Submerged macrophyte assessment in rivers: An automatic mapping method using Pléiades imagery

Diane Espel, Stéphanie Courty, Yves Auda, David Sheeren, Arnaud Elger

Submerged macrophyte assessment in rivers: an automatic mapping method using Pléiades imagery

Diane Espel[1,2], Stephanie Courty[2], Yves Auda[3], David Sheeren[4] and Arnaud Elger[1]

[1]EcoLab, Laboratoire Ecologie fonctionnelle et Environnement, Université de Toulouse, CNRS, Toulouse, France

[2]Adict Solutions, Toulouse, France

[3]G.E.T., Toulouse, France

[4]Université de Toulouse, INRAE, UMR DYNAFOR, Castanet-Tolosan, France

*Corresponding author phone: Diane Espel +33 629705145; e-mail: diane.espel@inp-toulouse.fr

Number of words

Abstract: 257

Abstract + keywords + text + acknowledgements: 7997

Number of tables: 2

Number of figures: 6

Number of references: 80

Supporting information: Yes

Abstract:

Submerged macrophyte monitoring is a major concern for hydrosystem management, particularly for understanding and preventing the potential impacts of global change on ecological functions and services. Macrophyte distribution assessments in rivers are still primarily realized using field monitoring or manual photo-interpretation of aerial images. Considering the lack of applications in fluvial environments, developing operational, low-cost and less time-consuming tools able to automatically map and monitor submerged macrophyte distribution is therefore crucial to support effective management programs. In this study, the suitability of very fine-scale resolution (50 cm) multispectral Pléiades satellite imagery to estimate submerged macrophyte cover, at the scale of a 1 km river section, was investigated. The performance of nonparametric regression methods (based on two reliable and well-known machine learning algorithms for remote sensing applications, Random Forest and Support Vector Regression) were compared for several spectral datasets, testing the relevance of 4 spectral bands (red, green, blue and near-infrared) and two vegetation indices (the Normalized Difference Vegetation Index, NDVI, and the Green-Red Vegetation Index, GRVI), and for several field sampling configurations. Both machine learning algorithms applied to a Pléiades image were able to reasonably well predict macrophyte cover in river ecosystems with promising performance metrics (R^2 above 0.7 and RMSE around 20 %). The Random Forest algorithm combined to the 4 spectral bands from Pléiades image was the most efficient, particularly for extreme cover values (0 % and 100 %). Our study also demonstrated that a larger number of fine-scale field sampling entities clearly involved better cover predictions than a smaller number of larger sampling entities.

Keywords: Aquatic vegetation; Remote sensing; Machine learning; Fluvial ecosystem; Random Forest; Support Vector Regression

1. Introduction

The essential role of macrophytes in freshwater ecosystems has been well demonstrated. They influence the physical, chemical and biological structures of hydrosystems and provide multiple ecosystem functions and services, such as water quality improvement, stabilization of streambeds and habitat provision (Carpenter and Lodge, 1986; Dennison *et al.*, 1993; Jeppesen *et al.*, 1998; Bornette and Puijalon, 2011; Choi *et al.*, 2014). However, their excessive development can have negative impacts on hydrosystem functioning, for instance, light attenuation, anoxia, reduction of flow velocity and increase of sedimentation (*e.g.* Caraco and Cole, 2002; Hussner *et al.*, 2017; Kagami *et al.*, 2019), modifying biotic interactions and disrupting community assembly (Santos *et al.*, 2011). Submerged species, especially, can also cause recurring problems for users and managers, *e.g.* inconvenience to water activities, olfactory nuisances and clogging of water intakes in power plants in case of uprooting (Jadhav and Buchberger, 1995; Bunn *et al.*, 1998; Sand-Jensen and Pedersen, 1999; Stephan and Gutknecht, 2002; Martin, 2017) which are still difficult to fully anticipate on large rivers. Although it is relatively well known that the patterns of spatial and temporal distribution of aquatic macrophytes result from the interactions between many environmental factors (hydrology, water temperature, light, nutrients, substrate, grazing), using these factors in statistical models remains insufficient to predict their distribution within large geographic areas as fluvial environments. Data on the current distribution of submerged macrophytes in the field are therefore pivotal, either for direct monitoring, as required by the EU Water Framework Directive to improve the ecological quality assessment of inland waters (WFD European Commission, 2000), but also for developing new distribution models.

Mapping and monitoring vegetation distribution are important technical tasks in sustainable management. Accordingly, numerous monitoring programs have focused on acquiring spatial information about species composition, maximum depth colonization, density, cover (*i.e.*

percentage of the horizontal surface occupied by vegetation), biomass and plant height (Johnson and Newman, 2011). Those programs have been reviewed in Stocks *et al.* (2019).

In fluvial environment, collecting data on submerged aquatic vegetation (SAV) that will sufficiently represent spatial variation along river reach is difficult, and requires labour-intensive, time-consuming and sometimes destructive fieldwork. Thus, SAV sampling will be more and more replaced by indirect mapping methods, especially thanks to remote sensing tools. Nowadays airborne or satellite sensors provide many observation opportunities at large scales at a given time, with relatively high spatial and temporal resolutions which are constantly improving. Actually, multispectral satellite data have been widely used to map the distribution of macrophytes over large areas (Gullström *et al.*, 2006; Nelson *et al.*, 2006; Dogan *et al.*, 2009; Tian *et al.*, 2010) and assess the spatio-temporal dynamics of aquatic vegetation (MacAlister and Mahaxay, 2009; Zhao *et al.*, 2012, 2013). However, many of these studies have focused on emergent or floating aquatic vegetation due to easier distinction of macrophyte spectral signal from water. Hyperspectral data have been less considered due to their limitations in terms of cost, availability, processing and high dimensionality of spectral data (Plaza *et al.*, 2009; Adam *et al.*, 2010; Mutanga *et al.*, 2012).

Recently, significant progress has been made in image processing for repetitive and automatic submerged macrophyte mapping over large areas combining punctual field data and various modelling methods. Some researchers have clearly demonstrated the feasibility of submerged macrophyte mapping using powerful machine learning algorithms (*e.g.* Artificial Neural Networks, Random Forest, Support Vector Machines or *K*-Nearest Neighbors) (*e.g.* Dogan *et al.*, 2009; Kotta *et al.*, 2013). The main characteristic of these supervised algorithms is to train a model on a part of the data and test the fitted model on the other part. Compared with other nonparametric methods, they also have no limitation for the number of independent variables (*i.e.* adapted to high dimensional data) and do not require normally-distributed variables. Most of these works have been conducted in marine environments, or were limited to large river basins, wetlands or lakes with the best satellite resolution limited to 2.41 m and various attempts have been made to resolve issues regarding submerged vegetation (Hedley *et al.*, 2012). However, the insufficient spatial resolution of image data, the spatial variability of depth and the strong attenuation of light in water are still limitations for remotely mapping SAV in fluvial environments (Marcus and Fonstad, 2008). To our knowledge, only one study has focused on submerged macrophytes in rivers combining ultra-light aircraft images to machine learning algorithms but the obtained map was limited to the presence/absence of canopy meadows (Durand *et al.*, 2016). Finally, a scientific issue is to develop machine learning models on high resolution satellite images with high potential in rivers to extend satellite remote sensing.

Additionally, human and financial resources allocated to acquire *in situ* aquatic vegetation data are generally limited. There is no standard sampling protocol while the quality of macrophyte cover prediction depends on the sampling strategy adopted. However, to our knowledge, there is no study in fluvial environments allowing the optimization of SAV monitoring methods, both in terms of sampling protocol (*i.e.* determining the spatial scale and the minimal number of sampling plots required) and prediction model choice to generate accurate and continuous map of riverine submerged macrophytes.

In that respect, a new mapping method was here investigated to monitor the distribution of submerged vegetation in fluvial environments at a very fine-scale resolution (50 cm) while limiting logistical, financial and human costs. The main objective of this study was to develop a cover prediction model combining machine learning algorithm, imagery spectral features and punctual field samples. In that respect, we investigated the performance of two machine learning regression models (Random Forest and Support Vector Regression) applied to high-resolution multispectral Pléiades satellite imagery, for automatically mapping macrophyte cover at the scale of a 1 km river section. Both algorithms are reliable and well-known in remote sensing; they proved to achieve satisfactory results for various and numerous remote sensing applications in ecology (*e.g*. Cutler *et al.*, 2007; Hunter *et al.*, 2010; Husson *et al.*, 2017; Traganos *et al.*, 2018; Villa *et al.*, 2018; Zafari *et al.*, 2019; Sabat-Tomala *et al.*, 2020). Support Vector Regression is more robust than Random Forest for a lower sample size but the latter is faster to compute (Mountrakis *et al.*, 2011; Belgiu and Drăguţ, 2016; Gholami and Fakhari, 2017).

In addition, we discussed several SAV cover sampling strategies involving various numbers of sampling entities with different sizes in order to guide managers optimizing their monitoring method. We assessed whether, for a given sampling effort, it is better to use a larger number of smaller sampling entities or a smaller number of larger entities for SAV cover estimation.

2. Materials and methods

2.1. Study area

The study was carried out during September 2017 in the Garonne River, a southwest French shallow river, approximately 15 km north of Toulouse (43°41'51"N, 1°22'09"E), next to the city of Seilh (Figure 1a).

The study area was about 1 km long and 110 m wide and it is a typical example of the mid-Garonne ecosystems with abundant macrophyte meadows which develop at low water depth, from the end of March to early October, and which are mainly constituted of submerged species, including two dominant species, such as *Myriophyllum spicatum* (Eurasian Water Milfoil, L. 1753) and *Ranunculus fluitans* (River Water-crowfoot, Lam. 1779). A dozen of aquatic macrophyte species occur at this site and are observed between 0.15 and 2 m depth.

The study area is composed of a shallow backwater downstream on the right bank and it is bordered by a pebble bed on its left bank. A natural weir marks the upstream limit of the site. Bedrocks, gravels and pebbles are the main substrates and different mixed sediments can be found on half of the study area. Previous *in situ* topography measurements during separate fieldwork campaigns (combining bathymetric data collected with a single beam echosounder and elevation measurements of the riverbed, banks and overflow areas) and 2D hydrodynamics simulations were used to determine water levels at the whole-site scale during the macrophyte sampling period (§2.2.). Water levels were relatively low, with average depths varying between 0.05 and 1.30 m over more than half of the site during this period. The downstream zone was deeper varying mainly between 1.30 and 3.50 m, and with a maximum depth of 4.05 m on the right side of the central channel.

2.2. Field data collection

Field survey of submerged macrophytes was conducted during the second half of September 2017 to get observed data of macrophyte covers at a period with available Pléiades images and a non-turbid water column. Numerical simulations based on a SAV model indicated that the biomass of macrophytes on this site and at that date was still very close to the peak biomass that occurred during the second half of August (unpublished results). A total of 55 sampling plots of 9 m² (*i.e.* 3 m side PVC square frames) were distributed all over the study site (Figure 2a). The sampling protocol focused on homogeneous areas of macrophyte meadows and was oriented in order to sample several combinations of substrate, depth and meadow abundance classes; it also included numerous open water plots (Table 1). Each plot (P) was divided into 16 quadrats (Q) of 0.75 m side (Figure 2b). Then the total cover (*i.e.* referring to the cover of all species) of SAV within each quadrat was estimated by two subaquatic observers. A total of 880 quadrats were thus sampled individually. Finally, we averaged the total cover from the 16 quadrats of each plot to get the total cover at the plot scale. The purpose here was to compare two field sampling levels: the plot level (*i.e.* 55 plots) versus the quadrat level (*i.e.* 880 quadrats).

The centre of each sampling plot was geolocated using a portable real time kinematic global positioning system (RTK-GPS) receiver (EMLID Reach RS™) with +/- 3 cm accuracy. The RTK-GPS reference receiver was located on field, close to the right river bank of the study area (*i.e.* <1 km from all sampling plots) and free of obstacles to ensure maximum exposure to radiometric signals. Both RTK-GPS receivers used GNSS signals from GPS and GLONASS satellites. All plots were North oriented. The geographical information system QGIS 2.18 was used to digitize the field data.

2.3. Remote sensing data acquisition into the Garonne River

Two types of optical measuring instruments were used to acquire remote sensing data: unmanned aerial vehicles (UAV) and satellite with high spatial resolutions. UAV imagery provided an overview of the distribution of submerged macrophytes meadows at the site scale whereas satellite image was used for the development of the automatic mapping method.

2.3.1. UAV imagery

Aerial photos were taken mid-September 2017 with a Phantom 4 Pro, a drone developed by DJI™ (DJI™, Shenzhen, China). We used a flying height of 30 m. The along- and across-track image covers were set to 70-80 %. The study site was surveyed in three flight blocks of 10-15 min (Figure 3a). Flight missions were programmed with the Litchi application software available for Android devices. Flight weather conditions were sunny with light wind at ground level. The camera used in this work was a three-band (red, green and blue) DJI™ digital camera, equipped with a 1" CMOS / 20-megapixel sensor camera, with a focal length of 24 mm, and a F-number of 2.8, providing an image size of 5472 × 3648 (columns × rows) (Figure 3b).
Mosaicking was processed using the commercial software Agisoft© Metashape Professional Edition (Saint Petersburg, Russia) with a spatial resolution of 1 cm (Figure 3c). Georeferencing was performed on QGIS 2.18 using seven ground control points which were taken *in situ* with the RTK-GPS with a linear transformation and nearest neighbors resampling. Then a map of

the total cover of submerged macrophytes was created on QGIS 2.18 by visual interpretation of the orthomosaic at each 3×3 m cell of an overlaid grid. This photo-interpreted cover map provided a field reference of the study site.

2.3.2. Satellite imagery

One Pléiades pan-sharpened image of 100 km² surrounding our study site (Figure 1b) was acquired on September 27, 2017, thanks to the “Initiative for Space Innovative Standards” (ISIS) program that results from a cooperation between Airbus Defence and Space and the French “Centre National d’Etudes Spatiales” (CNES). Pan-sharpening, resulting from a fusion process, corresponds to a multispectral image including 4 spectral bands from the visible (Red: 0.59-0.71 µm, Green: 0.50-0.62 µm, Blue: 0.43-0.55 µm or RGB) through the near-infrared (NIR: 0.74-0.94 µm) with the spatial resolution of panchromatic images (*i.e.* 50 cm) obtained by the Pléiades-1A satellites. The ortho-image was cloud free with minimal glint and corrected from acquisition and terrain off-nadir effects by the providers.

The raw Pléiades image was already ortho-rectified before delivery (with 2.5 m accuracy according to Grazzini *et al.* (2013). Its georeferencing has been improved using a linear transformation and nearest neighbors resampling. This process required 17 ground control points for which 3 cm accuracy GPS positions were determined with the RTK-GPS. This process allowed a new planimetric accuracy of the Pléiades image of *ca.* 0.35 m.

Then, radiometric corrections have included conversion of pixel digital number values (encoded on 16 bits) to Top of Atmosphere (TOA) reflectance for each spectral band (*i.e.* RGB and NIR) in order to normalize each spectral band in a a continuous range between 0 and 1. This process was made using the “Geosud TOA Reflectance” plugin available on QGIS 2.18.

Finally, a vector mask of the riparian area and dewatered banks was used to produce a minor riverbed-only image.

2.4. Modelling macrophyte cover using satellite imagery

Modelling the relationship between SAV cover and its spectral reflectance is essential for automatically mapping macrophytes. We used machine learning regression algorithms to predict macrophyte cover from high-resolution reflectance data based on Pléiades image. Figure 4 summarizes the main steps of our approach.

The predictive model was developed using free geomatics and statistics tools, such as QGIS 2.18 spatial analysis technologies to pre-process the remote sensing data and several R libraries, such as “rgdal”, “randomForest” and “e1071”, to develop the regression models and to generate the predicted distribution maps at the spatial resolution of the Pléiades images (*i.e.* 0.5 m of resolution). Finally, their resolution had been degraded to 3 m resolution for comparison with the photo-interpreted cover map (*i.e.* 3 m).

2.4.1. Defining predictive variables among spectral data

The use of red and especially of NIR band for submerged macrophyte detection is still debated because of their relatively fast absorption in water (Fyfe, 2003; Heege *et al.*, 2004; Dogan *et al.*, 2009; Chen *et al.*, 2018b). However, many submerged macrophytes reach the surface during low water period by creating a canopy having stronger contribution to the signal, potentially making these two bands suitable for the detection of aquatic vegetation. In addition, the use of normalized spectral indices could improve prediction models (Bradley and Fleishman, 2008). Different datasets, derived from the pre-processed satellite image, were therefore used to assess the potential of multispectral bands and vegetation indices to detect submerged macrophytes, and to examine the effect of spectral features on the quality of cover prediction. The four spectral reflectance datasets were the following:

(i) The first dataset was obtained from the Pléiades concatenated spectral data from the visible bands (red, green and blue bands (RGB)). This dataset was referred to as “dataset 0” (3 variables);

(ii) The second dataset ("dataset 1") combined RGB and near-infrared (NIR) bands (4 variables);

(iii) The third dataset ("dataset 2") included two widely-used vegetation indices reflecting the characteristic spectral signature of active vegetation: the Normalized Difference Vegetation Index, NDVI=(NIR-R)/(NIR+R), and the Green-Red Vegetation Index, GRVI=(G-R)/(G+R) (2 variables). According to Cho *et al.*, (2008) NDVI could only be useful in shallow clear waters (<0.5 m depth) because of fast absorption of the NIR band while GRVI, using red and green bands, decrease less rapidly depending on the depth and could perform better in detecting changes in canopy vegetation (Motohka *et al.*, 2010; Chen *et al.*, 2018a);

(iv) The last dataset ("dataset 3") included the RGB and NIR bands and the two GRVI and NDVI indices (6 variables).

2.4.2. Extraction of reflectance data based on sampling entities

Our field database was composed of 55 plots or 880 quadrats (each of them being characterized by a value of total macrophyte cover expressed in percentage). Each sampling entity (plot or quadrat) was composed of a given number *n* of pixels from the Pléiades image, each pixel being associated with a reflectance value in several spectral bands (or vegetation indices). The *n* values of reflectance of each spectral band are extracted using the *extract()* function from the R *raster* package. Then an area-weighted average reflectance of the *n* pixels belonging to each sampling entity was computed.

2.4.3. Machine learning regression models and SAV mapping

We tested the performance of two reliable machine learning algorithms to build a regression model that predicts SAV cover at the scale of the site: the Random Forest (RF) and Support Vector Regression (SVR) algorithms. For each model, the remote sensing variables (including vegetation indices for "dataset 2" and "dataset 3") were treated as independent variables and the total cover was treated as a dependent or response variable. Detailed reviews of SVR and RF in remote sensing can be found in Mountrakis *et al.* (2011) and Belgiu and Drăguţ (2016), respectively.

Regression analysis was carried out on the four datasets and at the two sampling levels (plots and quadrats). This procedure allows us to assess the effect of the spectral dataset and the relevance of the two field sampling levels.

We then explored the effects of the number and size of the sampling entities on prediction quality. Firstly, the effect of the sampling scale on cover prediction was addressed by comparing the results obtained using datasets with 55 entities but of different size: 0.75 m, 1.5 m (corresponding to a group of 4 quadrats per plot) and 3 m side. Secondly, the number of quadrats was gradually decreased from the initial dataset (880 quadrats) by randomly selecting 1, 2, 4 or 8 quadrat(s) per plot. This random selection of quadrats generated configurations with 55, 110, 220 or 440 quadrats. In addition, a subset of the plots was used to get a configuration with 15 and 30 quadrats. Thanks to an oriented random selection function, we ensured that the quadrat selection maintain a similar distribution of macrophyte abundance classes, as that of the initial 880 quadrats dataset (Table 1). The purpose here was to define the minimal number of sampling entities that allowed a satisfactory result. This was achieved by a learning curve analysis.

In total 72 models (combining 2 algorithms, 4 spectral datasets and 9 field sampling configurations) of cover prediction were run and compared with each other in order to determine the best one. The key steps of the method are detailed below.

- ***Tuning, training and testing the models***

In the case of a limited number of samples, the *k*-fold cross-validation technique is recognized as a valuable approach to split samples into *k* subsets of roughly equal size (Hastie *et al.*, 2009).

For a given combination of hyperparameters, *k*-1 partitions are used iteratively (*k* times) to train the model and to test it on the remaining partition. Thus, it potentially allows each sample to be used *k* times for multiple training or testing, with the purposes of (i) improving the learning process (fine tuning), (ii) using independent datasets for training and testing, and (iii) limiting overfitting (Anguita *et al.*, 2012; Ramezan *et al.*, 2019). The final model performance is calculated by averaging the *k* computed errors for the selected hyperparameters. Finally, the best prediction model corresponds to the model built with hyperparameters which generate the highest validation score (*i.e.* the lowest generalization error or test error). Additionally, a commonly-used method to reduce the variability in chosen parameters and the standard deviation of performance estimates of the tuned model, is to run a repeated *k*-fold cross-validation, named *J*-*k*-fold cross-validation (Moss *et al.*, 2018). We therefore used in this study a repeated *k*-fold cross-validation, *i.e.* with k fixed to 10, as is commonly chosen in the literature (Jung, 2017) and we set J to 10. Thus, in each round, a number of sampling entities was randomly selected, with 90 % used as training data and 10 % as test data.

During the learning step, the different combinations of hyperparameter values are examined and tuned for the calibration of the model. The combination of values that generate the lowest error on the test set is assumed to be optimal (Brenning, 2012; Cracknell and Reading, 2014; Sharma *et al.*, 2017). For the RF algorithm, the number of regression trees *ntree* ranged from 25 to 1000 (with a step of 25) to test the sensitivity of this parameter. The *ntree* values that yielded the lowest error were selected for each dataset. Due to conflicting reports in the literature concerning the potential influence of the *mtry* parameter (*i.e.* the number of predictive variables) on prediction performance (Cutler *et al.*, 2007; Strobl *et al.*, 2008), the *mtry* value was tested from 1 to the maximum number of spectral bands for each dataset (*e.g.* maximum of 6 for “dataset 3”). Concerning the SVR algorithm, the commonly used gaussian radial basis kernel function (RBF) was applied, given its traditional superior performance compared to other kernels (*i.e.* linear or polynomial) (Kavzoglu and Colkesen, 2009). The different parameters ranges were as follows: the regularization parameter C (*i.e.* cost) ranged from 0.1 to 1000 (by a factor of 10), ε was fixed to 0.1 and the width of the RBF kernel function γ ranged from 2^{-5} to 2^5 (by a factor of 2). Predictive variables were standardized.

- ***Model evaluation***

Two statistics were employed to evaluate the quality of the model predictions:

- The coefficient of determination (R^2) which varies between 0 and 1, to account for the goodness-of-fit between observations and predictions, and to define how much variance is explained by the model.
- The root-mean squared error (RMSE) to assess the predictive power of the model.

These statistics were computed at the scale of the field entities, as well as the scale of the site, using punctual visually-estimated covers for the former (depending on the configuration tested, for instance 55 plots or 880 quadrats), and the cover map obtained from drone imagery photointerpretation for the latter. However, only the metrics computed at the site scale will be discussed here to address the predictive quality of our models.

Difference maps between predicted and observed covers were also computed to pinpoint local prediction errors (expressed as cover percentages; a difference below or above 0 indicating an underprediction or an overprediction, respectively).

3. Results

3.1. Observed macrophyte distribution on the Garonne River

Monospecific and plurispecific meadows were observed in the field. Sampling entities with high cover values generally included the two dominant species (see §2.1) while the remaining species were observed at lower densities, except for *Potamogeton nodosus* and the *Elodea* species which were locally abundant, notably in the backwater. Species were distributed according to their ecological requirements (flow, turbidity). Table 1 summarizes the distribution of densities according to the adopted field sampling strategy. Table S1 (see supporting information) provides the database from field sampling.

The photointerpretation of the drone mosaic (*i.e.* the orthomosaic) (Figure 5a) highlighted the spatial distribution of macrophytes and revealed variability in the distribution of macrophyte meadows. These were found mainly upstream and halfway along riverbanks, but also at the backwater on the right bank. Overall, higher covers were found along the left bank, in areas of shallow depths (0.1-1 m) especially on mixed substrates. Lower densities were observed in deeper areas and/or on rock bed. Downstream meadows seemed to be patchier along the left bank. Rare meadows were present in the centre of the channel. Bare areas (*i.e.* with 0 % cover) were the most represented on the site, followed by areas with covers between 1 % and 10 %, especially present along the right bank, while areas with covers between 11-25 % and above 75 % were mostly located along the left bank. Based on this cover distribution and on point biomass measurements at given sampling plots (data not shown), the total biomass on the site was estimated to be 2.6 t of dry matter for a surface of *ca.* 10 ha.

3.2. Predicted cover of submerged macrophytes

All the machine learning results were analysed in terms of statistical performance and prediction quality (difference between predicted and observed covers). For each regression model, the R^2 and RMSE statistical results, as well as the range of the predicted cover values, are provided in the supporting information– Table S2. Figures S1 to S4 also group together the whole map results in the supporting information. Here we focus on the most relevant results.

3.2.1. Effect of the spectral features

Analysing the effect of spectral bands on macrophyte cover prediction has shown meaningful differences between the statistical metrics (R^2 and RMSE) of the different datasets tested, regardless of the sampling level (*i.e.* plot or quadrat) considered.

Prediction models built on "dataset 2" (*i.e.* NDVI and GRVI) obtained the least satisfactory mapping and statistical results, regardless of the algorithm or the sampling level. The determination coefficients (R^2) (*i.e.* similarity levels between predicted and photo-interpreted covers) were of 0.4-0.5 and the prediction errors (RMSE) ranged between 25 % and 27 % (Table S2). Overall, the bare areas (*i.e.* open water) were poorly predicted with important over-prediction, particularly with the Support Vector Regression (SVR) algorithm. Cover maps were improved to some extent with the Random Forest (RF) algorithm: boundaries of macrophyte patches were more clearly distinguished (Figures S1-S2). Moreover, the difference cover maps showed under-predicted areas, particularly for the meadows located along the left and right banks (Figures S3-S4).

In contrast, no meaningful difference with respect to the evaluation metrics, were found for the models based on the three remaining datasets and using either RF or SVR. All of them usually indicated high goodness-of-fit at the study scale (R^2), between 0.62 and 0.75, as well as a good prediction power with a RMSE of 21 % on average (maximum of 24 %) (Table S2).

However, regarding RF, a poorer prediction of the bare areas was noticed in case of "dataset 0", while the results obtained with datasets 1 and 3 were equivalent (see Figures S1, S3 and Table S2 for details). Adding NIR improved prediction quality in the bare areas, particularly at the quadrat sampling level.

For SVR, only the maps obtained with datasets 0 and 1 represented relatively well the spatial distribution of macrophyte meadows on the Garonne River (Figure S2). Indeed, with "dataset 3", high densities were better predicted (few under-prediction errors) than with "dataset 2" but bare areas still presented considerable over-prediction errors (differences up to 28 % between predicted and observed covers) (Figure S4). Furthermore, the predictive advantage of adding NIR ("dataset 1") was variable depending on the 2 sampling levels (*i.e.* plots or quadrats). For instance, improved results were noticed for models based on "dataset 0" (involving the 3 bands in the visible) and on quadrats: this dataset yielded the lowest difference between predicted and observed covers (Figure S4).

For the rest of result analysis, we excluded datasets 2 and 3 since they produce models that were poorly predictive, and we focused on models based either on RF and involving “dataset 1” or on SVR and involving datasets 0 and 1.

3.2.2. Effect of the learning algorithm

The two machine learning algorithms tested here and used on our two field datasets (55 plots and 880 quadrats) were able to predict the distribution and macrophyte cover on the Garonne River, reaching a maximum predicted cover value of 100 % (Table 2). Whether for plots or quadrats, no statistical pattern was highlighted through the comparison of the two algorithms. Regression models showed high R^2, oscillating between 0.67 and 0.74 and low prediction errors (RMSE around 19-21 %), even if SVR models reached slightly higher performance (Table 2).

Overall the predicted cover maps reproduced the observed spatial distribution of submerged macrophytes, with dense meadows along the left bank over more than half of the site (Figure 5c and Figures S1-S2). Nevertheless, there were two areas where the cover prediction error remained high (differences up to 100 % between the predicted and observed covers), namely in the backwater area, and downstream, within the dense meadow located along the left bank, where high densities (> 75 %) were detected as bare areas. These differences were identified by the dark red areas in the difference maps (*e.g.* Figure 5d).

Besides, detecting the patchier distributions downstream, within the backwater and along the right bank seems to be more or less difficult depending on the algorithm. In that respect, the prediction accuracy of macrophyte cover along the riverbank in the middle of the study site was lower using the SVR algorithm than using the RF algorithm (Figures S1-S2).

In addition, a visual comparison between observed and predicted cover data allowed to point at certain characteristics of the regression algorithms (see Figures S3-S5 in supporting information for details). If we focus on results based on SVR, at the entity scale, data were globally noisier especially with important discrepancies for extreme values of cover (*i.e.* 0 and 100 %) (Figure S5). In that respect, more prediction errors were observed at these cover densities at the site scale, in particular for meadows upstream and within the central channel (Figure S4).In addition, average densities (*i.e.* those between 25 and 75 % of cover) were hardly predicted with SVR models. Indeed, those areas were often underestimated by this algorithm, with differences between predicted and observed cover reaching a maximum of 50 %. On the other hand, the fit of the regression model with the observed covers was better using RF, compared to SVR (Figure 5b and Figure S5). The different covers seemed globally well predicted by RF, with the different classes of macrophyte abundance better discriminated, even if the highest covers (*i.e.* above 90 %) were often slightly under-predicted (maximum of difference under 28 %). Finally, the bare areas were better represented by this algorithm and, in case of over-prediction, the maximum difference was still limited to 20 %, particularly along the riverbeds (Figure 5d).

Consequently, the RF algorithm when associated with "dataset 1" was considered as the most convincing one for predicting the spatial distribution of submerged macrophyte cover.

3.2.3. Effect of field sampling strategy: 55 plots *vs* 880 quadrats

The comparison of the two sampling levels (plots *vs*. quadrats) highlighted clear differences between algorithms in their abilities to detect macrophyte patches. Whatever the algorithm, goodness-of-fits and spatial errors were not notably different between the two sampling entities although a model using 880 quadrats seemed to show performances somewhat higher than those obtained with 55 plots dataset (Table 2). Even if both of these sampling levels predicted a wide range of cover percentages, regression models based on quadrats predicted better the largest cover values, especially those above 98 % (Table 2). Furthermore, the limits of the meadow patches seemed less well-defined with a model using 55 plots, compared to 880 quadrats. Finally, the difference maps revealed that the differences between predicted and observed covers were reduced using 880 quadrats (differences less than 20 % on average) (Figure S3).

Consequently, although models based on the two field sampling levels were able to distinguish the spatial distribution of meadows, considering 880 quadrats instead of 55 plots showed better results. Overall, the best model for macrophyte cover prediction was obtained with RF using NIR band (§3.4.1 and 3.4.2) and based on field data sampled within 880 quadrats (Figure 5c). However, it remains to be seen whether these better performances are due to a finer size and/or to more numerous entities.

3.2.4. Effect of the size of sampling entity

The comparison of the models based on 3 different sizes (3 m, 1.5 m and 0.75 m side) of the sampling entities and with a fixed number of entities (n = 55) showed very clear differences regarding cover prediction accuracy.

Actually, even if no meaningful difference in model metrics (R^2 and RMSE) could be seen between the different scales, the RF model built on sampling entities of 3 m side presented slightly better R^2 and RMSE compared to models using smaller sampling entities (R^2 = 0.70, RMSE = 21.3 %) (Table S2). Boundaries between the different macrophyte patches were better defined too (Figure S1). The difference maps also showed the least spatial difference between predicted and observed covers, with average differences amounting to +8 % in bare areas and -40 % within macrophytes meadows. Spatial differences locally reached +12 % in bare areas and -72 % within meadows (Figure S3). Models based on quadrats of 0.75 m side showed the worst performances (R^2 = 0.62, RMSE = 23.7 %) (Table S2) and poor cover map results (with local prediction differences reaching +20 % in bare areas and -90 % within meadows) (Figure S3). With sampling entities of 1.5 m side, results were improved but still worse than a model based on 3 m side sampling entities. Similar results were also observed with the SVR algorithm (see Figures S2, S4 for details).

Therefore, for a limited number of entities (n = 55), predicting cover using 3 m side entities seemed more appropriate when addressing the spatial patterns of macrophytes. However as shown above, this sampling strategy was not fully effective for accurate prediction of cover, in particular compared to a sampling strategy with more entities.

3.2.5. Effect of the number of entities

Analysis of learning curves, which described the evaluation metrics of models based on different numbers of quadrats, showed that the maximum performance was reached for 110 quadrats (and above) whatever the spectral dataset and algorithm used (Figure 6 and see Figure S6 in supporting information for the SVR results). More precisely, the R^2 and RMSE remained stable from 110 quadrats for the RF models (Figure 6). Regarding the SVR, the same observations could be made, though the metrics were really stable from 220 quadrats for "dataset 1" (RGB NIR).

In terms of sampled surface, 220 quadrats correspond to 55 entities of 1.5 m side (*i.e.* 4 grouped quadrats of 0.75 m side each) and a 220 quadrats-based model has shown to be more performing than a 55 plots-based one (and, a fortiori, than a model based on 55 entities of 1.5 m) (Figure 6 and Figure S6). Consequently, for a same sampled surface, cover estimate will be better with models involving more but smaller entities than with models using a limited number of larger entities.

4. Discussion

Surveying submerged macrophytes using remote sensing is somewhat more difficult than surveying terrestrial vegetation because of water reflectance issues (linked to the strong reflection of the water surface and attenuation of light by the water column) and the limited spatial resolution of most sensors (Nelson *et al.*, 2006; Underwood *et al.*, 2006). In this study, we developed a new method for automatically mapping the cover of SAV on a river section during the peak of biomass. This method included very high-resolution (50 cm) spectral data from Pléiades pan-sharpened image combined with a machine-learning algorithm and an optimized method of macrophyte cover sampling.

4.1. Developing a remote sensing method for automatic mapping of vegetation cover

4.1.1. Choosing an appropriate machine learning algorithm

The achievement of this remote sensing method did not really depend on the two regression algorithms tested (RF and SVR). Between both of them statistical results were similar, with only small differences in the predicted cover maps. They showed reasonable fitting capacity (R^2 around 0.7) and relatively low prediction errors (RMSE < 25 %). In fact, a correct parameterization of different machine learning algorithms must lead to similar results. However, fewer local prediction errors were observed for the RF models. The spatial boundaries of the meadow located along the right bank were also better defined. Actually, SVR is more sensitive to parameter assignment than RF (among them the choice of the γ parameter which allows optimal data discrimination) during the tuning step (§2.4.3) that can significantly alter the performance of SVR algorithms (Brown *et al.*, 1999; Mountrakis *et al.*, 2011). Additionally, the RF algorithm is known to be effective in quickly handling high dimensional data and multicollinearity (Belgiu and Drăguţ, 2016). It was thus considered as the most appropriate machine learning algorithm for river macrophyte mapping.

4.1.2. Determining the best spectral features

For models based on the RF algorithm, the best predicted results were obtained using datasets combining visible spectral and NIR bands (vegetation indices such as NDVI and GRVI were of no use for cover prediction). Adding NIR to visible spectral bands also seemed to slightly benefit the prediction of low covers but did not systematically improve the quality of the prediction in the areas of higher expected covers. The advantage of adding NIR depended on the algorithm. Indeed, the use of NIR and the indices derived from it (NDVI) are still widely discussed in the aquatic remote sensing literature, because of the large absorption by water of red and, even more, NIR wavelengths (*e.g.* Pegau *et al.*, 1997; Fyfe, 2003; Cho *et al.*, 2008; Silva *et al.*, 2008). Moreover, even if the green wavelengths normally provide greater light penetration in turbid waters, green and red regions are considered as the best ones for submerged macrophyte sensing (Fyfe, 2003; Han and Rundquist, 2003; Williams *et al.*, 2003; Pinnel *et al.*, 2005). The use of GRVI did not yet compensate for the low performance of NDVI in our study. Besides, a study carried out on wetlands has shown that red and NIR regions could be saturated beyond a certain biomass density (Peñuelas *et al.*, 1993; Mutanga and Skidmore, 2004). Finally, Chen *et al.* (2018b) have clearly demonstrated that NIR should only be used in the case of very shallow aquatic environments (< 1 m depth). NIR could thus be useful for predicting macrophyte meadows along the riverbanks, or when submerged macrophytes form a canopy just below the water surface, which is often the case in eutrophic waters at the biomass peak period.

4.1.3. Optimizing field sampling

Training sample size also influenced the prediction accuracy of macrophyte cover. With a limited number of entities ($n = 55$), using 3 m side entities was more effective than using smaller entities. Indeed, as the geolocation of the Pléiades image was not perfect, it is likely that the small sampling entities did not exactly match with the satellite image. In this case, increasing the size of the sampling entity would solve this problem and could explain the better results for 3 m side size entities. However, for the RF (and SVR) regression models based on 55 plots, local prediction errors were larger, particularly for bare areas, than the model based on 880 quadrats. For a same surface sampled, the latter actually showed the best predictions, even if certain areas still presented local under-prediction errors. In fact, the bigger the training dataset the better the algorithm will learn. Predicted macrophyte meadows were particularly well distributed and, when present, local prediction errors were the lowest in comparison with other field sampling strategies. Both meadows along the riverbanks were well defined. Using a power relationship ($R^2 = 0.8$) between shoot biomass and cover previously established (unpublished study, see Figure S7 in supporting information), shoot biomass at the site scale was estimated to 2.3 t of dry matter for 10 ha. This is relatively consistent with the shoot biomass derived from photointerpretation (*i.e.* 2.6 t of dry matter) and confirmed the performance of our model.

Machine learning algorithms were also influenced by the number of training data and their distribution into the study site. Learning curves confirmed that up to 55 quadrats (*i.e.* entities of 0.75 m side) the number of data was insufficient to make valuable predictions. Considering a higher number of training data, particularly in areas where reflectance values were highly fluctuating, would improve prediction accuracy by accounting for larger reflectance variability. Our results showed that for models using 110 (and above) quadrats, predicted cover maps and model statistics were very similar to those based on 880 quadrats; the estimates with 220 quadrats were clearly improved in comparison to those obtained using 55 entities with higher sampling surface (1.5 or 3 m side), despite an equal or smaller total sampled surface. This study revealed the higher performance of sampling numerous (even smaller) entities. This criterion is particularly interesting regarding submerged macrophyte monitoring as the eye estimation of vegetation covers is relatively difficult at a 3 m scale, particularly in deep areas, due to the difficulty to get an overview of macrophyte cover on such a surface.

Compared to 55 entities of 3 m side (*i.e.* plots), the sampling surface is also reduced by 8 using 110 entities of 0.75 m side (*i.e.* quadrats). Even if using numerous smaller entities implies additional time in the field for movements in water and for GPS coordinate acquisition, this should be largely compensated by the time saved to estimate total cover within entities of smaller size , and by handling a more ergonomic sampling accessory (*i.e.* a smaller PVC frame). From our experience, the sampling time spent on an entity was almost linearly linked to its surface, because of the need to split large entities in smaller sub-entities to get accurate cover estimates. Thus, in comparison with 55 plots, the overall sampling effort should be significantly reduced when using 110 quadrats, while increasing prediction performance. Consequently, for future monitoring campaigns along a river section of 1 km long, we recommend estimating cover in 100 to 200 individual quadrats of a surface comprised between 0.5 and 1 m^2, which have shown to be optimal to obtain representative macrophyte mapping using Pléiades imagery. Despite our method performing reasonably well, macrophyte cover was still underestimated to some extent (maximum local difference between predicted and observed cover still reaching 28 %) and some bare areas were overestimated (maximum local difference reaching 22 %). Some studies reviewed in Guo *et al.* (2017) have discussed the effect of imbalanced data on the quality of machine learning classification. When some field attributes are infrequently present they can be most likely predicted as rare occurrences, undiscovered or ignored, or assumed as noise or outliers which results to more prediction errors of certain covers (Ali *et al.*, 2005). However, as pointed out by Visa and Ralescu (2005) perfect balanced training data is not a guarantee to improve a classifier performance. Field sampling has to be representative of the study site (Petersen *et al.*, 2005). Therefore, regardless of the distribution of our macrophyte abundance classes (Table 1), it is possible that better balancing the entity numbers, particularly for the extreme covers (*i.e*. 0% and 100 %), will improve mapping results in future investigations. Indeed, this sampling strategy would include more reflectance variability, as it is particularly observed in open water entities.

4.2. External factors influencing model quality

Isolating plant signal from the water column interference is still the main challenge of remote sensing of SAV, due to the low contrast (Williams *et al.*, 2003) and to the inherent difficulties in interpreting reflectance values of water (Peñuelas *et al.*, 1993; Lehmann and Lachavanne, 1997). Numerous studies have revealed that the spectral signal of SAV can also be limited by environmental and biological factors, such as water depth, turbidity/transparency, distance between vegetation canopies and water surface (Maritorena *et al.*, 1994; Han and Rundquist, 2003; Vis *et al.*, 2003; Dogan *et al.*, 2009; Liew and Chang, 2012). For instance, local overestimations of low covers were observed in the central channel or downstream of our study site, where water flow and depth are high or where suspended matter is highly concentrated with a thick layer of mud on the bottom. For future investigations it would be interesting to determine if predictions of macrophyte cover could be improved by including substrate types or measures of water clarity (*e.g.* Secchi depth, chlorophyll a content) in our models, as stated by Nelson *et al.* (2006).

Some of the 12 submerged species in our study site with low height (*e.g. Elodea canadensis*) could also be undetected. Several studies about submerged vegetation mapping have shown that non-canopy forming aquatic vegetation species generally lead to more detection errors (Vis *et al.*, 2003; Valta-Hulkkonen *et al.*, 2005; Wolter *et al.*, 2005). It has been reported that SAV can be remotely sensed to a maximum depth between 2 m and 3 m (Han and Rundquist, 2003; Sawaya *et al.*, 2003). Finally, low concentrations of some photosynthetic pigments in plant leaves, such as chlorophylls a and b, carotene and xanthophylls could also affect the spectral reflectance among vegetation (Kumar *et al.*, 2001).

5. Conclusion

Our results provided further evidence that macrophyte cover can be reasonably well predicted with automated regression procedures based on machine learning algorithms and a limited number of sampling entities.

Remote sensing of riverine submerged macrophytes by pansharpened Pléiades imagery associated to a Random Forest algorithm appeared to be a viable and valuable tool for estimating biophysical measures, such as macrophyte cover, at very high spatial resolution (50 cm) on a 1 km site on the Garonne River. Performance metrics were promising with R^2 above 0.7 and prediction error rates around 20 %. In this paper we provided a new, efficient and less time-consuming tool for monitoring SAV which should help steering environmental management actions such as SAV restoration projects or overgrowth management.

There is a significant opportunity for applying such a promising method to the multi-date monitoring of SAV in freshwater river environments. Indeed, the monitoring and mapping of macrophyte meadows over a range of spatial and temporal scales are of prime importance in assessing hydrosystem status. However, rivers are diverse and complex ecosystems, with significant variability of physical properties through both space and time. Future efforts involving detailed bathymetric data, light attenuation and water properties, may resolve depth-related confusion of SAV with substrate type, and are a prerequisite for multi-date vegetation monitoring, based on time series images.

Acknowledgements

The study was co-funded by the National Association for Technology Research (ANRT), the research and development company Adict Solutions and by the Adour-Garonne Water Agency. We wish to thank Olivier Berseille for field assistance, Perle Charlot and Tom Redouté for their contributions in the remote sensing analyses, the CNES Airbus Defence and Space for providing Pléiades images and the General Directorate of Civil Aviation (DGAC) for drone flight authorization.

References

Adam E, Mutanga O, Rugege D. 2010. Multispectral and hyperspectral remote sensing for identification and mapping of wetland vegetation: a review. *Wetlands Ecology and Management* 18: 281–296.

Ali A, Shamsuddin SM, Ralescu AL. 2005. Classification with class imbalance problem: A Review. *International Journal of Advances in Soft Computing and its Applications* 7: 176–205.

Anguita D, Ghio A, Oneto L, Ridella S. 2012. In-Sample and Out-of-Sample Model Selection and Error Estimation for Support Vector Machines. *IEEE Transactions on Neural Networks and Learning Systems* 23: 1390–1406.

Belgiu M, Drăguţ L. 2016. Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing* 114: 24–31.

Bornette G, Puijalon S. 2011. Response of aquatic plants to abiotic factors: a review. *Aquatic Sciences* 73: 1–14.

Bradley BA, Fleishman E. 2008. Can remote sensing of land cover improve species distribution modelling? *Journal of Biogeography* 35: 1158–1159.

Brenning A. 2012. Spatial cross-validation and bootstrap for the assessment of prediction rules in remote sensing: The R package sperrorest. 2012 IEEE International Geoscience and Remote Sensing Symposium, 5372–5375.

Brown M, Gunn SR, Lewis HG. 1999. Support vector machines for optimal classification and spectral unmixing. *Ecological Modelling* 120: 167–179.

Bunn SE, Davies PM, Kellaway DM, Prosser IP. 1998. Influence of invasive macrophytes on channel morphology and hydrology in an open tropical lowland stream, and potential control by riparian shading. *Freshwater Biology* 39: 171–178.

Caraco NF, Cole JJ. 2002. Contrasting Impacts of a Native and Alien Macrophyte on Dissolved Oxygen in a Large River. *Ecological Applications* 12: 1496–1509.

Carpenter SR, Lodge DM. 1986. Effects of submersed macrophytes on ecosystem processes. *Aquatic Botany* 26: 341–370.

Chen A, Orlov-Levin V, Meron M. 2018a. Applying High-Resolution Visible-Channel Aerial Scan of Crop Canopy to Precision Irrigation Management. *Proceedings* 2: 1–7.

Chen Q, Yu R, Hao Y, Wu L, Zhang W, Zhang Q, Bu X. 2018b. A New Method for Mapping Aquatic Vegetation Especially Underwater Vegetation in Lake Ulansuhai Using GF-1 Satellite Data. *Remote Sensing* 10: 1279.

Cho HJ, Kirui P, Natarajan H. 2008. Test of Multi-spectral Vegetation Index for Floating and Canopy-forming Submerged Vegetation. *International Journal of Environmental Research and Public Health* 5: 477–483.

Choi J-Y, Jeong K-S, La G-H, Joo G-J. 2014. Effect of removal of free-floating macrophytes on zooplankton habitat in shallow wetland. *Knowledge and Management of Aquatic Ecosystems* 11.

Cracknell MJ, Reading AM. 2014. Geological mapping using remote sensing data: A comparison of five machine learning algorithms, their response to variations in the spatial distribution of training data and the use of explicit spatial information. *Computers & Geosciences* 63: 22–33.

Cutler DR, Edwards TC, Beard KH, Cutler A, Hess KT, Gibson J, Lawler JJ. 2007. Random Forests for Classification in Ecology. *Ecology* 88: 2783–2792.

Dennison WC, Orth RJ, Moore KA, Stevenson JC, Carter V, Kollar S, Bergstrom PW, Batiuk RA. 1993. Assessing Water Quality with Submersed Aquatic Vegetation Habitat requirements as barometers of Chesapeake Bay health. *BioScience* 43: 86–94.

Dogan OK, Akyurek Z, Beklioglu M. 2009. Identification and mapping of submerged plants in a shallow lake using quickbird satellite data. *Journal of Environmental Management* 90: 2138–2143.

Durand A, Studer M, Marchand A-L, Mauris F, Richard N. 2016. Suivi environnemental des herbiers de rivière par imagerie acquise par ULM et drone : retour d'expérience et potentiel. *La Houille Blanche* 18–23.

Fyfe SK. 2003. Spatial and temporal variation in spectral reflectance: Are seagrass species spectrally distinct? *Limnology and Oceanography* 48: 464–479.

Gholami R, Fakhari N. 2017. Chapter 27 - Support Vector Machine: Principles, Parameters, and Applications. In: Samui P, Sekhar S, Balas VE, eds. Handbook of Neural Computation, Academic Press., 515–535.

Grazzini J, Lemajic S, Astrand J. 2013. External quality control of Pléiades orthoimagery - Part I: Geometric benchmarking and validation of Pléiades - 1A orthorectified data acquired over Maussane test site. Technical Report No. 82308JRC IES.

Gullström M, Lundén B, Bodin M, Kangwe J, Öhman MC, Mtolera MSP, Björk M. 2006. Assessment of changes in the seagrass-dominated submerged vegetation of tropical Chwaka Bay (Zanzibar) using satellite remote sensing. *Estuarine, Coastal and Shelf Science* 67: 399–408.

Guo H, Yijing L, Shang J, Mingyun G, Yuanyue H, Bing G. 2017. Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications* 73: 220–239.

Han L, Rundquist DC. 2003. The spectral responses of Ceratophyllum demersum at varying depths in an experimental tank. *International Journal of Remote Sensing* 24: 859–864.

Hastie T, Tibshirani R, Friedman J. 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer-Verlag New York Inc., New York, NY, 745 p.

Hedley JD, Roelfsema CM, Phinn SR, Mumby PJ. 2012. Environmental and Sensor Limitations in Optical Remote Sensing of Coral Reefs: Implications for Monitoring and Sensor Design. *Remote Sensing* 4: 271–302.

Heege T, Bogner A, Pinnel N. 2004. Mapping of submerged aquatic vegetation with a physically based process chain. Remote Sensing of the Ocean and Sea Ice 2003, International Society for Optics and Photonics., 43–50.

Hunter PD, Gilvear DJ, Tyler AN, Willby NJ, Kelly A. 2010. Mapping macrophytic vegetation in shallow lakes using the Compact Airborne Spectrographic Imager (CASI). *Aquatic Conservation: Marine and Freshwater Ecosystems* 20: 717–727.

Hussner A, Stiers I, Verhofstad MJJM, Bakker ES, Grutters BMC, Haury J, van Valkenburg JLCH, Brundu G, Newman J, Clayton JS, Anderson LWJ, Hofstra D. 2017. Management and control methods of invasive alien freshwater aquatic plants: A review. *Aquatic Botany* 136: 112–137.

Husson E, Reese H, Ecke F. 2017. Combining Spectral Data and a DSM from UAS-Images for Improved Classification of Non-Submerged Aquatic Vegetation. *Remote Sensing* 9: 247.

Jadhav RS, Buchberger SG. 1995. Effects of vegetation on flow through free water surface wetlands. *Ecological Engineering* 5: 481–496.

Jeppesen E, Sondergaard M, Sondergaard M, Christoffersen K. 1998. The structuring role of submerged macrophytes in lakes, New York, NY, 427 p.

Johnson JA, Newman RM. 2011. A comparison of two methods for sampling biomass of aquatic plants. *Journal of Aquatic Plant Management* 49: 1–8.

Jung Y. 2017. Multiple predicting K-fold cross-validation for model selection. *Journal of Nonparametric Statistics* 1–18.

Kagami M, Nishihiro J, Yoshida T. 2019. Ecological and limnological bases for management of overgrown macrophytes: introduction to a special feature. *Limnology* 20: 1–2.

Kavzoglu T, Colkesen I. 2009. A kernel functions analysis for support vector machines for land cover classification. *International Journal of Applied Earth Observation and Geoinformation* 11: 352–359.

Kotta J, Kutser T, Teeveer K, Vahtmäe E, Pärnoja M. 2013. Predicting Species Cover of Marine Macrophyte and Invertebrate Species Combining Hyperspectral Remote Sensing, Machine Learning and Regression Techniques. *PLOS ONE* 8: e63946.

Kumar L, Schmidt KS, Dury S, Skidmore AK, Meer FJ van der, Jong SMD. 2001. Review of hyperspectral remote sensing and vegetation science. Imaging Spectrometry: Basic Principles and Prospective Applications, Kluwer Academic Press, Dordrecht, The Netherlands., 111–155.

Lehmann A, Lachavanne J-B. 1997. Geographic information systems and remote sensing in aquatic botany. *Aquatic Botany* 58: 195–207.

Liew SC, Chang CW. 2012. Detecting submerged aquatic vegetation with 8-band WorldView-2 satellite images. 2012 IEEE International Geoscience and Remote Sensing Symposium, 2560–2562.

MacAlister C, Mahaxay M. 2009. Mapping wetlands in the Lower Mekong Basin for wetland resource and conservation management using Landsat ETM images and field survey data. *Journal of Environmental Management* 90: 2130–2137.

Marcus WA, Fonstad MA. 2008. Optical remote mapping of rivers at sub-meter resolutions and watershed extents. *Earth Surface Processes and Landforms* 33: 4–24.

Maritorena S, Morel A, Gentili B. 1994. Diffuse reflectance of oceanic shallow waters: Influence of water depth and bottom albedo. *Limnology and Oceanography* 39: 1689–1703.

Martin M. 2017. Toulouse : prolifération d'algues dans la Garonne. *France info Occitanie*.

Moss HB, Leslie DS, Rayson P. 2018. Using J-K fold Cross Validation to Reduce Variance When Tuning NLP Models. ssociation for Computational Linguistics, Santa Fe, New Mexico, USA., 2978–2989.

Motohka T, Nasahara KN, Oguma H, Tsuchida S. 2010. Applicability of Green-Red Vegetation Index for Remote Sensing of Vegetation Phenology. *Remote Sensing* 2: 2369–2387.

Mountrakis G, Im J, Ogole C. 2011. Support vector machines in remote sensing: A review. *ISPRS Journal of Photogrammetry and Remote Sensing* 66: 247–259.

Mutanga O, Skidmore AK. 2004. Narrow band vegetation indices overcome the saturation problem in biomass estimation. *International Journal of Remote Sensing* 25: 3999–4014.

Mutanga O, Adam E, Cho MA. 2012. High density biomass estimation for wetland vegetation using WorldView-2 imagery and random forest regression algorithm. *International Journal of Applied Earth Observation and Geoinformation* 18: 399–406.

Nelson SAC, Cheruvelil KS, Soranno PA. 2006. Satellite remote sensing of freshwater macrophytes and the influence of water clarity. *Aquatic Botany* 85: 289–298.

Pegau WS, Gray D, Zaneveld JRV. 1997. Absorption and attenuation of visible and near-infrared light in water: dependence on temperature and salinity. *Applied Optics* 36: 6035–6046.

Peñuelas J, Gamon JA, Griffin KL, Field CB. 1993. Assessing community type, plant biomass, pigment composition, and photosynthetic efficiency of aquatic vegetation from spectral reflectance. *Remote Sensing of Environment* 46: 110–118.

Petersen L, Minkkinen P, Esbensen KH. 2005. Representative sampling for reliable data analysis: Theory of Sampling. *Chemometrics and Intelligent Laboratory Systems* 77: 261–277.

Pinnel N, Heege T, Zimmermann S. 2005. Spectral Discrimination of Submerged Macrophytes in Lakes Using Hyperspectral Remote Sensing Data. 16.

Plaza A, Benediktsson JA, Boardman JW, Brazile J, Bruzzone L, Camps-Valls G, Chanussot J, Fauvel M, Gamba P, Gualtieri A, Marconcini M, Tilton JC, Trianni G. 2009. Recent advances in techniques for hyperspectral image processing. *Remote Sensing of Environment* 113: S110–S122.

Ramezan CA, Warner TA, Maxwell AE. 2019. Evaluation of Sampling and Cross-Validation Tuning Strategies for Regional-Scale Machine Learning Classification. *Remote Sensing* 11: 185.

Sabat-Tomala A, Raczko E, Zagajewski B. 2020. Comparison of Support Vector Machine and Random Forest Algorithms for Invasive and Expansive Species Classification Using Airborne Hyperspectral Data. *Remote Sensing* 12: 516.

Sand-Jensen K, Pedersen O. 1999. Velocity gradients and turbulence around macrophyte stands in streams. *Freshwater Biology* 42: 315–328.

Santos MJ, Anderson LW, Ustin SL. 2011. Effects of invasive species on plant communities: an example using submersed aquatic plants at the regional scale. *Biol Invasions* 13: 443–457.

Sawaya K E, Olmanson L G, Heinert N J, Brezonik P L, Bauer M E. 2003. Extending satellite remote sensing to local scales: land and water resource monitoring using high-resolution imagery. *Remote Sensing of Environment* 88: 144–156.

Sharma RC, Hara K, Hirayama H. 2017. A Machine Learning and Cross-Validation Approach for the Discrimination of Vegetation Physiognomic Types Using Satellite Based Multispectral and Multitemporal Data. *Scientifica* 8.

Silva TSF, Costa MPF, Melack JM, Novo EMLM. 2008. Remote sensing of aquatic vegetation: theory and applications. *Environmental Monitoring and Assessment* 140: 131–145.

Stephan U, Gutknecht D. 2002. Hydraulic resistance of submerged flexible vegetation. *Journal of Hydrology* 269: 27–43.

Stocks JR, Rodgers MP, Pera JB, Gilligan DM. 2019. Monitoring aquatic plants: An evaluation of hydroacoustic, on-site digitising and airborne remote sensing techniques. *Knowledge & Management of Aquatic Ecosystems* 420: 27.

Strobl C, Boulesteix A-L, Kneib T, Augustin T, Zeileis A. 2008. Conditional variable importance for random forests. *BMC Bioinformatics* 9: 307.

Tian YQ, Yu Q, Zimmerman MJ, Flint S, Waldron MC. 2010. Differentiating aquatic plant communities in a eutrophic river using hyperspectral and multispectral remote sensing. *Freshwater Biology* 55: 1658–1673.

Traganos D, Aggarwal B, Poursanidis D, Topouzelis K, Chrysoulakis N, Reinartz P. 2018. Towards Global-Scale Seagrass Mapping and Monitoring Using Sentinel-2 on Google Earth Engine: The Case Study of the Aegean and Ionian Seas. *Remote Sensing* 10: 1227.

Underwood EC, Mulitsch MJ, Greenberg JA, Whiting ML, Ustin SL, Kefauver SC. 2006. Mapping Invasive Aquatic Vegetation in the Sacramento-San Joaquin Delta using Hyperspectral Imagery. *Environmental Monitoring and Assessment* 121: 47–64.

Valta-Hulkkonen K, Kanninen AK, Ilvonen R. 2005. Assessment of aerial photography as a method for monitoring aquatic vegetation in lakes of varying trophic status. *Boreal Environment Research* 10: 57–66.

Villa P, Pinardi M, Bolpagni R, Gillier J-M, Zinke P, Nedelcuţ F, Bresciani M. 2018. Assessing macrophyte seasonal dynamics using dense time series of medium resolution satellite data. *Remote Sensing of Environment* 216: 230–244.

Vis C, Hudon C, Carignan R. 2003. An evaluation of approaches used to determine the distribution and biomass of emergent and submerged aquatic macrophytes over large spatial scales. *Aquatic Botany* 77: 187–201.

Visa S, Ralescu A. 2005. The Effect of Imbalanced Data Class Distribution on Fuzzy Classifiers - Experimental Study. IEEE, Reno, NV, USA., 749–754.

WFD European Commission. 2000. Directive 2000/06/EC of the European Parliament and of the Council of Europe (2000). Establishing a Framework for Community Action in the Field of Water Policy. Official Journal of the European Communities.

Williams DJ, Rybicki NB, Lombana AV, O'Brien TM, Gomez RB. 2003. Preliminary Investigation of Submerged Aquatic Vegetation Mapping using Hyperspectral Remote Sensing. *Environmental Monitoring and Assessment* 81: 383–392.

Wolter PT, Johnston CA, Niemi GJ. 2005. Mapping submergent aquatic vegetation in the US Great Lakes using Quickbird satellite data. *International Journal of Remote Sensing* 26: 5255–5274.

Zafari A, Zurita-Milla R, Izquierdo-Verdiguier E. 2019. Evaluating the Performance of a Random Forest Kernel for Land Cover Classification. *Remote Sensing* 11: 575.

Zhao D, Jiang H, Yang T, Cai Y, Xu D, An S. 2012. Remote sensing of aquatic vegetation distribution in Taihu Lake using an improved classification tree with modified thresholds. *Journal of Environmental Management* 95: 98–107.

Zhao D, Lv M, Jiang H, Cai Y, Xu D, An S. 2013. Spatio-Temporal Variability of Aquatic Vegetation in Taihu Lake over the Past 30 Years. *PLOS ONE* 8: e66365.

Figure 1. Study area. (a) Location of the study area in the southwest of France, north of Toulouse city on a raw Pléiades image; (b) Zoom in on a raw Pléiades image (50 cm of resolution) of the 1 km study site (Seilh).

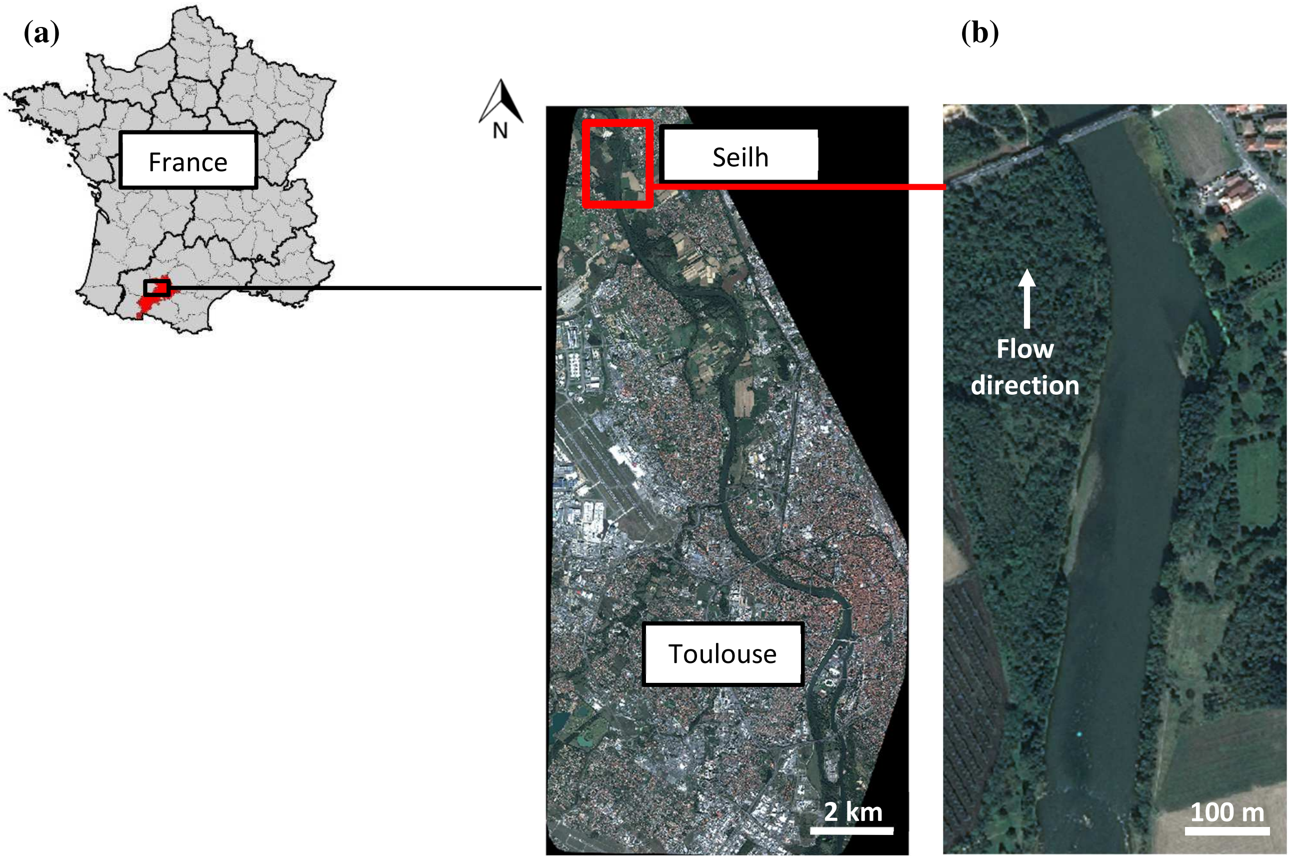
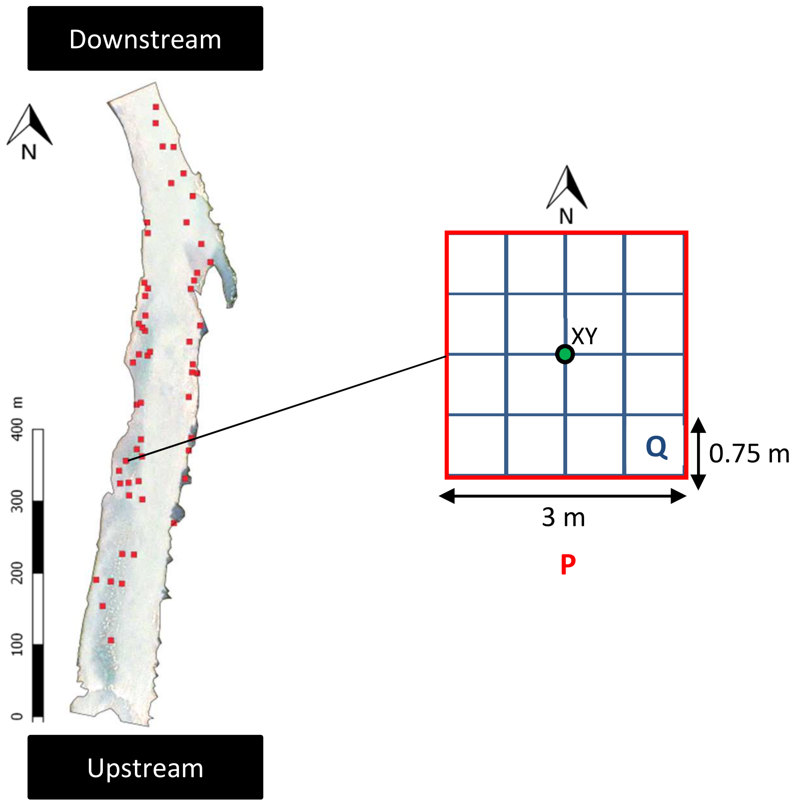


Figure 2. Sampling protocol for macrophyte monitoring.

(a) Location of the 55 sampling plots on the site (Seilh). The red squares indicate the geographical positions of the 55 field sampled plots (P).

(b) Diagram of a plot (P): the red square represents the sampling plot of a 9 m² surface and the blue squares correspond to quadrats (Q). The green "XY" dot symbolizes where GPS coordinates were taken.



(a) (b)

Figure 3. Drone images of the study site. (a) Image acquisition missions. Each point refers to a shooting, each colour corresponds to a mission; (b) Example of a raw drone image shot 30 m high above plurispecific macrophyte meadows; (c) Orthomosaic image fragment of the study site. The black rectangle represents the position of the individual shot in figure b.

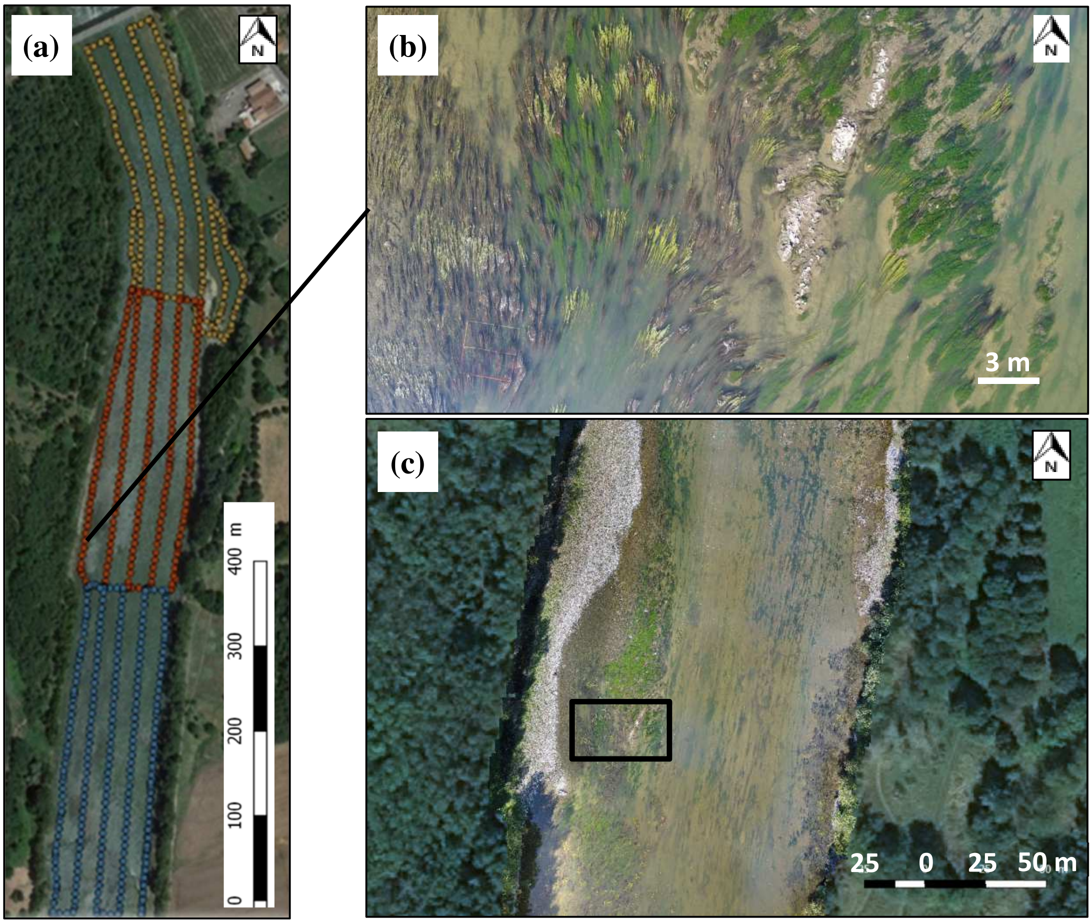


Figure 4. Flow chart illustrating the process to get total cover predictions from the Pléiades satellite image. Sampling entities can be 55 plots of 3 m side, 55 entities of 1.5 m side, or *n* quadrats of 0.75 m side (with *n* = 15, 30, 55, 110, 220, 440 or 880).

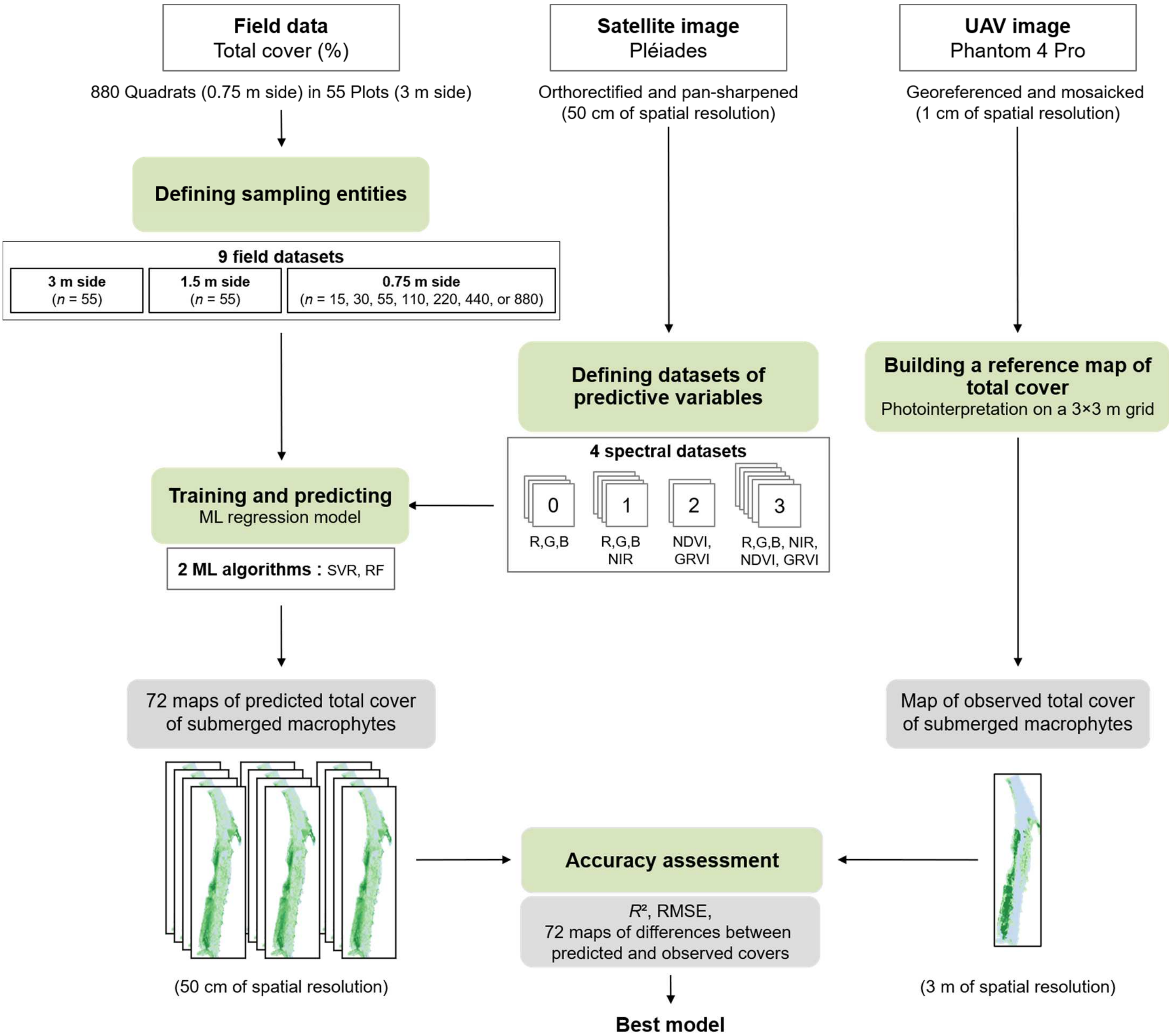
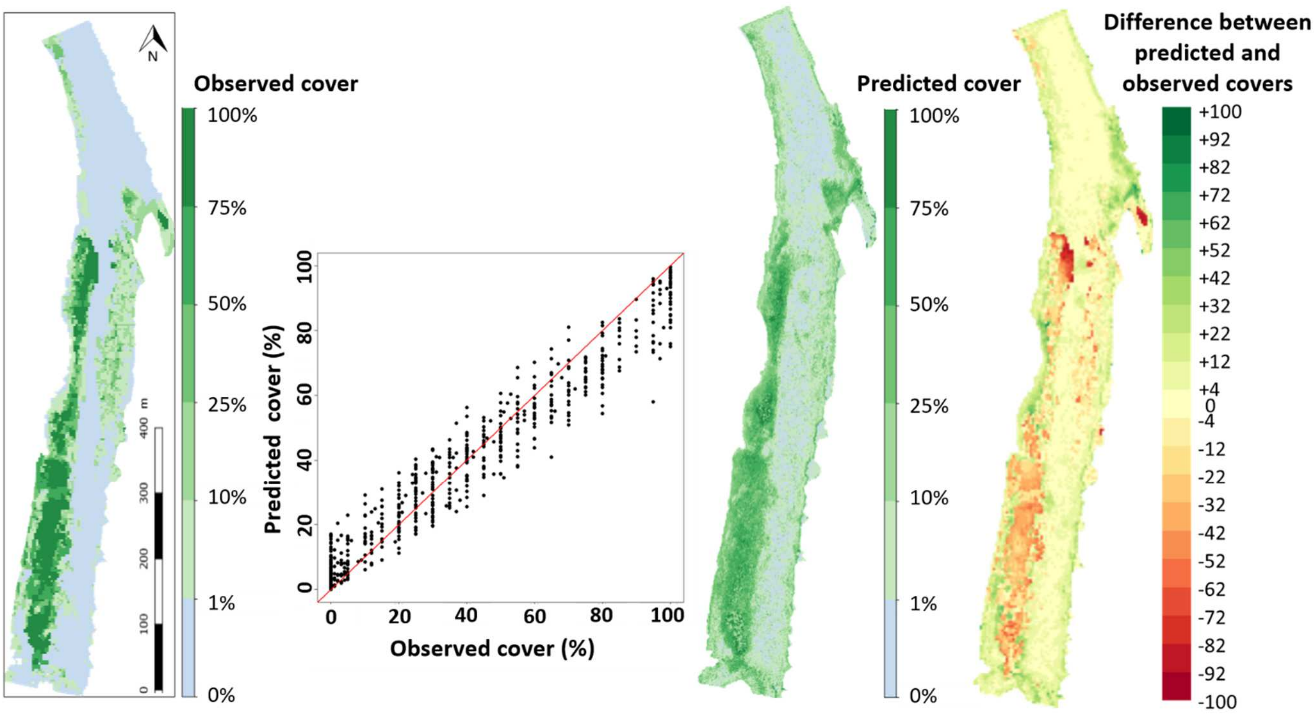


Figure 5. Results from the Random Forest regression model using “dataset 1” (RGB and NIR) and based on 880 quadrats. (a) Observed cover map obtained from drone image photointerpretation (b) Predicted cover as a function of observed cover within quadrats; (c) Predicted cover map and (d) Difference map between predicted and observed covers.



(a) (b) (c) (d)

Figure 6. Learning curves for the Random Forest regression model using datasets 0 (RGB) and 1 (RGB and NIR): RMSE as a function of the number of quadrats of 0.75 m side.

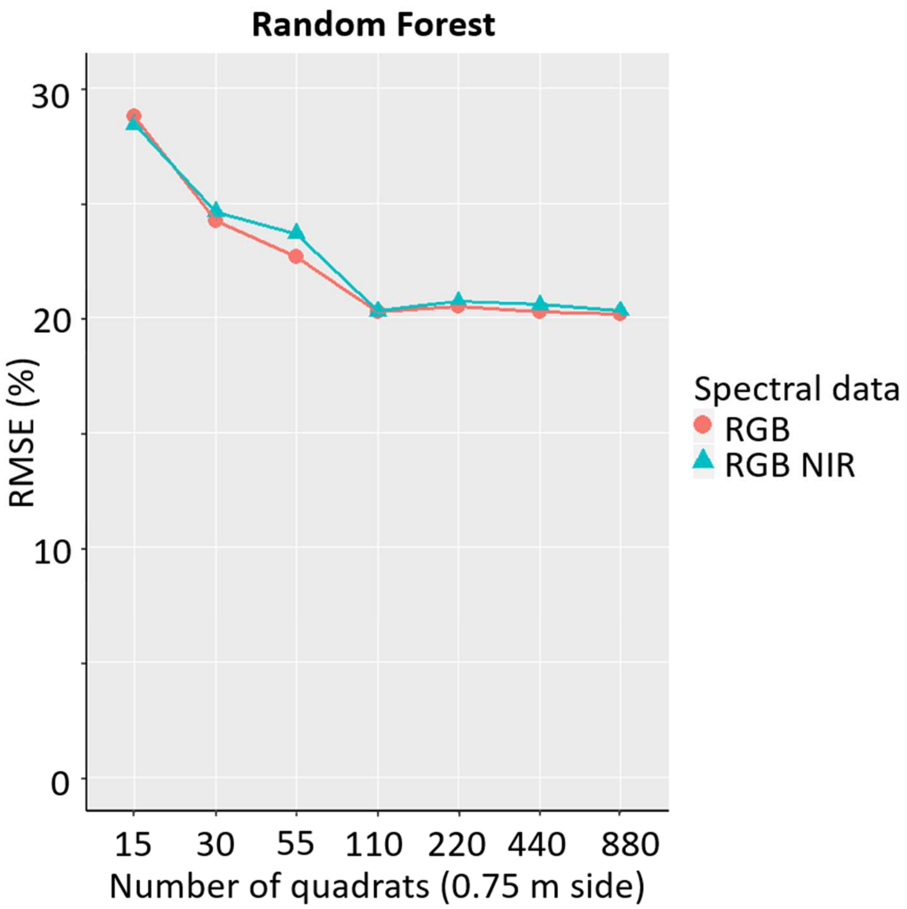


Table 1. Number of sampled entities according to macrophyte density ranges. Percentages in brackets represent the proportion of each cover group within the total number of samples.

Macrophyte cover	Plots (P)	Quadrats (Q)
0-1 %	16 (30 %)	359 (40 %)
1-10 %	6 (11 %)	65 (7 %)
11-25 %	7 (12 %)	99 (11 %)
26-50 %	12 (21 %)	160 (18 %)
51-75 %	10 (18 %)	100 (11 %)
>75 %	4 (7 %)	97 (11 %)
Total number of entities	**55**	**880**

Table 2. Prediction accuracy for different runs (differing by the algorithm, spectral dataset, or sampling level), assessed with the R^2 coefficient of determination and the RMSE; only results involving datasets 0 and 1 are presented. This table is an extract of the most relevant results of table S2 (see Supporting Information).

Algorithm	**Spectral data**	**Entity side size (m)**	**Number of entities**	**Maximum predicted cover (%)**	R^2	**RMSE (%)**
Random Forest	RGB NIR ("dataset 1")	3	55	87.3	0.69	21.28
		0.75	880	100.0	0.71	20.30
Support Vector Regression	RGB ("dataset 0")	3	55	99.0	0.72	20.37
		0.75	880	100.0	0.74	19.40
	RGB NIR ("dataset 1")	3	55	100.0	0.72	21.28
		0.75	880	100.0	0.67	21.29