

Article

A New Application of Random Forest Algorithm to Estimate Coverage of Moss-Dominated Biological Soil Crusts in Semi-Arid Mu Us Sandy Land, China

Xiang Chen ^{1,2,3}, Tao Wang ^{1,2}, Shulin Liu ^{1,*}, Fei Peng ^{1,4} , Atsushi Tsunekawa ³ ,
Wenping Kang ^{1,2}, Zichen Guo ^{1,2} and Kun Feng ^{1,2}

¹ Key Laboratory of Desert and Desertification, Northwest Institute of Eco-Environment and Resources, Chinese Academy of Sciences, Lanzhou 730000, China; chenxx1991@lzb.ac.cn (X.C.); wangtao@lzb.ac.cn (T.W.); pengfei@lzb.ac.cn (F.P.); laiwukang123@163.com (W.K.); guozichen15@lzb.ac.cn (Z.G.); fengkun@lzb.ac.cn (K.F.)

² University of Chinese Academy of Sciences, Beijing 100049, China

³ Arid Land Research Center, Tottori University, Tottori 680-0001, Japan; tsunekawa@alrc.tottori-u.ac.jp

⁴ International Platform for Dryland Research and Education, Tottori University, Tottori 680-0001, Japan

* Correspondence: liusl@lzb.ac.cn; Tel.: +86-0931-496-7573

Received: 5 May 2019; Accepted: 28 May 2019; Published: 30 May 2019



Abstract: Biological soil crusts (BSCs) play an essential role in desert ecosystems. Knowledge of the distribution and disappearance of BSCs is vital for the management of ecosystems and for desertification researches. However, the major remote sensing approaches used to extract BSCs are multispectral indices, which lack accuracy, and hyperspectral indices, which have lower data availability and require a higher computational effort. This study employs random forest (RF) models to optimize the extraction of BSCs using band combinations similar to the two multispectral BSC indices (Crust Index-CI; Biological Soil Crust Index-BSCI), but covering all possible band combinations. Simulated multispectral datasets resampled from in-situ hyperspectral data were used to extract BSC information. Multispectral datasets (Landsat-8 and Sentinel-2 datasets) were then used to detect BSC coverage in Mu Us Sandy Land, located in northern China, where BSCs dominated by moss are widely distributed. The results show that (i) the spectral curves of moss-dominated BSCs are different from those of other typical land surfaces, (ii) the BSC coverage can be predicted using the simulated multispectral data (mean square error (MSE) < 0.01), (iii) Sentinel-2 satellite datasets with CI-based band combinations provided a reliable RF model for detecting moss-dominated BSCs (10-fold validation, $R^2 = 0.947$; ground validation, $R^2 = 0.906$). In conclusion, application of the RF algorithm to the Sentinel-2 dataset can precisely and effectively map BSCs dominated by moss. This new application can be used as a theoretical basis for detecting BSCs in other arid and semi-arid lands within desert ecosystems.

Keywords: moss-dominated biological soil crusts (BSCs); random forest (RF) algorithm; in-situ hyperspectral dataset; multispectral remote sensing; Mu Us Sandy Land

1. Introduction

Biological soil crusts (BSCs) containing microphytic communities (i.e., cyanobacteria, lichens, liverworts, and mosses), grow within or directly on top of soil [1]. BSCs are the primary producers, sinks of carbon and nitrogen, and soil stabilizers, and they mainly exist in arid and semi-arid areas that cover over 35% of global land surfaces [1–4]. BSCs are a top management priority in desertified lands because of their extreme vulnerability to disturbances from human activities and climate change, which have recently been shown to negatively affect such areas [1]. It is essential to obtain accurate

information about the spatial distribution of BSCs and the associated temporal changes, to enable the assessment and protection of such ecosystems [5].

Remote sensing can be used to map BSCs [6], and considerable research has focused on recognizing the particular spectral features of BSCs [7–9]. Furthermore, the differences between spectra relating to BSC, vegetation, and bare soil have been analyzed to enable the effective determination of BSCs [9–12] and to quantitatively predict their relative cover [5]. Based on these efforts, several BSC indices have been developed using optical reflectivity. The Crust Index (CI) [10] and the Biological Soil Crust Index (BSCI) [11] were employed to identify BSCs using multispectral optical information obtained from Landsat Thematic Mapper (TM) and Landsat Enhanced Thematic Mapper Plus (ETM+) images, respectively. Specifically, the CI was proposed to extract BSCs dominated by cyanobacteria, based on the interpretation that phycobilin of cyanobacteria increases reflectivity in the blue band [10]. The BSCI was employed in the slope between the green and red band to extract BSCs dominated by lichen [11]. However, satisfactory results cannot be obtained when applying the CI and BSCI in regions covered by a mixture of photosynthetic and non-photosynthetic vegetation, bare sand, rocks, and BSCs [8,13] because it is difficult to extract the subtle spectral characteristics of BSCs [14]. In addition, there are no BSC indices for detecting moss-dominated BSCs.

Optical hyperspectral-based processing, such as the Continuum Removal Crust Identification Algorithm (CRCIA) [8] and the Crust Development Index (CDI) [9] have been proposed for identifying the subtle spectral features of BSCs. These two approaches highlight the mathematical capability of decision trees in recognizing specific spectral characteristics of BSCs. They also proved that multispectral indices have limitations when used to extract BSCs. Although the performance of hyperspectral remote sensing data is superior to that of multispectral data, hyperspectral data have lower data availability and require higher computational efforts. It is necessary to use multispectral data and develop a method for mapping BSCs that not only captures their subtle spectral features, but also enables them to be mapped conveniently and efficiently.

Thus, the extraction of BSCs is sophisticated and needs to be simplified by selecting extremely subtle and comprehensive spectral characteristics. The random forest (RF) algorithm is an ensemble machine learning technique with data mining capabilities [15], and has been used with feature selection approaches to extract tiny spectral differences [16,17]. The contribution of BSCs to the spectral surface characteristics of the soil depends not only on their existence, but also on their level of coverage [5,6]. In this study, a new application of the ensemble of stochastic regression trees of the RF algorithm [18] is proposed to map the relative coverage of BSCs. The RF algorithm performs well when using high dimensional input variables (multispectral data are high dimensional) and limited training samples (in-situ measurements of BSC coverage are limited) for the output variables [18]. Moreover, RF is a powerful method that can cope with missing observations and an unbalanced dataset [16] (the BSC coverage dataset is an unbalanced dataset).

The present study aims to use multispectral satellite images in the quantification of BSC coverage by applying the RF algorithm to improve the accuracy and efficiency for extracting BSCs. The specific objectives are (i) to observe the spectral differences between BSCs and other typical land surfaces using in-situ hyperspectral data, (ii) to determine whether RF models can be used to obtain the subtle spectral differences required to predict BSC coverage via simulated multispectral data, and (iii) to obtain a reliable RF model for mapping BSCs in Mu Us Sandy Land, by comparing data sources and band combinations.

2. Materials and Methods

2.1. Study Area

This study was conducted in Mu Us Sandy Land in northern China (37°28′–39°49′N, and 106°57′–110°37′E), which covers an area measuring approximately 4×10^4 km² and an elevation ranging from 875 m to 1685 m (Figure 1). Administratively, the area lies within the southern part of Ejin

Horo Banner, the northern part of the Yuyang sandy area of Yulin County, and the northeastern part of Yanchi County. The annual mean temperature ranges from 6.0 °C to 8.5 °C [19]; precipitation occurs mainly in July and September (particularly during August) and this accounts for 60–75% of the annual total precipitation. The potential annual evaporation is 2300 mm, which is six times that of annual precipitation on average. Northwest winds prevail in winter, spring, and autumn, and southeast winds prevail during summer [19]. The soil type is loose aeolian sandy soil, and the land is barren and vulnerable to wind erosion. More than 80% of sandy areas are covered by sandy grassland, and the dominant species is *Artemisia ordosica* [20]. Moss-dominated BSC is widespread within the *A. ordosica* community and is a potent indicator of the fixation phase of sand dunes [20]. Other natural vegetation types, including steppe, meadow, and shrub exist within Mu Us Sandy Land, and farmlands are distributed along the river or scattered in the sandy grasslands, artificial forests, and shrubs [21]. Mu Us Sandy Land is one of the 12 sandy zones in China, but is the only one located in an intermediate region between the typical steppe and the desert. As it belongs to semi-arid continental climate, Mu Us Sandy Land is sensitive to climate change as well as changes in land utilization [21]. In this study, three field campaigns were undertaken in 2017 and 2018 within Mu Us Sandy Land during the late growing season to determine the growth peak of BSCs.

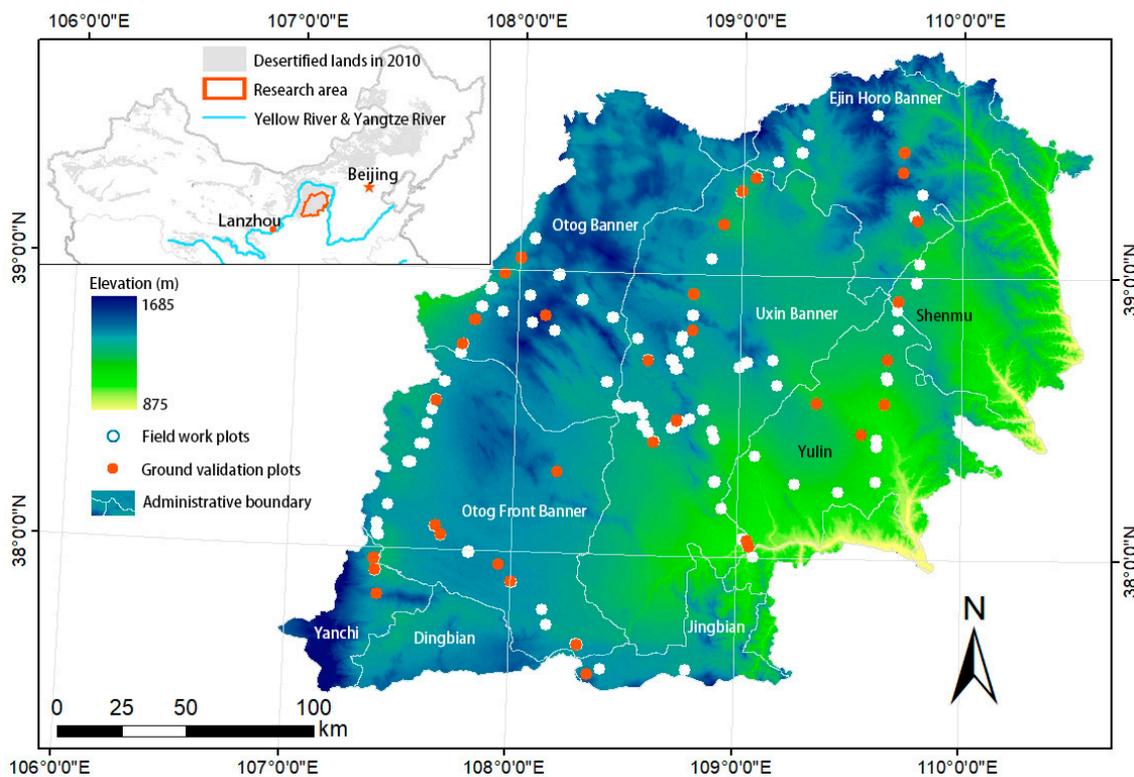


Figure 1. Elevation map of study area and locations of quadrat surveys (white and red points).

2.2. Datasets of Spectra and Coverage of BSCs on a Hoop Scale

2.2.1. In-Situ Hyperspectral Dataset

A field survey was conducted between 28 June and 4 July, 2017. A portable spectrometer (ASD Field Spec Handheld 2) was used to measure in-situ spectral reflectance at 352 points (Figure 2a) on mixed and typical land surfaces such as BSCs, bare sand, and different types of plants. The handheld instrument was used to obtain measurements at wavelength increments of 2 nm between 325 and 1075 nm, with a 15° field of view (FOV), at a height of approximately 1 m above the ground. Spectral measurements were taken under bright and sunny conditions from 10:00 to 15:00 Chinese standard time (UTC+8). Ten spectral curves were measured at each point and the calculated mean values were

taken as the final reflectance spectra. In all, BSC spectra were obtained from 138 plots. Furthermore, spectral data of other ground objects were from 214 plots.

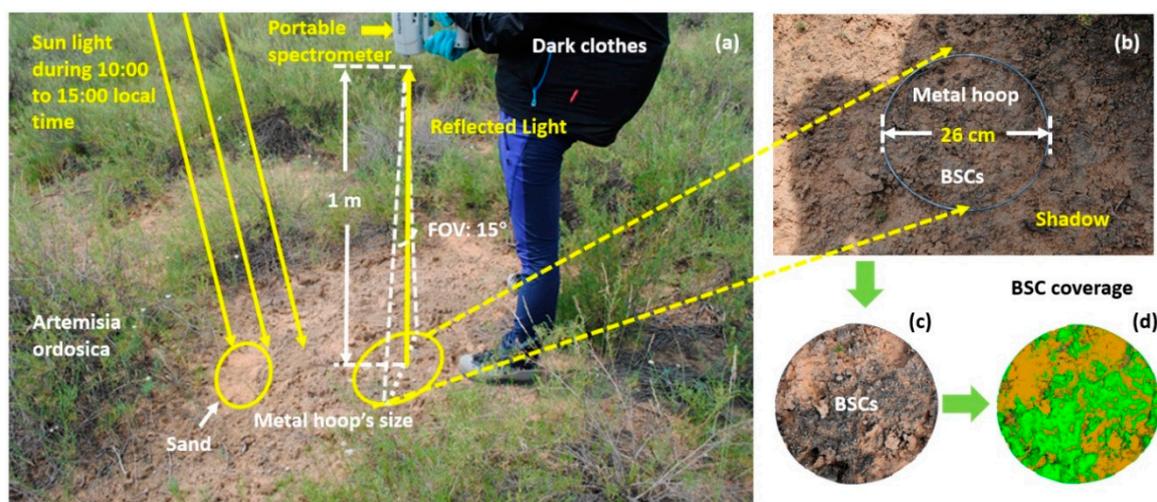


Figure 2. (a) Undertaking ground spectral measurements. (b) Digital photo of a metal hoop in the shade (for digital photo analysis of biological soil crust (BSC) coverage). (c) Clipped photo of a metal hoop. (d) Analysis of BSC cover.

2.2.2. Simulated Multispectral Dataset

To determine the subtle spectral features that could be used to estimate different BSC coverage from multispectral satellite images, in-situ hyperspectral data were firstly resampled using multispectral Landsat-8 and Sentinel-2 channels (see Table 1) by employing the spectral response functions of the respective sensors to simulate the satellite dataset [22,23]. As the distance between the handheld spectrometer and the objects on the ground was short, the BSC coverage was recorded in real-time (by analyzing the instantaneous digital photos (Figure 2c,d; see details in Section 2.2.3)). The simulated dataset was not influenced by issues, such as atmospheric effects or the delay between the field survey and remote sensing data acquisition, which occur in remote sensing images [17,24]. As wavelengths in the 325–400 nm and 900–1075 nm ranges were affected by severe noise, only the 400–900 nm wavelength range was considered in the analysis (Table 1).

Table 1. Overview of spectral bands of Landsat-8 OLI (Operational Land Imager) and Sentinel-2 MSI (Multispectral Instrument) data from 400 to 900 nm.

Landsat-8 OLI			Sentinel-2 MSI		
Band	Range/nm	Resolution/m	Band	Range/nm	Resolution/m
Band1 (Coastal)	430–450	30	Band2 (Blue)	457–523	10
Band2 (Blue)	450–510	30	Band3 (Green)	543–578	10
Band3 (Green)	530–590	30	Band4 (Red)	653–683	10
Band4 (Red)	640–670	30	Band5 (Red Edge)	698–713	20
Band5 (Near-infrared, NIR)	850–880	30	Band6 (Red Edge)	732–748	20
			Band7 (Red Edge)	773–793	20
			Band8A (NIR)	855–875	10

2.2.3. BSC Coverage on a Hoop Scale obtained from Digital Photos

After conducting ground spectral measurements, a metal hoop was constructed with a diameter of 26 cm (Figure 2b) to ensure that each sample would have 100% FOV coverage by the portable spectrometer sensor. Digital photos of BSCs entirely within the hoop were taken (D3000 Digital camera, Nikon, Bangkok, Thailand) in a shadow to avoid the influence of plants' shadow when classifying

BSCs (Figure 2b). Adobe Photoshop CC 2018 and CAN_EYE software programs were used to clip the metal hoop range and extract the BSC coverage (Figure 2c,d).

2.3. Datasets of Spectra and BSC Coverage on a “Pixel Scale”

2.3.1. Satellite Multispectral Dataset

To enable a comparison with the simulated multispectral dataset, bands of satellite images ranging from 400 to 900 nm were considered (Table 1). These geometrically corrected sensor data are available for Landsat-8 as Surface Reflectance (SR) images defined in the Worldwide Reference System (WRS) path/row coordinate system [25,26] and for Sentinel-2 as Level-2A products defined in Bottom-Of-Atmosphere (BOA) granules, also called tiles, which are $100 \times 100 \text{ km}^2$ ortho-images in UTM (Universal Transverse Mercator) WGS84 (World Geodetic System 1984) projection [27]. To validate their applicability to BSC mapping, 5 scenes of Landsat-8 OLI image data and 15 scenes of Sentinel-2 MSI image data covering the entire study area under cloudless conditions were acquired and analyzed (Table 2).

Table 2. Overview of satellite scenes applied in this study.

Landsat-8 OLI		Sentinel-2 MSI	
Path/Row	Acquisition Date (y-m-d)	Tiles	Acquisition Date (y-m-d)
128/33, 128/34	2018-10-04	48SYG	2018-09-29
129/32	2018-10-11	49SBD	2018-10-04
129/33, 129/34	2018-10-27	49TDE	2018-10-06
		48SXH, 48SYH, 49SBB, 49SBC	2018-10-09
		49SDC	2018-10-11
		49SCB, 49SCC, 49SCD, 49SDD, 49TCE	2018-10-26
		48SXG, 48SYG	2018-10-29

The Landsat-8 has a swath of approximately 185 km (15° FOV from a height of 705 km) and offers global area coverage every 16 days [25]. The SR products have a 30 m spatial resolution and are produced at the Earth Resources Observation and Science (EROS) Center using the Landsat Surface Reflectance Code (LaSRC) [22].

The Sentinel-2 swath is approximately 290 km (20.6° FOV from a height of 786 km) and offers global area coverage every 10 days [23,27]. The Sen2Cor tool serves to generate and format the Sentinel-2 Level 2A product to correct for the atmosphere, terrain, and cirrus from Top-Of-Atmosphere (TOA) Level 1C input data [28]. The Level 2A BOA product includes three different resolutions of 60, 20, and 10 m [27,28]. The resolutions of Sentinel-2 dataset used were converted to 20 m.

2.3.2. BSC Coverage on Pixel Scale obtained from Quadrat Survey Data

During the second and third fieldwork conducted on 16–28 August and 23–29 October, 2018, the species composition of plants and general BSC coverage were obtained from a total of 319 surveyed units (Figure 1). A rope square measuring $30 \text{ m} \times 30 \text{ m}$ was arranged in the field within each survey unit (based on the spatial resolution of the Landsat-8 image is 30 m) to simulate one pixel of remote sensing data. To avoid corresponding problems between the image pixel and ground pixel, the survey pixel located the middle of the area with uniform and consistent landscape in a scale of approximately 500×500 area was chosen. To calculate the area covered by BSCs, a $1 \text{ m} \times 1 \text{ m}$ wire square was employed on foot within each rope square [29]. Quadrat surveys were assisted by an unmanned aerial vehicle (UAV), and the UAV was flown at a height of 8 m to obtain a detailed coverage of BSCs (Figure 3). Furthermore, the coordinates of each survey unit were recorded using a GNSS receiver (Garmin GPS 72, $\pm 15 \text{ m}$) to validate the BSC distribution maps via remote sensing data (Figure 1).

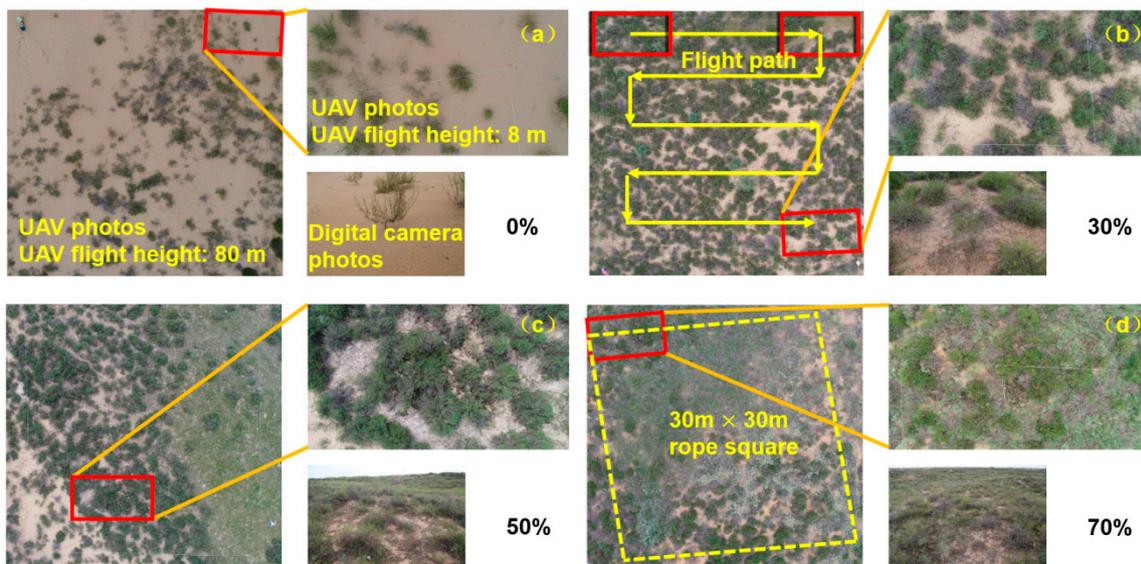


Figure 3. Quadrat survey units of (a) 0%, (b) 30%, (c) 30%, (d) 70% BSC coverage.

2.4. Methods

2.4.1. Band Combinations from BSC Indices

To capture the subtle spectral features of BSCs, reflectance feature spaces composed of multispectral indices similar to the CI [10] and BSCI [11], but covering all possible combinations of bands, were calculated to build models using the following:

$$CI = 1 - (R_{red} - R_{blue}) / (R_{red} + R_{blue}) \tag{1}$$

$$BSCI = (1 - L \times |R_{red} - R_{green}|) / R_{GRNIR}^{mean} \tag{2}$$

$$CI_{\lambda_i \lambda_j} = 1 - (R_{\lambda_i} - R_{\lambda_j}) / (R_{\lambda_i} + R_{\lambda_j}) \tag{3}$$

$$BSCI_{\lambda_i \lambda_j \lambda_k} = (1 - L \times |R_{\lambda_i} - R_{\lambda_j}|) / R_{\lambda_i \lambda_j \lambda_k}^{mean} \tag{4}$$

where R is the reflectance spectra at a wavelength of λ_i , λ_j , and λ_k ; and $R_{\lambda_i \lambda_j \lambda_k}^{mean}$ is the mean reflectance of R_{λ_i} , R_{λ_j} , and R_{λ_k} . To amplify the absolute difference between R_{λ_i} and R_{λ_j} , L was set to 2 as an adjustment parameter, based on observations of Chen et al. [11]. Different bands were recombined to generate both the CI-based band combinations (for Landsat data, there are $C_5^2 = 10$ conditions; for Sentinel data, there are $C_7^2 = 21$ conditions), and the BSCI-based band combinations (for Landsat data, there are $C_5^3 = 10$ conditions; for Sentinel data, there are $C_7^3 = 35$ conditions).

2.4.2. Random Forest (RF) Regression Models

The RF models are generated from an association between the bagging method and randomized subspace method [15]. Every decision tree grows until it reaches a predefined minimum node (nodesize) via a random feature selection in the training dataset. In this study, the number of trees (ntree) was set to 10,000. The size of the variable’s subset (mtry) and the nodesize was set to 5 [16]. The optimization of the parameters was conducted using the randomForest package based on R Version 3.5 [15]. The randomForest() function was used to set ntree. The tuneRF() function was used to set mtry. The treesize() function was used to set nodesize. RF is good at measuring the importance of variables (i.e., the importance of every variable to the performance of a model) [18], which can assist in ranking the useful spectral band combinations from the multispectral data employed to estimate BSC coverage. The first measure of variable importance is calculated from permuting out-of-bag (OOB)

data (%IncMSE), and the mean square error (MSE) for each tree on the OOB portion of the dataset is computed. After permuting each predictor variable, the MSE is computed again, and the mean value (the difference between the two MSEs among all the trees) is then calculated, and normalized by the standard deviation of the differences [18]. The second measure of variable importance is the residual sum of squares for the regression of the total decrease in node impurities from splitting the variables, and this measure is also averaged over all trees (IncNodePurity) [18]. For different combinations of bands based on CI and BSCI indices, RF is iteratively fitted, that is, at each iteration, new forests are developed in the model one after another (starting with the most important ones) [30]. The rfcv function in the R package randomForest is used to show the cross-validated prediction performance (error.cv: corresponding vector of MSEs at each step) of models with a descending number of predictors (n.var: ranked by variable importance) based on a nested cross-validation procedure [31]. Therefore, the nested subset of the combination of bands in the IncNodePurity ranking that had the lowest error rate was used as the optimal band combinations for predicting the coverage of BSCs.

2.4.3. Accuracy Assessment

Four evaluation parameters appropriate for the continuous model were selected: Coefficient of Determinant (R^2), Mean Absolute Error (MAE), Mean Square Error (MSE) and Normalized Mean Square Error (NMSE) [31], as follows:

$$R^2 = \frac{\sum_i (\text{predicted} - \text{mean}(\text{observed}))^2}{\sum_i (\text{observed} - \text{mean}(\text{observed}))^2} \quad (5)$$

$$\text{MAE} = \frac{\sum_{i=1}^n |\text{predicted} - \text{observed}|}{n} \quad (6)$$

$$\text{MSE} = \frac{\sum_{t=1}^n (\text{observed}_t - \text{predicted}_t)^2}{n} \quad (7)$$

$$\text{NMSE} = \frac{\text{mean}((\text{predicted} - \text{observed})^2)}{\text{mean}((\text{mean}(\text{observed}) - \text{observed})^2)} \quad (8)$$

For models on a hoop scale, a 10-fold cross-validation was chosen, as it is one of the most preferred techniques used to evaluate models and is acknowledged to be better than the use of residually based metrics [31,32].

For models on a pixel scale, the 319 quadrat survey plots previously mentioned were separated into two parts: 269 plots were randomly chosen as training data for calibrating the model by 10-fold cross-validation (white points in Figure 1) and the remaining 50 plots were selected as testing data for ground validation (red points in Figure 1).

The methods described above were conducted using R Version 3.5 (Figure 4). The band combination steps using BSCs indices were implemented using the R package, hsdar [24,33]. The R packages, caret [34] and randomForest [18], was used for training and testing the RF model, and the rgdal package [35] was used for processing geospatial data.

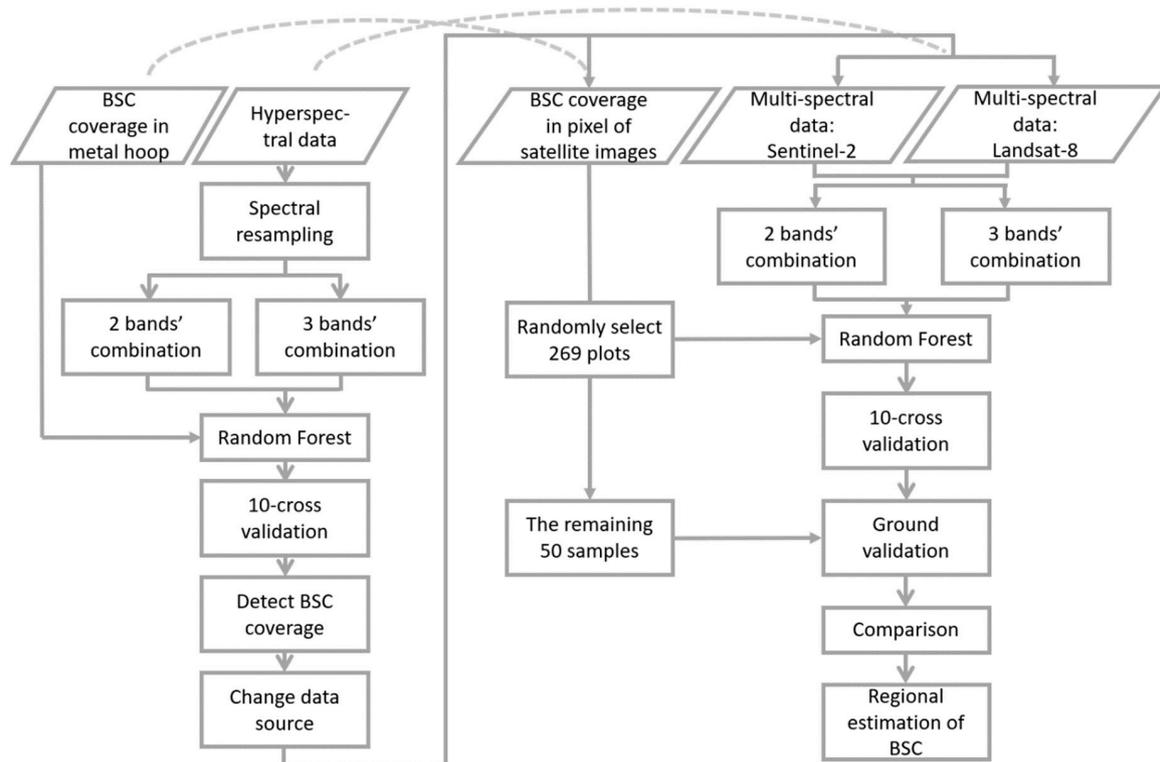


Figure 4. Flow chart of the study process.

3. Results

3.1. BSC Reflectance Features from In-Situ Spectral Measurements

The reflectance curves revealed the spectral features of BSCs, bare sand, and plants (Figure 5). The reflectance of plants showed distinctive features with a maximum value in the green band, an absorption maximum in the red band, and a notable soar from 700 to 800 nm (the vegetation red edge band). All plant species generally had the lowest values of reflectance in the blue band and an absorption of approximately 500 nm. Spectral reflectance values of physical crusts closely resembled those of bare sand, with an intersect occurring around the visible red band. The spectral curves of varying BSC coverage showed intermediate reflectance values compared with other ground objects throughout the visible spectrum. BSCs also showed absorption in both the blue and red bands, but this was much weaker than plants.

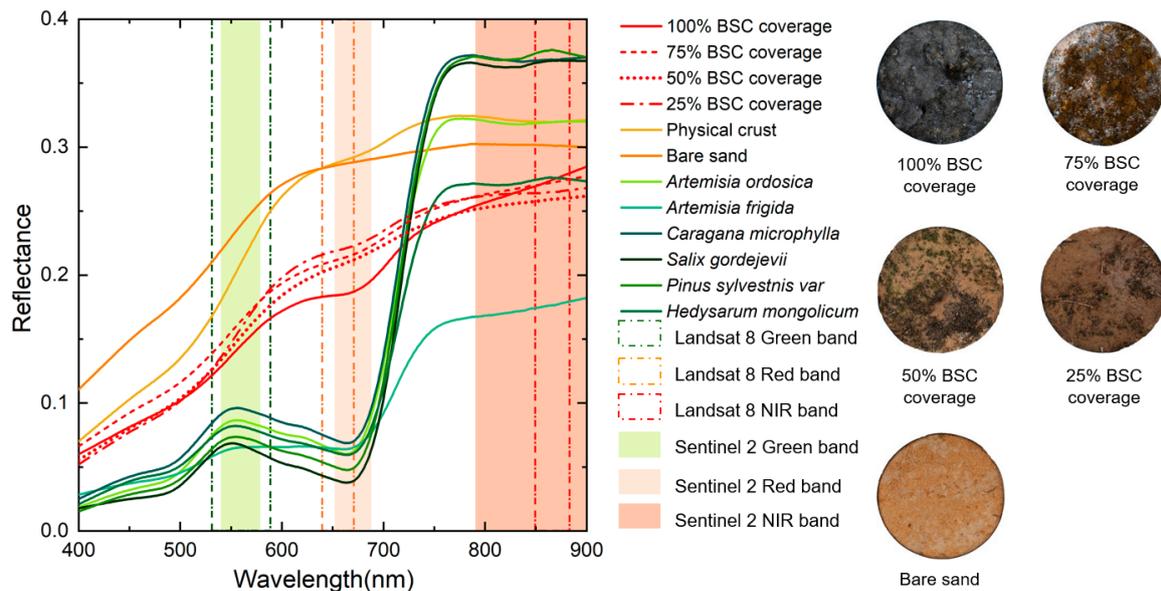


Figure 5. Reflectance spectra of different ground objects and different BSC coverage. Yellow lines are physical crust and bare sand. Green lines are vegetations. Red lines are different coverage of BSCs.

3.2. Implementation of RF Model with the Simulated Multispectral Dataset

The band combinations were ranked according to their importance in estimating BSC coverage as determined by the RF algorithm (Figure 6e–h). The simulated Landsat dataset, which used combinations of bands based on CI, had the lowest error rate when the top five important band combinations were employed (Figure 6a). The green-red bands were the most important band combination (Figure 6e). The red-NIR combination of bands, which represent the band combination of classical vegetation index (the Normalized Difference Vegetation Index (NDVI)), was ranked in second place. The RF model, using the Sentinel dataset with band combinations based on BSCI, achieved the best result among the top 18 important band combinations (Figure 6d), of which the blue-green-red bands, blue-red-red edge bands, and blue-green-red edge bands were the top three important performing combinations (Figure 6h). The best results for other models were obtained when all combinations of bands were selected (Figure 6b,c). The sensitive bands of the top three BSCI-based band combinations, which used the simulated Landsat data, were bands 1 to 4, without the NIR band (Figure 6f). Similarly, the NIR band did not rank among the top three most important band combinations when using both the CI- and BSCI-based formula with the simulated Sentinel models (Figure 6g,h). In contrast, the blue, green, red, and vegetation red edge bands were found to be important for predicting BSC coverage using the simulated Sentinel-2 dataset.

Figure 7 shows the corresponding 10-fold cross-validation results of estimated coverage versus measured coverage of BSCs on a hoop scale. The models using simulated multispectral data for predicting BSC coverage exhibited high performance of $R^2 > 0.950$ and $MSE < 0.010$. There were no significant differences between different datasets with different band combinations with respect to their ability to estimate BSC coverage.

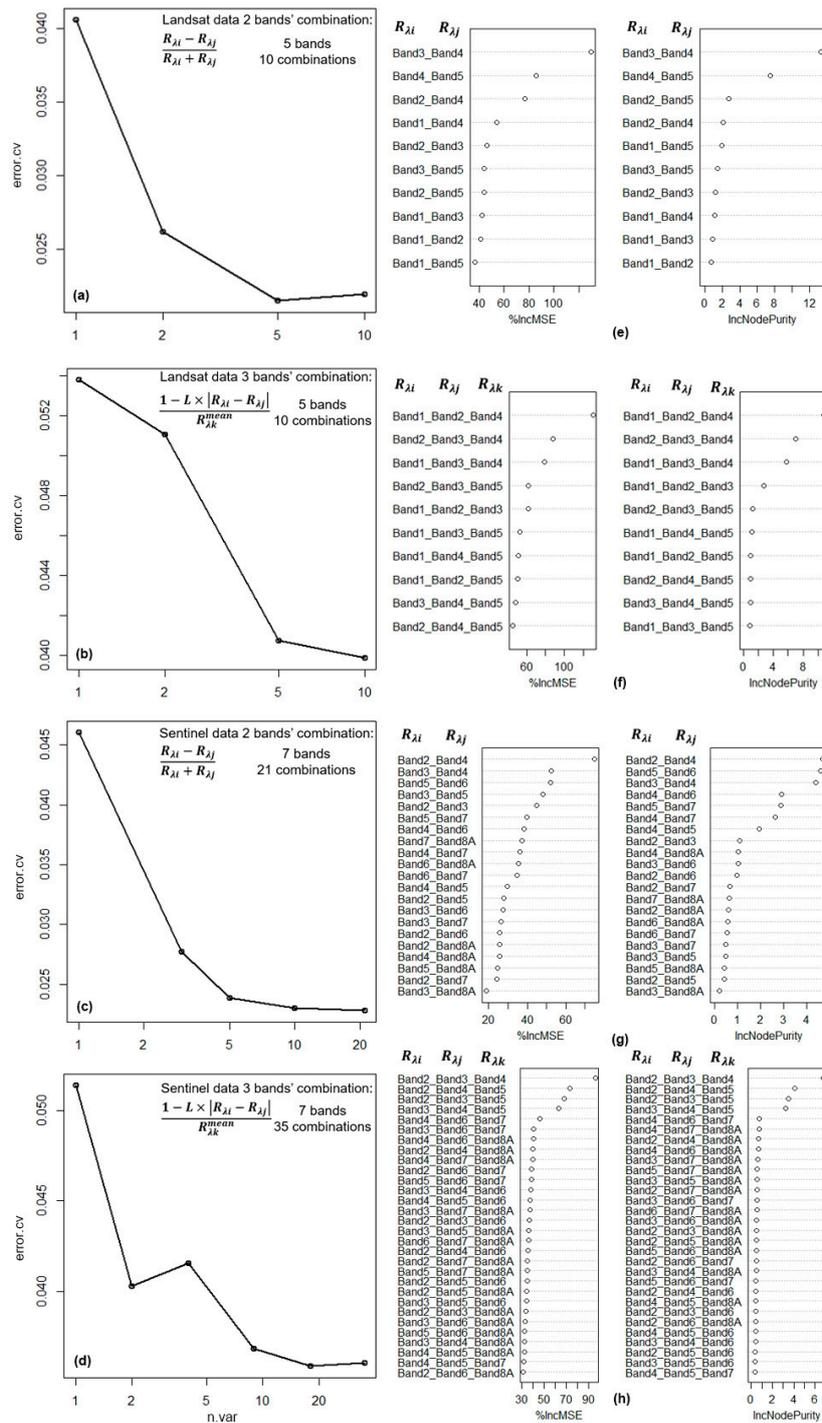


Figure 6. Variable selection of random forest (RF) models for the simulated multispectral data of (a,e) Landsat channel with band combinations based on Crust Index (CI); (b,f) Landsat channel with band combinations based on Biological Soil Crust Index (BSCI); (c,g) Sentinel channel with band combinations based on CI; (d,h) Sentinel channel with band combinations based on BSCI. n.var is the vector of number of variables used in each step; “error.cv” is the corresponding vector of mean square errors (MSEs) in each step; %IncMSE is the standard deviation of the difference between MSE for each tree and MSE after permuting each predictor variable and then averaging over all trees; IncNodePurity is the residual sum of squares from the total decrease in node impurities from splitting of variable and then averaged over all trees. BandX_BandY means Band X combined with BandY (see more details of Band number in Table 1).

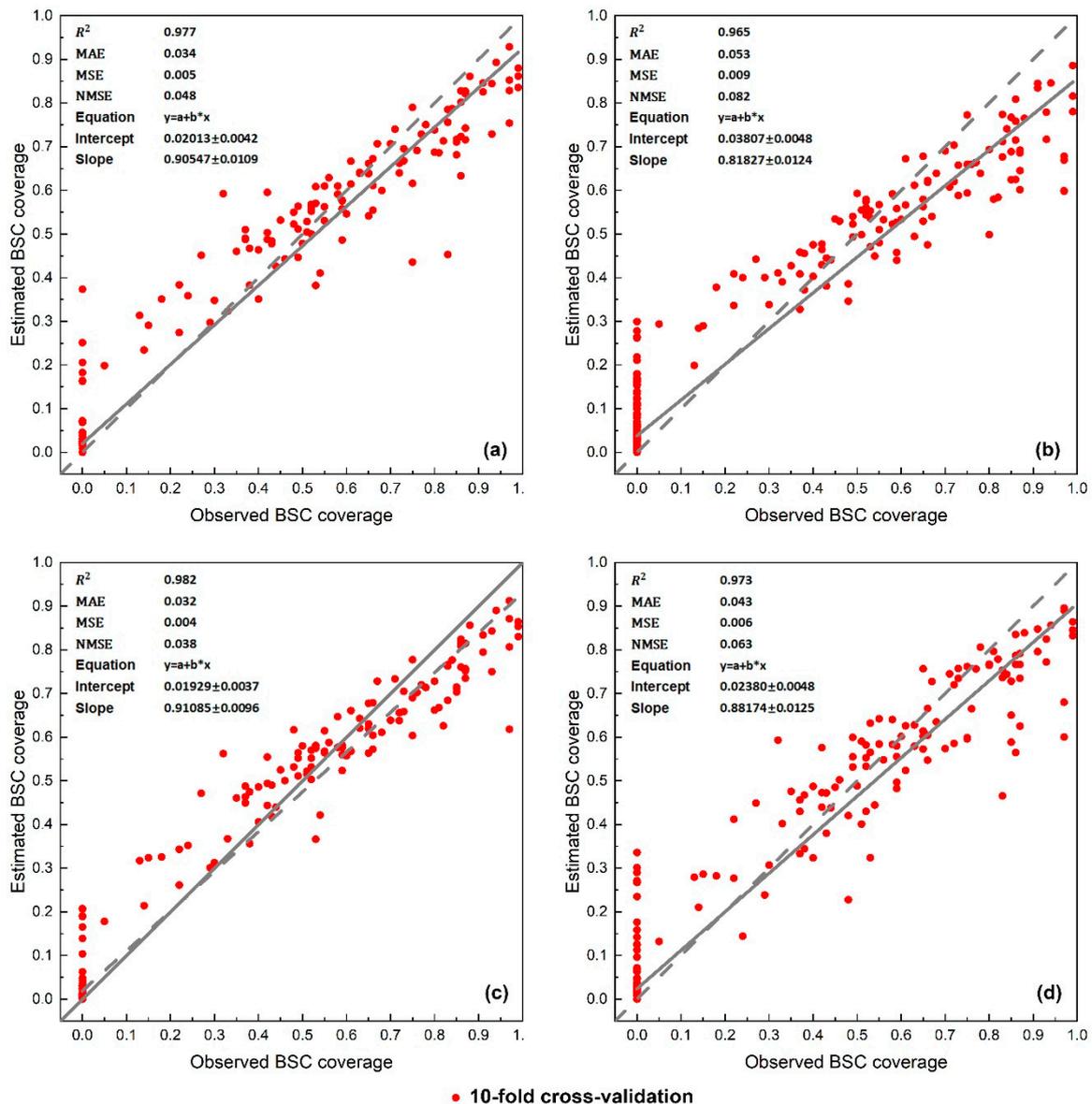


Figure 7. Plots of estimated coverage versus measured coverage of BSCs on hoop scale by: (a) Landsat data with band combinations based on CI; (b) Landsat data with band combinations based on BSCI; (c) Sentinel data with band combinations based on CI; (d) Sentinel data with band combinations based on BSCI. Solid line plot function is $y = x$. Dashed lines are linear fittings of predicted BSC coverage. Red dots represent results from 10-fold cross-validation of models.

3.3. Quantification of BSC Surface Cover in Mu Us Sandy Land

Compared to the simulated multispectral data, there was a decrease in the model performance with satellite images (Figure 7 vs. Figure 8), and the performance of 10-fold cross-validation (shown as red points in Figure 8) dropped on average, but did not significantly decrease ($R^2 = 0.974$ vs. $R^2 = 0.944$).

For ground validation (shown as gray triangles in Figure 8), BSC coverage was predicted with R^2 equal to 0.557 (CI-based band combination) and 0.588 (BSCI-based band combination) using Landsat-8 images. These performances were significantly inferior to those of models using the simulated Landsat dataset (Figure 7a vs. Figure 8a, Figure 7b vs. Figure 8b). The BSC coverage was predicted with $R^2 = 0.906$ (CI-based band combination) and $R^2 = 0.899$ (BSCI-based band combination) using Sentinel-2 scenes (Figure 8c,d). The ground validation of models showed high performance for pixels

with low BSCs coverage, whereas pixels with high BSC coverage (observed BSCs coverage > 0.125) were underestimated by models.

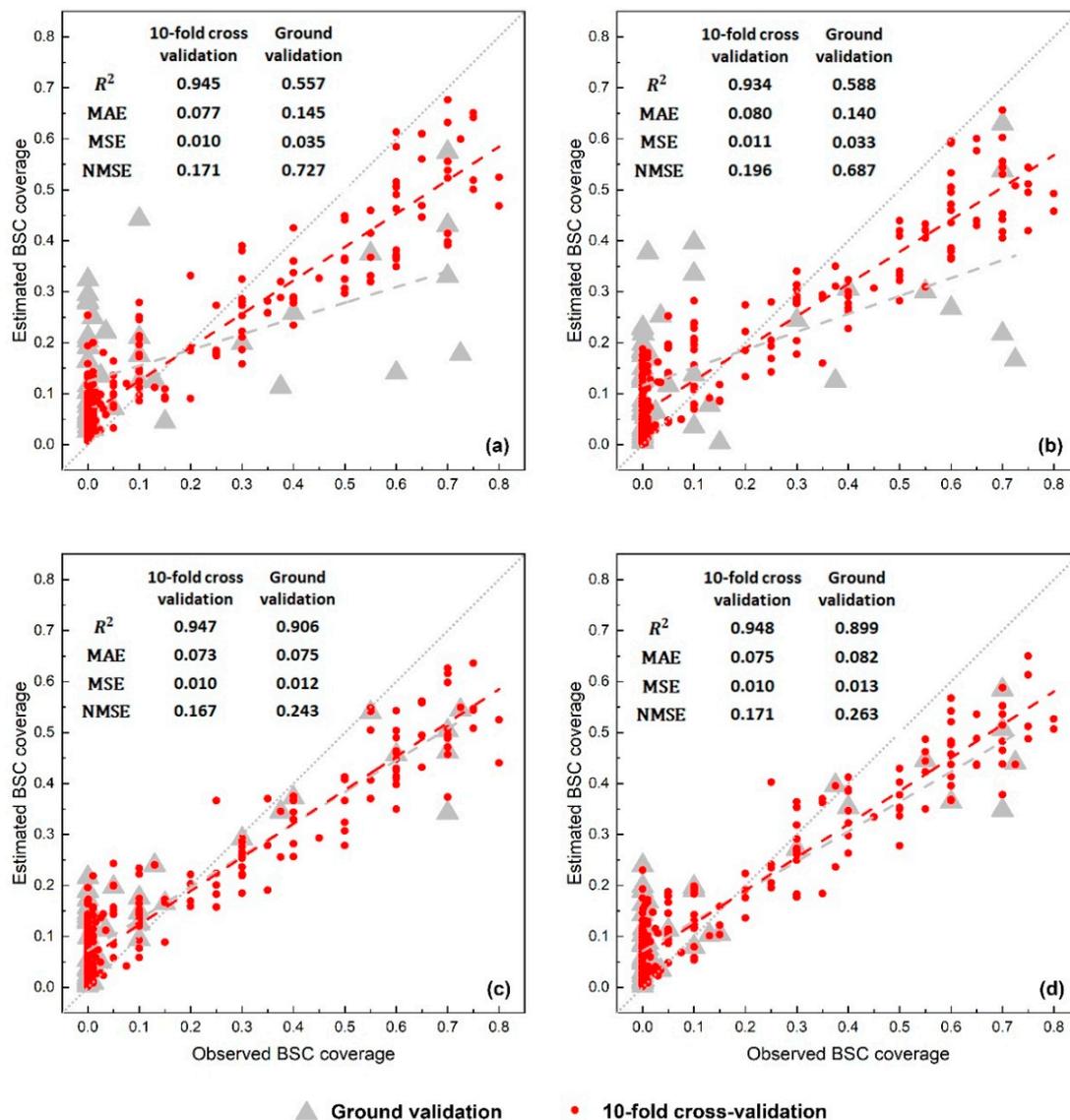


Figure 8. Plots of estimated coverage versus measured coverage of BSCs on a pixel scale by: (a) Landsat data with band combinations based on CI; (b) Landsat data with band combinations based on BSCI; (c) Sentinel data with band combinations based on CI; (d) Sentinel data with band combinations based on BSCI. Dot line plot function is $y = x$. Red dashed lines are linear fittings of 10-fold cross-validation; grey dashed lines are linear fittings of ground validation; red points represent results from 10-fold cross-validation of models; gray triangles represent results of ground validation.

The general BSC distribution was roughly similar for the four models (Figure 9). All models showed BSC widely distributed in the northeastern and southeastern corner of Mu Us Sandy land and sparse distribution on sand dunes in the southwest of Otog Front Banner (administrative regions marked in Figure 1) (Figure 9a–d). In Ejin Horo Banner, Jingbian County, Dingbian County, and Yanchi County, the models using Landsat data with band combinations based on CI showed the highest distribution of BSC coverage (Figure 9a). There was a distinct gradient distribution of BSCs from south to north in Uxin Banner from the Sentinel data models using CI-based band combinations (Figure 9c).

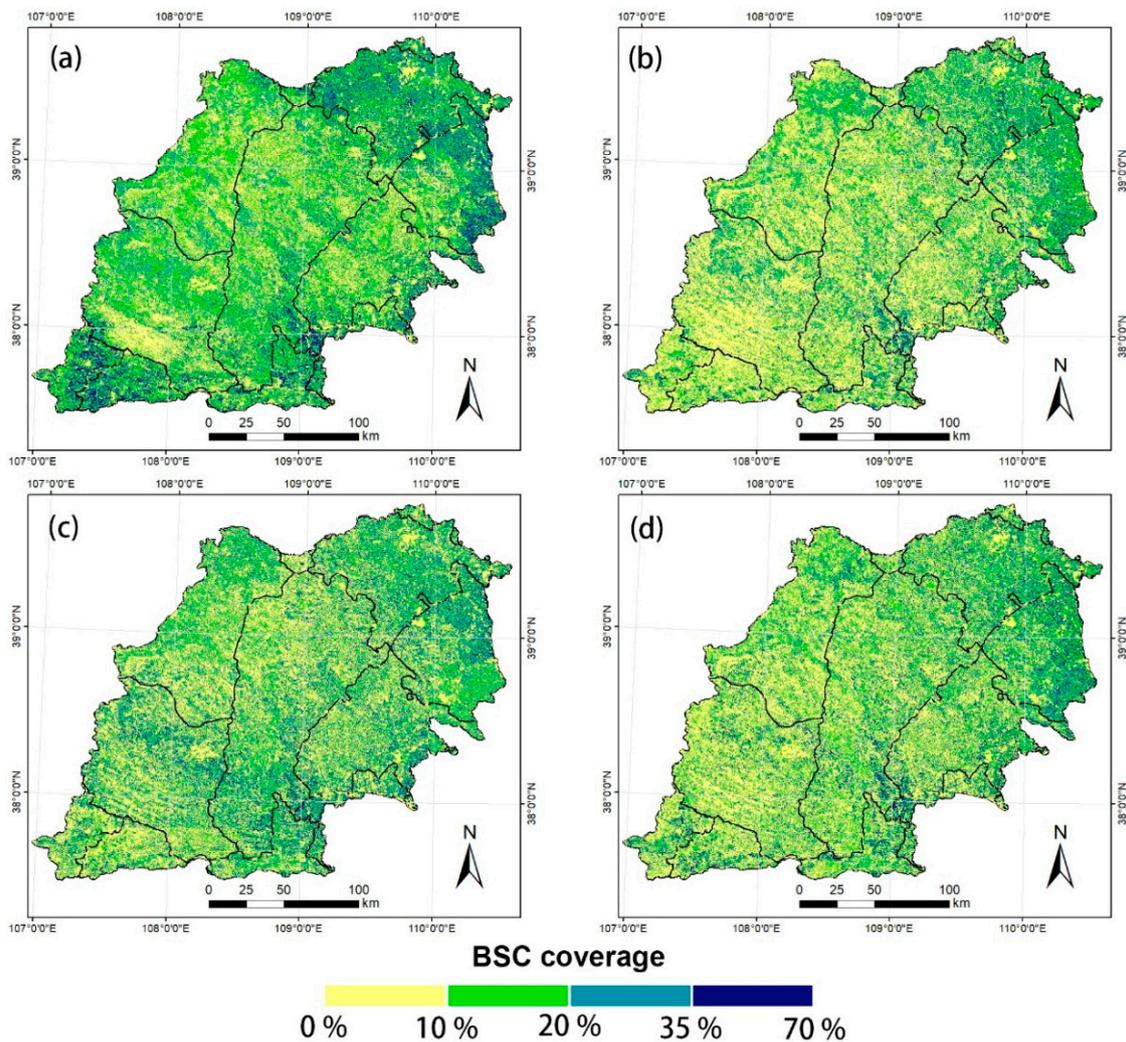


Figure 9. Distribution of BSC coverage in Mu Us Sandy Land by: (a) Landsat data with band combinations based on CI; (b) Landsat data with band combinations based on BSCI; (c) Sentinel data with band combinations based on CI; (d) Sentinel data with band combinations based on BSCI.

4. Discussion

4.1. Reflectance Features of BSCs

Two primary absorption characteristics of BSCs of approximately 520 and 680 nm, which have been described in previous studies, are believed to be related to the existence of carotenoids and chlorophyll a, respectively [8,9,36]. Our study also detected these absorptions, which were relatively weak compared to those of plants (Figure 5). The increase in BSC reflectance within the green band can distinguish plants from BSCs. Moreover, bare sand showed no absorption around 680 nm, whereas BSC showed weak absorption in the red band. This can be used to differentiate between the BSCs and the bare sand. However, when the pixel filled only with plants and soil, without BSCs, the model might confuse the existence of BSCs and recognize plants as BSCs.

Nevertheless, indices calculated using single band combination such as CI [10], may not enable the precise detection of BSCs. Linear regressions between the band combination based on CI and the coverage of BSCs showed that the most sensitive band combination was band 4 (red) with band 5 (NIR) for both the simulated Landsat-8 dataset (Figure 10a) and actual Landsat-8 images (Figure 10c). This red-NIR combination is also recognized as the band combination of the NDVI. For Sentinel-2 channel, Band 7 (vegetation red edge) with band 8A (NIR) and band 6 (vegetation red edge) with band

8A (NIR) were the most sensitive band combination in the simulated dataset (Figure 10b) and real images (Figure 10d), respectively. It appears that the vegetation red edge with the NIR band is sensitive to BSCs in Sentinel-2 channels. However, all these combinations are also sensitive to vegetation. Our research found that a single pair of bands is unsuitable for spectrally discriminating BSCs, due to the strong spectral characteristics of vegetation, and thus we investigated the deeper detection of sensitive spectral information using the machine learning method.

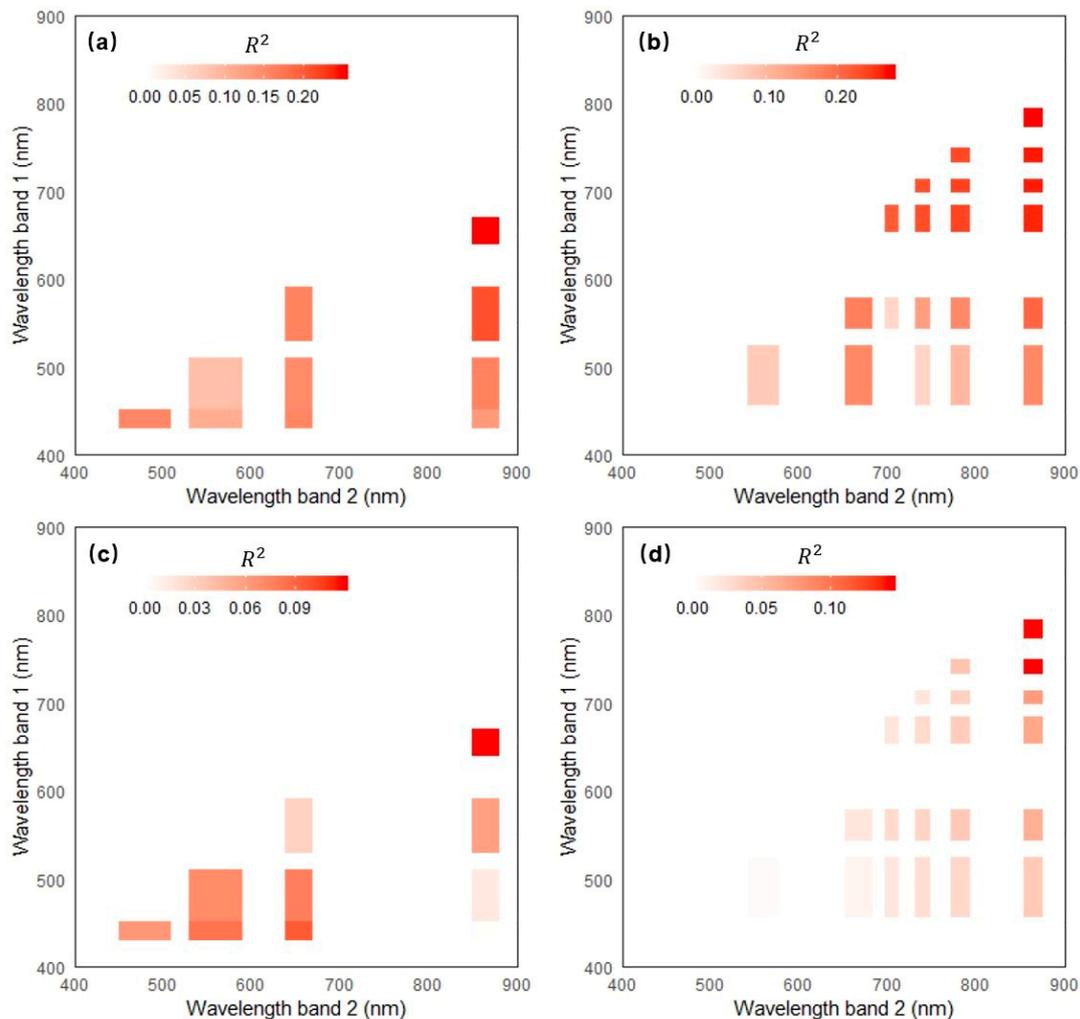


Figure 10. Correlograms depicting R^2 values (red color) of linear regressions between CI-based band combinations computed by the value of reflectance at x- and y-axes and coverage of BSCs using: (a) simulated multispectral data of Landsat-8 channel; (b) simulated multispectral data of Sentinel-2 channel; (c) Landsat-8 images; and (d) Sentinel-2 images.

4.2. Implementation of RF Model with the Simulated Multispectral Dataset

Karnieli [10] found that BSCs on a bare substrate have higher values of reflectance in the blue band than bare soil. However, this spectral feature was not determined in our study (Figure 5), nor in the results of Chen et al. [11] and Weber et al. [8]. Weber et al. [8] believed that spectral information of minerals with strong absorption features in the blue band, may interfere with spectral measurements in the field. The training plots investigated in the field were mixed by BSCs, plants and soil, because this is the actual landscape. However, as with the combination of red and blue bands employed by CI [10], our models also found that the combination of red and blue bands is important in CI-based RF models (Figure 6a,c). Another two important CI-based band combinations in our models were the green-red bands and the red-NIR bands (Figure 6a,c) where the combination of green-red bands

separated BSCs from plants, and the combination of red-NIR bands (which is the band combination of the NDVI) separated BSCs from bare sand. Although the BSCI proposed by Chen et al. [11] included the NIR band, the NIR band was not included in the top four most sensitive band combinations in our models using the BSCI-based formula, whereas the frequency of the NIR band that appeared in our models was the highest (Figure 6b,d). These results indicate that important band combinations used in our models are able to detect BSCs information and are superior to the multispectral BSC indices that combine only two or three bands.

The high capability of the RF algorithm in detecting BSC information from multispectral channels is seen in Figure 7. Since Pirotti et al. [37] have proved the best performance of random forest for classification, the standard deviation (SD) of Mean Square Errors (MSEs) was calculated in each fold validation of 10-fold validation (Table 3). The SDs are all < 0.01. This further proved the powerful ability of the RF algorithm for regression. The hyperspectral dataset requires high computational efforts and has difficulties in large-scale data acquisition. Therefore, use of the RF model with multispectral remote sensing data to detect BSCs is more convenient and efficient than using hyperspectral datasets such as studies of Weber et al. [8], Chamizo et al. [9], and Rodríguez-Caballero et al. [5]. This has already been proven while estimating BSCs using the CI [10] or BSCI [11]. In addition, the use of regression models in our study enabled extraction of BSCs and the estimation of quantitative BSC coverage.

Table 3. Standard deviation (SD) and Mean Square Error (MSE) in each fold validation of 10-fold cross validation.

10-Fold Cross Validation on “Hoop Scale”				10-Fold Cross Validation on “Pixel Scale”			
Dataset	Band Combination	SD	MSE	Dataset	Band Combination	SD	MSE
Landsat-8	CI	0.0074	0.005	Landsat-8	CI	0.0047	0.010
Landsat-8	BSCI	0.0035	0.009	Landsat-8	BSCI	0.0031	0.011
Sentinel-2	CI	0.0044	0.004	Sentinel-2	CI	0.0056	0.010
Sentinel-2	BSCI	0.0012	0.006	Sentinel-2	BSCI	0.0025	0.010

4.3. Quantification of BSC Surface Cover in Mu Us Sandy Land

The models trained by the simulated multispectral dataset on a hoop scale for both Landsat and Sentinel channels showed high performance (Figure 7). However, the performance reduced when the models were applied to remote sensing data (Figure 8), which could be attributed to the time gaps between the ground survey and acquisition of satellite images [17]. The model on a hoop scale provided high performance as no time gaps exists, because it was built using the data of the simulated multispectral data (resampled by in-situ hyperspectral data) for the precise plot where BSC coverage had been determined (by analyzing the instantaneous digital photos). The issues with time gap problems for models on a pixel scale posed no problem with our research, but they pose a difficulty for all research applying ground surveys to remote sensing dataset.

Our results emphasize that the RF algorithm provides the highest estimation of BSC coverage when using Sentinel-2 satellite sensors. Models using the direct Sentinel-2 dataset performed almost as well as the simulated Sentinel-2 dataset. It seems the multispectral Sentinel-2 sensors provide a better spatial resolution than Landsat-8 sensors. Landsat-8 had a lower spatial resolution and its performance in ground validation was inferior (Figure 8a,b). In addition, higher spectral resolution and band setting of Sentinel-2 (three vegetation red edge bands) might be one of the reasons for the superior performance to estimate BSC coverage. Future satellite missions may offer better-suited data sources to enable BSC mapping.

Furthermore, our best result of BSC distribution (Figure 9c) was generally matched with the results of aboveground biomass (AGB) distribution [38], vegetation coverage distribution [39], and sand dune distribution [40] in Mu Us Sandy Land. Moreover, some researchers believe that moss-dominated BSCs have a positive correlation with perennial plant coverage and soil organic matter [41,42]. Our method overestimated BSCs under the coverage of 30% (Figures 7 and 8). The main reasons might be the

influence of vegetation and unbalanced training datasets that were used. As we discussed in Section 4.1, when the pixel mixed only by plants and soil, without BSCs, the model might recognize plants as BSCs. One of the most common landscapes in Mu Us Sandy Land, however, is sparsely distributed with some vascular plants without BSCs. Therefore, the training datasets including plants without BSCs may lead to overestimating results of BSCs. Our future work could focus on the spatial analysis of the relationships between BSC, vegetation, and bare sand to quantify the impact of plants and soil on the determination of BSCs. The seasonality and different dry-wet conditions of BSCs have been studied earlier [43,44], which highlight the seasonal changes of BSCs in arid and semi-arid land. The BSC coverage predicted by our research provides a snapshot of Mu Us Sandy Land at the end of the growing season in 2018. However, our study did not make an assessment of trends and phenological changes in the region. BSCs are poikilohydric plants that lie dormant when dry [44]. Soil moisture and precipitation can increase the CI, BSCI, and NDVI of moss-dominated BSC [43]. The training data on a pixel scale, which was investigated on the ground, sometimes were collected when it was raining. However, it is difficult to obtain satellite images with good quality during the rainy seasons. These artificial errors might be one of the reasons for underestimations of high BSC coverage in our models. Further research, consequently, needs to focus on improving the BSC monitoring by considering the season, weather, and soil moisture content. Finally, it is recommended to unify season and weather conditions while collecting spectral information or BSC coverage to build BSC RF models.

5. Conclusions

In this study, a new application of RF was proposed to quantitatively detect moss-dominated BSCs. This application not only can attain more accurate results than multispectral indices, but also is more efficient than hyperspectral methods. A spectral analysis of the main ground objects in Mu Us Sandy Land was initially conducted, which provided sufficient information to distinguish moss-dominated BSCs. However, using a simple band combination proved difficult in discriminating between plants and BSCs. Thus, we implemented the RF algorithm to analyze the simulated multispectral dataset, which provided promising results. The Sentinel-2 dataset was shown to be suitable for use in training reliable RF models that can predict BSC coverage using band combinations based on the CI. The ultimate aim of this study was to derive regional scale maps of BSC in Mu Us Sandy Land, which are urgently required to obtain accurate spatial information relating to desertification. Such applications are essential for local people and politicians in maintaining ecosystem services, and the methods used in this study can help map BSC coverage in other arid and semi-arid areas.

Author Contributions: Conceptualization, T.W. and S.L.; Validation, X.C., Z.G. and K.F.; Formal Analysis, X.C.; Investigation, X.C., W.K., Z.G. and K.F. and S.L.; Resources, F.P.; Data Curation, X.C.; Writing—Original Draft Preparation, X.C.; Writing—Review & Editing, X.C., A.T., S.L., F.P. and W.K.; Visualization, X.C.; Supervision, T.W. and A.T.; Project Administration, T.W., A.T. and S.L.; Funding Acquisition, T.W. and A.T.

Funding: This research was funded by Project of National Key Research and Development Program of China, grant number 2016YFC0500902; the China Scholarship Council, grant number 201704910491; and the Joint Research Program of Arid Land Research Center, Tottori University.

Acknowledgments: The authors thank Che Tao for valuable academic suggestions in our study, and Qian Tana, Gou Xiaowei, Wu Jing, Gao Yang for their technical assistance in part of our study. The authors also thank Liu Wenli for English corrections. The first author expresses special thanks to Liu Jia for help while studying in Japan.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Belnap, J.; Lange, O.L. *Biological Soil Crusts: Structure, Function, and Management*, 2nd ed.; Springer: Berlin, Germany, 2003; pp. 401–471.
2. Evans, R.D.; Lange, O.L. Biological Soil Crusts and Ecosystem Nitrogen and Carbon Dynamics. In *Biological Soil Crusts: Structure, Function, and Management*; Baldwin, I.T., Caldwell, M.M., Eds.; Springer: Berlin, Germany, 2001; Volume 150, pp. 263–280.

3. Belnap, J.; Gillette, D.A. Disturbance of biological soil crusts: Impacts on potential wind erodibility of sand desert soils in Southeastern Utah. *Land Degrad. Dev.* **1997**, *8*, 355–362.
4. Li, X. Biological soil crust as a bio-mediator alters hydrological processes in stabilized dune system of the Tengger Desert, China. In Proceedings of the EGU General Assembly Conference Abstracts, Vienna, Austria, 17–22 April 2016.
5. Rodríguez-Caballero, E.; Escribano, P.; Cantón, Y. Advanced image processing methods as a tool to map and quantify different types of biological soil crust. *ISPRS J. Photogramm. Remote Sens.* **2014**, *90*, 59–67. [[CrossRef](#)]
6. Weber, B.; Büdel, B.; Belnap, J. *Biological Soil Crusts: An Organizing Principle in Drylands*, 1st ed.; Springer: New York, NY, USA, 2016; pp. 37–236.
7. Karnieli, A.; Shachak, M.; Tsoar, H.; Zaady, E.; Kaufman, Y.; Danin, A.; Porter, W. The effect of microphytes on the spectral reflectance of vegetation in semiarid regions. *Remote Sens. Environ.* **1996**, *57*, 88–96. [[CrossRef](#)]
8. Weber, B.; Olehowski, C.; Knerr, T.; Hill, J.; Deuschewitz, K.; Wessels, D.C.; Eitel, B.; Büdel, B. A new approach for mapping of Biological Soil Crusts in semidesert areas with hyperspectral imagery. *Remote Sens. Environ.* **2008**, *112*, 2187–2201. [[CrossRef](#)]
9. Chamizo, S.; Stevens, A.; Cantón, Y.; Miralles, I.; Domingo, F.; Van Wesemael, B. Discriminating soil crust type, development stage and degree of disturbance in semiarid environments from their spectral characteristics. *Eur. J. Soil Sci.* **2012**, *63*, 42–53. [[CrossRef](#)]
10. Karnieli, A. Development and implementation of spectral crust index over dune sands. *Int. J. Remote Sens.* **1997**, *18*, 1207–1220. [[CrossRef](#)]
11. Chen, J.; Zhang, M.Y.; Wang, L.; Shimazaki, H.; Tamura, M. A new index for mapping lichen-dominated biological soil crusts in desert areas. *Remote Sens. Environ.* **2005**, *96*, 165–175. [[CrossRef](#)]
12. Rozenstein, O.; Karnieli, A. Identification and characterization of Biological Soil Crusts in a sand dune desert environment across Israel–Egypt border using LWIR emittance spectroscopy. *J. Arid. Environ.* **2015**, *112*, 75–86. [[CrossRef](#)]
13. Rodríguez-Caballero, E.; Escribano, P.; Olehowski, C.; Chamizo, S.; Hill, J.; Cantón, Y.; Weber, B. Transferability of multi- and hyperspectral optical biocrust indices. *ISPRS J. Photogramm. Remote Sens.* **2017**, *126*, 94–107. [[CrossRef](#)]
14. Escribano, P.; Palacios-Orueta, A.; Oyonarte, C.; Chabrilat, S. Spectral properties and sources of variability of ecosystem components in a Mediterranean semiarid environment. *J. Arid. Environ.* **2010**, *74*, 1041–1051. [[CrossRef](#)]
15. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [[CrossRef](#)]
16. Wang, J.; Ding, J.; Abulimiti, A.; Cai, L. Quantitative estimation of soil salinity by means of different modeling methods and visible-near infrared (VIS-NIR) spectroscopy, Ebinur Lake Wetland, Northwest China. *PeerJ* **2018**, *6*, e4703. [[CrossRef](#)] [[PubMed](#)]
17. Meyer, H.; Lehnert, L.W.; Wang, Y.; Reudenbach, C.; Nauss, T.; Bendix, J. From local spectral measurements to maps of vegetation cover and biomass on the Qinghai-Tibet-Plateau: Do we need hyperspectral information? *Int. J. Appl. Earth Obs. Geoinf.* **2017**, *55*, 21–31. [[CrossRef](#)]
18. Liaw, A.; Wiener, M. Classification and regression by randomForest. *R News* **2002**, *2*, 18–22.
19. Zhang, J.; Wu, B.; Li, Y.; Yang, W.; Lei, Y.; Han, H.; He, J. Biological soil crust distribution in Artemisia ordosica communities along a grazing pressure gradient in Mu Us Sandy Land, Northern China. *J. Arid Land* **2013**, *5*, 172–179. [[CrossRef](#)]
20. Cheng, X.; An, S.; Liu, S.; Li, G. Micro-scale spatial heterogeneity and the loss of carbon, nitrogen and phosphorus in degraded grassland in Ordos Plateau, northwestern China. *Plant Soil* **2004**, *259*, 29–37. [[CrossRef](#)]
21. Wu, B.; Ci, L.J. Landscape change and desertification development in the Mu Us Sandland, Northern China. *J. Arid Environ.* **2002**, *50*, 429–444. [[CrossRef](#)]
22. Landsat 8 Surface Reflectance Code LaSRC Product Guide. Available online: <https://www.usgs.gov/media/files/landsat-8-surface-reflectance-code-lasrc-product-guide> (accessed on 17 December 2018).
23. Sentinel-2 User Handbook. Available online: https://sentinels.copernicus.eu/web/sentinel/user-guides/document-library/-/asset_publisher/xslst4309D5h/content/sentinel-2-user-handbook (accessed on 24 July 2015).
24. Meyer, H. Data-Driven Model Development in Environmental Geography. Ph.D. Thesis, The Philipps-University of Marburg, Marburg, Germany, 17 July 2017.

25. Irons, J.R.; Dwyer, J.L.; Barsi, J.A. The next Landsat satellite: The Landsat Data Continuity Mission. *Remote Sens. Environ.* **2012**, *122*, 11–21. [[CrossRef](#)]
26. Vermote, E.; Justice, C.; Claverie, M.; Franch, B. Preliminary analysis of the performance of the Landsat 8/OLI land surface reflectance product. *Remote Sens. Environ.* **2016**, *185*, 46–56. [[CrossRef](#)]
27. Drusch, M.; Del Bello, U.; Carlier, S.; Colin, O.; Fernandez, V.; Gascon, F.; Hoersch, B.; Isola, C.; Laberinti, P.; Martimort, P.; et al. Sentinel-2: ESA's Optical High-Resolution Mission for GMES Operational Services. *Remote Sens. Environ.* **2012**, *120*, 25–36. [[CrossRef](#)]
28. Sen2Cor Configuration and User Manual. Available online: <http://step.esa.int/main/third-party-plugins-2/sen2cor/> (accessed on 6 April 2018).
29. Elzinga, C.L.; Salzer, D.W.; Willoughby, J.W. *Measuring & Monitoring Plant Populations*; U.S. Bureau of Land Management: Lincoln, NE, USA, 1998.
30. Abdel-Rahman, E.M.; Ahmed, F.B.; Ismail, R. Random forest regression and spectral band selection for estimating sugarcane leaf nitrogen concentration using EO-1 Hyperion hyperspectral data. *Int. J. Remote Sens.* **2012**, *34*, 712–728. [[CrossRef](#)]
31. Ramasubramanian, K.; Singh, A. *Machine Learning Using R*, 1st ed.; Apress: New York, NY, USA, 2016; pp. 297–329.
32. Kohavi, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI), Montreal, QC, Canada, 20–25 August 1995.
33. Lehnert, L.W.; Meyer, H.; Obermeier, W.A.; Silva, B.; Regeling, B.; Bendix, J. Hyperspectral Data Analysis in R: The hsdar Package. *J. Stat. Softw.* **2019**, *89*, 877. [[CrossRef](#)]
34. Kuhn, M. Building predictive models in R using the caret package. *J. Stat. Softw.* **2008**, *28*, 1–26. [[CrossRef](#)]
35. Package 'rgdal'. Available online: <https://cran.r-project.org/web/packages/rgdal/index.html> (accessed on 14 March 2019).
36. Weber, B.; Hill, J. Remote Sensing of Biological Soil Crusts at Different Scales. In *Biological Soil Crusts: An Organizing Principle in Drylands*, 1st ed.; Weber, B., Büdel, B., Belnap, J., Eds.; Springer: New York, NY, USA, 2016; Volume 226, pp. 215–234.
37. Pirotti, F.; Sunar, F.; Piragnolo, M. Benchmark of Machine Learning Methods for Classification of a Sentinel-2 Image. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2016**, *41*, 335–340. [[CrossRef](#)]
38. Yan, F.; Wu, B.; Wang, Y. Estimating spatiotemporal patterns of aboveground biomass using Landsat TM and MODIS images in the Mu Us Sandy Land, China. *Agric. For. Meteorol.* **2015**, *200*, 119–128. [[CrossRef](#)]
39. Zichen, G.; Shulin, L.; Wenping, K.; Xiang, C.; Xueqin, Z. Change Trend of Vegetation Coverage in the Mu Us Sandy Region from 2000 to 2015. *J. Desert Res.* **2018**, *38*, 1099–1107.
40. Wang, T. *Atlas of Sandy Desert and Aeolian Desertification in Northern China*, 1st ed.; Science Press: Beijing, China, 2014; pp. 182–187.
41. Li, X. *Eco-Physiology of Biological Soil Crusts in Desert Regions of China*, 1st ed.; Higher Education Press: Beijing, China, 2016; pp. 1–51.
42. Danin, A.; Ganor, E. Trapping of airborne dust by mosses in the Negev Desert, Israel. *Earth Surf. Process. Landf.* **1991**, *16*, 153–162. [[CrossRef](#)]
43. Fang, S.; Yu, W.; Qi, Y. Spectra and vegetation index variations in moss soil crust in different seasons, and in wet and dry conditions. *Int. J. Appl. Earth Obs. Geoinf.* **2015**, *38*, 261–266. [[CrossRef](#)]
44. Karnieli, A.; Kokaly, R.F.; West, N.E.; Clark, R.N. Remote Sensing of Biological Soil Crusts. In *Biological Soil Crusts: Structure, Function, and Management*, 2nd ed.; Belnap, J., Lange, O.L., Eds.; Springer: Berlin, Germany, 2003; Volume 150, pp. 431–455.

