Remote Sens. **2014**, *6*, 3427-3445; doi:10.3390/rs6043427

OPEN ACCESS

remote sensing

ISSN 2072-4292

www.mdpi.com/journal/remotesensing

Article

Estimating Fractional Shrub Cover Using Simulated EnMAP Data: A Comparison of Three Machine Learning Regression Techniques

Marcel Schwieder *, Pedro J. Leitão, Stefan Suess, Cornelius Senf and Patrick Hostert

Geography Department, Humboldt-Universität zu Berlin, Unter den Linden 6, D-10099 Berlin, Germany; E-Mails: p.leitao@geo.hu-berlin.de (P.J.L.); stefan.suess@geo.hu-berlin.de (S.S.); cornelius.senf@geo.hu-berlin.de (C.S.); patrick.hostert@geo.hu-berlin.de (P.H.)

* Author to whom correspondence should be addressed; E-Mail: marcel.schwieder@geo.hu-berlin.de; Tel.: +49-30-2093-6877; Fax: +49-30-2093-6848.

Received: 16 January 2014; in revised form: 11 April 2014 / Accepted: 14 April 2014 / Published: 22 April 2014

Abstract: Anthropogenic interventions in natural and semi-natural ecosystems often lead to substantial changes in their functioning and may ultimately threaten ecosystem service provision. It is, therefore, necessary to monitor these changes in order to understand their impacts and to support management decisions that help ensuring sustainability. Remote sensing has proven to be a valuable tool for these purposes, and especially hyperspectral sensors are expected to provide valuable data for quantitative characterization of land change processes. In this study, simulated EnMAP data were used for mapping shrub cover fractions along a gradient of shrub encroachment, in a study region in southern Portugal. We compared three machine learning regression techniques: Support Vector Regression (SVR); Random Forest Regression (RF); and Partial Least Squares Regression (PLSR). Additionally, we compared the influence of training sample size on the prediction performance. All techniques showed reasonably good results when trained with large samples, while SVR always outperformed the other algorithms. The best model was applied to produce a fractional shrub cover map for the whole study area. The predicted patterns revealed a gradient of shrub cover between regions affected by special agricultural management schemes for nature protection and areas without land use incentives. Our results highlight the value of EnMAP data in combination with machine learning regression techniques for monitoring gradual land change processes.

Keywords: EnMAP; hyperspectral; land cover; partial least squares regression; random forest regression; shrub encroachment; Portugal; sub-pixel mapping; support vector regression

1. Introduction

Human-induced alterations of terrestrial ecosystems have far reaching impacts on the Earth system [1,2]. They threaten fragile ecosystems and related provision of services, which are of great importance for human well-being. Vegetation plays a pivotal role in this context as it influences various ecosystem functions, such as climate regulation, carbon storage, sustaining biodiversity, and protecting fertile soils [3]. Changing terrestrial vegetation thus has far-reaching impacts on many subsystems of the earth.

One example of these ecosystem alterations is the abandonment of former agricultural land, usually followed by natural succession of the vegetation, such as the shrub encroachment on formerly cultivated land [4]. Shrub encroachment is not a regional phenomenon, but is observed in various Mediterranean landscapes [5–7], as well as in other arid to semi-arid regions around the globe [4,8], which cover approximately 40% of the Earth's terrestrial surface [9]. It is, hence, of special importance to monitor and manage these areas in a sustainable way.

Even though shrub encroachment is often referred to as an indicator of desertification processes and related ecosystem degradation [10], it can also enhance ecosystem functions by improving water quality, storing carbon, and by the protection of soils [11]. Furthermore, shrub encroachment has an impact on the local biodiversity and fire regimes [12]. Effects of increasing shrub cover are therefore dependent on various environmental conditions [4,11]. Maestre *et al.* [6] pointed out that an increase in shrubby vegetation might even be able to reverse desertification processes in Mediterranean regions, depending on the shrub traits and the prevalent cultural and environmental conditions. These sometimes controversial findings stress the importance of an accurate shrub encroachment monitoring through spatially and temporally seamless mapping of shrub cover, helping to better understand impacts on ecosystems in a regional context and to develop appropriate management strategies.

Monitoring shrub encroachment is not a trivial task, mostly because of the high costs of field surveys, which are therefore often restricted to small extents with scarce data availability [13]. Remote sensing has proven its potential to overcome these limitations in many ecological applications, as it enables the acquisition of affordable data over large extents and in remote areas [14]. Depending on their spatial, spectral and temporal resolution, remote sensing data allow for the classification and quantification of ecosystem characteristics and the estimation of related biophysical parameters at different scales [15]. For the monitoring of gradual processes, e.g., shrub encroachment in heterogeneous landscapes, sub-pixel analysis is valuable approach. However, sub-pixel mapping is challenging using classical multi-spectral remote sensing data, as different vegetation often has similar spectral features, especially in the visible infrared (VIS) and near infrared (NIR) region of the electromagnetic spectrum [16].

Hyperspectral remote sensing data, which consist of a nearly-contiguous spectral signal, have proven the capability to overcome some limitations of classical multi-spectral remote sensing.

Airborne (e.g., AISA, AVIRIS, HyMAP) and spaceborne (e.g., CHRIS, Hyperion) hyperspectral data were used in several studies, demonstrating the ability of high-dimensional data for sub-pixel mapping (e.g., [17,18]). Despite these recent advances, there is still a lack of reliable satellite-based hyperspectral data with an adequate temporal resolution, which restricts most of regional analyses to one point in time. Upcoming satellite missions, such as the German EnMAP (Environmental Mapping and Analysis Program) bear the potential to make use of the advantages of broad-scale hyperspectral data. The instrument will provide over 200 spectral bands covering the visible near infrared (VNIR) to shortwave infrared (SWIR) (420 to 2450 nm) with a ground sampling distance (GSD) of 30 m × 30 m. The sensor's pointing capability (±30°) will substantially shorten the repeat cycle of 21 days to a target revisit time of 4 days [19–21].

To handle the high dimensionality of hyperspectral data, methods dealing with multi-collinear datasets are required [22]. State of the art machine learning algorithms like Support Vector Regression (SVR [23]), Random Forests (RF; [24]) or Partial Least Squares Regression (PLSR; [25]) have proven to perform adequately in various studies (e.g., [26–29]). They are widely applied in many remote sensing related studies and have shown robust and reliable regression or classification results [29].

To investigate the potential of EnMAP data for estimating fractional shrub cover in heterogeneous landscapes, we developed a methodological framework to compare the prediction performance of three state-of-the-art regression algorithms (SVR, RF, and PLSR). Furthermore, we tested the algorithms sensitivity to trainings sample size, giving recommendations for future applications in characterizing heterogeneous environments. The study was performed for a test site in southern Portugal, affected by shrub encroachment after agricultural land abandonment in the past decades.

2. Data and Methods

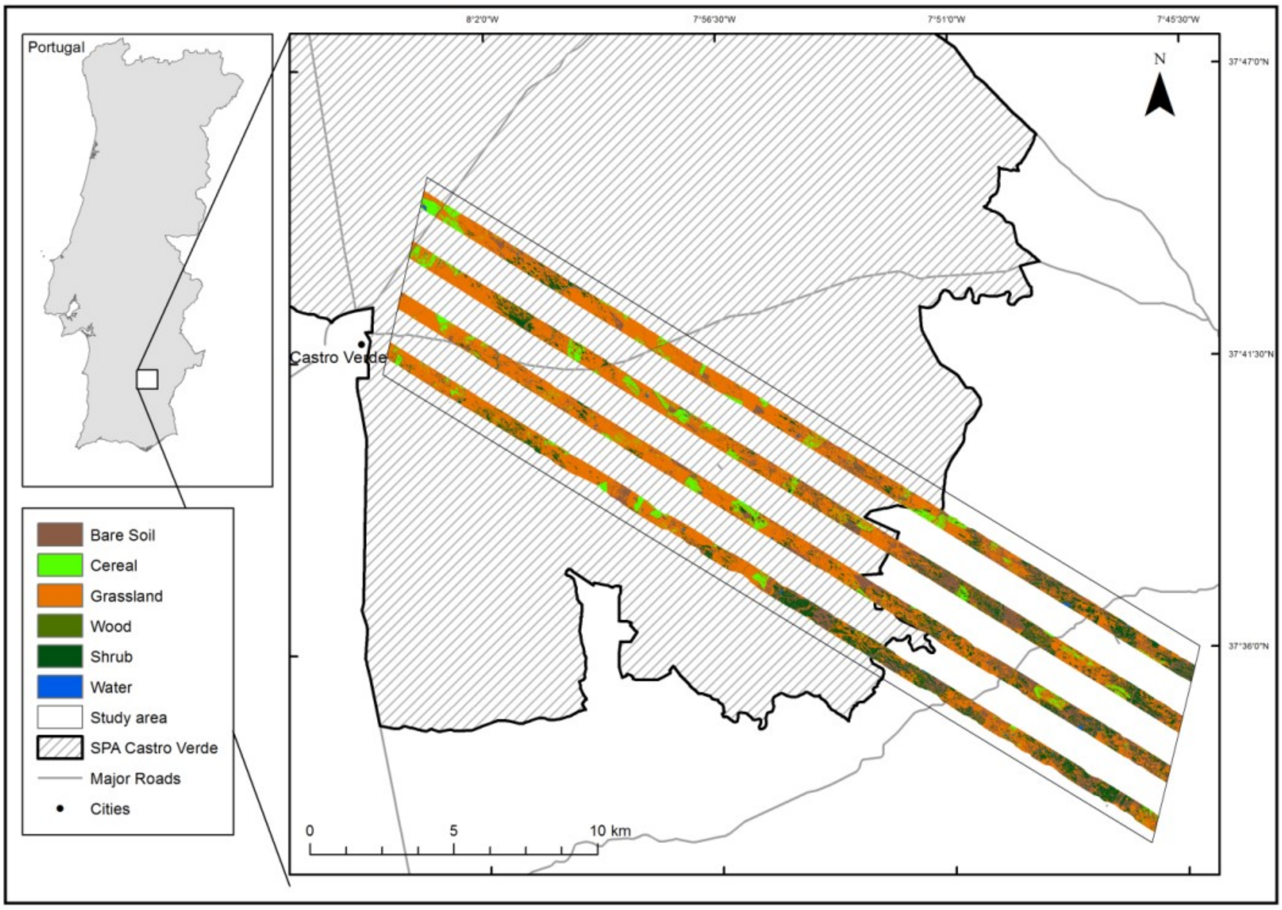
2.1. Study Area

The study site is located in southeastern Portugal within the Baixo Alentejo region (Figure 1). It is dominated by rolling hills forming a gentle terrain (100–300 m) [30] dominated by poor litho-soils of non-calcareous schist [31]. Very hot summers with maximum temperatures of around 30 °C–40 °C and a yearly precipitation of 500–650 mm, primarily in the mild winter months, are predominant [32].

During the so-called "Wheat Campaign" in the early 20th century in Portugal strove for self-sufficiency in agricultural productivity. Land was cultivated even under unfavorable biophysical conditions, which subsequently led to soil exhaustion and impoverishment. This development resulted in the application of extensive farming practices following a rotational scheme, in which land was alternately cultivated or left fallow [31]. Decades of these traditional agricultural practices, led to a characteristic landscape mosaic of cereal, fallow, and ploughed fields, often referred to as pseudo-steppe (or cereal steppe), which provides habitats of national and international importance, e.g., for several endangered bird species [33]. This typical landscape is complemented by "*montado*" woodlands, which are agricultural areas characterized by scattered cork or holm oak trees (*Quercus suber* and *Quercus rotundifolia*) and an understory mostly used as pasture or for cereal cultivation.

The growing competition on global agricultural markets and agricultural subsidies, did not allow farmers keeping the traditional agricultural management schemes. This led to an intensification of agricultural production on fertile soils and to the abandonment of formerly cultivated poor soils [34,35]. Shrub encroachment on the abandoned farmland threatens the fragile regional ecosystem [12] and leads to heterogeneous patterns of shrub cover densities dominated by rockroses (C*istus spp.*), which are common pioneer shrubs in that region of the Mediterranean [13,31,36]. To counteract this threat, the Special Protection Area (SPA) for birds of Castro Verde was established in 1995 within the European Union's Birds Directive (79/409/EEC, later replaced by 2009/147/EC). This SPA covers an area of approximately 80,000 ha, including around 65,000 ha of pseudo-steppe (Figure 1) [37]. Farmers within the SPA receive subsidies to keep management schemes (including the traditional extensive farming on a rotation basis) that maintain the key human-made habitats necessary to conserve the local biodiversity [38].

Figure 1. Study area in southern Portugal and classified land cover. The area covers partly the Special Protected Area Castro Verde.



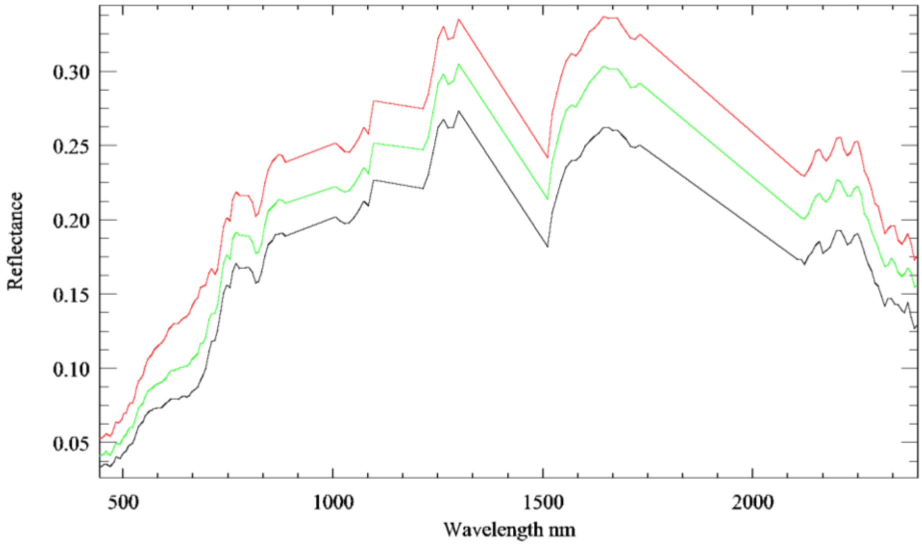
We chose the test site to cover areas inside and outside the SPA (Figure 1), following the observed gradient of (rockrose) shrub cover densities from low cover inside the SPA, to high cover outside. The extent of the analyzed study area is approximately 30 × 7 km^2 and oriented from Northwest to Southeast.

2.2. Data

Two data acquisition campaigns were carried out in April and August 2011, in Southern Portugal. During both campaigns airborne hyperspectral data were acquired with the AISA Eagle (400–970 nm;

3.3 nm max. spectral resolution) and Hawk (970–2450 nm; 8 nm max. spectral resolution) sensors from two different altitudes (1500 m and 4500 m) following the same nadir lines in four flight stripes at each altitude (Figure 1). All data went through radiometric and atmospheric correction using the radiative transfer model ATCOR4 for airborne atmospheric correction [39] with subsequent spectral polishing (Savitzky-Golay filtering) and nadir normalization to correct for across-track illumination differences [40]. This was followed by a geometric correction and subsequent data stacking and mosaicking of the 4500 m data [41], resulting in four image stripes with a GSD (ground sampling distance) of 1.8 m (600 m swath in each flight line; 1500 m altitude) for April and August and two image mosaics with 5.4 m GSD (app. 7000 m mosaic swath; 4500 m altitude) for April and August. The 5.4 m GSD mosaics were further used to generate synthetic EnMAP data, by aggregating the spatial resolution (from 5.40 m to 30 m), simulating EnMAP spectra with the respective spectral response curves and adding noise with the EnMAP-end-to-end Simulation Tool (EeteS) [42]. The resulting EnMAP image had a spatial resolution of 30 m and covered the spectral range from 423 to 2439 nm. After excluding spectral regions close to atmospheric absorption and extrapolated bands (during the simulation), reflectance values in 146 bands remained. Due to requirements of the applied algorithms we scaled the data between 0 and 1 by dividing the reflectance trough 10,000 using ENVI version 4.8 (*Exelis Visual Information Solutions, Boulder, Colorado*).

Figure 2. Exemplary spectral signatures of 30 × 30 m simulated EnMAP shrub pixel with below 10% (**red**), approximately 50% (**green**) and above 90% shrub fraction (**black**).



Based upon local knowledge, the hyperspectral signatures of the AISA images, aerial photo interpretation and pictures taken during several field campaigns, we derived reference data to produce a discrete land cover map from the 1.8 m GSD image from April by applying a Support Vector Classification [23]. In the first step six classes were distinguished (Bare Soil, Cereal, Grassland (Fallow vegetation), Woodlands, Shrublands (*Cistus*), and Water; see Figure 1). The reference data for these six classes were then reclassified to a Cistus-Shrub class and a Non-Cistus-Shrub class including all other land cover features, *i.e.*, annual vegetation, trees, bare soil, water, impervious areas, and other shrub species. The binary reference data were then used to derive a shrub map from the August 1.8 m

GSD image, as in the summer less other photosynthetic active vegetation is present that could influence the spectral signal of shrub pixel [13]. The binary shrub map had an area adjusted overall accuracy [43] of 94.2% (*cf.* supplementary material 1). It was aggregated to 30 m GSD to fit the spatial resolution of the EnMAP data, which resulted in shrub cover fractions per pixel ranging from 0 (0% cover) to 1 (100% cover). Three examples of spectral signatures of shrub cover with varying fractional coverage are shown in Figure 2. The fractional shrub cover map was used as reference data for training and validating the regression models.

2.3. Data Analysis

To estimate fractional shrub cover from EnMAP data, we define a functional relationship between image spectra and shrub fractions as:

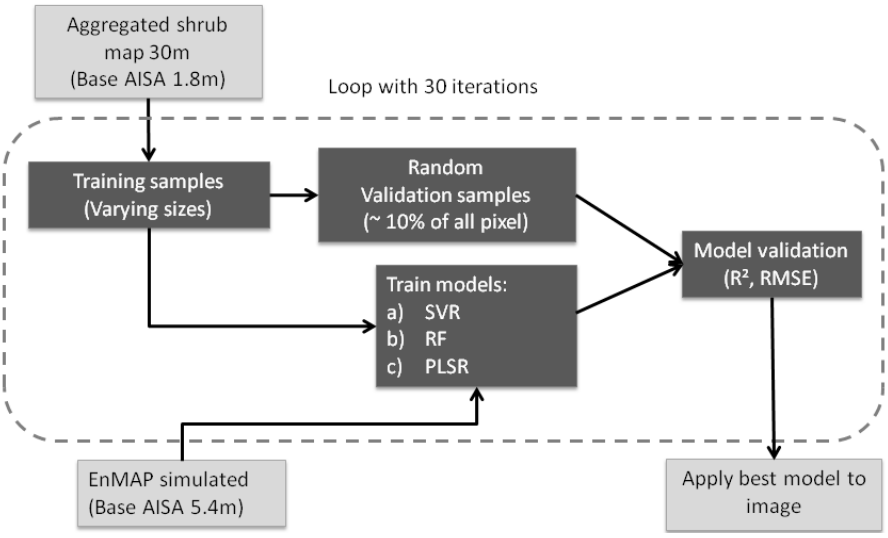
$$y_i = \delta(X_i^p); i \in \{1, \dots, n\}$$

where δ is an unknown functional relationship (*i.e.*, the regression model), y_i the observed shrub fractions and X_i^p the p-dimensional spectral feature space at point i. Based on a known set of y_i and X_i^p, δ can be defined (*i.e.*, training process); and unknown shrub fractions y_{new} can be estimated using a new set of spectra X_{new}^p, *i.e.*, for a whole image (*i.e.*, application process). As regression model, we use three state-of-the-art algorithms (SVR, RF, and PLSR), described in detail in the following sub-sections. To test the sensitivity of the algorithms to sample size variations during the training process, five different sample sizes (100, 200, 500, 700, 1000 pixel) were drawn from the reference map using a random sampling design. Each regression algorithm was trained with the same training samples and validated using 7000 independently drawn samples (approximately 10% of the total number of pixels). To assess the accuracy of each single model for the prediction of fractional shrub cover we calculated the coefficient of determination (R^2). To account for model uncertainty related to random effects, we performed 30 iterations on all datasets and all algorithms. Based on the highest value of R^2 (after 30 iterations) we have chosen the "best" models for each algorithm, which were then applied to predict continuous shrub cover maps for the whole study area. Additionally, we calculated the rooted mean square error (RMSE) for each model, in order to assess the actual error, which can be expected when the models are applied (Figure 3). In order to validate the spatial accuracy of the predictions, the predicted shrub cover fraction maps were compared to the reference map by calculating the absolute error and the mean absolute error (MAE) for each map. All analysis steps were performed in the statistical software R [44] using the packages described in the respective sub-sections below.

2.3.1. Support Vector Regression

Support Vector Machines is a non-parametric machine learning algorithm for classification and regression purposes [45]. In the case of regression, the underlying principle is to build a linear model based on a subset of high dimensional prediction variables to predict one-dimensional response values, in this case fractional shrub cover. Therefore, the input data are transformed by using an appropriate Kernel function so that a linear model can be fitted. The Gaussian Radial Basis Kernel (RBF) was used, as it is known to show reliable results without prior knowledge to the data distribution [46]. The Kernel radius is defined by the parameter γ. Once the training data are mapped into hyperspace, a linear model is solved while minimizing Vapnik's ε-insensitive cost function to reduce model complexity. Samples within a ε-defined margin are ignored, while the influence of the samples outside this margin is penalized. The parameter C regularizes the trade-off between model complexity and error frequency [45,47,48]. Thus, only a few training data (i.e., support vectors) are used to define the regression model, which increases the SVR processing performance in terms of computational time. For the implementation the *R*-package *e1071* [49] was used, which is based on the LIBSVM library [50]. To avoid model over-fitting, the best kernel-parameter combination of C and γ was selected through a heuristic grid search. During the grid search the model performance was optimized and evaluated by the mean squared error in a 10-fold cross-validation. The ε value was not adjusted and set to 0.1.

Figure 3. Workflow of the analysis. Light grey boxes contain input and output products, dark grey boxes show steps during the analysis. Steps within the broken line were repeated 30 times.


Aggregated shrub map 30m (Base AISA 1.8m)
Loop with 30 iterations
Training samples (Varying sizes)
Random Validation samples (~ 10% of all pixel)
Model validation (R², RMSE)
Train models:
a) SVR
b) RF
c) PLSR
EnMAP simulated (Base AISA 5.4m)
Apply best model to image

2.3.2. Random Forest Regression

Random Forest (RF) is an ensemble algorithm that is based on the Classification And Regression Tree algorithm (CART) [51]. The general principle is to combine many decision trees, as each alone is assumed to be prone to errors, to a single strong decision tree ensemble. The results of all trees are averaged (regression) or selected by a majority vote (classification), respectively [24]. To individualize each tree in the "forest", the trees are grown with a random subset drawn with replacement from the original training samples, and only a defined subset of input variables is used to find the best split criteria at each node. The remaining training samples are used to determine the performance of each tree, known as the out-of-bag error estimate (OOB). Aggregated over all trees in the ensemble, the OOB can be used as independent measure of classification or regression accuracy. RF has been shown to be very cost-efficient and to handle large data sets easily [27]. The *R*-package *randomForest* [52] was used in this study, which allows to tune the input parameters—*i.e.*, the number of trees and the number of features to be drawn from at each node with respect to the OOB. In this study, the initial RF model was built using 100 trees and the feature subset was defined as one third of the overall number of features. The size of the feature subset was iteratively increased or decreased, respectively, by the factor of 1.5 until there was no improvement in OOB greater than 0.02.

2.3.3. Partial Least Squares Regression

PLSR is based on the partial least squares approach introduced by Herman Wold in the 1970's and it found many applications in a wide range of research fields that are dealing with high-dimensional and particular collinear datasets [25]. The underlying concept of PLSR is to construct new predictor components, based on the covariance between the predictor and response variables with respect to collinearity [29]. The components are then used to build a linear regression model. Mevik *et al.* [53] developed the *R*-package *pls* and the package *autopls* [54] provides a *pls* based wrapper for a backward selection of significant components. We used the *autopls* procedure, during which the prediction model is internally validated using a 10-fold cross-validation and the components are selected by an optimization procedure based on variable significance and autocorrelation [54].

3. Results

3.1. Regression Performance

Predictions of the SVR showed higher accuracies with increasing sample size. Mean R^2 values of 30 iterations continuously increased from 0.50 (100 samples) to 0.64 (1000 samples) and the mean RMSE values declined from 0.15 to 0.12 (Table 1; Figure 4), respectively. A certain degree of saturation of the gain in prediction accuracy was observed between 700 and 1000 samples, as the respective mean R^2 and RMSE values were equal (R^2 = 0.64; RMSE = 0.12). The variation in the accuracy measures between iterations, as expressed by the standard deviation, decreased with large samples (Figure 4).

Table 1. Descriptive statistics of the accuracy measures after 30 iterations for all sample sizes.

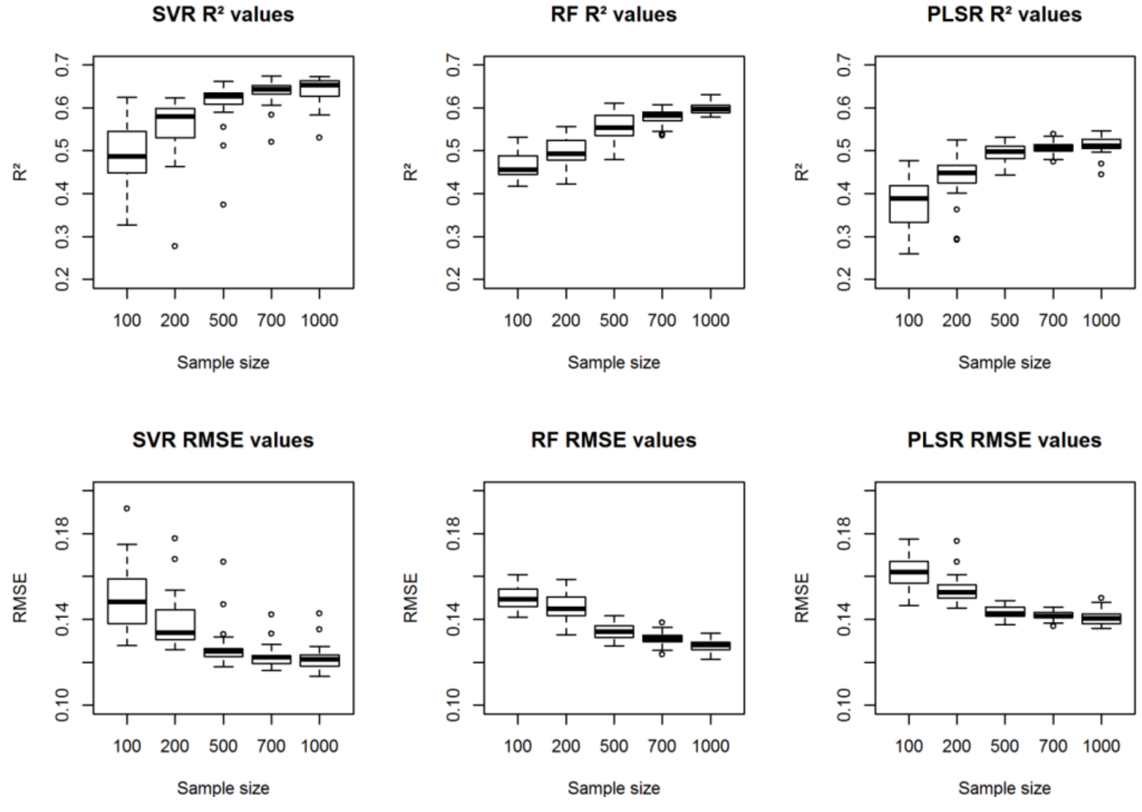
		R^2					RMSE				
	Sample Size	**100**	**200**	**500**	**700**	**1000**	**100**	**200**	**500**	**700**	**1000**
SVR	Mean	0.50	0.56	0.61	0.64	0.64	0.15	0.14	0.13	0.12	0.12
	Std.	0.07	0.07	0.05	0.03	0.03	0.02	0.01	0.01	0.01	0.01
RF	Mean	0.47	0.50	0.56	0.58	0.60	0.15	0.15	0.13	0.13	0.13
	Std.	0.03	0.03	0.03	0.02	0.01	0.01	0.01	0.01	0.01	0.01
PLSR	Mean	0.38	0.44	0.50	0.51	0.51	0.16	0.15	0.14	0.14	0.14
	Std.	0.05	0.05	0.02	0.02	0.02	0.01	0.01	0.01	0.01	0.01

RF predictions increased in accuracy from R^2 of 0.47 (100 samples) to 0.60 (1000 samples) and the RMSE respectively decreased from 0.15 (100 samples) to 0.13 (1000 samples) (Table 1; Figure 4). RF predictions still gained in accuracy between the large sample sizes, no saturation could be observed. The standard deviation of the prediction accuracy within 30 iterations was small, especially with large sample sizes (Table 1).

Model predictions based on the PLSR algorithm resulted in increasing accuracies from R^2 of 0.38 (100 samples) to 0.51 (1000 samples) and a decreasing RMSE from 0.16 (100 samples) to 0.14 (1000 samples) (Table 1; Figure 4). Accuracy measures saturated between the large sample sizes of 500, 700, and 1000 at R^2 values of 0.5–0.51 and a RMSE value of 0.14 (Figure 4).

As the processing times may vary strongly with different processor performances, our observations only indicate an approximate trend, rather than absolute reference values. Nevertheless, as all models were built and validated with the same input data and under equal conditions, the experimental set-up allows a valid comparison of processing times. The mean processing time of the three regression techniques grew along with the sampling size. RF was the fastest regression technique with processing times of approximately one minute, even with sizes of 1000 samples. SVR average processing time grew from below one minute (100 samples) up to five minutes (1000 samples). However, PLSR with backward selection (*autoPLS*) needed an average of approximately two minutes for models with 100 samples and constantly slowed down to between 35 and 40 min mean processing time per model with 1000 samples (*cf.* supplementary material 2).

Figure 4. Boxplots of R^2 and RMSE values for all algorithms and sample sizes after 30 iterations.



3.2. Spatial Pattern of Fractional Shrub Cover

The best results were observed when using SVR with the biggest sample size (1000 samples; Table 1). To achieve comparable results, we used the best models of each algorithm trained with 1000 samples for the spatial prediction of shrub cover fractions (Figure 5). The mean absolute errors of the predictions were respectively 0.06, 0.07, and 0.10 for SVR, RF, and PLSR, when compared with the observed data. In general, all maps showed similar patterns of shrub cover distribution (Figures 5 and 6, however, the RF and PLSR based maps showed shrub overestimations (leading to higher errors) especially in non-shrub areas (Figures 5 and 6). Furthermore, it can be observed that all algorithms showed errors in the dense shrub areas with high cover fractions. Scatterplots between predicted and observed shrub fractions (Figure 7) revealed that these errors are mainly due to shrub cover underestimations of all algorithms, but especially, RF and PLSR tend to underestimate the high shrub cover fractions.

Figure 5. (**a**) Shrub cover maps based on SVR, RF and PLSR models using 1000 sample points. Blue colors indicate low, green to yellow colors high shrub cover fractions; (**b**) Error maps between the observed and predicted shrub cover fractions for every algorithm. Yellow colors indicate low errors, red high errors. Close ups A and B show details of the spatial prediction accuracy in regions with sparse (**A**) and dense (**B**) shrub cover. In all maps the white areas indicate regions of missing data.

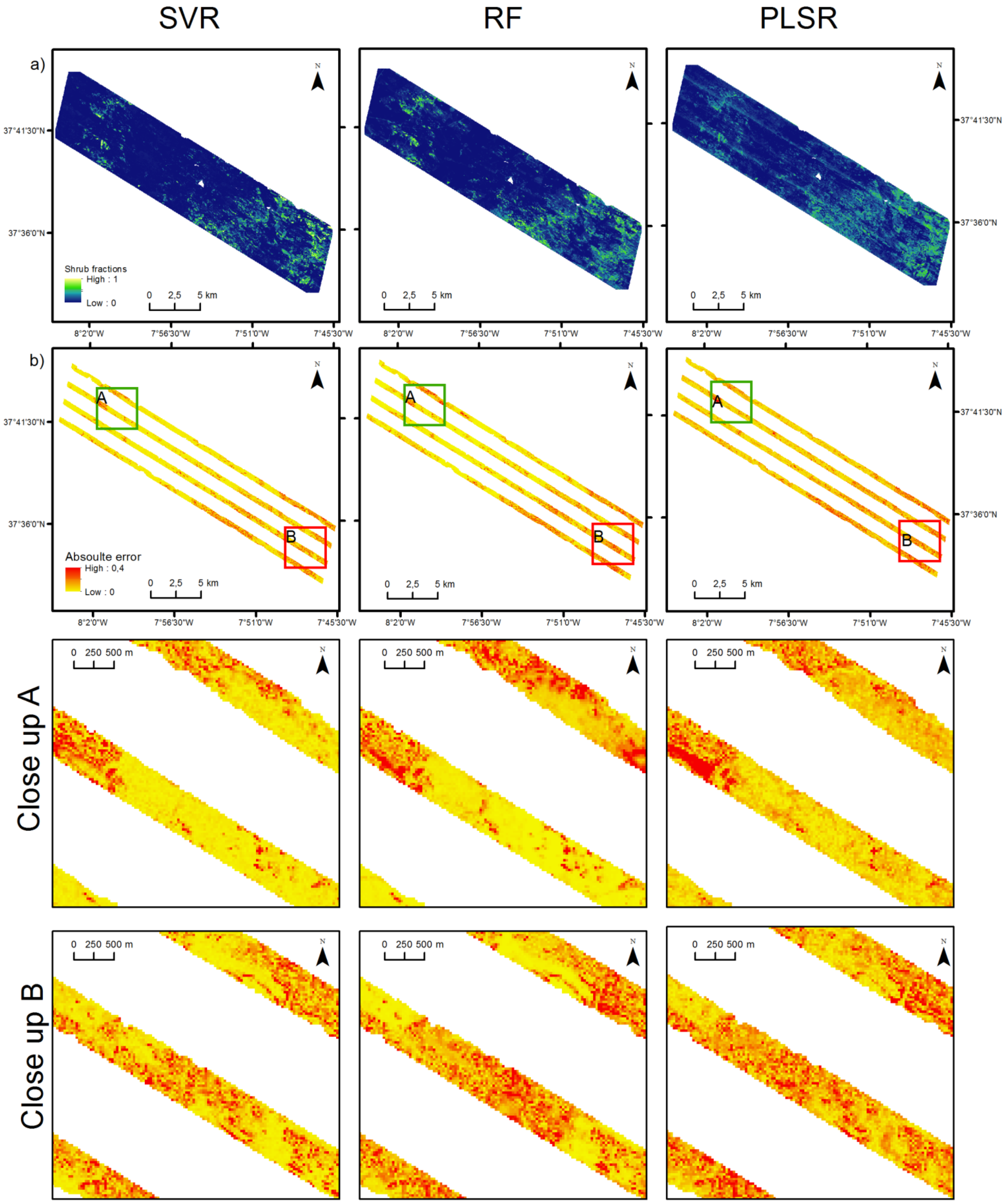
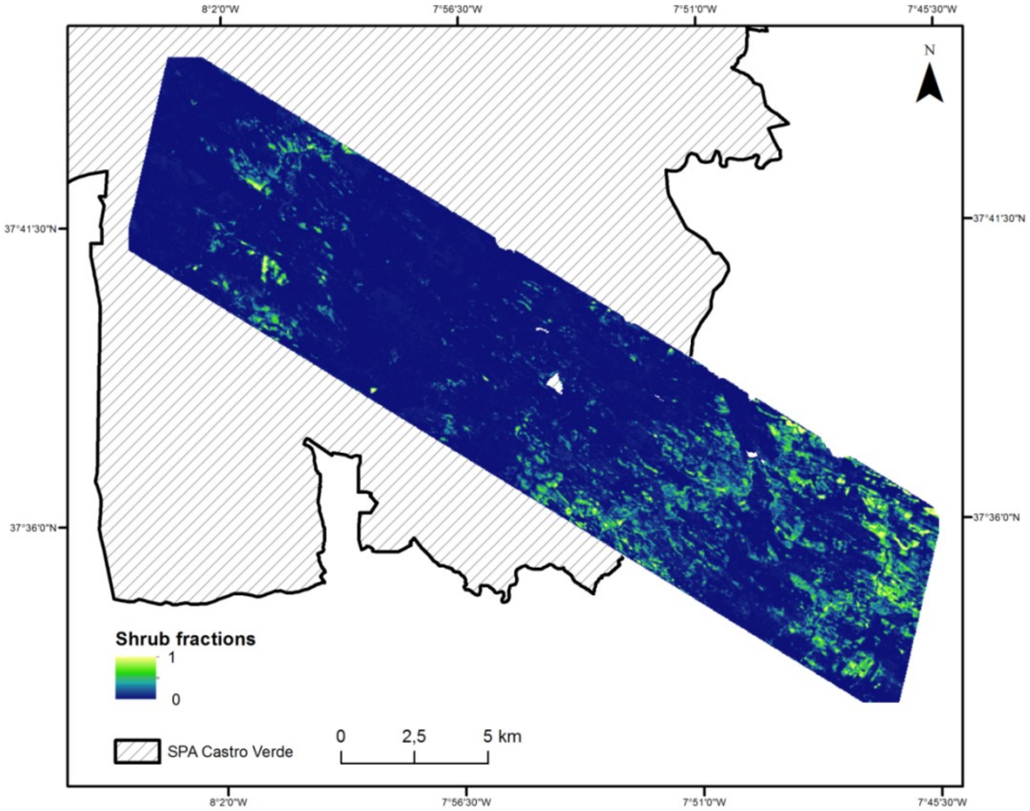
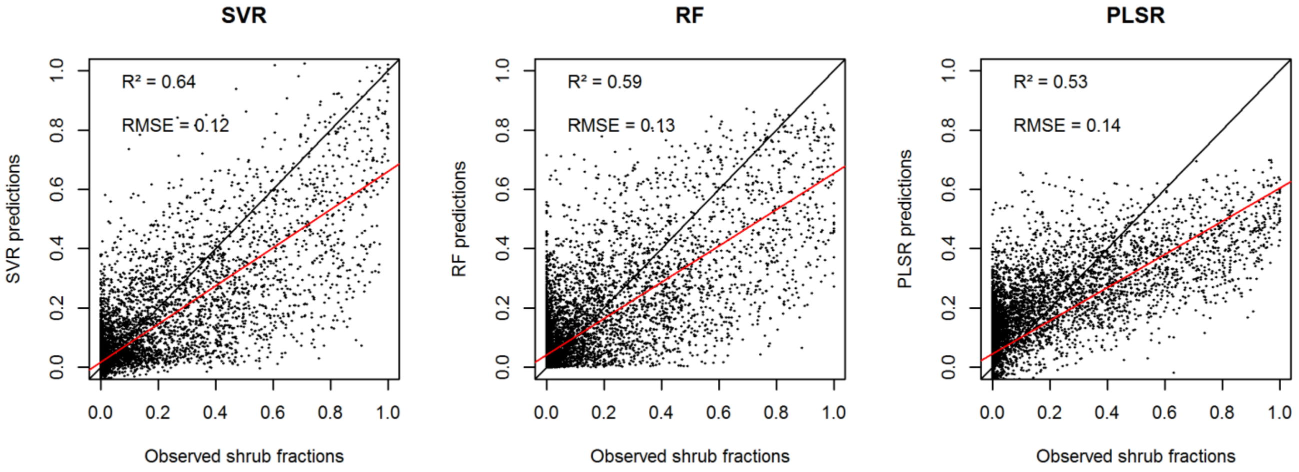


Figure 6. Fractional shrub cover map based on the SVR model trained with 1000 random samples. Yellow to green areas indicate high shrub cover fractions, blue colors low cover fractions.



The general spatial pattern revealed a gradient of shrub cover in the study area from the northwest, inside the SPA (Figure 6), with only a few sparse shrub patches and average cover fraction of 3.8%, to large dense shrub cover areas in the southeast, outside the SPA, with average cover fraction of 12.3%.

Figure 7. Scatterplots of the observed and predicted shrub cover fractions in the three shrub cover maps (based on approximately 10,000 pixels). The black line shows the optimal 0–1 line and the red line the actual regression model between the observed and predicted cover fractions. The related accuracy measures are shown in the respective plots. Negative outliers of SVR and PLSR are not shown.



4. Discussion

4.1. Comparison of Regression Algorithms

Three state of the art machine learning techniques were applied to quantify fractional shrub cover in a study area in southern Portugal. Their predictive performances as well as their sensitivity to training sample size were compared. The results revealed that all algorithms were able to predict the general shrub cover patterns, while SVR models outperformed RF and PLSR on all cases. Although the latter two algorithms reached lower accuracies than the SVR models, their predictions were still reasonably good with mean RMSE values of 0.13 (RF) and 0.14 (PLSR).

Even though comparison among studies in different areas and with different validation procedures is difficult, our results showed a good prediction accuracy compared to other studies predicting fractional vegetation cover with earth observation data (e.g., [27,28,55]). Tuia *et al.* [55] compared the performance of a SVR and a Multioutput-SVR algorithm for fractional vegetation cover estimation with hyperspectral CHRIS/PROBA data and achieved mean RMSE values between 0.25 and 0.33 for their SVR predictions. They also analyzed the influence of sampling size by incrementally splitting their input data into training and validation datasets. Analogous to our findings, their predictions gained in accuracy when using a larger sampling size for training.

Walton [27] achieved high accuracies in predicting fractional urban forest canopy cover by using a SVR model. The model was trained with reflectance data and tasseled cap values derived from a multi-seasonal Landsat-7 stack. The influence of training sample size was tested by varying the training sample sizes according to increasing percentages of the total number of pixels in the image (1%, 2%, 4%, 8%, and 16%). Walton's comparison of observed and predicted values revealed MAE's of around 0.12 and, in line with our results, a gain in accuracy with larger sampling sizes. While comparison among different studies is critical due to potential differences in the study design, our agreement between predicted and observed pattern reaches an MAE of 0.06. The results might be influenced by our experimental setup in which we used simulated spectral data and reference data from the same acquisition date, but they also highlight the benefits of EnMAP data, as we were able to accurately quantify shrub cover in a heterogeneous and complex ecosystem from a single time step.

Nonetheless, multi-seasonal spectral information have led to improvements in land cover classifications of our study area in other studies [30,56] and are furthermore necessary to monitor shrub encroachment and phenological change [57]. With the upcoming EnMAP mission, hyperspectral data will be available at multi-seasonal resolution (with a revisit cycle of up to four-days), likely further improving the prediction of fractional shrub cover using hyperspectral data and enabling a time consistent monitoring of shrub encroachment.

4.2. Training Sample Size

In all models, the prediction accuracy increased with larger training sample sizes, which corresponds to findings in other studies (see Section 4.1). The gain between the larger sample sizes was lower than that between the smaller sample sizes, indicating some degree of saturation in accuracy, most particularly in the SVR and PLSR models. Both algorithms showed only a marginal increase in accuracy between 500 and 1000 training samples (Table 1). The SVR models trained with

500 samples already yielded higher accuracies than the RF and PLSR models trained with 1000 samples. By applying 30 permutations of each test, we estimated the mean and standard deviation of accuracy for each algorithm. The standard deviation generally declined with larger training sample sizes, indicating that larger sample sizes lead to more robust models. This was especially the case for SVR and PLSR models (Table 1; Figure 4). The small training samples do not necessarily cover the complete spectral variability in the study area and thus might lead to inaccurate predictions by PLSR and SVR. RF in turn, seems to be more robust in terms of training sample size, showing less variation in accuracy throughout all sample sizes.

Especially for the operational use of future EnMAP data for fractional shrub cover mapping, sample sizes are an important issue, as it is often not feasible to gather huge amounts of ground-truth data. In our case sampling sizes of approximately 1% of the total number of pixel were sufficient to produce a SVR model with an accuracy of 0.64 (R^2) leading to a continuous shrub map with an MAE of 0.06.

4.3. Uncertainty and Error Propagation

Although, we tried to standardize all tests and rule out random effects by the permutation test, some sources of error remain. One point to be considered is uncertainty in the input data. The spectral and shrub fractional cover information resulted from two over-flights, acquired on the same date with the same instruments, but from different elevations (1500 and 4500 m) and thus resulting in different GSD. Despite the geometrical correction of both images, there is still some sub-pixel uncertainty in their spatial fit among each other. Both images were further processed and spatially aggregated to 30 m, which might compensate some error at the fine spatial resolution, but geometric error effects were not tested in this study.

The image with 1.8 m GSD was used to produce the binary shrub map, which was subsequently used to develop the fractional shrub cover reference data at 30 m spatial resolution. Even though the binary shrub cover map was validated with a high overall accuracy, it is still a potential source of error. As the map resulted from a hard classification with discrete classes, pixel holding shrub fractions lower than 50% were most probably assigned to the non-shrub class. These shares of false quantifications add up over the whole study area and, thus, bear a potential source of error. These errors were consequently also present in the aggregated 30 m fractional shrub map and may have influenced the results. Moreover, it cannot be completely ruled out that vegetation other than *Cistus* shrubs might be captured within the shrub class. Nonetheless, as *Cistus* is the major shrub species in the test site and reference samples were collected during the dry season, we assume the influence of other shrubs on our results to be negligible.

The EnMAP image was simulated using EeteS, which despite reliable results in other case studies [42], remains with certain limitations as it is a simulation of what EnMAP spectra could look like under specific constraints (e.g., atmospheric influences, noise, *etc.*). Nonetheless, the simulated image approximates spectral, radiometric and spatial characteristics of future EnMAP data and thus gives a broad insight in the usefulness of EnMAP hyperspectral image data for future applications.

4.4. Predicted Spatial Patterns of Shrub Cover

The general shrub patterns within the study area were well captured in all predicted maps. However, RF and PLSR led to substantial errors in both dense shrub cover and no shrub cover regions. The scatterplots (Figure 6) revealed that all three algorithms overestimated shrub cover fractions close to zero and underestimated high shrub cover fractions. This was particular the case for the RF and PLSR models, with many regions with no shrub cover being predicted with cover fractions of up to 40%, whereas high cover fractions were mostly underestimated (*i.e.*, maximum shrub cover fractions of 90% and 70% for RF and PLSR models, respectively). These effects might be explained by our random sampling strategy, which represents the statistical distribution of shrub pixel in the reference image. The distribution is right skewed, showing many non-shrub pixels, which include aggregated *Cistus* shrub pixel with less than 50% cover of rockrose shrubs in the 1.80 m input classification (*cf.* Section 4.3). This could explain the overestimations of the low cover fractions. Furthermore, could the relatively small amount of trainings pixels with high shrub cover fractions not be sufficient for accurate training of the regression algorithms. A potential solution to this problem could be an adjusted, data driven sampling strategy, better representing high shrub fractions in the training sample while still covering all spectral variation in the non-shrub regions. However, such a sampling strategy would require *a priori* knowledge on the statistical target data distribution within the reference image, which is rarely the case in non-experimental setups.

The continuous shrub map predicted by the overall best model (SVR trained with 1000 samples) has a spatial accuracy of 0.06 (MAE) when compared to the input data (Figure 5; Figure 7). The shrub cover patterns predicted seem to indicate a positive effect of the land use incentives within the SPA. Here agricultural land abandonment—and, thus, shrub encroachment—is prevented, which would be disadvantageous for the local biodiversity and the species that are adapted to the human-made landscape of the pseudo-steppe [12]. This effect is particularly evident along the southern border of the SPA (Figure 7), where the spatial patterns alter from sparse to dense shrub cover areas. Since these areas outside the SPA are not affected by special land use incentives that promote extensive agriculture, but are widely left unmanaged, shrub encroachment is unbridled. Nevertheless, following the findings of Maestre *et al.* [6] this development must not necessarily be related to ecosystem degradation, but might be interpreted as natural succession on poor soils, after many years of exploitation, thus indicating ecosystem recovery. These observations highlight the ambiguity of shrub encroachment processes and stress the value of our results for the detailed mapping and monitoring of shrub cover. Future EnMAP hyperspectral imagery in combination with SVR regression thus offers unique opportunities for monitoring gradual processes over large extents and can likely facilitate the understanding of complex land use change processes in natural or semi-natural landscapes.

5. Conclusions

For the first time simulated EnMAP data were used to predict shrub cover fractions in a study area in southern Portugal. The study area is affected by different agricultural management schemes, ranging from areas with strong land use incentives within the Special Protected Area for Birds to areas without particular management. In order to accurately quantify the shrub cover in our study area, a methodological framework was developed to compare the performance of three regression algorithms (Support Vector Regression, SVR; Random Forest, RF; Partial Least Squares Regression, PLSR), along with the influence of the sampling size on the prediction accuracy.

All regression algorithms showed good prediction accuracies if initialized with large training samples, but SVR outperformed the others (Mean $R^2 = 0.64$; Mean RMSE = 0.12). In our study area, SVR performed best with a training sample size of approximately 1% (N = 700) of the total number of pixels. The fractional shrub map derived from the SVR model had the best agreement with the reference map (MAE = 0.06) and revealed a comprehensible gradient of shrub cover in the study area, likely driven by local land use management strategies. The assessment of these observations in a local context stresses the ambiguity of shrub encroachment processes relative to the prevalent environmental and cultural conditions.

Our results suggest that the upcoming EnMAP mission, in conjunction with SVR, bears great potential for the quantification of fractional vegetation cover and for monitoring gradual land use change processes in a consistent manner. This is of special importance for conservation research, the support of landscape management and policy decisions making processes, as highlighted in our study area. Furthermore, we suggest analyzing the potential of regression techniques for the quantification of vegetation in complex ecosystems, with emphasis on the benefits of multi-seasonal hyperspectral data.

Acknowledgments

The authors would like to thank Karl Segl from the GFZ German Research Center for Geosciences, for his contribution to this research by simulating the EnMAP imagery with EeteS. This research is part of the research activities of the EnMAP Core Science Team (ECST), and was funded by the German Aerospace Centre (DLR)—Project Management Agency, granted by the Ministry of Economics and Technology (BMWi; grant nr.50EE0949). This work was also partly funded by the European Facility for Airborne Research (EUFAR) with the HyMedEcos-Gradients project (EUFAR 11-04) with support from the Airborne Research and Survey Facility (ARSF), Geodetic Equipment Facility (GEF) and Field Spectroscopy Facility (FSF) of the UK's Natural Environmental Research Council (NERC). During the field campaigns in Portugal, support was given by the Liga para a Proteção da Natureza (LPN).

Author Contributions

Marcel Schwieder: Contributed to designing the research; performed the research; analyzed the data; wrote the paper. Pedro Leitão: Designed the research; contributed to performing the research; contributed to writing the paper. Stefan Suess: Contributed to data analysis; contributed to writing the paper. Cornelius Senf: Contributed to data analysis; contributed to writing the paper. Patrick Hostert: Contributed to performing the research; contributed to writing the paper.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Vitousek, P.M. Beyond global warming: Ecology and global change. *Ecology* **1994**, *75*, 1861–1876.
2. Rey Benayas, J.; Martins, A.; Nicolau, J.M.; Schulz, J.J. Abandonment of agricultural land: An overview of drivers and consequences. *CAB Rev.: Perspect. Agric. Vet. Sci. Nutr. Nat. Resour.* **2007**, *2*, 1–14.
3. Millenium Ecosystem Assessment. *Ecosystems and Human Well-Being: Synthesis*; Island Press: Washington, DC, USA, 2005; p. 137.
4. Eldridge, D.J.; Bowker, M.A.; Maestre, F.T.; Roger, E.; Reynolds, J.F.; Whitford, W.G. Impacts of shrub encroachment on ecosystem structure and functioning: Towards a global synthesis. *Ecol. Lett.* **2011**, *14*, 709–722.
5. Hill, J.; Hostert, P.; Tsiourlis, G.; Kasapidis, P.; Udelhoven, T.; Diemer, C. Monitoring 20 years of increased grazing impact on the Greek Island of Crete with earth observation satellites. *J. Arid Environ.* **1998**, *39*, 165–178.
6. Maestre, F.T.; Bowker, M.A.; Puche, M.D.; Belén Hinojosa, M.; Martínez, I.; García-Palacios, P.; Castillo, A.P.; Soliveres, S.; Luzuriaga, A.L.; Sánchez, A.M.; *et al.* Shrub encroachment can reverse desertification in semi-arid Mediterranean grasslands. *Ecol. Lett.* **2009**, *12*, 930–941.
7. Shoshany, M.; Karnibad, L. Mapping shrubland biomass along Mediterranean climatic gradients: The synergy of rainfall-based and NDVI-based models. *Int. J. Remote Sens.* **2011**, *32*, 9497–9508.
8. Van Auken, O.W. Shrub invasions of North American semiarid grasslands. *Annu. Rev. Ecol. Syst.* **2000**, *31*, 197–215.
9. Van Auken, O.W. Causes and consequences of woody plant encroachment into western North American grasslands. *J. Environ. Manag.* **2009**, *90*, 2931–2942.
10. Millenium Ecosystem Assessment. A Report of the Millennium Ecosystem Assessment. In *Ecosystems and Human Well-Being: Desertification Synthesis*; Island Press: Washington, DC, USA, 2005; p. 26.
11. Fonseca, F.; Figueiredo, T.; Bompastor Ramos, M.A. Carbon storage in the Mediterranean upland shrub communities of Montesinho Natural Park, Northeast of Portugal. *Agrofor. Syst.* **2012**, *86*, 463–475.
12. Moreira, F.; Russo, D. Modelling the impact of agricultural abandonment and wildfires on vertebrate diversity in Mediterranean Europe. *Landsc. Ecol.* **2007**, *22*, 1461–1476.
13. Calvão, T.; Palmeirim, J.M. A comparative evaluation of spectral vegetation indices for the estimation of biophysical characteristics of Mediterranean semi-deciduous shrub communities. *Int. J. Remote Sens.* **2011**, *32*, 2275–2296.
14. Lambin, E.F.; Geist, H. *Land-Use and Land-Cover Change: Local Processes and Global Impacts*; Springer: Berlin, Germany, 2006.
15. DeFries, R.; Pagiola, S.; Adamowicz, W.L.; Akcakaya, R.H.; Arcenas, A.; Babu, S.; Balk, D.; Confalonieri, U.; Cramer, W.; Falconi, F.; *et al.* Analytical Approaches for Assessing Ecosystem Condition and Human Well-Being. In *Ecosystems and Human Well-Being: Current States and Trends*; Hassan, R.M., Ed.; Island Press: Washington, DC, USA, 2005; pp. 37–71.

16. Lawrence, R.L.; Wood, S.D.; Sheley, R.L. Mapping invasive plants using hyperspectral imagery and Breiman Cutler classifications (RandomForest). *Remote Sens. Environ.* **2006**, *100*, 356–362.
17. Thenkabail, P.S.; Enclona, E.A.; Ashton, M.S.; van der Meer, B. Accuracy assessments of hyperspectral waveband performance for vegetation analysis applications. *Remote Sens. Environ.* **2004**, *91*, 354–376.
18. Chopping, M.; Su, L.; Laliberte, A.; Rango, A.; Peters, D.P.C.; Kollikkathara, N. Mapping shrub abundance in desert grasslands using geometric-optical modeling and multi-angle remote sensing with CHRIS/PROBA. *Remote Sens. Environ.* **2006**, *104*, 62–73.
19. Stuffler, T.; Förster, K.; Hofer, S.; Leipold, M.; Sang, B.; Kaufmann, H.; Penné, B.; Mueller, A.; Chlebek, C. Hyperspectral imaging—An advanced instrument concept for the EnMAP mission (environmental mapping and analysis programme). *Acta Astronaut.* **2009**, *65*, 1107–1112.
20. Kaufmann, H. *Science Plan of the Environmental Mapping and Analysis Program (EnMAP).* Scientific Technical Report; Deutsches Geoforschungszentrum GFZ: Potsdam, Germany, 2012; p. 65.
21. Storch, T.; Bachmann, M.; Eberle, S.; Habermeyer, M.; Makasy, C.; Miguel, A.; Mühle, H.; Müller, R. Enmap Ground Segment Design: An Overview and Its Hyperspectral Image Processing Chain. In *Earth Observation for Global Change. Lecture Notes in Geoinformation and Cartography*; Krisp, J.M., Meng, L., Pail, L., Stilla, U., Eds.; Springer: Heidelberg, Germany, 2013; pp. 49–62.
22. Dormann, C.F.; Elith, J.; Bacher, S.; Buchmann, C.; Carl, G.; Carré, G.; Marquéz, J.R.G.; Gruber, B.; Lafourcade, B.; Leitão, P.J.; *et al.* Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecography* **2013**, *36*, 27–46.
23. Schölkopf, B.; Smola, A.J.; Williamson, R.C.; Bartlett, P.L. New support vector algorithms. *Neural Comput.* **2000**, *12*, 1207–1245.
24. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32.
25. Wold, S.; Sjöström, M.; Eriksson, L. PLS-regression: A basic tool of chemometrics. *Chemom. Intell. Lab. Syst.* **2001**, *58*, 109–130.
26. Srivastava, A.N.; Oza, N.C.; Stroeve, J. Virtual sensors: Using data mining techniques to efficiently estimate remote sensing spectra. *IEEE Trans. Geosci. Remote Sens.* **2005**, *43*, 590–599.
27. Walton, J.T. Subpixel urban land cover estimation: Comparing Cubist, Random Forests, and support vector regression. *Photogramm. Eng. Remote Sens.* **2008**, *74*, 1213–1222.
28. Verrelst, J.; Muñoz, J.; Alonso, L.; Delegido, J.; Rivera, J.P.; Camps-Valls, G.; Moreno, J. Machine learning regression algorithms for biophysical parameter retrieval: Opportunities for Sentinel-2 and -3. *Remote Sens. Environ.* **2012**, *118*, 127–139.
29. Bajwa, S.G.; Kulkarni, S.S. Hyperspectral Data Mining. In *Hyperspectral Remote Sensing of Vegetation*; Thenkabail, P.S., Lyon, J.G., Huete, A., Eds.; CRC Press: Boca Raton, FL, USA, 2012; pp. 93–120.
30. Leitão, P.J.; Moreira, F.; Osborne, P.E. Breeding habitat selection of steppe birds in Castro Verde: A remote sensing and advanced statistics approach. *Ardeola* **2010**, *2010*, 93–116.
31. Doorn, A.M.; Pinto Correia, T. Differences in land cover interpretation in landscapes rich in cover gradients: Reflections based on the montado of South Portugal. *Agrofor. Syst.* **2007**, *70*, 169–183.

32. Delgado, A.; Moreira, F. Bird assemblages of an Iberian cereal steppe. *Agric. Ecosyst. Environ.* **2000**, *78*, 65–76.
33. Moreira, F.; Beja, P.; Morgado, R.; Reino, L.; Gordinho, L.; Delgado, A.; Borralho, R. Effects of field management and landscape context on grassland wintering birds in southern Portugal. *Agric. Ecosyst. Environ.* **2005**, *109*, 59–74.
34. Pereira, H.M., Domingos, T., Vicente, L., Eds. *Portugal Millennium Ecosystem Assessment: State of the Assessment Report*; Centro de Biologia Ambiental, Faculdade de Ciências da Universidade de Lisboa, 2004; p. 68. Available online: http://ecossistemas.org (accessed on 17 April 2014).
35. Catry, I.; Franco, A.M.A.; Rocha, P.; Alcazar, R.; Reis, S.; Cordeiro, A.; Ventim, R.; Teodósio, J.; Moreira, F.; Chiaradia, A. Foraging habitat quality constrains effectiveness of artificial nest-site provisioning in reversing population declines in a colonial cavity nester. *PLoS One* **2013**, *8*, 1–10.
36. Moreira, F. Relationships between vegetation structure and breeding bird densities in fallow cereal steppes in Castro Verde, Portugal. *Bird Study* **1999**, *46*, 309–318.
37. Moreira, F.; Leitão, P.J.; Morgado, R.; Alcazar, R.; Cardoso, A.; Carrapato, C.; Delgado, A.; Geraldes, P.; Gordinho, L.; Henriques, I.; *et al.* Spatial distribution patterns, habitat correlates and population estimates of steppe birds in Castro Verde. *Airo* **2007**, 5–30.
38. Marta-Pedroso, C.; Domingos, T.; Freitas, H.; Groot, R.S. Cost-benefit analysis of the zonal program of Castro Verde (Portugal): Highlighting the trade-off between biodiversity and soil conservation. *Soil Tillage Res.* **2007**, *97*, 79–90.
39. Richter, R.; Schläpfer, D. Geo-atmospheric processing of airborne imaging spectrometry data. Part 2: Atmospheric/topographic correction. *Int. J. Remote Sens.* **2002**, *23*, 2631–2649.
40. Richter, R.; Schläpfer, D. *Atmospheric/Topographic Correction for Airborne Imagery, Atcor-4 User Guide*, Version 4.2.; Remote Sensing Data Center, German Aerospace Center (DLR): Oberpfaffenhofen, Germany, 2007. Available online: http://dlr.de/ (accessed on 17 April 2014).
41. Schläpfer, D.; Richter, R. Geo-atmospheric processing of airborne imaging spectrometry data. Part 1: Parametric orthorectification. *Int. J. Remote Sens.* **2002**, *23*, 2609–2630.
42. Segl, K.; Guanter, L.; Rogass, C.; Kuester, T.; Roessner, S.; Kaufmann, H.; Sang, B.; Mogulsky, V.; Hofer, S. Eetes—The EnMAP end-to-end simulation tool. *IEEE J. Select. Top. Appl. Earth Obs. Remote Sens.* **2012**, *5*, 522–530.
43. Olofsson, P.; Foody, G.M.; Stehman, S.V.; Woodcock, C.E. Making better use of accuracy data in land change studies: Estimating accuracy and area and quantifying uncertainty using stratified estimation. *Remote Sens. Environ.* **2013**, *129*, 122–131.
44. R Development Core Team. R: A Language and Environment for Statistical Computing. Available online: http://www.R-project.org (accessed on 17 April 2014).
45. Brereton, R.G.; Lloyd, G.R. Support vector machines for classification and regression. *Analyst* **2010**, *135*, 230–267.
46. Karatzoglou, A.; Meyer, D.; Hornik, K. Support vector machines in R. *J. Stat. Softw.* **2006**, *15*, 2–28.
47. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297.
48. Smola, A.J.; Schölkopf, B. A tutorial on support vector regression. *Stat. Comput.* **2004**, *14*, 199–222.

49. Hornik, K.; Leisch, F.; Meyer, D.; Weingessel, A.; Dimitriadou, E. *E1071: Misc Functions of the Department of Statistics (e1071)*; TU Wien: Vienna, Austria, 2010.
50. Chang, C.-C.; Lin, C.-J. LIBSVM. *ACM Trans. Intell. Syst. Technol.* **2011**, *2*, 1–27.
51. Breiman, L.; Friedman, J.; Olshen, R.; Stone, C. *Classification and Regression Trees*; Wadsworth and Brooks: Belmont, CA, USA, 1984; p. 358.
52. Liaw, A.; Wiener, M. Classifiaction and regression by randomForest. *R News* **2002**, *2*, 18–22.
53. Mevik, B.-H.; Wehrens, R. The PLS package: Principal component and partial least squares regression in R. *J. Stat. Softw.* **2007**, *2*, 1–24.
54. Schmidtlein, S.; Oldenburg, C.; Feilhauer, H. *R-Package AutoPLS: PLS Regression with Backward Selection of Predictors*, Version 1.2-6; 2013. Available online: http://cran.r-project.org/ (accessed on 17 April 2014)
55. Tuia, D.; Verrelst, J.; Alonso, L.; Perez-Cruz, F.; Camps-Valls, G. Multioutput support vector regression for remote sensing biophysical parameter estimation. *IEEE Geosci. Remote Sens. Lett.* **2011**, *8*, 804–808.
56. Leitão, P.J. Improving Species Distribution Models to Describe Steppe Bird Occurrence Patterns and Habitat Selection in Southern Portugal. Ph.D. Thesis, University of Southampton, Southampton, UK, 2008.
57. Oldeland, J.; Dorigo, W.; Wesuls, D.; Jürgens, N. Mapping bush encroaching species by seasonal differences in hyperspectral imagery. *Remote Sens.* **2010**, *2*, 1416–1438.