/10.3390/rs11091117

Zhu, B., Lui, N., Irvin, J., Le, J., Tadwalkar, S., Wang, C., Ouyang, Z., Liu, F. Y., Ng, A. Y., & Jackson, R. B. (2022). METER-ML: A Multi-Sensor Earth Observation Benchmark for Automated Methane Source Mapping. *arXiv*. https://doi.org/10.48550/arxiv.2207.11166

Annex A:

Figure A1: WWTP infrastructures vary wildly in size and morphology, depending on the underlying technology and the type of treatment they perform. However, they generally stand out for the circular or rectangular structures of settlers, tanks, and trickling filters. These very high-resolution orthophotos from the Spanish National Geographic Institute (IGN) are displayed at the same scale to make their differences more evident.

Annex B:

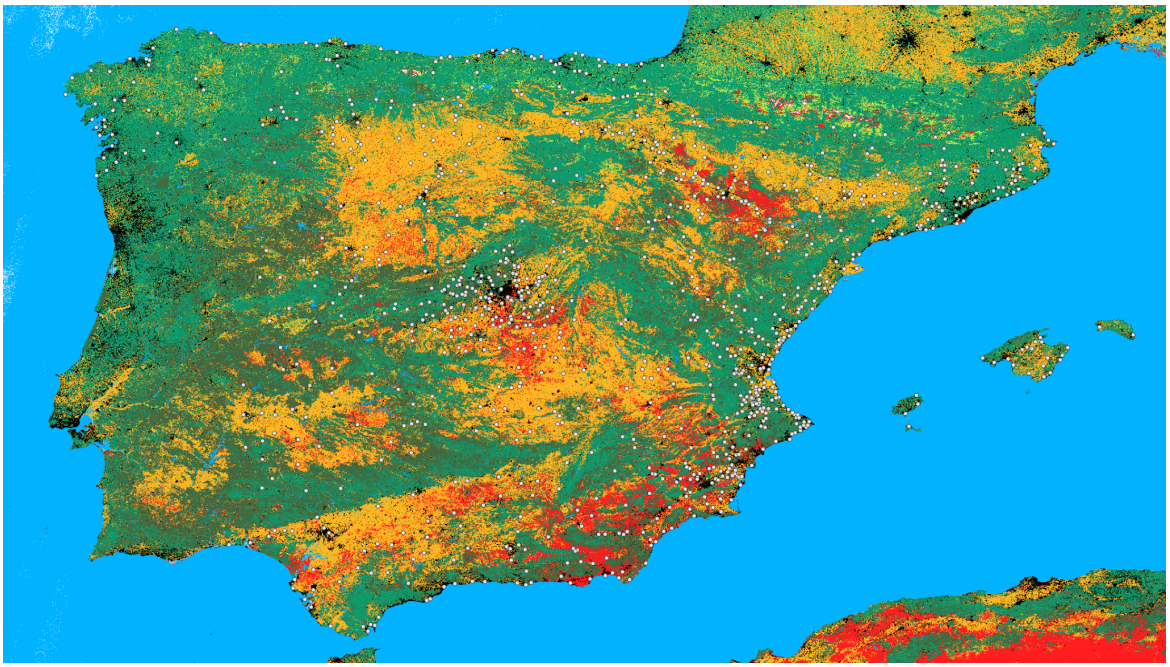

Figure B1: Map of the labeled WWTPs over the Dynamic World composite. Here, we show the nine land-use classes, although we usually only employ the water (0-class) one as a mask.

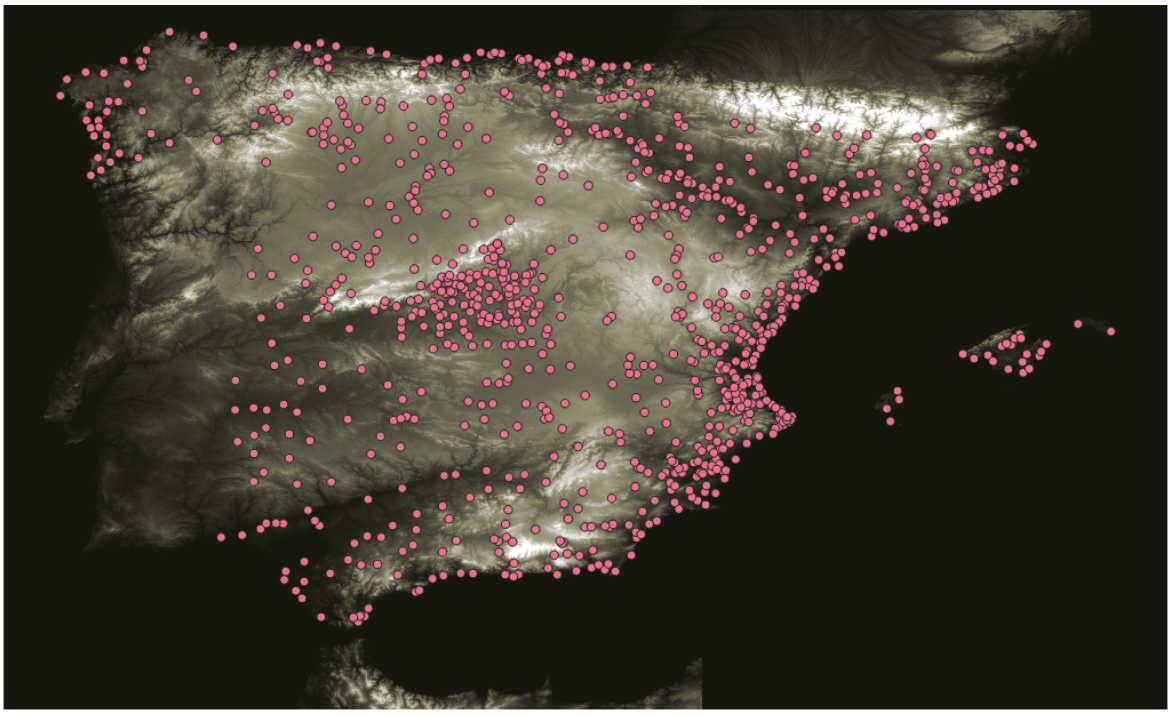

Figure B2: Map of the labeled WWTPs over the ASTER Digital Elevation Model. We can observe that plants are usually located in valleys or the sea-side.

Annex C:

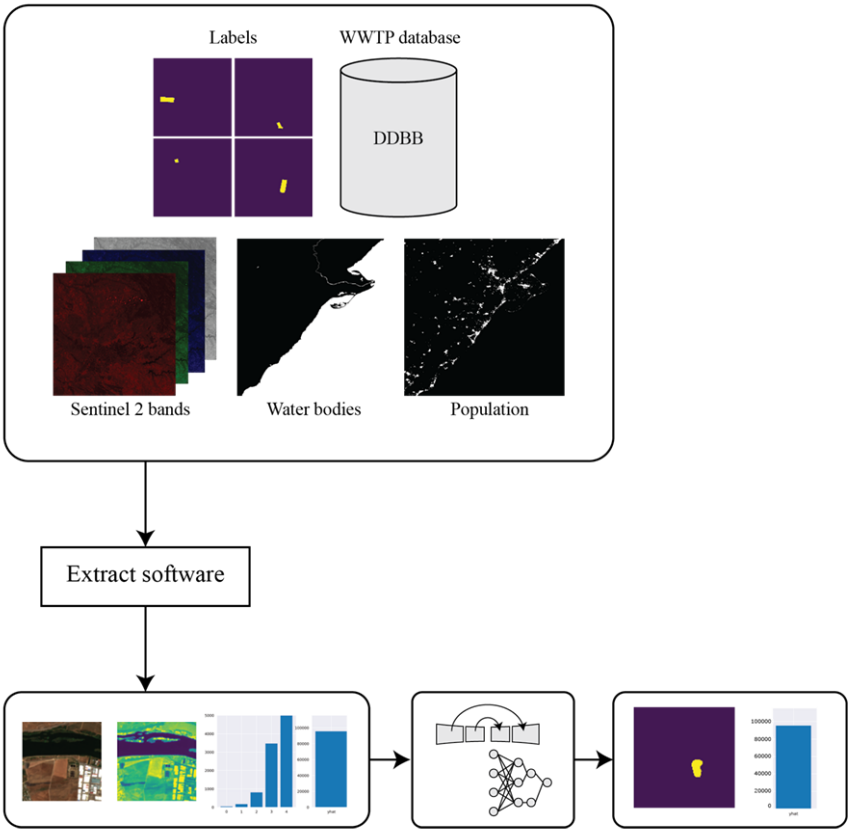

Figure C1: Data extraction and training pipeline. The extract software creates datasets from Sentinel-2 multispectral imagery, a set of labels, a WWTP database, and, optionally, additional data sources. The model uses the multidimensional raster and vector data dataset to learn to predict segmentation labels and capacity.

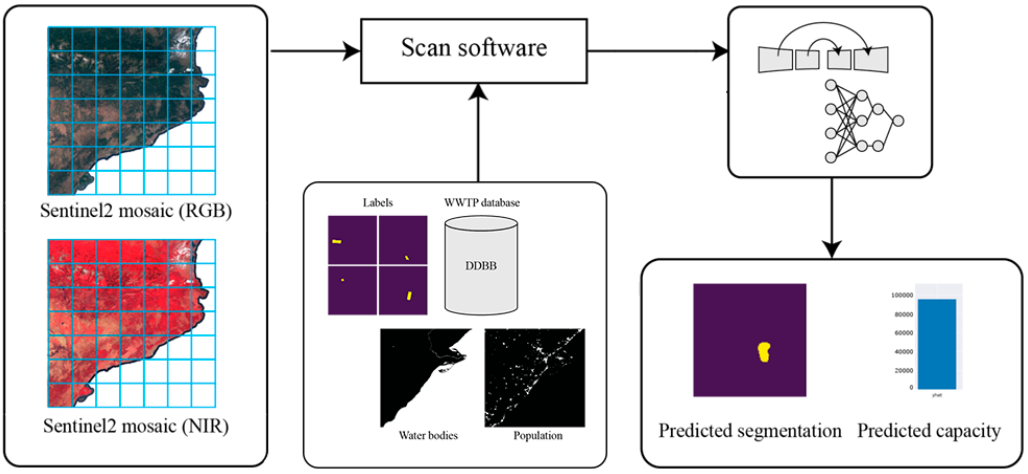

Figure C2: Scanning pipeline. The scan software sweeps the mosaic, combining RGB and NIR images and additional data, and inputs that data to a trained model to obtain predictions.

Annex D:

Dataset (max shift)	2,000	4,000	8,000	16,000
RGB - centered	0.28 / 0.35	0.28 / 0.34	0.28 / 0.34	0.27 / 0.32
RGB+NIR - centered	0.33 / 0.41	0.37 / 0.53	0.35 / 0.47	0.34 / 0.49
RGB (50 px)	0.12 / 0.17	0.1 / 0.13	0.1 / 0.12	0.11 / 0.14
RGB+NIR (50 px)	0.19 / 0.3	0.24 / 0.33	0.23 / 0.32	0.22 / 0.31

Table D1: Impact on IoU / recall of different cutoff values for image rescaling, for datasets made of RGB and RGB+NIR images, centered or with shifts of up to 50 px. Using cross-entropy loss.

Dataset (max shift)	2.000		4.000		8.000		16.000		22.560
	8 bit	16 bit	8 bit	16 bit	8 bit	16 bit	8 bit	16 bit	16 bit
RGB centered	0.28 0.35	0.26 0.36	0.28 0.34	0.27 0.37	0.28 0.34	0.28 0.35	0.27 0.32	0.26 0.34	0.25 0.37
RGB+NIR centered	0.33 0.41	0.34 0.46	0.37 0.53	0.37 0.47	0.35 0.47	0.35 0.44	0.34 0.49	0.34 0.45	0.38 0.49
RGB (50 px)	0.12 0.17	0.13 0.17	0.1 0.13	0.11 0.16	0.1 0.12	0.1 0.13	0.11 0.14	0.09 0.12	0.07 0.05
RGB+NIR (50 px)	0.19 0.3	0.19 0.28	0.24 0.33	0.21 0.32	0.23 0.32	0.23 0.3	0.22 0.31	0.23 0.3	0.21 0.32

Table D2: Impact on IoU / recall of the different cutoff values for image rescaling, and image depth, for datasets made of RGB and RGB+NIR images, centered or with shifts of up to 50 px. Using cross-entropy loss.

Dataset (max shift x nº shifts)	No water-bodies		ASTER		GFCC30	
	4000 NPY	22560 NPY	4000 NPY	22560 NPY	4000 NPY	22560 NPY
RGB+NIR (50 px)	0.24 / 0.3	0.23 / 0.34	0.27 / 0.39	0.26 / 0.38	0.23 / 0.34	0.21 / 0.32
RGB+NIR (96 px)	0.1 / 0.17	0.1 / 0.17	0.14 / 0.3	0.12 / 0.22	0.11 / 0.23	0.09 / 0.18
RGB+NIR (96px x2)	0.15 / 0.24	0.12 / 0.2	0.19 / 0.37	0.19 / 0.36	0.16 / 0.26	0.13 / 0.23
RGB+NIR (96 px x3)	0.18 / 0.29	0.17 / 0.29	0.19 / 0.33	0.2 / 0.3	0.17 / 0.27	0.15 / 0.25

Table D3: Impact on IoU / recall of the different cutoff values for image rescaling, and image depth, for datasets made of 4-d RGB+NIR and 5-d RGB+NIR+water-bodies images, with shifts of up to 50 and 96 px, and additional (x2, x3) shifts to increase the number of negative instances (as image size is 128px, the edge is 64px from center). Using cross-entropy loss.

Dataset (max shift x nº shifts)	No population		100m		250m		1000m	
	4000 NPY	22560 NPY	4000 NPY	22560 NPY	4000 NPY	22560 NPY	4000 NPY	22560 NPY
RGB+NIR (50)	0.24 0.3	0.23 0.34	0.24 0.35	0.23 0.32	0.22 0.35	0.22 0.33	0.21 0.29	0.23 0.32
RGB+NIR (96)	0.1 0.17	0.1 0.17	0.12 0.26	0.08 0.13	0.1 0.18	0.09 0.19	0.07 0.13	0.1 0.2
RGB+NIR (96x2)	0.15 0.24	0.12 0.2	0.13 0.22	0.15 0.26	0.15 0.25	0.14 0.24	0.12 0.21	0.16 0.26
RGB NIR (96x3)	0.18 0.29	0.17 0.29	0.17 0.3	0.15 0.23	0.17 0.29	0.15 0.25	0.18 0.31	0.15 0.24

Table D4: Impact on IoU / recall of the different cutoff values for image rescaling, and image depth, for datasets made of 4-d RGB+NIR and 5-d RGB+NIR+population images, with shifts of up to 50 and 96 px, and additional (x2, x3) shifts to increase the number of negative instances (as image size is 128px, the edge is 64px from center). Using cross-entropy loss.

Dataset (max shift x nº shifts) & model	No water-bodies	ASTER		GFCC30		Dynamic World	
		Mask	Full	Mask	Full	Mask	Full
RGB+NIR (50,96,192)	0.09 0.33 0.48	0.09 0.43 0.42	0.1 0.42 0.45	0.09 0.3 0.38	0.09 0.3 0.38	0.09 0.3 0.35	0.05 0.21 0.31
RGB+NIR (50,96,192) znorm	0.1 0.3 0.47	0.09 0.47 0.34	0.09 0.34 0.48	0.1 0.32 0.39	0.1 0.28 0.4	0.1 0.33 0.35	0.08 0.28 0.4
RGB+NIR (50,96,192) + 1 layer	0.08 0.38 0.41	0.12 0.46 0.46	0.11 0.39 0.47	0.1 0.39 0.38	0.09 0.3 0.37	0.09 0.26 0.34	0.08 0.32 0.37
RGB+NIR (50,192x4)	0.08 0.28 0.46	0.1 0.36 0.43	0.07 0.38 0.4	0.08 0.27 0.41	0.09 0.31 0.45	0.07 0.19 0.38	0.07 0.26 0.45
RGB+NIR (50,192x4) znorm	0.08 0.28 0.5	0.09 0.32 0.44	0.07 0.3 0.42	0.08 0.25 0.4	0.08 0.24 0.44	0.08 0.28 0.35	0.08 0.28 0.48
RGB+NIR (50,192x4) + 1 layer	0.09 0.36 0.4	0.1 0.33 0.46	0.08 0.32 0.42	0.08 0.23 0.4	0.1 0.32 0.46	0.07 0.21 0.43	0.07 0.25 0.47

Table D5: Impact of stacking a water-bodies (5-d RGB+NIR+water bodies) channel, standardizing on training, and increasing the model depth. Mask: channel as a binary mask. Full: raw values. Values: IoU, recall, and precision. Using cross-entropy loss.

Dataset (max shift x nº shifts) & model	Cross-entropy	Dice's
RGB+NIR (50,96,192)	0.11 / 0.38 / 0.46	0.12 / 0.17 / 0.13
+ znorm	0.11 / 0.41 / 0.45	0.1 / 0.49 / 0.14
+ znorm + layer	0.11 / 0.39 / 0.44	0.12 / 0 / 0
+ znorm + layer + bnorm	0.12 / 0.44 / 0.51	0.16 / 0.52 / 0.39
RGB+NIR (50,192x4)	0.11 / 0.33 / 0.47	0.12 / 0 / 0
+ znorm	0.12 / 0.37 / 0.53	0.15 / 0.15 / 0.2
+ znorm + layer	0.11 / 0.38 / 0.51	0.12 / 0 / 0
+ znorm + layer + bnorm	0.12 / 0.43 / 0.46	0.18 / 0.54 / 0.43
RGB+NIR (50,128x8)	0.12 / 0.33 / 0.6	0.12 / 0 / 0
+ znorm	0.12 / 0.37 / 0.56	0.12 / 0 / 0
+ znorm + layer	0.13 / 0.34 / 0.59	0.12 / 0 / 0
+ znorm + layer + bnorm	0.12 / 0.37 / 0.61	0.18 / 0.5 / 0.53

Table D6: Impact of the loss function on models trained with different datasets and accumulative architecture modifications (z-score normalization, additional conv. layer, batch normalization). Enclosed between parenthesis: number of shifts. Values: IoU, recall, and precision.

Dataset (max shift x nº shifts) & model - Dice's'	No WB data
RGB+NIR (50,96,192)	0.10 / 0.42 / 0.36
+ znorm	0.12 / 0.44 / 0.41
+ znorm + layer	0.12 / 0.50 / 0.40
+ znorm + layer + bnorm	0.09 / 0.48 / 0.47

Table D7: Impact of using the alternative Dice's loss (with squared terms in the denominator) on models trained with different accumulative architecture modifications (z-score normalization, additional conv. layer, batch normalization). Enclosed between parenthesis: number of shifts. Values: IoU, recall, and precision.

Dataset (max shift x nº shifts) & model - Dice's	No WB data
RGB+NIR (50,96,192)	0.12 / 0.17 / 0.13
+ znorm + layer + bnorm	0.17 / 0.56 / 0.38
+ znorm + bnorm	0.14 / 0.42 / 0.45
+ znorm + layer + bnorm + batch_16	0.15 / 0.46 / 0.47
+ znorm + layer + bnorm (modified)	0.17 / 0.50 / 0.47

Table D8: Impact of using different accumulative architecture modifications (z-score normalization, additional conv. layer, batch normalization, batch normalization positioned after the layer's activation function) and change in the batch size. Enclosed between parenthesis: number of shifts. Values: IoU, recall, and precision.

Dataset (max shift x nº shifts) & model - Dice's	No elevation data	Elevation data
RGB+NIR (50,192x4)	0.19 / 0.51 / 0.50	0.17 / 0.52 / 0.46
RGB+NIR (50,128x8)	0.19 / 0.53 / 0.49	0.19 / 0.56 / 0.38

Table D9: Impact of stacking an elevation channel (5-d RGB+NIR+Elev) on the dataset images (4-d RGB+NIR). Enclosed between parenthesis: number of shifts. Values: IoU, recall, and precision.

Data (shifts, nº shifts) / model	No water-bodies	ASTER		GFCC30		Dynamic World	
		Mask	Full	Mask	Full	Mask	Full
RGB+NIR (50,192x4)	0.19 0.51 0.50	0.19 0.52 0.44	0.16 0.59 0.38	0.19 0.50 0.54	0.18 0.51 0.47	0.17 0.52 0.46	0.17 0.51 0.44
RGB+NIR+Elev. (50,192x4)	0.17 0.52 0.46	0.18 0.53 0.47	0.19 0.52 0.51	0.18 0.52 0.46	0.18 0.51 0.46	0.18 0.53 0.46	0.19 0.52 0.44
RGB+NIR (50,128x8)	0.19 0.53 0.49	0.19 0.52 0.52	0.17 0.57 0.39	0.19 0.55 0.47	0.19 0.54 0.47	0.18 0.52 0.49	0.19 0.51 0.44
RGB+NIR+Elev (50,128x8)	0.19 0.56 0.38	0.17 0.53 0.48	0.18 0.56 0.49	0.18 0.54 0.46	0.18 0.48 0.51	0.19 0.54 0.47	0.19 0.53 0.48

Table D10: Impact of stacking water-bodies, and elevation data (5-d RGB+NIR+water-bodies, 5-d RGB+NIR+elevation, 6-d RGB+NIR+water-bodies+elevation) on top of 4-d RGB+NIR images. Mask: channel as a binary mask. Full: raw values. Values: IoU, recall, and precision

#I	#N	Shift	NR	IoU / recall / prec.
1	4k	50	0	0.15 / 0.48 / 0.24
2	8k	50	0	0.16 / 0.45 / 0.32
3	12k	50	0	0.16 / 0.42 / 0.30
4	16k	50	0	0.17 / 0.50 / 0.29
5	20k	50	0	0.17 / 0.52 / 0.30
6	24k	50	0	0.18 / 0.43 / 0.33

Table D11: Impact of the training dataset size on the test results. I: number of instances; N: number of images. NR: negative instances ratio.

#I	#N	Shift	NR	IoU / recall / prec.
4	16k	50-96	0.25	0.16 / 0.46 / 0.34
6	24k	50-96	0.25	0.18 / 0.44 / 0.38
8	32k	50-96	0.25	0.16 / 0.44 / 0.32
10	40k	50-96	0.25	0.18 / 0.43 / 0.39
12	48k	50-96	0.25	0.16 / 0.44 / 0.30
6	24k	50-192	0.44	0.17 / 0.42 / 0.43
9	36k	50-192	0.44	0.17 / 0.43 / 0.44
12	48k	50-192	0.44	0.18 / 0.40 / 0.43

Table D12: Impact of the training dataset size and negative instances ratio on the test results. I: number of instances; N: number of images; NR: negative instances ratio.

#I	#N	Shift	NR	IoU / recall / prec.
2	8k	50-192	0.44	0.17 / 0.43 / 0.40
3	12k	50-192	0.57	0.17 / 0.39 / 0.41
4	16k	50-192	0.65	0.15 / 0.42 / 0.35
5	20k	50-192	0.69	0.17 / 0.41 / 0.48
6	24k	50-192	0.72	0.18 / 0.42 / 0.43
7	28k	50-192	0.74	0.15 / 0.41 / 0.41

Table D13: Impact of the training dataset size and negative instances ratio on the test results. I: number of instances; N: number of images; NR: negative instances ratio.

#I	#N	Shift	NR	IoU / recall / prec.
3	12k	50-192	0.29	0.17 / 0.45 / 0.37
4	16k	50-192	0.43	0.17 / 0.44 / 0.41
5	20k	50-192	0.51	0.18 / 0.45 / 0.45
6	24k	50-192	0.58	0.18 / 0.47 / 0.42
7	28k	50-192	0.62	0.17 / 0.40 / 0.43
8	32k	50-192	0.65	0.17 / 0.40 / 0.53
9	36k	50-192	0.67	0.18 / 0.40 / 0.46
10	40k	50-192	0.7	0.18 / 0.39 / 0.49
11	44k	50-192	0.71	0.13 / 0.50 / 0.30
12	48k	50-192	0.72	0.18 / 0.44 / 0.41

Table D14: Impact of the training dataset size and negative instances ratio on the test results. I: number of instances; N: number of images; NR: negative instances ratio.

Data & augment.	# N	NR	IoU / recall / precision
50,192x5	20k	0.51	0.20 / 0.54 / 0.43
+1 F	24k	0.43	0.21 / 0.57 / 0.39
+1 FR			0.21 / 0.60 / 0.36
+1 FZ			0.20 / 0.55 / 0.43
+1 FZR			0.21 / 0.59 / 0.38
+2 F	28k	0.37	0.21 / 0.57 / 0.43
+2 FR			0.21 / 0.53 / 0.44
+2 FZ			0.22 / 0.60 / 0.41
+2 FZR			0.21 / 0.60 / 0.41
+3 F	32k	0.43	0.21 / 0.63 / 0.35
+3 FR			0.22 / 0.58 / 0.41
+3 FZ			0.16 / 0.64 / 0.32
+3 FZR			0.22 / 0.62 / 0.41

Table D15: Impact of adding different number and types of augmentations. N: number of images; NR: negative instances ratio; F: flip; R: rotation; Z: zoom.

Input data	# N	NR	IoU / recall / precision
+4 F	36k	0.48	0.23 / 0.58 / 0.47
+4 FR			0.22 / 0.58 / 0.45
+4 FZ			0.23 / 0.58 / 0.44
+4 FZR			0.22 / 0.62 / 0.41
+5 F	40k	0.52	0.22 / 0.59 / 0.42
+5 FR			0.23 / 0.58 / 0.46
+5 FZ			0.23 / 0.60 / 0.46
+5 FZR			0.22 / 0.59 / 0.42

Table D16: Impact of adding different number and types of augmentations. N: number of images; NR: negative instances ratio; F: flip; R: rotation; Z: zoom.

Instances + augment. instances + vector data	# train images	Neg. ratio train.	IoU / recall / precision / MAE
5 + 2 (flip + rot)’	28k	0.37	0.23 / 0.50 / 0.44 / 0.019
5 + 2 (flip + rot)’’	28k	0.37	0.23 / 0.48 / 0.47 / 0.019
5 + 2 (flip + rot)’’’	28k	0.37	0.22 / 0.52 / 0.44 / 0.018
5 + 2 (flip + rot) + 100m	28k	0.37	0.23 / 0.49 / 0.47 / 0.016
5 + 4 (flip + rot) + 100m	36k	0.48	0.23 / 0.51 / 0.41 / 0.016
5 + 2 (flip + rot) + DW	28k	0.37	0.23 / 0.48 / 0.49 / 0.02
5 + 4 (flip + rot) + DW	36k	0.48	0.24 / 0.49 / 0.48 / 0.017
5 + 2 (flip + rot) + GFCC30	28k	0.37	0.23 / 0.47 / 0.47 / 0.016
5 + 4 (flip + rot) + GFCC30	36k	0.48	0.23 / 0.49 / 0.51 / 0.013

Table D17: Impact of the supplementary data (no data, WorldPop 100m, Dynamic World or GFCC30) on the capacity regression error (mean absolute error). All datasets have two instances shifted by up to 50 px (positive instance) and three instances shifted by up to 96 px (50% chance of a negative instance) plus two or four augmented instances (flips and rotations) per WWTP and tile.

Instances + augment. instances + vector data	TPR (labeled)	FPR (labeled)	TPR (all)	FPR (all)	Capacity
5 + 2 (flip + rot) + 100m	0.2	0.014	0.13	0.014	6.30E+07
5 + 4 (flip + rot) + 100m	0.11	0.007	0.08	0.007	1.90E+07
5 + 2 (flip + rot) + GFCC30	0.09	0.004	0.05	0.004	3.90E+07
5 + 4 (flip + rot) + GFCC30	0.09	0.005	0.07	0.004	4.20E+07
5 + 2 (flip + rot) + DW	0.22	0.017	0.16	0.017	1.30E+08
5 + 4 (flip + rot) + DW	0.17	0.014	0.11	0.014	8.50E+07
5 + 2 (flip + rot)’	0.17	0.12	0.01	0.01	8.3e+09
5 + 2 (flip + rot)’’	0.18	0.1	0.01	0.01	9.8e+07
5 + 2 (flip + rot)’’’	0.28	0.19	0.05	0.05	4.2e+08

Table D18: Impact of the supplementary data (no data, WorldPop 100m, Dynamic World or GFCC30) on the true positives, false positives and predicted capacity when scanning a mosaic of the Catalonia region, concerning the labeled WWTPs, and all WWTPs in the UWWTD dataset. All datasets have two instances shifted by up to 50 px (positive instance) and three instances shifted by up to 96 px (50% chance of a negative instance) plus two or four augmented instances (flips and rotations) per WWTP and tile. The model results from training over 80% of the dataset (20% kept for validation and testing). Ground truth capacity: 1.81e+07.

Instances + augment. instances + vector data	TPR (labeled)	FPR (labeled)	TPR (all)	FPR (all)	Capacity
5 + 2 (flip + rot) + 100m	0.19	0.02	0.12	0.02	1.00E+08
5 + 4 (flip + rot) + 100m	0.13	0.009	0.09	0.008	3.60E+07
5 + 2 (flip + rot) + GFCC30	0.22	0.07	0.15	0.07	6.30E+08
5 + 4 (flip + rot) + GFCC30	0.1	0.003	0.05	0.003	2.70E+07
5 + 2 (flip + rot) + DW	0.26	0.03	0.16	0.03	2.47E+08
5 + 4 (flip + rot) + DW	0.12	0.01	0.08	0.01	9.50E+07

Table D19: Impact of the supplementary data (no data, WorldPop 100m, Dynamic World or GFCC30) on the true positives, false positives and predicted capacity when scanning a mosaic of the Catalonia region, concerning the labeled WWTPs, and all WWTPs in the UWWTD dataset. All datasets have two instances shifted by up to 50 px (positive instance) and three instances shifted by up to 96 px (50% chance of a negative instance) plus two or four augmented instances (flips and rotations) per WWTP and tile. The model results from training over 90% of the dataset (10% kept for validation). Ground truth capacity: 1.81e+07.

Data + augment. / architecture	#Par.	No data	WoldPop 100m	Dynamic World
5+2 / 16,32,64,128,256,512	8M	0.23 / 0.50 / 0.44 / 0.019	0.23 / 0.49 / 0.47 / 0.016	0.23 / 0.47 / 0.47 / 0.016
5+2 / 32,64,128,256,512	10M	0.23 / 0.44 / 0.54 / 0.019	0.23 / 0.52 / 0.46 / 0.016	0.23 / 0.51 / 0.50 / 0.019
5+2 / 64,128,256,512	16M	0.25 / 0.49 / 0.52 / 0.018	0.26 / 0.51 / 0.56 / 0.018	0.24 / 0.49 / 0.53 / 0.014
5+2 / 64,128,256,512,512	22M	0.25 / 0.50 / 0.46 / 0.017	0.25 / 0.50 / 0.51 / 0.017	0.25 / 0.48 / 0.52 / 0.015
5+2 / 64,128,256,256,512,512	23M	0.26 / 0.55 / 0.49 / 0.019	0.24 / 0.54 / 0.46 / 0.024	0.25 / 0.50 / 0.46 / 0.016
5+2 / 128,256,256,512,512	24M	0.25 / 0.50 / 0.46 / 0.017	0.25 / 0.50 / 0.51 / 0.017	0.25 / 0.48 / 0.52 / 0.015
5+2 / 128,256,512,512	28M	0.25 / 0.50 / 0.54 / 0.015	0.25 / 0.53 / 0.50 / 0.016	0.26 / 0.47 / 0.56 / 0.017
5+2 / 128,256,512	41M	0.24 / 0.46 / 0.55 / 0.022	0.25 / 0.49 / 0.51 / 0.017	0.24 / 0.49 / 0.45 / 0.019
5+4 / 16,32,64,128,256,512	8M	0.24 / 0.47 / 0.50 / 0.016	0.23 / 0.51 / 0.41 / 0.016	0.23 / 0.49 / 0.51 / 0.013
5+4 / 64,128,256,512	22M	0.26 / 0.49 / 0.52 / 0.019	0.26 / 0.55 / 0.45 / 0.015	0.25 / 0.50 / 0.52 / 0.013
5+4 / 128,256,512,512	28M	0.25 / 0.50 / 0.55 / 0.015	0.24 / 0.50 / 0.48 / 0.016	0.28 / 0.56 / 0.56 / 0.018

Table D20: Impact of different architectures and supplementary data (no data, WorldPop 100m, Dynamic World or GFCC30) on the segmentation and capacity regression performance. All datasets have two instances shifted by up to 50 px (positive instance) and three instances shifted by up to 96 px (50% chance of a negative instance) plus two or four augmented instances (flips and rotations) per WWTP and tile. Architecture is given as a sequence of layers with values indicating the number of convolutional kernels. Results are given in segmentation IoU / segmentation recall / segmentation precision / regression mean absolute error.

Instances + augment. instances + vector data	TPR (labeled)	FPR (labeled)	TPR (all)	FPR (all)	Capacity
5+2 (DW) / 16,32,64,128,256,512	0.22	0.16	0.017	0.017	1.3e+8
5+2 (DW) / 64,128,256,512	0.09	0.04	0.001	0.001	1.9+e6
5+2 (DW) / 128,256,512,512	0.1	0.05	0.001	0.001	1.2e+7
5+2 (100m) / 16,32,64,128,256,512	0.2	0.13	0.014	0.014	6.3e+7
5+2 (100m) / 64,128,256,512	0.06	0.04	0.011	0.011	5.4e+6
5+2 (100m) / 128,256,512,512	0.11	0.07	0.002	0.001	8.2+6
5+2 (100m) / 16,32,64,128,256,512	0.17	0.12	0.01	0.01	8.3e+9
5+2 (100m) / 64,128,256,512	0.03	0.02	0.0004	0.0004	4e+6
5+2 (100m) / 128,256,512,512	0.05	0.03	0.0003	0.0003	3.5e+6
5+4 (DW) / 16,32,64,128,256,512	0.17	0.11	0.014	0.014	8.5e+7
5+4 (DW) / 64,128,256,512	0.02	0.010	0.0001	0.0001	1e+4
5+4 (DW) / 128,256,512,512	0.08	0.04	0.001	0.001	1.2e+7
5+4 (100m) / 16,32,64,128,256,512	0.11	0.08	0.007	0.007	1.9e+7
5+4 (100m) / 64,128,256,512	0.05	0.02	0.001	0.001	6.1e+6
5+4 (100m) / 128,256,512,512	0.03	0.01	0.002	0.002	1.1e+6
5+4 / 64,128,256,512	0.07	0.03	0.001	0.001	1.2e+7
5+4 / 128,256,512,512	0.1	0.08	0.006	0.006	5.7e+7

Table D21: Impact of the supplementary data (Dynamic World, WorldPop 100m) and architecture on the true positives, false positives and predicted capacity when scanning a mosaic of the Catalonia region, concerning the labeled WWTPs, and all WWTPs in the UWWTD dataset. All datasets have two instances shifted by up to 50 px (positive instance) and three instances shifted by up to 96 px (50% chance of a negative instance) plus two or four augmented instances (flips and rotations) per WWTP and tile. Architecture is given as a sequence of layers whose values indicate the number of conv. kernels. The model results from training over 80% of the dataset (20% kept for validation and testing). Ground truth capacity: 1.81e+07.

Annex E:

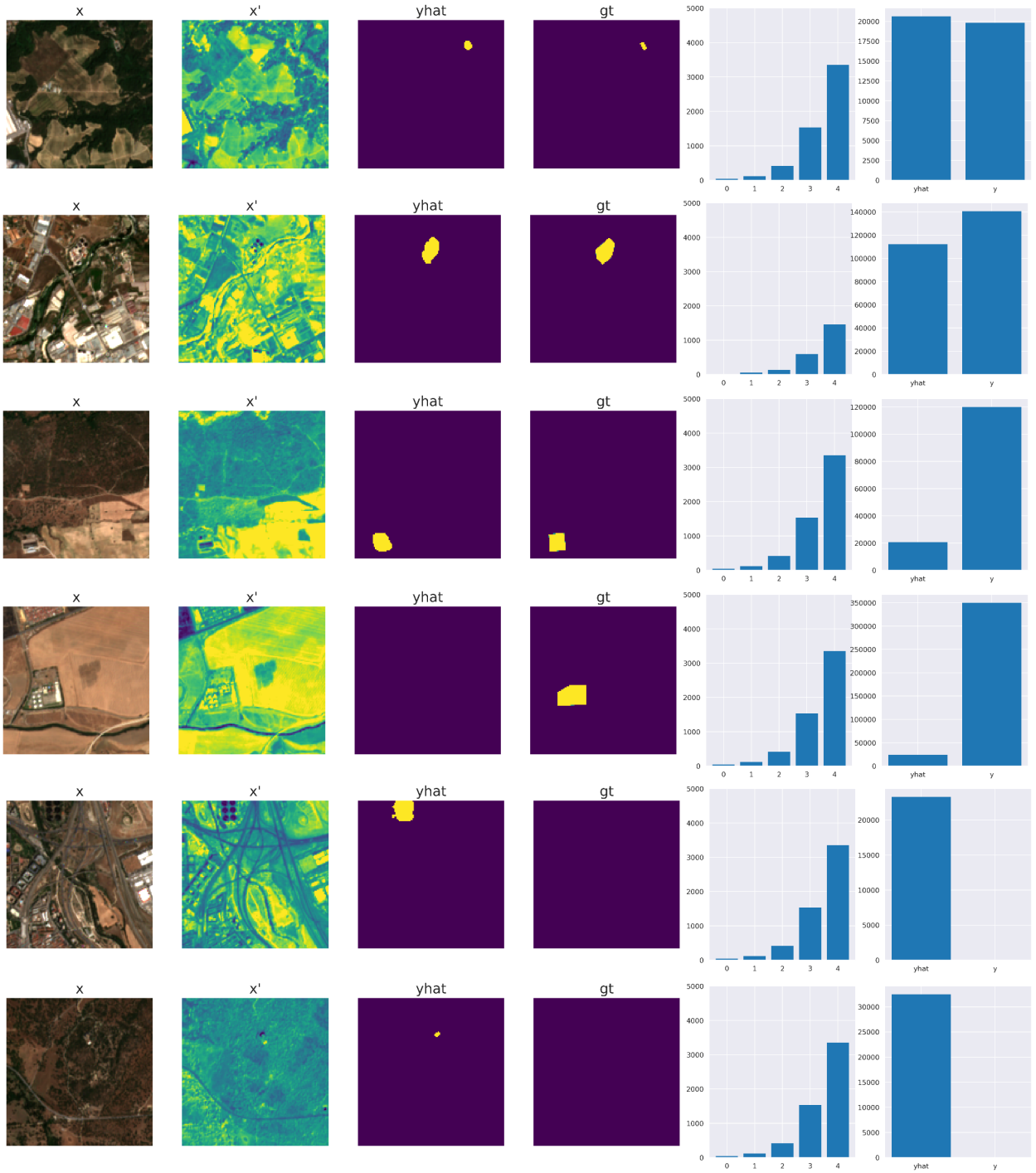

Figure E1: Predictions examples when using water-bodies or population data to infer capacity. From left to right: RGB input, NIR input, segmentation prediction, segmentation ground truth, vector input, capacity prediction and ground truth. On the two top rows, examples of true positives with fair capacity prediction; on the third row, a true positive with underestimation of the capacity; on the fourth row, a false negative; on the fifth and six rows, false positives.

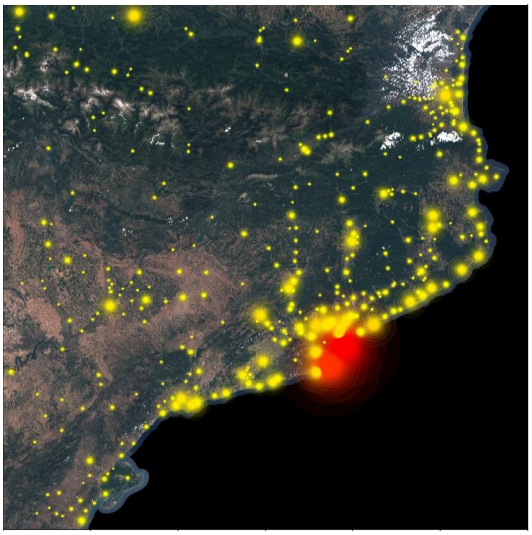

Figure E2: Mapping of WWTP over Catalonia. The color indicates absolute capacity (from yellow to red), while the size shows normalized capacity (concerning the rest of the displayed WWTPs).

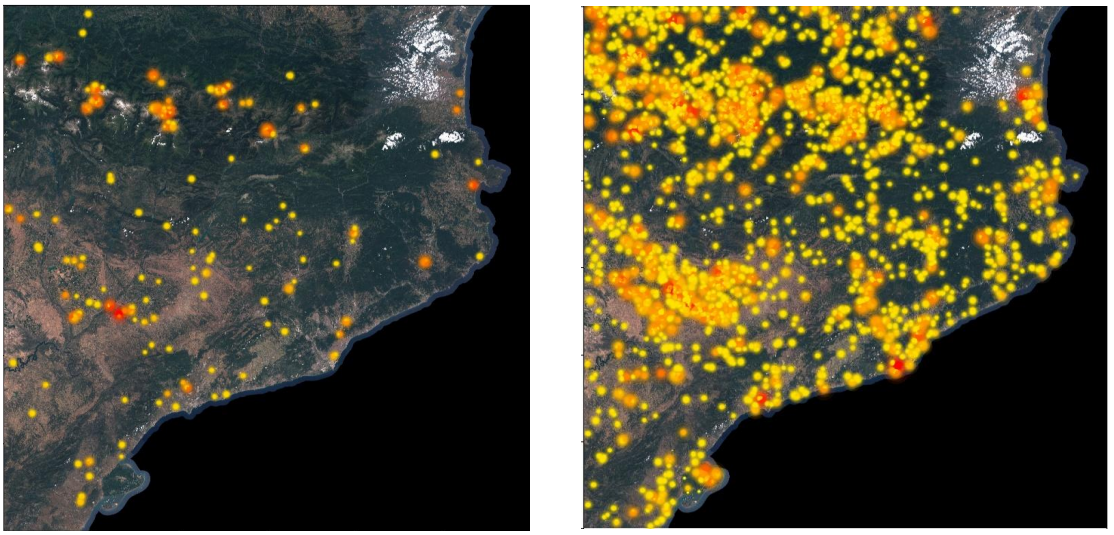

Figure E3: WWTP predictions mapped over Catalonia from models using water-bodies (Dynamic World) data. The color indicates absolute capacity (from yellow to red), while the size shows normalized capacity (concerning the rest of the displayed WWTPs). These models are better at inferring capacity over the river basins.

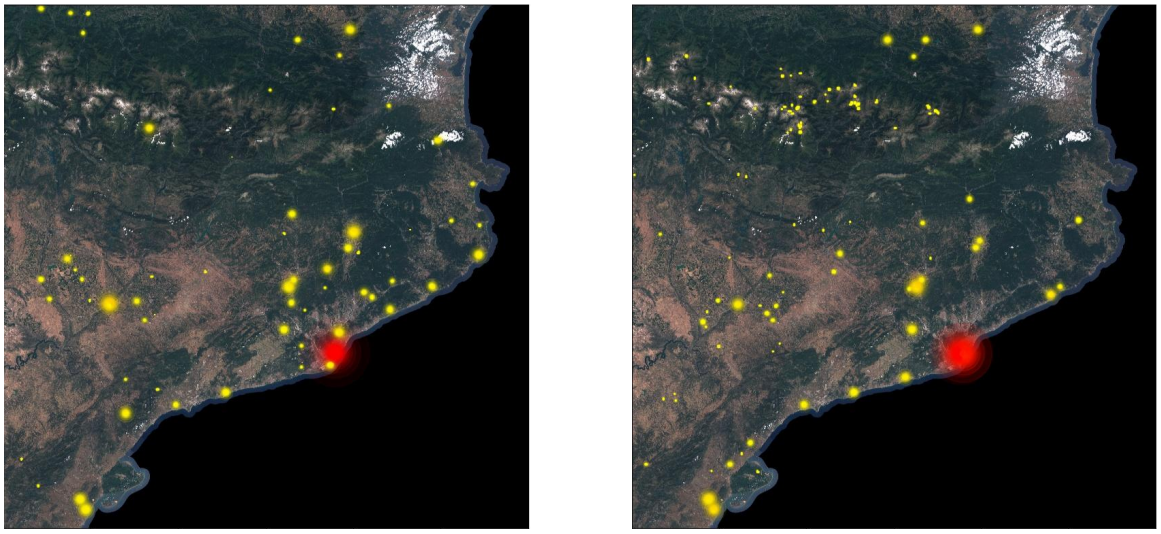

Figure E4: WWTP predictions mapped over Catalonia from models using population (WorldPop 100m) data. The color indicates absolute capacity (from yellow to red), while the size shows normalized capacity (concerning the rest of the displayed WWTPs). These models are better at inferring capacity in the surroundings of inhabited areas.

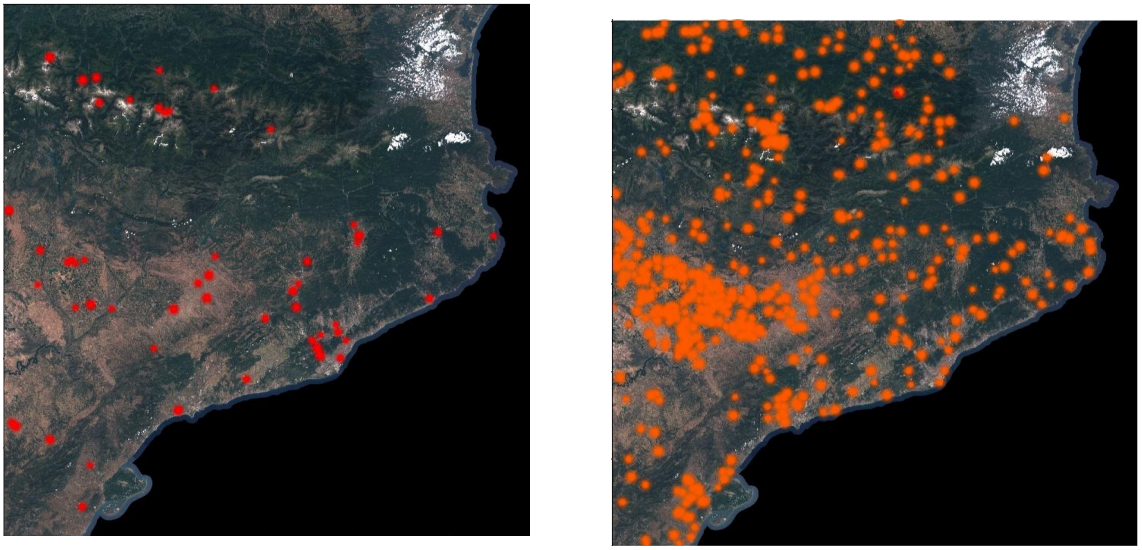

Figure E5: WWTP predictions mapped over Catalonia from models using no additional data. The color indicates absolute capacity (from yellow to red), while the size shows normalized capacity (concerning the rest of the displayed WWTPs). These models show low skill for discriminating capacity.