European Journal of Remote Sensing

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/tejr20

Biomass retrieval based on genetic algorithm feature selection and support vector regression in Alpine grassland using ground-based hyperspectral and Sentinel-1 SAR data

Eugenia Chiarito, Francesca Cigna, Giovanni Cuozzo, Giacomo Fontanelli, Abraham Mejia Aguilar, Simonetta Paloscia, Mattia Rossi, Emanuele Santi, Deodato Tapete & Claudia Notarnicola

To cite this article: Eugenia Chiarito, Francesca Cigna, Giovanni Cuozzo, Giacomo Fontanelli, Abraham Mejia Aguilar, Simonetta Paloscia, Mattia Rossi, Emanuele Santi, Deodato Tapete & Claudia Notarnicola (2021) Biomass retrieval based on genetic algorithm feature selection and support vector regression in Alpine grassland using ground-based hyperspectral and Sentinel-1 SAR data, European Journal of Remote Sensing, 54:1, 209-225, DOI: 10.1080/22797254.2021.1901063

To link to this article: https://doi.org/10.1080/22797254.2021.1901063

Published online: 21 Mar 2021.

EUROPEAN JOURNAL OF REMOTE SENSING
2021, VOL. 54, NO. 1, 209–225
https://doi.org/10.1080/22797254.2021.1901063

Taylor & Francis
Taylor & Francis Group

Biomass retrieval based on genetic algorithm feature selection and support vector regression in Alpine grassland using ground-based hyperspectral and Sentinel-1 SAR data

Eugenia Chiarito [a,b], Francesca Cigna[c], Giovanni Cuozzo[a], Giacomo Fontanelli[b], Abraham Mejia Aguilar[a], Simonetta Paloscia[b], Mattia Rossi[a,b], Emanuele Santi[b], Deodato Tapete[c] and Claudia Notarnicola[a]

[a]Institute for Earth Observation, EURAC Research, Bolzano, Italy; [b]Earth Observation, Institute of Applied Physics - National Research Council of Italy (IFAC–CNR), Sesto Fiorentino, Italy; [c]Italian Space Agency (ASI), Rome, Italy

ABSTRACT

A general framework for the integration of multi-sensor data for dry and fresh biomass retrieval is proposed and tested in Alpine meadows and pastures. To this purpose, hyperspectral spectroradiometer (as simulation of hyperspectral imagery) and biomass samples were collected in field campaigns and Copernicus Sentinel-1 Interferometric Wide (IW) swath SAR backscattering coefficients were used. First, a genetic algorithm feature selection was performed on hyperspectral data, and afterwards the resulting most sensitive bands where combined with SAR data within a support vector regression (SVR) model. The most sensitive hyperspectral bands were mainly located in different regions of the SWIR range for both fresh and dry biomass, and in the red and near-infrared regions mainly for dry biomass, but with less influence for fresh biomass. The R^2 correlation values between the sampled and the estimated biomass range from 0.24 to 0.71. The relatively low performances are mainly related to the saturation effect in the optical bands, as well as to the paucity of points for high values of biomass. The methodology allows a better understanding of the interaction between grassland systems and the electromagnetic spectrum by offering a model with a reduced number of narrow bands in the context of a multi-sensor integration.

ARTICLE HISTORY
Received 1 October 2020
Revised 28 February 2021
Accepted 6 March 2021

KEYWORDS
Biomass retrieval; feature selection; support vector regression; hyperspectral and SAR data integration; Alpine environment; grassland

Introduction

The study of grasslands in Alpine regions is of great interest as they play a key role in mountain environments dynamics such as water storage, biodiversity conservation, slope stabilizers, carbon sinks, and fodder for livestock (Kumar & Mutanga, 2017; Rossi et al., 2019; J. Wang et al., 2019). Grassland biomass assessment allows for monitoring of vegetation productivity and health (Cho et al., 2007), and works as an indicator of climate change, when potential earlier green-up phases or elevated exposure to drought events are detected (Stendardi et al., 2019). To improve the monitoring of biomass, frequent temporal monitoring at a regional scale is required (Kumar & Mutanga, 2017).

The most traditional and widely used techniques for biomass estimation are based on in situ destructive sampling with the removal of vegetation of a predefined sample unit of the grassland and laboratory analysis. This methodology shows strong limitations to provide detailed temporal and spatial information and is highly conditioned by its costs (L. Chen et al., 2018; J. Wang et al., 2019). Remote sensing techniques are a cost-effective alternative, especially with the recent launch of new satellites (e.g. Copernicus Sentinel fleet) and availability of more frequent and better-detailed imaging (Kumar & Mutanga, 2017; J. Wang et al., 2019), with increasing influence in farm management (Stendardi et al., 2019).

For biomass retrieval, different approaches are proposed depending on the data source and availability. Mainly they can be divided into parametric and non-parametric, depending on the applied technique, by using optical, synthetic aperture radar (SAR) or multi-source data as input information. In the case of optical data, two different sensors typologies can be identified, depending on their bandwidth, and therefore on the number of bands within the optical domain: multispectral broad bands and hyperspectral narrow bands sensors. In the analysed papers, multispectral processed data are mainly related to space-borne sensors (e.g. NASA/USGS Landsat 8 OLI and Copernicus Sentinel-2; Chang & Shoshany, 2016; L. Chen et al., 2018; Naidoo et al., 2019; Stendardi et al., 2019; J. Wang et al., 2019). Due to the limited availability of hyperspectral satellite data, most studies are based on ground and airborne data collected with spectroradiometers or on simulated data.

Vegetation indexes (VIs) and single-band ratios (SRs) are the main explored features, especially the normalized difference vegetation index (Chang &

CONTACT Eugenia Chiarito eugechiarito@gmail.com Instituto Geográfico Nacional (IGN), Buenos Aires, Argentina

Shoshany, 2016; L. Chen et al., 2018; Naidoo et al., 2019; J. Wang et al., 2019). Other less widespread models include single-band data (L. Chen et al., 2018; Naidoo et al., 2019), visible and near-infrared (NIR) band ratios (Naidoo et al., 2019), and other indices such as enhanced vegetation index (EVI), land surface water index (Chang & Shoshany, 2016; J. Wang et al., 2019), or biophysical variables such as leaf area index (LAI) (L. Chen et al., 2018) or fraction vegetation cover (Chang & Shoshany, 2016; L. Chen et al., 2018; Shimabukuro et al., 1998). The methodologies to build the retrieval models include parametric techniques such as multiple linear regression models (MLR) (Chang & Shoshany, 2016; Psomas et al., 2011; J. Wang et al., 2019) and geographically weighted regression (L. Chen et al., 2018), or non-parametric techniques, mainly artificial neural networks (L. Chen et al., 2018), support vector regression (SVR) (L. Chen et al., 2018; J. Wang et al., 2019), or random forest (RF) (L. Chen et al., 2018; Naidoo et al., 2019; J. Wang et al., 2019). Factors such as light spectra combinations, resolution, instrument, and platform specifications play an important role to define the mathematical expression that best quantifies vegetation parameters like biomass (Xue & Su, 2017). Relationships between biomass and other vegetation properties and spectral reflectance in grasslands were found in literature and define functional relationships between absorption features and wavelength (Adam et al., 2014; Fava et al., 2009; Naidoo et al., 2019; Psomas et al., 2011; Pullanagari et al., 2018). However, when the growth processes become discontinuous and the senescent or reproductive material is included, it is difficult to describe biomass by using simple relationships (Darvishzadeh et al., 2008). In some cases, more than one VI can be needed to describe the whole phenological behaviour (J. Wang et al., 2019). For these reasons, training customized algorithms can be useful to try to reach an optimal exploitation of the available (remote and in situ) data, especially for non-linear complex relationships (Pullanagari et al., 2018).

Very few studies were found where hyperspectral selected features are combined with SAR information to retrieve biomass in Alpine grasslands. Results show that combining the texture characteristics and volume information of SAR data with the vegetation biophysical status of optical data can result in more sensitive (and accurate) estimation, especially for wavelength ranges where optical signal saturates during the advanced phases of plant growth and thus a high amount of biomass (Kumar & Mutanga, 2017; J. Wang et al., 2019). Performances improve when multi-sensor data are fused, with similar accuracy for different machine learning techniques (J. Wang et al., 2019). Deeper analysis is necessary for narrow bands sensitivity to biomass. Models based on hyperspectral and SAR data fusion are still barely explored, but literature suggests that they will improve the predictive accuracy compared to the current state of the art (Fava et al., 2009).

Specialized bibliography showed that combining optical and SAR data provides more robust models, but narrow hyperspectral bands were not fused to SAR bands for grassland monitoring yet. Nevertheless, it is expected to become of greater use in the future thanks to the increasing availability of satellite SAR and hyperspectral data. It is worthwhile mentioning some satellite missions such as the radar constellations Copernicus Sentinel-1, the Italian Space Agency's (ASI) COSMO-SkyMed (Caltagirone et al., 2014), and ASI's Hyperspectral Precursor of the Application Mission (PRISMA) launched in March 2019 (e.g., Loizzo et al., 2016), as well as other planned hyperspectral missions, such as the German Aerospace Center's (DLR) EnMAP (Guanter et al., 2015).

In this context, the main objective of this paper is to implement a robust and flexible algorithm, defined considering the latest and new generation of SAR and hyperspectral data, and based on a two-step procedure:

- A feature selection obtained by means of a genetic algorithm to analyse which portions of the spectrum from hyperspectral data contain relevant information for biomass estimation, also considered in the perspective of the new satellite mission PRISMA. The choice of a genetic algorithm is due to its capability of selecting bands with the highest information content, taking also into account their adjacency relationship (Leardi, 2000).
- A regression model based on SVR to test in Alpine grasslands, where well-assessed functional relationships can be faced with the difficulties described before by using both the hyperspectral bands derived from the previous analysis and SAR data to estimate biomass. SVR approach was selected for its capability of generalization of the extracted model also when the number of available training samples is limited (Pasolli et al., 2011).

By combining hyperspectral and SAR data, this synergetic approach can contribute to a better understanding of grassland systems interaction with different portions of the electromagnetic spectrum and provide a more robust and accurate retrieval model.

State of the art

Hyperspectral sensors with narrower bands provide more detailed information about vegetation dynamics and are used to develop new semi-empirical indices

(Cho et al., 2007; Darvishzadeh et al., 2008), as well as create more sensitive models. Specifically, the hyperspectral features can overcome optical multiband sensors saturation issues when biomass exceeds certain thresholds (J. Chen et al., 2009). Well-known spectral features for biomass have been identified in literature such as the red edge (L. Chen et al., 2018; Clevers et al., 2007; Fava et al., 2009; Psomas et al., 2011; C. Wang et al., 2017) and the water absorption bands (Psomas et al., 2011; C. Wang et al., 2017). Notwithstanding the interaction between spectral measurements and vegetation has been widely studied, when hyperspectral data are acquired under specific field conditions, many external factors affect the spectral response (Hong et al., 2020), such as instrument characteristics, soil and plant conditions, and land cover heterogeneity. To this purpose, a more general and flexible approach is needed to consider the ecosystem complexity and variability in mountain environment. It is the case of SVR models, known because of their generalization ability for a given limited data set (De Almeida et al., 2019; Monnet et al., 2011). It shall be mentioned that the applicability of the remote sensing products relies heavily on the instruments and platforms used in each particular issue (Xue & Su, 2017), the seasonal vegetation stage, and regional climate constraint (X. Wang et al., 2020). Hyperspectral imaging provides unique information to build the spectral signature of the observed surface. This configures a complex, high-dimensionality data set, where only a small number of bands among the hundreds registered may be related to the desired traits (Moghimi et al., 2018; Pal & Foody, 2010). The research included different successful techniques to identify the EP (Cho et al., 2007), the red edge inflection point, soil-adjusted VI, band depth indices (J. Chen et al., 2009), and first-order derivative reflectance and regression models, like partial least squares (J. Chen et al., 2009; Cho et al., 2007; Darvishzadeh et al., 2008). However, the tested VIs models show an underestimation of high biomass values, given the signal saturation in these phases. An improvement is found when red edge bands are involved in place of red bands (Cho et al., 2007). When comparing with multispectral data, hyperspectral sensors outperform the broader bands sensors, attaining higher coefficients of determination and lower root mean square error (RMSE) (Fava et al., 2009; Naidoo et al., 2019). Finally, shifting the band centre was found to affect the model accuracy (Fava et al., 2009).

Considering the high number of hyperspectral bands available, models based on few parameters (such as VIs and single bands) may not entirely exploit all the available information (Moghimi et al., 2018; Xue & Su, 2017). However, if a high number of bands are analysed, multicollinearity effect is likely to occur (Cho et al., 2007). To reduce the overfitting risk, computational cost, and storage optimization, feature selection preprocessing is desired to achieve a robust retrieval model (Pullanagari et al., 2018; Rasel et al., 2019; C. Wang et al., 2017). This allows gaining predictive ability and retaining meaningful features with respect to a given task (Bradley et al., 2018; Xue & Su, 2017). Moreover, some studies showed that SVM approaches are sensitive to data set dimensionality reduction (Gidudu & Heinz, 2007; Jain, 1997; Pal & Foody, 2010).

VIs and SR sensitivity were tested (Fava et al., 2009), combining bands between 400 and 1000 nm, reaching over 180,000 parameters, using ordinary least squares linear regression. Some of the analysed methods include interactive variables selection, elimination of uninformative variables, iterative predictor weighting, and genetic algorithms (GA) (Leardi, 2000). Other possible approaches are stepwise forwards selection, SVM band shaving (Clevers et al., 2007), and RF approach (Rasel et al., 2019). However, even though RF is useful for sensitivity analysis, this approach does not provide information about the optimal number of bands, and in addition it needs to be complemented with other procedures such as recursive feature elimination (Rasel et al., 2019). Another stepwise approach involving multiple linear regression (SMLR) was tested (C. Wang et al., 2017), which combines a forward selection and a backward elimination: the independent variables are imported one at a time based on their significance on the regression model. All the involved variables are analysed to test if any will be removed using a significant criterion until no more variables can be imported or eliminated. Another methodology proposes a combination between partial least squares regression models and RMSE leave-one-out cross-validation technique (J. Chen et al., 2009). Given the high-dimensionality problem, feature selection based on an exhaustive search method is not possible: the number of possible bands combination would lead to an unfeasible approach. The random search method is used instead as it was performed in the above-mentioned publications. To reduce the convergence times to the optimal solution, the random search can be modified into a guided random search, in which attention is adaptively increased to focus on the band combinations with the most promising results (Gidudu & Heinz, 2007). It is the case for GA feature selection, combined to SVR regression models. All the former mentioned techniques, when applied to hyperspectral bands, normally behave in such a way as to have the selected bands spread along the whole spectrum, while the GA has the capability of selecting less dispersed bands and producing more interpretable results (Leardi, 2000). This model ability for feature selection was tested previously, and a special approach was proposed to

minimize the risk of random correlations (Leardi & Lupáñez González, 1998). A methodology to account for the autocorrelation among adjacent bands, i.e. if one of them is selected as relevant, the ones surrounding it should also have a high probability of being selected in the final model (Leardi, 2000). However, to our knowledge, no study proved to be globally effective for biomass retrieval, and a small number of studies explored these advanced feature selection techniques based on GA in hyperspectral data to create an empirical reliable model afterwards.

To overcome the saturation problem that optical data present for high biomass values (Chang & Shoshany, 2016; Kumar & Mutanga, 2017), complementary information was found in the microwave region of the spectrum, where SAR data play an important role by offering vegetation structure/density volumetric information (Chang & Shoshany, 2016; Naidoo et al., 2019). SAR sensors sensitivity to topography remains the main limitation, but the integration of both radar and optical data proved to be useful, strengthening the retrieval performances, both for vegetation and soil moisture estimation (L. Chen et al., 2018; Ghasemi et al., 2011; Naidoo et al., 2019; Stamenkovic et al., 2017; J. Wang et al., 2019). Although studies on biomass using SAR data are mostly focused on forests, some recent research showed Sentinel-1 C-band capability to detect grassland phenology stages (Stendardi et al., 2019) and reduce the estimation errors, compared to optical models (J. Wang et al., 2019). The data extracted from SAR sensors for grassland studies are based on backscattering coefficients σ0. Sentinel-1 provides data in two polarizations: VV and VH. Although the latter is usually more sensitive to biomass (Chang & Shoshany, 2016), it is also more affected by topographic effects. To overcome this issue, the intensity ratio ($\bar{\sigma}^0_{VH}/\bar{\sigma}^0_{VV}$) was also tested (Kumar & Mutanga, 2017; Naidoo et al., 2019).

Materials and methods

Study area

The study area is part of the Long-Term Socio-Ecological Research (LT(S)ER) sites and is located in Matsch/Mazia Valley (Central Alps) in South Tyrol, Italy. The LT(S) ER project was launched with the aim to monitor ecological and climatic developments. The valley extent is of around 9000 ha, with elevation going from 920 m a.s.l. (Sluderno) to 3738 m a.s.l. (Palla Bianca). The main land use is agriculture, with a strong presence of intensively managed meadows (Bertoldi et al., 2010; Pasolli et al., 2011).

Figure 1(d) shows the study area, located in Muntatschinig, at 1450 m a.s.l. approximately. Both meadow and pasture show similar climate and soil characteristics as they are located close to each other. The mean annual temperature is around 6.5°C, and

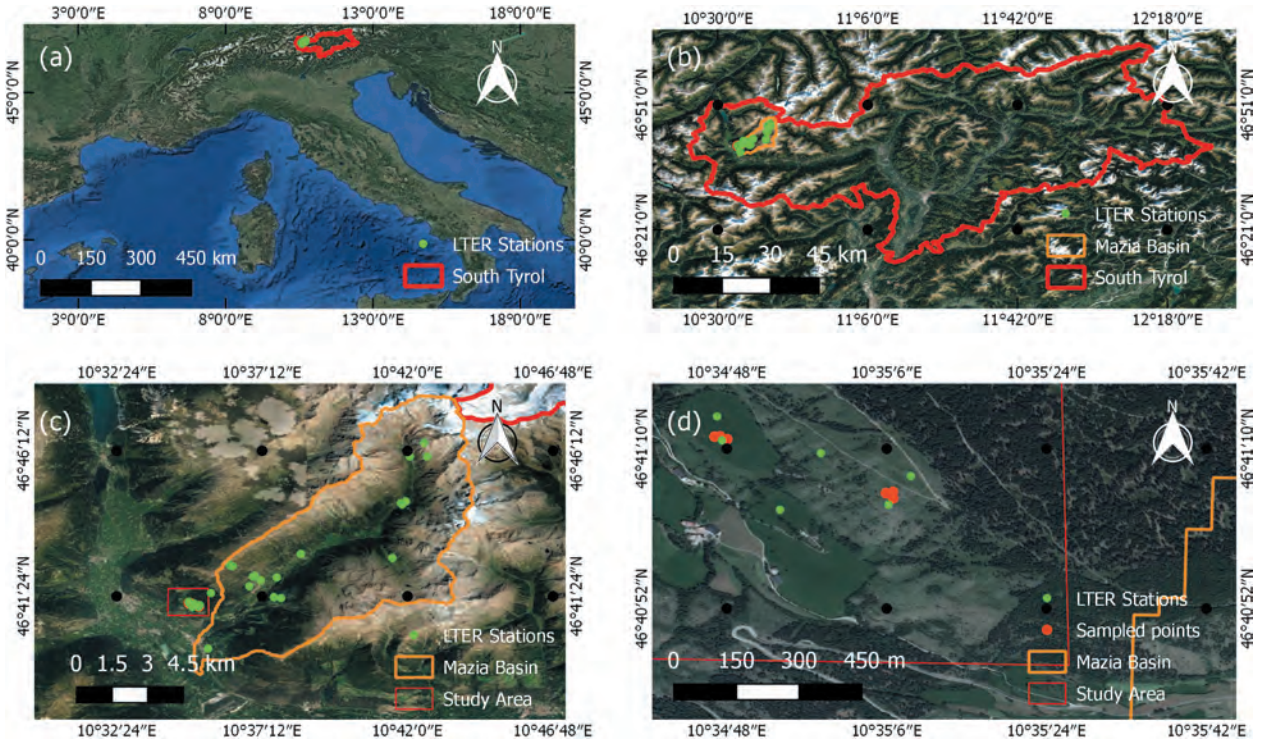

Figure 1. (a) Location of South Tyrol province in the northeast of Italy. (b) South Tyrol province in red, Mazia Valley in orange. (c) LTSER climate stations in Mazia Valley. The red square in the south west contains the studied area. (d) Zoom to the study area: in red the sampled spots (in the east the pasture, in the west the meadow); in green the LTSER stations. The orange line belongs to Mazia Valley basin boundary.

the mean annual precipitation is 525 mm (in Mazia 1580 m a.s.l.), configuring the area as one of the driest areas in the Alpine region. While the pasture grows naturally and has no management, the meadow has an artificial irrigation system that is actioned every 10 days approximately depending on the weather, is fertilized mechanically in early spring and late autumn and episodically by livestock, and is harvested up to three times each season as a source of fodder (Rossi et al., 2019).

Data acquisition and processing

Field campaign data from three consecutive years (2017–2019) from May to October were analysed in the current study, with sampling frequency of 15 days (subjected to weather conditions), between 10:00, 14:00 CEST. In total 26 campaigns were carried out: 68 samples were collected in meadow and 73 in pasture fields. The procedure consisted in collecting contemporary hyperspectral reflectance data and biomass sampling. The data set was completed with Sentinel-1 IW swath SAR backscattering coefficient data acquired over the same period of the field campaigns. The workflows of each component of the methodology are detailed in the following sections.

It is worth mentioning that the methodology was developed with the aim of exploiting in the future PRISMA hyperspectral and COSMO-SkyMed StripMap PingPong dual-polarization images. These were tasked to be acquired over the study region and, in particular, the PingPong data as part of a tailored experimental monitoring campaign carried out since mid-2019 in the framework of the project ALGORITMI (Tapete et al., 2020). While building an image stack via these bespoke campaigns, as the first approach the methodology was tested using spectroradiometer in situ measurements and Sentinel-1 C-band data acquired in 2017–2019, with the aim to develop the methodology and run initial tests, in view of future experiments with PRISMA and COSMO-SkyMed PingPong imagery.

Biomass sampling

Each campaign consisted in sampling both meadow and pasture fields. Three to four different spots were selected each time for each vegetation type, and a destructive method was used for biomass sampling within a georeferenced 50 × 50cm plot. The vegetation samples were removed at ground level and fully stored. The fresh samples were weighted and dried in laboratory oven at 85°C and weighted again to measure both dry weight and plant water content. Temporal trend of biomass mean sampled values for meadow and pasture for the 2019 campaign are displayed in Figure 2.

Hyperspectral data acquisition and pre-processing

Within each plot three to four spectral signatures were registered (depending on the plot heterogeneity) to reduce sources of error and uncertainties before the destructive plant biomass samples collection. The spectroradiometer measurements were averaged per vegetation type. The used spectroradiometer was an SVC HR-1024i (Spectra Vista Corporation, New York, NY, USA) with 1024 spectral bands ranging from 338 to 2500 nm. The measurements were collected with the instrument mounted on a tripod at around 1.2 m above ground level. A reference measurement was made at the beginning of each plot registration

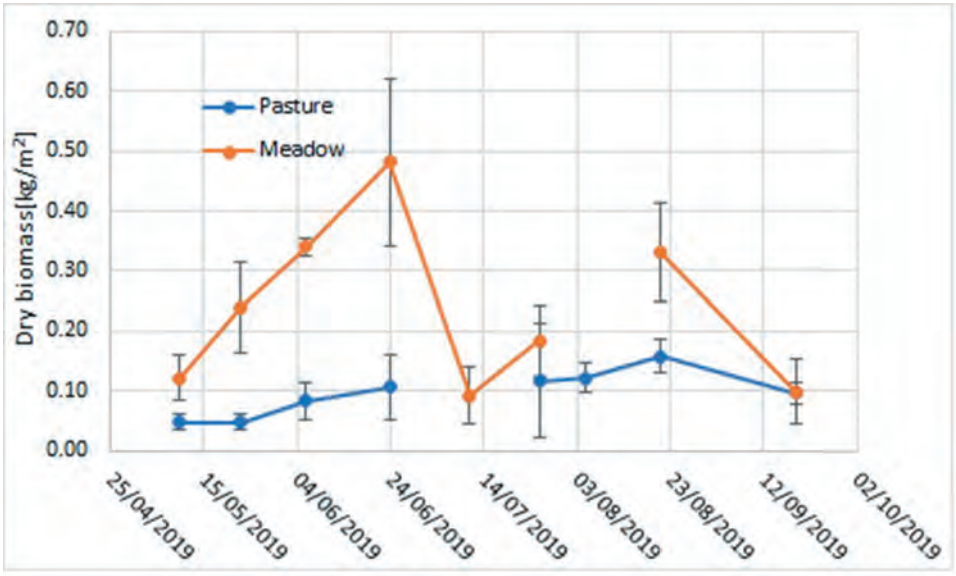

Figure 2. Temporal trend of dry biomass field campaigns conducted in pasture and meadow in 2019. Note: the points are the real collected samples (daily average), and the line between them is set just to provide an idea of the phenological trend. No real interpolation is intended.

using a white reflectance panel (Spectralon) to acquire reflectance conditions. The main aim of using a field spectrometer was to simulate the PRISMA bands: to this purpose, all the 1024 bands were resampled to match PRISMA bands. The spectra integration into wider bands was performed considering PRISMA bandwidth as an input. Since the spectral response distribution was not available at the moment of the current study because the mission was in the commissioning phase, the mean value within each band range was calculated. To perform this processing, R *hsdar* package was implemented (Lehnert et al., 2018). This step addresses two goals: first, to reduce the number of bands (reaching a more suitable correlation between the number of biomass samples compared to the number of hyperspectral bands); and most importantly, to emulate the PRISMA specifications.

SAR data acquisition and pre-processing

Sentinel-1 (S1) images were considered by selecting acquisition dates to minimize the time lapse between field campaigns and satellite overpasses. Table 1 contains a summary of both dates. The time lag varies between 0 and 2 days in all cases but the first three campaigns, with the maximum distance of 5 days.

SAR Single Look Complex products were downloaded from the Copernicus Open Access Hub https://scihub.copernicus.eu/dhus/#/home, and pre-processing was performed by using the SNAP (Sentinel Application Platform) software: http://step.esa.int/main/toolboxes/snap/. Pre-processing steps included calibration, multilooking, filtering, and geometric correction, with layover and shadowing analysis. The result was a set of geocoded intensity images that were finally converted to backscattering coefficients.

Table 1. Field campaign and Sentinel-1 (S1) satellite overpass dates.

Field campaign	S1 overpass
07/08/2017	02/08/2017
21/08/2017	26/08/2017
13/09/2017	07/09/2017
11/10/2017	13/10/2017
09/05/2017	10/05/2017
08/05/2018	07/05/2018
17/05/2018	18/05/2018
01/06/2018	31/05/2018
09/06/2018	11/06/2018
30/06/2018	30/06/2018
13/07/2018	12/07/2018
10/05/2019	08/05/2019
23/05/2019	25/05/2019
06/06/2019	06/06/2019
24/06/2019	25/06/2019
28/06/2019	30/06/2019
11/07/2019	12/07/2019
26/07/2019	25/07/2019
05/08/2019	05/08/2019
21/08/2019	18/08/2019

Retrieval model

A single data set was built by merging biomass, hyperspectral, and SAR data for each date and for each plot, including all three-year campaigns (Figure 3). Meadow and pasture data were processed separately to analyse whether the differences observed in biomass temporal trends and spectral response ranges had any influence on the feature selection and model regression. All three yearly campaigns were compared to each other to assure that they can be merged in a single data set. This procedure was performed for all dates for hyperspectral and biomass data. Figure 3 shows graphically the data for the three acquisition years.

In the case of meadow, it was observed that the rate between the number of samples collected shortly before harvest compared to the total amount of samples is much lower, thus misrepresenting this particular vegetation condition. Simulated data were added to the data set to evaluate if any improvement is reflected in the results. Daily average and standard deviation for the date before harvest were used as input to randomly create 12 new samples within one standard deviation threshold.

The flow chart in Figure 4 summarizes the complete procedure for biomass retrieval using hyperspectral and SAR data. Each step is explained in detail in the following sections. The data set was split into two sets, using 80% for training and 20% for validation. Feature selection is applied to identify the most relevant spectral bands by using the GA approach. The following step is the training of an SVR approach to create a regression model based on the selected features.

Feature selection

To better understand hyperspectral bands sensitivity to biomass estimation, a GA approach was applied, based on Leardi (2000). The proposed methodology aims at identifying which regions of the spectrum are more significant for the regression model. The advantage GA offers is the possibility of assigning a high probability of selection to a given band if an adjacent band was highlighted as relevant during the algorithm run. The outcome is a feature selection grouped in some regions of the spectrum instead of a sparser band's selection.

GA technique is inspired by natural evolutionary dynamics turning feature selection procedure into a combinatory and optimization problem. It is focused on finding the best genes (variables) out of a random limited number of individuals (population). Each individual represents a possible solution that combines some of the available features in a boolean vector, indicating if each band shall be considered or not. The GA fitness evaluation is performed iteratively, considering a new population

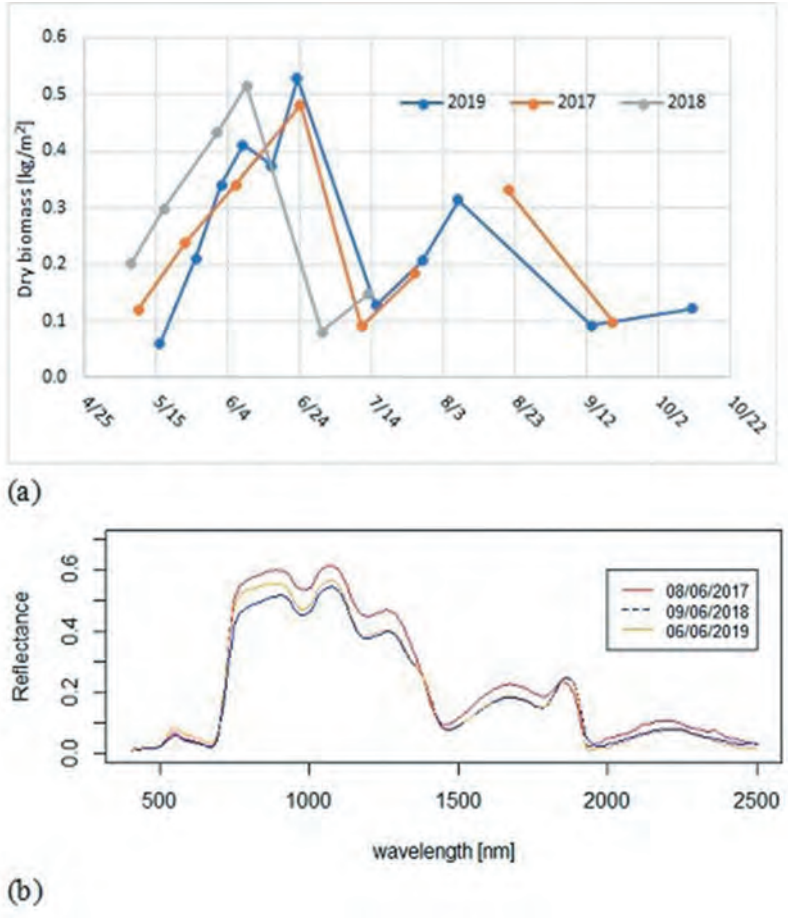

Figure 3. (a) Daily averaged temporal trends of dry biomass for meadow, overlapping all three campaigns. Note: the points are the real collected samples (daily average), and the line between them is set just to provide an didea of the phenological trend. No real interpolation is intended. (b) Spectral response range for dry biomass for meadow from in situ spectroradiometer measurement for a specific date: the same analysis was performed for every available group of dates to assure all three years could be merged in a single data set.

generation after generation, creating more fitted individuals after each run (Leardi, 2000; Leardi & Lupáñez González, 1998).

The methodology's main weakness is noise-modelling risk since a high number of models are tested, thus the best performance may be related to random instead of physical correlations. Cross-validation can partially reduce this risk, but the problem persists as the objects used to test the model performance are the same used for feature selection (Leardi, 2000). The proposed alternative is to perform several independent GA short runs, training the model with information from previous runs instead of performing a single long GA run.

The model parameters are defined based on indications about suitable set-up in the specialist literature (Leardi & Lupáñez González, 1998) as follows:

- Population size: 30 chromosomes
- Number of genes (considered bands) in the initial population: five
- Model fitting function: RF
- Method: repeated cross-validation (five folds)
- Response: minimizing RMSE
- Probability of mutation: 1%
- Number of runs: 50
- Window size smoothing: three bands wide

The rationale behind this parameterization is to achieve a fast increase in the response performance at an early stage of the process by allowing high elitism, a limited population size, and high probability of mutation in order to reduce the overfitting risk due to the evaluation of a high number of models (Leardi & Lupáñez González, 1998).

The modifications to the original algorithm (Leardi, 2000) create a vector on how likely each band may be selected in the final model. The algorithm starts creating a random population, and at the end of each individual run, an optimal solution is given. Initially the probability that a given band is selected is the same for all bands and equal to the ratio between the number of bands in

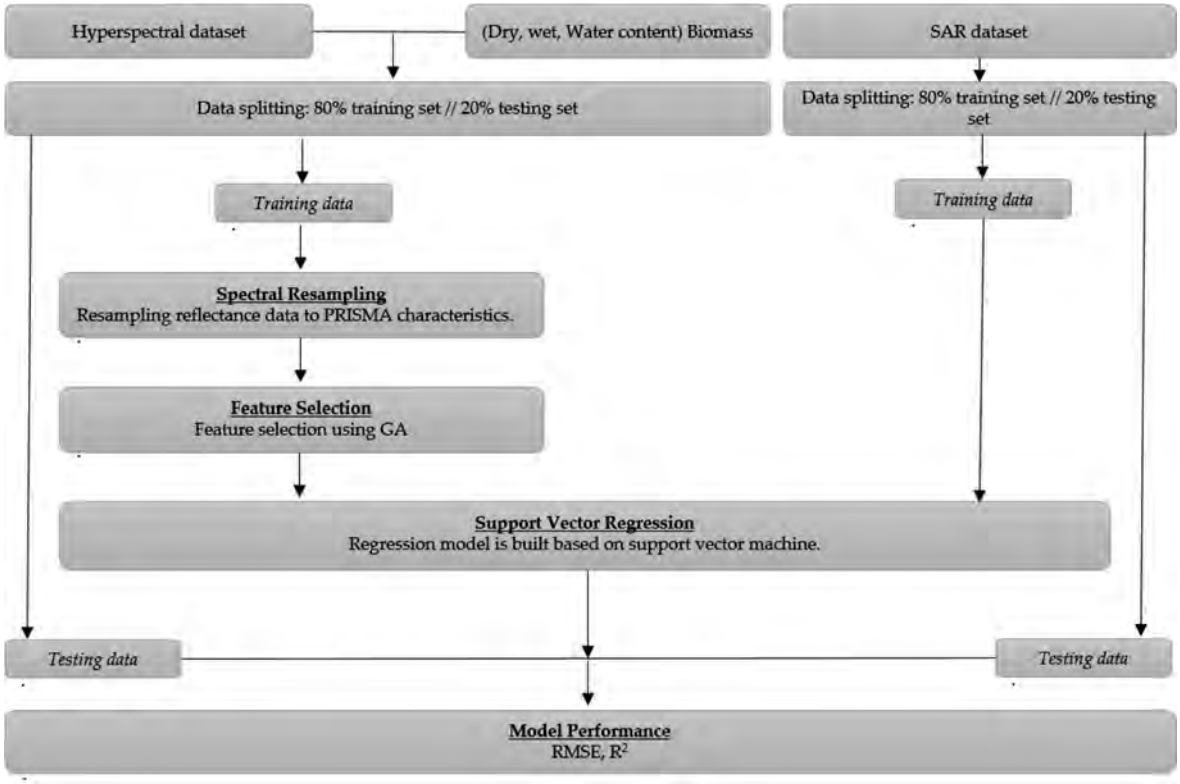

Figure 4. General processing flow chart containing the algorithm main steps to achieve a biomass prediction model from a hyperspectral and SAR data set.

a chromosome and the total number of bands. A simplified flow chart is displayed in Figure 5.

The frequency with which each band is selected after a run is of great utility to evaluate the band overall sensitivity. Therefore, information about the previous run outcome is introduced as an input, increasing the probability of a previously selected band to be selected again in the future analysis. The relevance given to this information increases along with the number of performed runs.

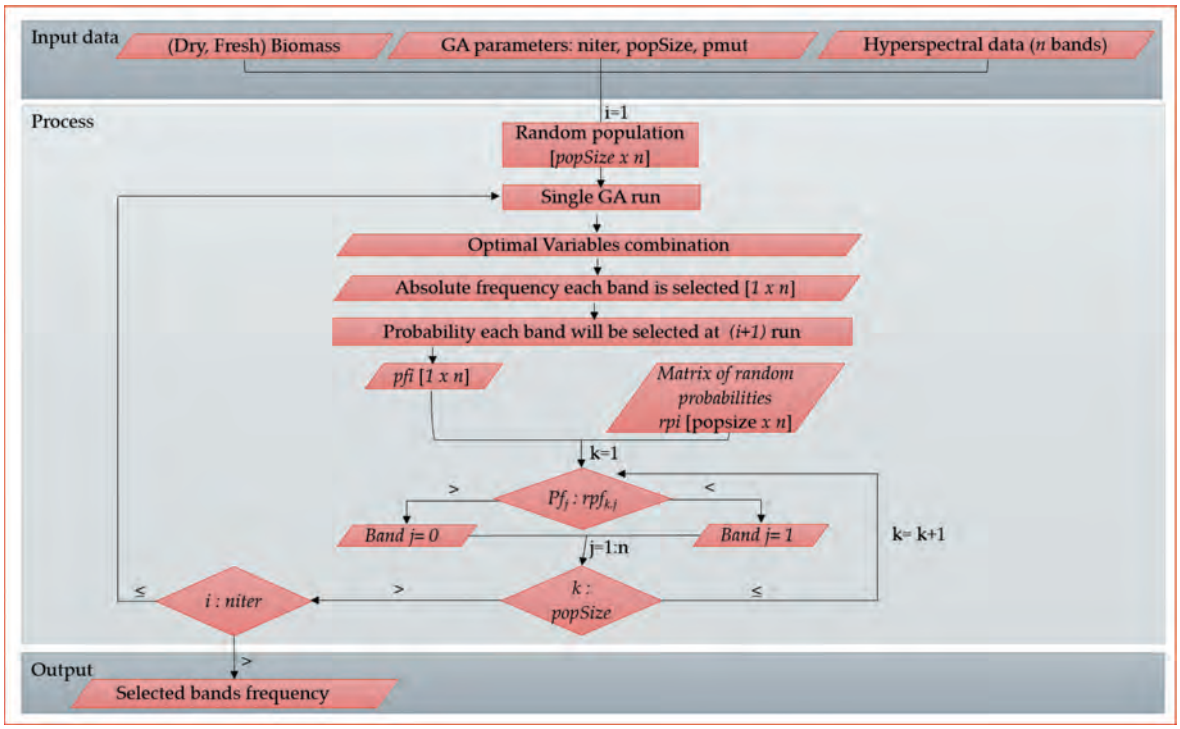

Figure 5. Feature selection workflow procedure based on Leardi (2000). It ranks the band's sensitivity to biomass retrieval depending on the number of times each one is selected in the GA optimal combination.

Another hypothesis considered is that, if a given band proves to be sensitive for biomass estimation, it is likely that the adjacent bands will have relevant information as well. Smoothing the vector of probabilities is then desirable. Window of size three is used in this paper.

The final probability p_i of band i to be present in a chromosome is a weighted combination of the previous probabilities, where initial uniform probability is more important in the first runs (and guarantees that no band has ever null probability) and loses relevance as the number of runs increases.

At the beginning of each individual run, a random probability value is assigned to the initial population and compared to the vector of probabilities. For each band, these two numbers are compared, and if the random probability of a given band is lower, then its value is set to 1 (band present), otherwise it is set to 0. The higher the probability number pi, the more likely the band will be present in the chromosome. The output is a vector accounting for how frequently each band was selected after 50 individual runs.

Support vector regression

After the feature selection is performed, an inversion model is built based on a machine learning approach. The best model is considered to be the one with the best compromise between prediction performance and number of variables. This condition is reached by maximizing the model R^2 performance, while minimizing the number of bands. The input data set is completed with SAR backscattering coefficients with VV and VH polarization.

The non-parametric technique SVR is the regression version ofSVM. The goal of SVR is to find a function relating the input features (selected hyperspectral bands and backscattering coefficients) to a target variable (biomass) (De Almeida et al., 2019). The objective is to create a model, with an error lower than a given ε threshold, by identifying the hyperplane which maximizes the margin and tolerating a part of error budget. This function will at most have ε deviation from the obtained targets y_i for all the training data, and at the same time will be as flat as possible. For SVR training and validation the R programming language was used with package *e1071* (Meyer et al., 2019).

The flow chart in Figure 6 shows how the regression model is derived by using hyperspectral and SAR data, and its overall performance is evaluated by estimating biomass values that will be compared to the testing data set.

Results

Band selection from GA approach

The band selection with GA was performed on meadow and pasture separately, and for fresh and dry biomass data. The selection frequency is the GA output: this information tells the times each hyperspectral reflectance band was selected in the final solution after all the runs were performed. It provides an indication

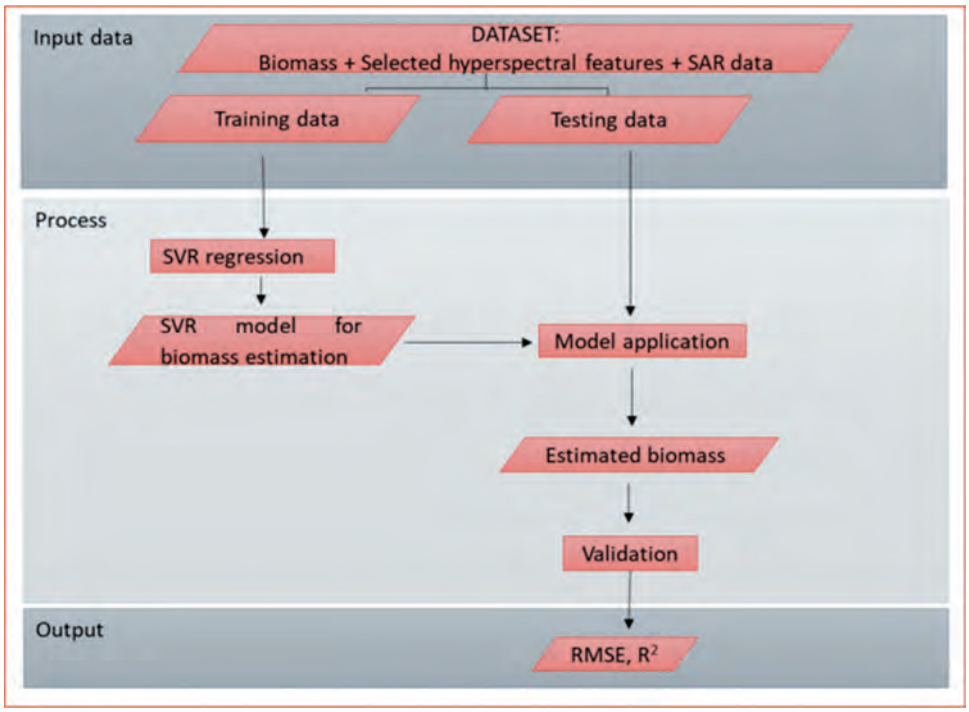

Figure 6. Flow chart biomass estimation based on SVR. After the feature selection is performed, the SVR model is run on a training set, adding one optical band at a time, starting from the most relevant one. The output is a table with the performance of each regression model, evaluated on the testing set.

about the relative importance of each band. Figure 7 summarizes how frequently the bands were selected in each case for dry biomass, overlapped with vegetation spectral signature to clearly indicate the more sensitive wavelengths position.

As general results, both meadow GA feature selections (either for plain field data set and combined field with simulated data) have analogous behaviours, with negligible influence from the added simulated data. Some influence of soil signal to reflectance in pasture response can be noted mainly in the Short Wave Infrared (SWIR) and visible regions. The bare soil typical spectral response tends to be more homogeneous in these regions compared to vegetation signature. Therefore, the water absorption band found in vegetation signatures can be partially masked by the soil and vegetation mix signals. In the case of pasture, because of partially covered soils, the registered response contains information of both soil and vegetation, while meadow grassland normally fully covers the soil. This can be a reason why some bands (in the blue and green regions as well as SWIR region around 1700 nm) were frequently selected in the case of meadows, while in the case of pasture they were found

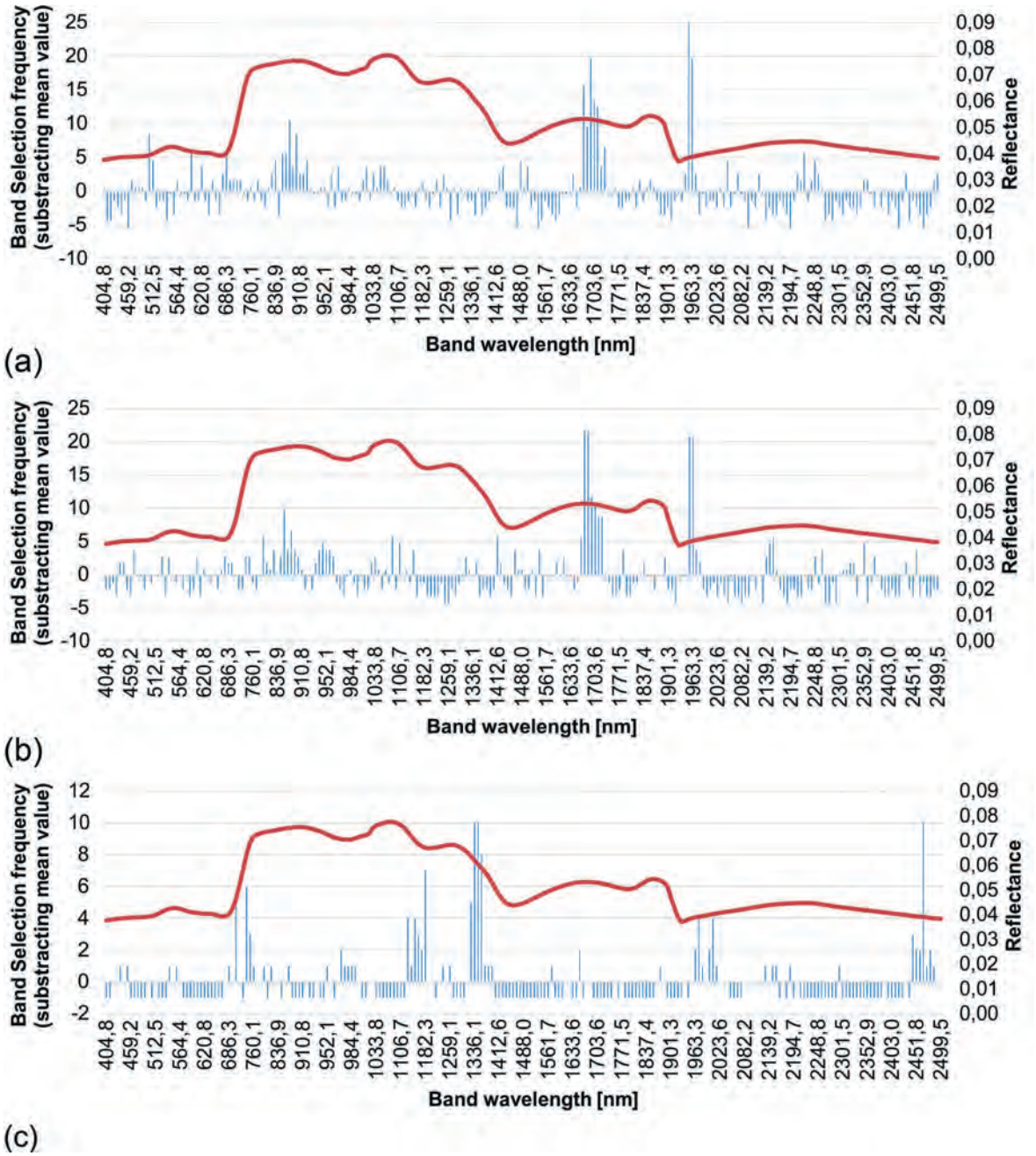

Figure 7. Band selection frequency in the dry biomass case. To make the result visually clearer, the total mean value was subtracted. (a) Meadow study case, (b) meadow study case, adding simulated data to increase the number of values within the maximum of sampled biomass, and (c) pasture study case. Note: the vegetation spectral signature (continuous line in orange) belongs to meadow and is set as reference. The spectral signature used in each figure as guideline may differ as they may belong to different records.

negligible. This means that introducing such bands in the former regression model did not alter significantly the overall performance. Finally, band selection frequency for pasture was lower than for meadow: while the highest frequency for dry biomass was 11 times above the mean value, the highest frequency for meadow was 25 times above the mean value.

To analyse the resulting vector of frequencies in each case, statistics quartiles frequency values were calculated, and focus was made on bands with frequency values above the third quartile (Q3).

In the visible range (i) around 550 nm, (ii) in the green region, and (iii) in the red edge between 680 and 700 nm, bands with high frequency in all three studied cases were highlighted. However, while in the case of pasture some bands in the blue region show high selection rate (440 nm), in the case of meadow this happens for some bands in the red region (between 600 and 620 nm).

The NIR region from 820 to 920 nm and around 1000 nm shows some of the most frequently selected bands in the whole spectrum either for meadow data set with and without simulated data. In the case of pasture, the strongest response is focused by the end of the NIR range at 970 nm up to 1000 nm, with some sparse bands in the rest of the region. In the SWIR range, either for pasture and meadow, reflectance in bands from 1940 to 1980 nm was selected frequently, but while for pasture bands from 1300 to 1400 nm and from 2450 to 2500 nm was frequently selected, for both meadow data sets highlighted the region from 1650 to 1730 nm.

Analogous analysis is performed in Figure 8 for fresh biomass.

In general, the most sensitive bands show similar selection frequency values: 26 and 19 times above their averages for meadow and pasture, respectively. However, the selected bands are more concentrated in narrower ranges and have a stronger signal, showing the frequency of selection is higher in contrast to dry biomass cases. This can be observed by comparing Figures 7 and 8: the bands that were most frequently selected in the fresh biomass analysis are less disperse along the whole spectrum.

Reflectance most selected bands in the visible range show no similarities between both vegetation types, up to the red and red-edge region, with sensitive bands between 630 and 675 nm. In the NIR shoulder region, bands right after visible range between 750 and 830 nm proved to be relevant for biomass retrieval in both cases, but only for meadow this relevance is extended almost within the whole NIR range, up to 990 nm, almost all over the NIR plateau.

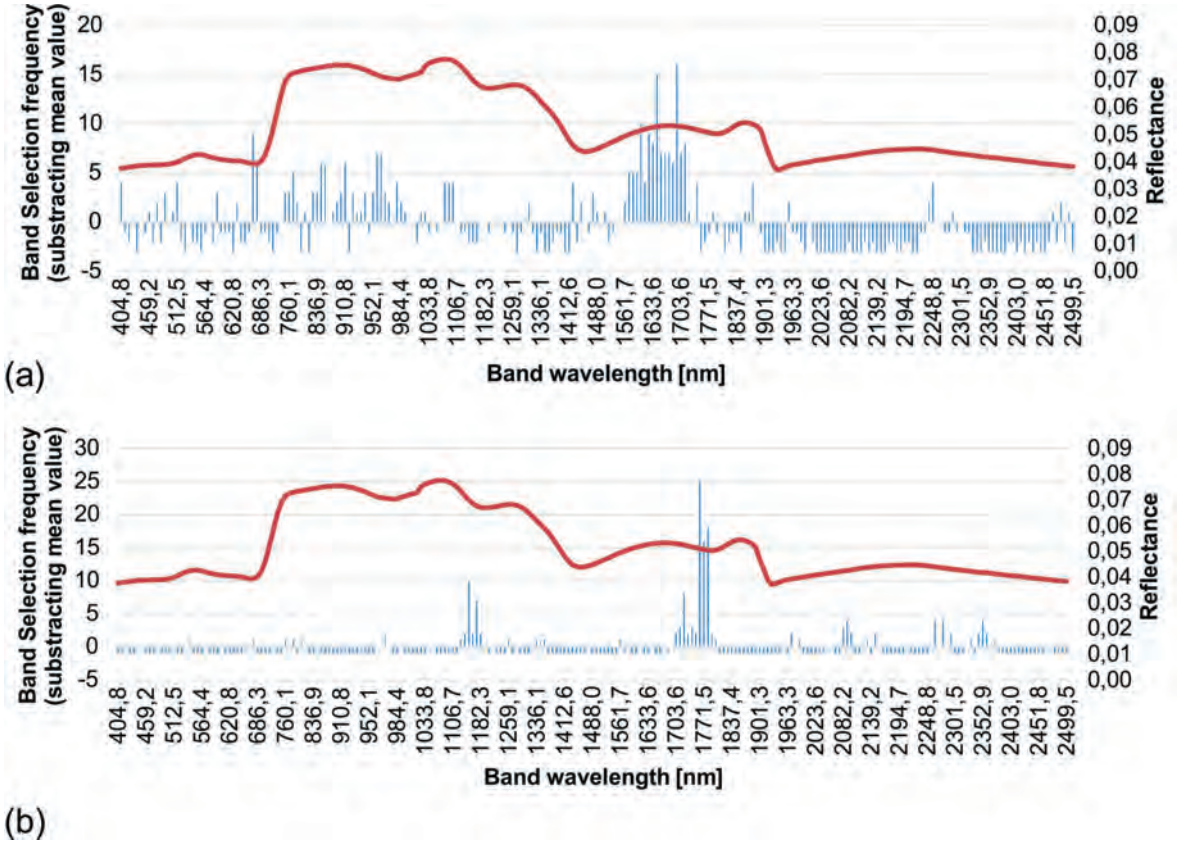

Figure 8. Fresh biomass band selection frequency. To make the result visually clearer, the total mean value was subtracted. (a) Meadow study case and (b) pasture study case. Note: the vegetation spectral signature (continuous orange line in the graph) belongs to meadow's reference. The spectral signature used in each figure as guideline may differ as they may belong to different records.

The SWIR range does not show overlapped regions of relevant bands for both vegetation types, but instead a clear shift can be seen: for meadow the most significant bands go from 1500 to 1700 nm, with some minor relevance in the absorption band around 1400 nm, while the pasture's most relevant range goes from 1700 to 1800 nm. Some other sparse bands are also relevant between 1900 and 2500 nm, with similar distribution compared to biomass output.

When the analysis is performed within the same vegetation type, but comparing dry and fresh biomass, many coincidences are found in the NIR and SWIR regions, especially before the first absorption band in the NIR region, and right after the first water absorption band in the SWIR region in the case of meadow. In the case of pasture, coincidences begin in the red–red edge region, with a lower range by the end of the NIR region, and two ranges in the SWIR region: at the beginning of the SWIR range and before the first water absorption band. Within the NIR region, interaction occurs at cellular level and coincidences are expected, while the SWIR band interacts with plant water content, which is the component that distinguishes fresh and dry biomass.

Biomass estimation from SVR approach

Once the feature selection was completed, regression models were built by using an SVR approach. The first implemented model contained the most relevant band from the hyperspectral data set and both backscattering coefficients from the SAR data set. One band at the time was added, and performance was evaluated using RMSE and R^2 statistics. The most suitable model was the one with the best compromise between retrieval accuracy and the smallest number of bands possible. Table 2 summarizes the results.

The final model was built with the following bands, as shown in Table 3.

In the case of meadows when the regression model is performed using all the bands, R^2 reaches the value of 0.65, and, using 5% of the available bands, the performance slightly improves up to R^2 = 0.71. The 5% of the bands were considered because the highest R^2 value was reached with the lowest number of bands, and including further features had a negligible impact on the model accuracy. The comparison between ground reference and estimated biomass values are reported in Figure 9.

In the case of pasture, the coefficient of determination R^2 using all the bands is f 0.19, slightly lower than the performance using 2.5% of the bands, which shows a correlation R^2 of 0.24. However, the overall result is still not reliable and needs to be improved especially considering a high number of samples. As can be seen in Figure 10, the model showed low accuracy when biomass reaches its highest values. It is worthwhile mentioning that a high number of samples need to be considered to consolidate this conclusive remark.

The same analysis was performed on fresh biomass, with similar correlation coefficients found for both pasture and meadow and using 54% and 28% of the available bands, reaching the optimal compromise with the lowest number of bands available, as summarized in Table 4.

The bands considered in the final model are listed in Table 5.

Table 2. Number of bands from feature selection and statistical regression model performance for dry biomass.

Dry biomass	Number of bands	R^2	RMSE (g/cm^2)
Meadow (field data)	13	0.71	0.077
Meadow (field + simulated data)	9	0.71	0.077
Pasture	6	0.24	0.078

Table 3. Band's wavelength of the optimal model achieved for dry biomass estimation.

Meadow (field data)	Meadow (field + simulated data)	Pasture
505.1	858.5	738.4
595.7	1673.8	1182.3
686.3	1683.8	1336.1
879.7	1693.7	1347.1
900.5	1703.6	1358.1
1673.8	1713.4	2465.5
1683.8	1723.2	738.4
1693.7	1954.6	1182.3
1703.6	1963.3	1336.1
1713.4		
1732.9		
1954.6		
1963.3		

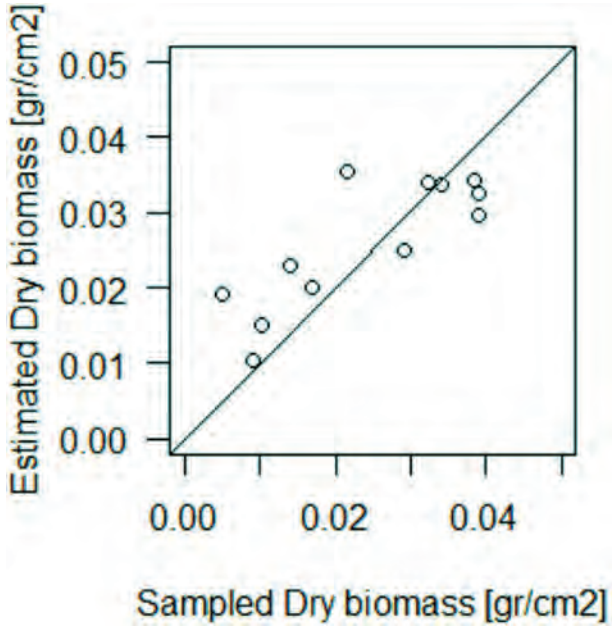

Figure 9. Correlation between sampled and estimated dry biomass for meadow (R^2 = 0.71).

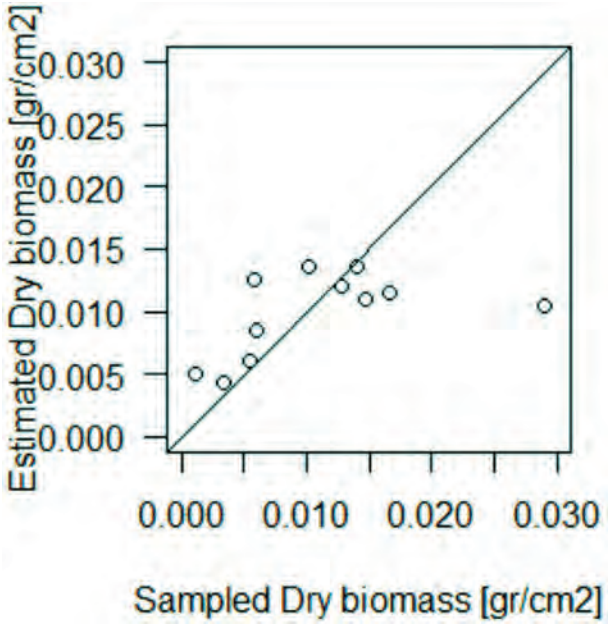

Figure 10. Correlation between sampled and estimated dry biomass for pasture (R^2 = 0.24).

Table 4. Number of bands from feature selection and statistical regression model performance for fresh biomass.

Fresh biomass	Number of bands	R^2	RMSE (g/cm^2)
Meadow (field data)	127	0.60	0.653
Pasture	67	0.47	0.152

Table 5. Wavelength of the optimal model selected thetop 20 bands for fresh biomass estimation.

Meadow (field data)	Pasture
666.6	973.4
676.3	1138.8
771.1	1149.6
847.7	1160.5
858.5	1171.4
910.8	1182.3
955.9	1703.6
962.6	1713.4
1572.1	1723.2
1603.0	1732.9
1623.5	1742.6
1633.6	1752.3
1643.7	1761.9
1653.8	1771.5
1663.9	1781.0
1673.8	2082.2
1683.8	2090.4
1693.7	2264.0
1703.6	2279.1
1713.4	2352.9

When all the bands were considered, the R^2 statistic performance was 0.51 for meadow and 0.46 for pasture. This analysis confirms the underestimation of high values of biomass, as indicated in Figure 11.

Regression models built without SAR data showed slightly poorer performance in all cases. For dry biomass in meadow, R^2 parameter was 0.63, while the RMSE value was 0.013 kg/m^2 when 13 bands were considered, while it was 0.20 for R^2 and RMSE was 0.091 kg/m^2 in pasture with six bands involved in the regression. When fresh biomass regression model was run without integrating SAR data, the performance in meadow using 127 bands presented an R^2 value of 0.57, slightly lower than the 0.60 value when SAR data were included, and RMSE was 0.580 kg/m^2. Finally, the R^2 found for fresh pasture biomass was 0.45 and RMSE equal to 0.135 kg/m^2 when only 67 hyperspectral bands were considered in the model. When only SAR data were introduced in the SVR model, the regression accuracy dropped significantly, with RMSE values of 0.162 and 0.863 kg/m^2, and R^2 of 0.034 and 0.017 for dry and fresh meadow biomass, respectively. For pasture, the retrieved values provided similar performance: RMSE was 0.076 and 0.152 kg/m^2, and R^2 resulted in 0.11 and 0.055 for dry and fresh biomass. SAR stand-alone data performances are introduced here as a reference with respect to the models that use both optical and SAR data. However, low performances were expected when only backscattering values are used in the regression, without complementary information about the vegetation status, soil moisture availability, surface roughness, and other target properties that introduce additional ambiguities and non-linearity in the retrieval process (Ghasemi et al., 2011; Pasolli et al., 2011).

Discussion

The optimal compromise between the number of bands and model performance in an SVR approach was evaluated by using R^2 and RMSE as reference performance indicators.

During feature selection, as expected both green and red edge regions were selected as they provide relevant information, in agreement with the specialized literature (L. Chen et al., 2018; Clevers et al., 2007; Fava et al., 2009; Psomas et al., 2011; C. Wang et al., 2017). The red edge portion of the spectrum is where the maximum slope change occurs in the vegetation reflectance spectral signature and contains information about vegetation stress, chlorophyll, and nitrogen status (Psomas et al., 2011). Blue bands were also found to be relevant, especially in the meadow, and are related to chlorophyll *b* absorption (J. Chen et al., 2009). The selected SWIR region contains information about leaf water content and leaf mass, and therefore is correlated to LAI and biomass (Psomas et al., 2011).

In the case of dry biomass analysis, all the cases highlight bands with frequency in the visible range within the green region (around 550 nm), in agreement with Fava et al. (2009), who found the region between 535 and 565 nm to be sensitive. In the case of pasture, reflectance of some bands in the blue region (between 435 and 490 nm) shows high performance (J. Chen et al., 2009; Fava et al., 2009). In the case of the meadow, the same occurs for some bands in the red

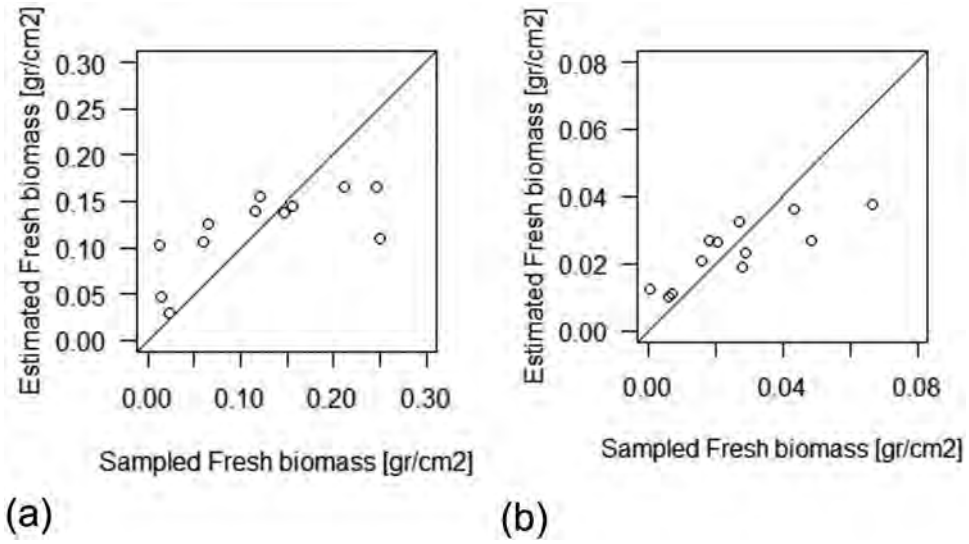

Figure 11. Correlation between observed and estimated dry biomass for (a) meadow (R^2 = 0.60) and (b) pasture (R^2 = 0.47).

region between 600 and 670 nm. In all the cases we also found sensitive bands in the red edge region from 690 to 740 nm, in agreement with the revised bibliography (J. Chen et al., 2009) (between 697 and 723 nm) and Fava et al. (2009). Clevers et al. found that 61% of the grassland biomass variation was explained by combining two bands between 859 and 1006 nm in the NIR region, and one in the red edge region between 668 and 776 nm, in a normalized VI (Clevers et al., 2007). Another study by Adam et al. (2014) found that the best band combinations for EVI index using narrow bands for papyrus swamp biomass were 445, 682, and 829 nm; 497, 676, and 1091 nm; and 495, 678, and 1120 nm. Even though the studied type of vegetation is not grassland, the highlighted bands match with the ones selected in our feature selection. The NIR plateau is very frequently selected in both cases of meadow (with and without simulated data) and has some of the strongest responses in the GA output, in the ranges from 830 to 930 nm and from 950 to 990 nm, in agreement with Chen's feature selection (J. Chen et al., 2009). The latter study highlighted the range between 760 and 950 nm, where leaf internal scattering is stronger in the NIR. In the case of pasture, even when some bands in the same range are selected, the strongest response is concentrated by the end of the NIR range from 970 to 990 nm, with some sparse bands from 760 to 820 nm. These regions were mentioned above in combination with visible range bands as relevant (Fava et al., 2009), ranging from 770 to 940 nm and from 960 to 940 nm. Finally, high performance is highlighted in all cases for the absorption water band in the SWIR range around 1900 nm. Even though with slightly different bands, Psomas et al. propose single-band models based on bands 1710 or 1699 nm (Psomas et al., 2011). When more bands are included in MLR models, blue bands at 468 and 478 nm are combined with SWIR band at 1780 nm or green band at 518 nm is combined with SWIR bands at 1205, 1215, 1225, 1235, 1710, and 1720 nm in linear equations using 3–4 bands.

On the other hand, when feature selection was performed on the fresh biomass samples, the highlighted bands in the red and red edge range for both meadow and pasture were comparable with the bands that were found to be sensitive from 668 to 1006 nm using SVM band shaving (Clevers et al., 2007) and 724 nm using stepwise forward regression, and from 697 to 723 nm for fresh biomass (in this case forest biomass; L. Chen et al., 2018). The more extended range of bands selected in the case of meadow, compared to pasture, can be explained because the plants show healthier and more vigorous status in the irrigated field, and therefore a stronger signal. The water bands at 937 and 1135 nm (Clevers et al., 2007) based on stepwise forward regression, band 895 from SMLR (C. Wang et al., 2017), and the NIR plateau from 765 to 863 nm (L. Chen et al., 2018) are pointed as relevant too. In the SWIR range, some agreement is found for crop biomass retrieval using bands 1285 and 1320 to 1365 nm (C. Wang et al., 2017).

From a computational point of view, the genetic algorithm approach is highly time-consuming, even though independent short runs (instead of one single long run) improve the processing duration and allow groups of bands to be found within regions of the spectrum, rather than sparse bands across the whole spectrum range. The results of the regression model confirm that feature selection successfully identified the most sensitive bands, leaving out features with little or no significant information for biomass retrieval. The performance improvement in the case of dry biomass for meadow (with and without simulated data) does not show a trend clear enough to make reliable conclusions with this specific number of simulated data. The lack of accuracy found for the

highest biomass values may be related to the signal saturation, especially in the red edge region. The most sensitive bands were in the range of those used for VIs calculation in the visible and NIR (VNIR) and SWIR range (Naidoo et al., 2019; C. Wang et al., 2017). The overall performance of this study was comparable to the performance found in literature, and the processing showed to be highly dependent on the data set dimensions. In this regard, further improvement is expected as a high number of field campaigns may be included.

The model performance improved in all cases when SAR and hyperspectral data were integrated, even if the improvement was relatively small. Even though the trend was always towards improvement, further efforts should be taken to achieve better results, and other SAR information should be considered to be tested. To this aim, further tests will be carried out in the future considering other information such as data from COSMO-SkyMed PingPong dual-polarization. In this case, specifically, the use of an additional frequency may allow to have complementary information due to the different penetration capabilities in the canopy (Notarnicola & Posa, 2007).

To this aim, further field campaigns are planned in the future in the framework of the ongoing ALGORITMI project (Tapete et al., 2020) to increase the data set size, as well as the addition of simulated data combining PROSPECT-PROSAIL models and PRISMA imagery for hyperspectral data, and Water Cloud Model for SAR data, combined with COSMO-SkyMed single- and dual-polarization X-band SAR images. These additional data sets and their ingestion into the methodology will allow for a larger training set, including additional information from a yet unexplored SAR band and multiple polarizations, and are expected to provide a more accurate prediction model. The resulting final model is expected to be capable of biomass retrieval by integrating multisensory data. Moreover, being the model not based on predefined indices and bands, it can be applied in different vegetation conditions.

Conclusions

This paper presented a general and flexible methodology for biomass estimation based on SVR approach by integrating biomass field measurements, hyperspectral spectroradiometer data, and SAR imaging in Alpine pastures and meadows. The methodology consisted in performing a feature selection on the hyperspectral bands to reduce the model complexity and collinearity problems. An optimized genetic algorithm approach was applied to this scope to select the most informative bands from the hyperspectral data set. A multi-sensor synergetic model using both hyperspectral and SAR information is explored, with the aim to combine information about plant biochemical and structural properties, and density/volumetric information. A generic framework is proposed that starting from the band selection (with a GA) and the retrieval approach (with the machine learning method based on SVR) can develop a more generic model useful for the integration of multiple data sets, merging of bands from hyperspectral sensors with SAR data.

The overall accuracy resulted in a coefficient of determination R^2 of 0.71 in the best case, for dry biomass in meadow, and of 0.24 in the worst case for dry biomass in pasture. The poor accuracy may be ascribed to several factors, such as the paucity of ground data used for training and validation, possible saturation effect, and low representativeness in the sample data of high values of biomass.

Narrow optical band data combined with microwave data remain as a partially unexplored field and need to be further investigated. The combined use of these bands through data integration resulted in a model that optimizes the available data. However, the selected techniques are sensitive to the data set size, which is a constraint if the input data set is limited. Further investigations in this regard will point to data simulation techniques to expand the input data set, aiming to achieve an enhanced hybrid algorithm. The innovation here presented copes the increasing availability of data sets with narrower bands in the optical domain, with SAR data that provide detail on other aspects of vegetation in terms of geometrical properties and volume contribution. The machine learning approach offers the possibility to build this stand-alone model based only on a large and multi-sensor data set and to be easily adapted to different environmental conditions. Since the explored methodology did not reach operational level yet, performance improvement is still expected.

In this regard, the model accuracy can be improved by extending the data set through the addition of new field campaigns, simulated hyperspectral, and SAR data, as well as other features of interest, such as LAI, VIs or bands single ratios, and the relation between both SAR backscattering coefficients. Moreover, testing the algorithm on PRISMA and COSMO-SkyMed StripMap PingPong will allow the definition of models that use hyperspectral information together with SAR data at two frequencies, thus interacting with different parts of the canopy. Thanks to the use of raster images, it will be possible to derive classified images with biomass levels.

Acknowledgments

This work was carried out by EURAC, IFAC/CNR, and ASI in the framework of the ALGORITMI project. Copernicus

Sentinel-1 SAR data were made available through the Copernicus Open Access Hub.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the Italian Space Agency (ASI) in the framework of the 2019-2021 project "Development of algorithms for estimation and monitoring of hydrological parameters from satellite and drone", under ASI-CNR /IFAC grant agreement n.2018-37-HH.0. The EURAC team was also co-funded by the European Regional Development Fund (ERDF), Operational Program Investment for growth and jobs ERDF 2014-2020 under Project Number ERDF1094, Data Platform and Sensing Technology for Environmental Sensing LAB – DPS4ESLAB.

ORCID

Eugenia Chiarito http://orcid.org/0000-0003-1468-7205

References

Adam, E., Mutanga, O., Abdel-Rahman, E. M., & Ismail, R. (2014). Estimating standing biomass in papyrus (Cyperus papyrus L.) swamp: Exploratory of in situ hyperspectral indices and random forest regression. *International Journal of Remote Sensing*, *35*(2), 693–714. https://doi.org/10.1080/01431161.2013.870676

Bertoldi, G., Della Chiesa, S., Niedrist, G., Rist, A., Tasser, E., & Tappeiner, U. (2010). Space-time evolution of soil moisture, evapotranspiration and snow cover patterns in a dry alpine catchment: An interdisciplinary numerical and experimental approach. *EGU General Assembly Conference abstracts*, 12109. https://doi.org/http://adsabs.harvard.edu/abs/2010EGUGA.1212109B

Bradley, P. E., Keller, S., & Weinmann, M. (2018). Unsupervised feature selection based on ultrametricity and sparse training data: A case study for the classification of high-dimensional hyperspectral data. *Remote Sensing*, *10*(10), 1564. https://doi.org/10.3390/rs10101564

Caltagirone, F., Capuzi, A., Coletta, A., De Luca, G. F., Scorzafava, E., Leonardi, R., Rivola, S., Fagioli, S., Angino, G., Labbate, M., Piemontese, M., Zampolini Faustini, E., Torre, A., De Libero, C., & Esposito, P. G. (2014). The COSMO-SkyMed dual use earth observation program: Development, qualification, and results of the commissioning of the overall constellation. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *7*(7), 2754–2762. https://doi.org/10.1109/JSTARS.2014.2317287

Chang, J., & Shoshany, M. (2016). Mediterranean shrublands biomass estimation using Sentinel-1 and Sentinel-2. *International Geoscience and Remote Sensing Symposium (IGARSS), 2016-Novem(July 2016)* (pp. 5300–5303). https://doi.org/10.1109/IGARSS.2016.7730380

Chen, J., Gu, S., Shen, M., Tang, Y., & Matsushita, B. (2009). Estimating aboveground biomass of grassland having a high canopy cover: An exploratory analysis of in situ hyperspectral data. *International Journal of Remote Sensing*, *30*(24), 6497–6517. https://doi.org/10.1080/01431160902882496

Chen, L., Ren, C., Zhang, B., Wang, Z., & Xi, Y. (2018). Estimation of forest above-ground biomass by geographically weighted regression and machine learning with sentinel imagery. *Forests*, *9*(10), 1–20. https://doi.org/10.3390/f9100582

Cho, M. A., Skidmore, A., Corsi, F., Van Wieren, S. E., & Sobhan, I. (2007). Estimation of green grass/herb biomass from airborne hyperspectral imagery using spectral indices and partial least squares regression. *International Journal of Applied Earth Observation and Geoinformation*, *9*(4), 414–424. https://doi.org/10.1016/j.jag.2007.02.001

Clevers, J. G. P. W., Van der Heijden, G. W. A. M., Verzakov, S., & Schaepman, M. E. (2007). Estimating grassland biomass using SVM band shaving of hyperspectral data. *Photogrammetric Engineering and Remote Sensing*, *73*(10), 1141–1148. https://doi.org/10.14358/PERS.73.10.1141

Darvishzadeh, R., Skidmore, A., Schlerf, M., Atzberger, C., Corsi, F., & Cho, M. (2008). LAI and chlorophyll estimation for a heterogeneous grassland using hyperspectral measurements. *ISPRS Journal of Photogrammetry and Remote Sensing*, *63*(4), 409–426. https://doi.org/10.1016/j.isprsjprs.2008.01.001

De Almeida, C. T., Galvão, L. S., De Aragão, L. E. O. C. E., Ometto, J. P. H. B., Jacon, A. D., De Pereira, F. R. S., Sato, L. Y., Lopes, A. P., De Graça, P. M. L. A., De Silva, C. V. J., Ferreira-Ferreira, J., & Longo, M. (2019). Combining LiDAR and hyperspectral data for aboveground biomass modeling in the Brazilian Amazon using different regression algorithms. *Remote Sensing of Environment*, *232*(March), 111323. https://doi.org/10.1016/j.rse.2019.111323

Fava, F., Colombo, R., Bocchi, S., Meroni, M., Sitzia, M., Fois, N., & Zucca, C. (2009). Identification of hyperspectral vegetation indices for Mediterranean pasture characterization. *International Journal of Applied Earth Observation and Geoinformation*, *11*(4), 233–243. https://doi.org/10.1016/j.jag.2009.02.003

Ghasemi, N., Sahebi, M. R., & Mohammadzadeh, A. (2011). A review on biomass estimation methods using synthetic aperture radar data. *International Journal Ofgeomatics and Geosciences*, *1*(4), 776–788.

Gidudu, A., & Heinz, R. (2007). Comparison of feature selection techniques for SVM classification. *10th international symposium on physical measurement and signature in remote sensing, ISPMSRS* (pp. 1–6).

Guanter, L., Kaufmann, H., Segl, K., Foerster, S., Rogass, C., Chabrillat, S., Kuester, T., Hollstein, A., Rossner, G., Chlebek, C., Straif, C., Fischer, S., Schrader, S., Storch, T., Heiden, U., Mueller, A., Bachmann, M., Mühle, H., Müller, R., Ohndorf, A., … Sang, B. (2015). The EnMAP spaceborne imaging spectroscopy mission for earth observation. *Remote Sensing*, *7*(7), 8830–8857. https://doi.org/10.3390/rs70708830

Hong, Y., Chen, S., Chen, Y., Linderman, M., Mouazen, A. M., Liu, Y., Guo, L., Yu, L., Liu, Y., Cheng, H., & Liu, Y. (2020). Comparing laboratory and airborne hyperspectral data for the estimation and mapping of topsoil organic carbon: Feature selection coupled with random forest. *Soil and Tillage Research*, *199*(January), 104589. https://doi.org/10.1016/j.still.2020.104589

Jain, A. (1997). Feature selection: Evaluation, application, and small sample performance. *IEEE Transactions on*

Pattern Analysis and Machine Intelligence, *19*(2), 153–158. https://doi.org/10.1109/34.574797

Kumar, L., & Mutanga, O. (2017). Remote sensing of above-ground biomass. *Remote Sensing*, *9*(9), 1–8. https://doi.org/10.3390/rs9090935

Leardi, R. (2000). Application of genetic algorithm-PLS for feature selection in spectral data sets. *Journal of Chemometrics*, *14*(5–6), 643–655. https://doi.org/10.1002/1099-128X(200009/12)14:5/6<643::AID-CEM621>3.0.CO;2-E

Leardi, R., & Lupáñez González, A. (1998). Genetic algorithms applied to feature selection in PLS regression: How and when to use them. *Chemometrics and Intelligent Laboratory Systems*, *41*(2), 195–207. https://doi.org/10.1016/S0169-7439(98)00051-3

Lehnert, L., Meyer, H., Obermeier, W. A., Silva, B., Regeling, B., & Bendix, J. (20189). Hyperspectral data analysis in R: The hsdar package. *J. Stat. Softw., 89*(12), 1-23. https://doi.org/10.18637/jss.v089.i12

Loizzo, R., Ananasso, C., Guarini, R., Lopinto, E., Candela, L., & Pisani, A. R. (2016). The PRISMA hyperspectral mission. *Proceedings of ESA living planet symposium 2016, Prague, Czech Republic, SP-740* (pp. 9–13).

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., Chang, -C.-C., & Lin, -C.-C. (2019). Package 'e1071': Misc functions of the department of statistics, probability theory group (Formerly: E1071), TU Wien. *R package version 1.7-3*. https://cran.r-project.org/package=e1071

Moghimi, A., Yang, C., & Marchetto, P. M. (2018). Ensemble feature selection for plant phenotyping: A journey from hyperspectral to multispectral imaging. *IEEE Access*, *6*, 56870–56884. https://doi.org/10.1109/ACCESS.2018.2872801

Monnet, J. M., Chanussot, J., & Berger, F. (2011). Support vector regression for the estimation of forest stand parameters using airborne laser scanning. *IEEE Geoscience and Remote Sensing Letters*, *8*(3), 580–584. https://doi.org/10.1109/LGRS.2010.2094179

Naidoo, L., Van Deventer, H., Ramoelo, A., Mathieu, R., Nondlazi, B., & Gangat, R. (2019). Estimating above ground biomass as an indicator of carbon storage in vegetated wetlands of the grassland biome of South Africa. *International Journal of Applied Earth Observation and Geoinformation*, *78*(January), 118–129. https://doi.org/10.1016/j.jag.2019.01.021

Notarnicola, C., & Posa, F. (2007). Inferring vegetation water content from C- and L-Band SAR images. *IEEE Transactions on Geoscience and Remote Sensing*, *45*(10), 3165–3171. https://doi.org/10.1109/TGRS.2007.903698

Pal, M., & Foody, G. M. (2010). Feature selection for classification of hyperspectral data by SVM. *IEEE Transactions on Geoscience and Remote Sensing*, *48*(5), 2297–2307. https://doi.org/10.1109/TGRS.2009.2039484

Pasolli, L., Notarnicola, C., Bruzzone, L., Bertoldi, G., Della Chiesa, S., Hell, V., Niedrist, G., Tappeiner, U., Zebisch, M., Del Frate, F., & Vaglio Laurin, G. (2011). Estimation of soil moisture in an Alpine catchment with RADARSAT2 Images. *Applied and Environmental Soil Science*, *2011*, 1–12. https://doi.org/10.1155/2011/175473

Psomas, A., Kneubühler, M., Huber, S., Itten, K., & Zimmermann, N. E. (2011). Hyperspectral remote sensing for estimating aboveground biomass and for exploring species richness patterns of Grassland habitats. *International Journal of Remote Sensing*, *32*(24), 9007–9031. https://doi.org/10.1080/01431161.2010.532172

Pullanagari, R. R., Kereszturi, G., & Yule, I. (2018). Integrating airborne hyperspectral, topographic, and soil data for estimating pasture quality using recursive feature elimination with random forest regression. *Remote Sensing*, *10*(7), 1117. https://doi.org/10.3390/rs10071117

Rasel, S. M. M., Chang, H. C., Ralph, T. J., Saintilan, N., & Diti, I. J. (2019). Application of feature selection methods and machine learning algorithms for saltmarsh biomass estimation using Worldview-2 imagery. *Geocarto International*, Advance online publication,1–25. https://doi.org/10.1080/10106049.2019.1624988

Rossi, M., Niedrist, G., Asam, S., Tonon, G., Tomelleri, E., & Zebisch, M. (2019). A comparison of the signal from diverse optical sensors for monitoring alpine grassland dynamics. *Remote Sensing*, *11*(3), 6–8. https://doi.org/10.3390/rs11030296

Shimabukuro, Y. E., Batista, G. T., Mello, E. M. K., Moreira, J. C., & Duarte, V. (1998). Using shade fraction image segmentation to evaluate deforestation in landsat thematic mapper images of the Amazon region. *International Journal of Remote Sensing*, *19*(3), 535–541. https://doi.org/10.1080/014311698216152

Stamenkovic, J., Guerriero, L., Ferrazzoli, P., Notarnicola, C., Greifeneder, F., & Thiran, J. P. (2017). Soil moisture estimation by SAR in Alpine fields using gaussian process regressor trained by model simulations. *IEEE Transactions on Geoscience and Remote Sensing*, *55*(9), 4899–4912. https://doi.org/10.1109/TGRS.2017.2687421

Stendardi, L., Karlsen, S. R., Niedrist, G., Gerdol, R., Zebisch, M., Rossi, M., & Notarnicola, C. (2019). Exploiting time series of Sentinel-1 and Sentinel-2 imagery to detect meadow phenology in mountain regions. *Remote Sensing*, *11*(5), 1–24. https://doi.org/10.3390/rs11050542

Tapete, D., Cigna, F., Paloscia, S., Santi, E., Pettinato, S., Fontanelli, G., Chiarito, E., Notarnicola, C., Cuozzo, G., Jacob, A., De Gregorio, L., & Rossi, M. (2020). Development of algorithms for the estimation of hydrological parameters combining COSMO-SkyMed and Sentinel time series with in situ measurements. *IEEE Mediterranean and Middle-East Geoscience and Remote Sensing Symposium 2020 (M2GARSS 2020)* (pp. 53–56). https://doi.org/10.1109/M2GARSS47143.2020.9105313

Wang, C., Feng, M., Yang, W., Ding, G., Xiao, L., Li, G., & Liu, T. (2017). Extraction of sensitive bands for monitoring the winter wheat (Triticum aestivum) growth status and yields based on the spectral reflectance. *PLoS ONE*, *12*(1), 1–16. https://doi.org/10.1371/journal.pone.0167679

Wang, J., Xiao, X., Bajgain, R., Starks, P., Steiner, J., Doughty, R. B., & Chang, Q. (2019). Estimating leaf area index and aboveground biomass of grazing pastures using Sentinel-1, Sentinel-2 and Landsat images. *ISPRS Journal of Photogrammetry and Remote Sensing*, *154*(June), 189–201. https://doi.org/10.1016/j.isprsjprs.2019.06.007

Wang, X., Dannenberg, M. P., Yan, D., Jones, M. O., Kimball, J. S., Moore, D. J. P., Leeuwen, W. J. D., Didan, K., & Smith, W. K. (2020). Globally consistent patterns of asynchrony in vegetation phenology derived from optical, microwave, and fluorescence satellite data. *Journal of Geophysical Research: Biogeosciences*, 125(7), 1–15. https://doi.org/10.1029/2020jg005732

Xue, J., & Su, B. (2017). Significant remote sensing vegetation indices: A review of developments and applications. *Journal of Sensors*, *2017*, 1–17. https://doi.org/10.1155/2017/1353691