remote sensing

MDPI

Article

High-Resolution Vegetation Mapping Using eXtreme Gradient Boosting Based on Extensive Features

Heng Zhang [1], Anwar Eziz [1,*], Jian Xiao [1], Shengli Tao [2], Shaopeng Wang [1], Zhiyao Tang [1], Jiangling Zhu [1] and Jingyun Fang [1]

[1] College of Urban and Environmental Sciences, Peking University, Beijing 100871, China; heng.zhang@pku.edu.cn (H.Z.); xiaojian2017@pku.edu.cn (J.X.); shaopeng.wang@pku.edu.cn (S.W.); zytang@urban.pku.edu.cn (Z.T.); jlzhu@urban.pku.edu.cn (J.Z.); jyfang@urban.pku.edu.cn (J.F.)

[2] Laboratoire Évolution et Diversité Biologique, UMR 5174 (CNRS/IRD/UPS), 31062 Toulouse Cedex 9, France; shengli.tao@univ-tlse3.fr

* Correspondence: anwareziz@pku.edu.cn

Received: 27 May 2019; Accepted: 18 June 2019; Published: 25 June 2019

check for updates

Abstract: Accurate mapping of vegetation is a premise for conserving, managing, and sustainably using vegetation resources, especially in conditions of intensive human activities and accelerating global changes. However, it is still challenging to produce high-resolution multiclass vegetation map in high accuracy, due to the incapacity of traditional mapping techniques in distinguishing mosaic vegetation classes with subtle differences and the paucity of fieldwork data. This study created a workflow by adopting a promising classifier, extreme gradient boosting (XGBoost), to produce accurate vegetation maps of two strikingly different cases (the Dzungarian Basin in China and New Zealand) based on extensive features and abundant vegetation data. For the Dzungarian Basin, a vegetation map with seven vegetation types, 17 subtypes, and 43 associations was produced with an overall accuracy of 0.907, 0.801, and 0.748, respectively. For New Zealand, a map of 10 habitats and a map of 41 vegetation classes were produced with 0.946, and 0.703 overall accuracy, respectively. The workflow incorporating simplified field survey procedures outperformed conventional field survey and remote sensing based methods in terms of accuracy and efficiency. In addition, it opens a possibility of building large-scale, high-resolution, and timely vegetation monitoring platforms for most terrestrial ecosystems worldwide with the aid of Google Earth Engine and citizen science programs.

Keywords: vegetation mapping; XGBoost; simplified field survey; Dzungarian Basin; New Zealand

1. Introduction

Earth's diverse ecosystems provide generous resources for human populations. Accurate mapping of the diverse ecosystems or communities could facilitate environmental planning, resource management, and social wellbeing [1–4]. Since the rapid growth of the human population and climate warming is accelerating the pace of vegetation dynamics [5,6], timely information on the distribution of vegetation over regional to global scales has growing significance for a wide range of end users.

Mapping vegetation at the initial stage mainly depends on experts' empirical knowledge in defining boundaries between vegetation classes [7]. Even though fairly accurate at a small regional scale, this method is not only time-consuming but also has limited extendibility. Use of remote sensing (RS) technology has greatly increased the efficiency and accuracy of the mapping [8]. Availability of large stacks of various RS data, such as the Moderate Resolution Imaging Spectroradiometer (MODIS), Landsat, Polarization Synthetic Aperture Radar (PolSAR), Light Detection And Ranging (LiDAR), and aerial photography, enable mappers to produce vegetation maps from regional to global scales with considerable accuracy [9–11]. However, it still remains challenging to increase the mapping accuracy in mosaic and heterogeneous environments due to (a) incapability of traditional methods in differentiating subtle differences of vegetation classes based on spectral features [12], and (b) paucity of vegetation field survey data due to costly and time-consuming fieldwork [6].

Unlike conventional land cover mapping, with fewer classes that are comparatively easy to distinguish through spectral differences and objective-based methods, mapping vegetation needs to consider biotic and abiotic factors that affect the distribution of vegetation. In this sense, adding various environmental data—climatic, edaphic, geographic features, etc.—to the variable or feature set could be helpful in increasing classification performances [13]. With the aid of the Google Earth Engine (GEE) platform, a cloud-based platform for planetary-scale geospatial analysis, accessing a massive amount of RS data and preprocessing have become unprecedentedly convenient, which make a room for use of extensive RS features as well as biotic and abiotic factors [14].

Machine learning (ML), known for its outstanding capacity in identifying complex or nonlinear processes, has become widely used in ecological applications such as vegetation mapping in recent years [13,15–20]. The decision tree (DT)-based ML algorithms (such as extremely randomized trees, gradient boosting machine, random forest) with flexibility in data types, resilience to noise, and excellent performance in prediction, are believed to be effective in tackling qualitative and classification tasks. Among them, extreme gradient boosting (XGBoost), a scalable tree boosting method, is known for outstanding performances and state-of-the-art results in many research areas, as well as in ML competitions [21]. By virtue of improvements of the algorithm level, XGBoost is more efficient and robust to noise, class imbalance and uneven distribution, which is always the case in fieldwork data due to the complexity and even inaccessibility of the natural world. Additionally, XGBoost is also available for parallel and heterogeneous computing on both central processing units (CPUs) and graphics processing units (GPUs), which can drastically accelerate training and predicting speed [21,22]. In the field of RS, XGBoost also exhibits promising performances on classification tasks, and outperforms various benchmark classifiers such as support vector machine (SVM), and k-nearest neighbor (kNN) in many circumstances [23–26]. Specifically, in the cases like land cover mapping with numerous classes or crop classification tasks in precision agriculture, which have similarities with the vegetation mapping task, XGBoost was reported for ideal results [27,28]. Although incorporating extensive features will raise computing demands to a great extent, the parallel and heterogeneous computing in XGBoost could accelerate the training process. In this sense, XGBoost is the one of the suitable ML algorithms for vegetation mapping.

Generally, two approaches can be used to predict the spatial distribution of vegetation. The first one, known as the niche-based modelling approach (predicting first and assembling later), by mapping each species individually and assembling later according to specific vegetation classification rules and interactions of different species (like interspecific competition), requires ground truth observations to cover all the actual niches of the studied species, which is not the case in most of the studies at small to medium scales. Otherwise, the mapping results would be biased or misleading and often suffer from over-prediction in spite of its efficiency in computing [29,30]. The alternative approach is known as the cluster-based method (assembling first and predicting later), in which species are first categorized into different groups (vegetation classes) based on their similarities of absence and presence in observed sites. Compared with the first approach, this strategy may be powerful in detecting shared patterns of environmental response for infrequently recorded species. In addition, it could be more reliable and convenient in evaluations. By using multiclass classifiers and mapping the vegetation classes directly, it can avoid merging the results from various different species distribution models that are not comparable in results. Thus, in terms of the scale of our study cases, the cluster-based approach is believed to be more feasible than the other [29–31].

With the above considerations, we presented a workflow to produce high-resolution multiclass vegetation maps by adopting XGBoost based ML techniques based on extensive features and vegetation data, and we tested it on two strikingly different cases: (1) the Dzungarian Basin (DzB), China, and (2) New Zealand (NZ). To tackle the second problem mentioned above (i.e., the paucity of vegetation field survey data), we advocated simplified field survey procedures. That is, a dominant species-oriented field survey, which can reduce the heavy fieldwork intensity so that a study can cover a larger area at a lower cost in a shorter amount of time without sacrificing the quality.

We mainly want to address this question: Could the XGBoost-based ML workflow produce a high-resolution multiclass vegetation map with relatively high accuracy, based on a combination of extensive features (RS features, environmental data) and field survey data?

2. Materials and Methods

The workflow of this study included three linked sections, denoted as data preparation, modeling and mapping, and evaluation as presented in Figure 1.

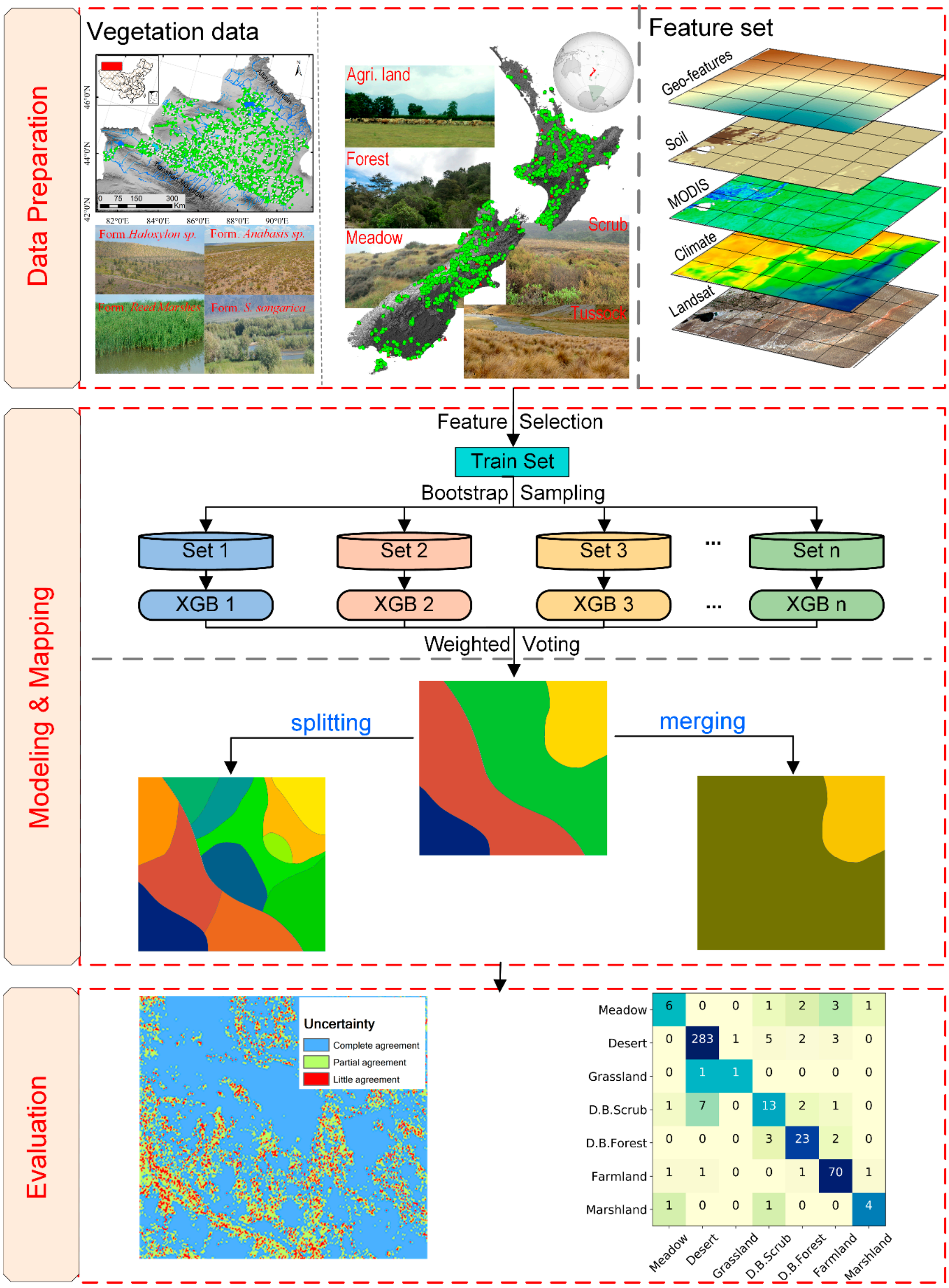

Figure 1. Workflow of vegetation mapping.

To test the extendibility of our method to different ecosystems, we selected DzB and NZ as our study cases, due to the striking differences in main vegetation types (desert vs. grassland and forest), climate (dry vs. wet), location (northern vs. southern hemisphere) and topography (plain vs. mountain). Moreover, DzB and NZ are representative places that include most of the major terrestrial habitat types including forests, grasslands, shorelines, wetlands, and deserts so that the extendibility of our method could be better substantiated.

2.1. Data Preparation

2.1.1. Vegetation Data

DzB: Simplified Field Survey Data

The simplified field survey we proposed is based on the reasoning that characteristics of dominant and subdominant species of each synusia (upper canopy, lower canopy, shrub layer, and herbaceous layer) determine the vegetation types [32,33]. During implementation, one should make sure that the following information is recorded with precision: (1) geographical coordinates, (2) dominant and subdominant plant species in each synusia, and (3) their relative coverages (RLCs). In addition, ground photographs should be taken with special care. At each site, more than two photographs should be taken, with at least one wide-angle or close-range photo (should be taken perpendicular to the ground), respectively. After the fieldwork, the wide-angle photos were helpful for confirming the accuracy of the information of dominant or subdominant species. The close-range photos should be calibrated according to view angle, enhanced, and then segmented to correct the initial RLC records. Sampling routes should be designed based on the main habitat distribution in the mapping region, according to previous research or RS images. For mapping accuracy, we recommend the sample size of each type to be more than 10.

Based on the simplified field survey procedures, we conducted extensive field sampling in DzB, which is located in northwestern China and covers about a 200,000 km^2 area, about 1.5 times larger than England, from mid-May 2015 to late August 2016 (summers only). With the small amount of rainfall and long hot summer, the vegetation mainly consists of xerophytes such as *Haloxylon* spp., *Calligonum* spp., *Agriophyllum squarrosum* and *Erodium oxyrrhynchum*. We took between two and four photos within every 1–2 km distance point using a Canon digital camera. Overall 3006 points (506 farmland and 2500 natural vegetation points) with the key information (two plant species for both woody and herb layers) were photographed. Plant species that were difficult to recognize were collected and sent back to the laboratory for further identification. To ensure the accuracy of information of plant species, all the ground photographs were reassessed and corresponding corrections were made. We preferred to assemble the species matrix into community-level entries before model generation due to possible niche overlaps among observed species and the spatial scale of the study areas [30]. According to the RLC of the plants, only the points of natural vegetation were divided into woody (2026 points with >5% RLC of woody or semi-woody plants) and herbaceous plant groups (474 points with <5% RLC of woody or semi-woody plants) before vegetation classification. After that, we classified the plots into 32 (27 shrub and 5 tree layer classes) woody and 17 herbaceous plant groups (Figures S1 and S2) by adopting Ward's minimum variance hierarchical clustering method based on a Mantel correlation test, by which the optimal number of classes for dividing the cluster tree were determined [34]. Based on the Chinese vegetation classification system and the classification results [35], we developed a hierarchical vegetation classification system (HVCS) of DzB, composed of seven vegetation types, 20 subtypes, and 50 associations (including farmland) (Table S1).

NZ: Simulated Survey Data

The second case, NZ, is an island country located southeast of Australia in the southern hemisphere. It has a temperate climate with plentiful rainfall, which makes NZ one of world's biodiversity hotspots [36]. Natural vegetation in NZ is mainly composed of grasslands, tussocks, and complex temperate forest consisting of *Pinaceae Lindl.*, Fagales, giant tree Pteridiaceae, and creepers [37,38].

Because a large part of NZ consists of agricultural types such as croplands and pastures, which are comparatively easy for classification but have less significance in vegetation mapping than natural vegetation types (forests, scrubs, etc.), evaluations on totally randomly chosen points would be less convincing. To tackle this problem, we collected 11,234 vegetation survey plots between 1978 and 2017 from the Global Biodiversity Information Facility (GBIF) [39] and added 1035 randomly chosen points of agricultural areas and glaciers, defined based on the field photographs from EOMF field photo library (http://www.eomf.ou.edu/photos/), by singling out one point every 1 km. Then, the vegetation types were extracted from the latest-available vegetation cover map of NZ (VCMNZ, 2015 updated) [40]. Finally, classes with less than eight occurrences were filtered out, and a total of 41 vegetation classes remained (Table S2). To test the performance of the XGBoost model in a different number of classes, we also predicted a map of habitats (10 classes available from GBIF), including the original field survey types, agricultural areas, and glaciers.

2.1.2. Features

RS Data

RS data used in this study includes features computed on MODIS and Landsat 8 Operational Land Imager (OLI) images, which were preprocessed on GEE. The details are addressed below.

(1) MODIS data. MODIS products included snow cover (MCD43D40) [41], percent non-vegetated, percent tree cover (vegetation continuous fields, MOD44B) [42], normalized difference vegetation index or enhanced vegetation index (NDVI or EVI, MOD13Q1) [43], leaf area index (LAI, MOD15A2H) [44], and land surface temperature of day and night (LST Day & Night, MOD11A2) [45] (Table S2). Before computing, we employed median or mean filters over the last 10 years for each season to obtain high-quality images. In accordance with soil and vegetation conditions in study areas, the modified soil adjusted vegetation index (MSAVI) [46], temperature vegetation dryness index (TVDI) [47], soil bio crust index (BCI) [48], soil moisture index (SMI) [49], soil relative salinity index (SRSI) [50], and vegetation cover index (FVC) [51] were computed based on cloud-free MODIS surface reflectance products (MOD09A1). The seasonal values and the seasonal changes of all the MODIS features (MCD43D40 and MOD44B, not included) were calculated.

(2) Landsat data. Landsat 8 OLI image collections of tier 1 calibrated top-of-atmosphere (TOA) reflectance in the last 10 years during summer were selected and then filtered for a cloud-free image via GEE [52]. Based on the cloud-free image, spectral variables, textures, color spaces, and moving window statistical variables were computed.

(i) Spectral variables, including combinations of surface reflectance bands, can be more closely related with certain vegetation characteristics than original data [53,54]. The spectral variables used in this research consist of original bands and derived indices, including RVI (band5/band4), DVI (band5-band4), SAVI ((band5–band4)/(band5 + band4 + L)*(1 + L), L = 0.5) [55], NDSI ((band4–band5)/(band4 + band5)) [56] and other band combinations (e.g., (band4–band3)/band5, band5/band3, band7/band4 and (band3 × band4)/band2)).

(ii) Textures are widely used to denote the differences in objects with similar reflectance, since they can provide the essential information about spatial orders of color patterns or reflectance intensity of the object [57,58]. In this study, 13 out of the 14 textures (maximal correlation coefficient not included) based on gray level co-occurrence matrix (GLCM) were computed with compute unified device architecture (CUDA) programming [59]. We calculated a set of textures for both the panchromatic band (band 8) and the above spectral variables by setting the window size to 13 × 13 for band 8 and to 11 × 11 for the spectral variables, and by setting shift distance to center to 1. As suggested, the value for every texture in each pixel was calculated as the average in four directions [57]. Then, principle component analysis (PCA) and linear discriminant analysis (LDA) were optionally adopted for dimensionality reduction (DR).

(iii) Color spaces, specific organizations of colors based on the object color profiling, are useful when some typical colors cannot be effectively expressed in red–green–blue (RGB) channels [60]. Hence, we applied conversions from RGB to HSV, YCbCr color spaces with band combinations {band4, band3, band2}, {band5, band4, band3}, {band6, band5, band2}, {band7, band5, band3}, {band7, band5, band4}, {band6, band5, band4}, and {band5, band7, band1} (setting band collections to RGB channels) using MATLAB (version R2018a). DR by PCA or LDA were also optionally conducted.

(iv) Moving window statistical variables, including coefficient of variance (CV), standard deviation (Std.), skewness, and kurtosis, can be beneficial for discriminating objects that have special spectral frequency distributions [61,62]. Hence, we calculated moving window CV, Std., skewness, and kurtosis values with a 9 × 9 moving window for each band and spectral variable, with CUDA programming on GPUs.

Because 30 m resolution (15 m resolution for band 8) Landsat 8 images contained roads and buildings irrelevant to vegetation mapping, filtering them out was a necessary step to ensure mapping accuracy. First, all of the calculated features were down-sampled to 120 m resolution via the nearest neighbor method. Then, median filter with size 1320 m (11 pixel width × 120 m/pixel width), Gaussian filter with sigma 3.5 along with mean filter with size 1080 m (9 pixel width × 120 m/pixel width) were employed in sequence. Finally, the resulting images were down-sampled to 1 km resolution by the nearest neighbor method.

Ancillary Data

(1) Climatic data. Bioclimatic variables associated with plant growth [63,64] were collected from WorldClim on GEE [65]. In addition, warmth index (WI), coldness index (CI), potential evapotranspiration (PET), actual evapotranspiration (AET), and water deficit (WD) were calculated according to original meteorological data [66–69].

(2) Edaphic data. Edaphic data for DzB and NZ, including soil organic carbon (SOC), pH, soil type, cation exchange capacity (CEC), bulk density (BLKD), chemical limitation to plant growth, mean annual soil temperature, soil drainage, induration soil hardness, and water yield were obtained from ISRIC world soil information [70] (download website: https://soilgrids.org) or the Landcare Research LRIS portal v1.0 (download website: http://lris.scinfo.org.nz/).

(3) Geographic data. Distances ($d(x)$) to the saline (DzB) and fresh waterbody (DzB and NZ) and sea shoreline (NZ) were computed through the following eikonal equation (Equation (1)) solved by the fast marching method [71], where high-resolution water masks were collected from the JRC Global Surface Water Mapping Layers, v1.0 on GEE [72].

$$\left|\nabla d(x)\right| = 1, \quad \text{s.t. } d(x)\big|_{\overline{\Omega}} = 0,\ \Omega = \text{water mask}. \tag{1}$$

Furthermore, slope, aspect, altitude, and hill-shade were calculated using ArcGIS software (version 10.3, ESRI®), based on the Shuttle Radar Topography Mission (SRTM) 30 m digital elevation model (DEM) obtained from the United States Geological Survey (USGS) Earth Resources Observation and Science (EROS) data center [73].

Finally, all the features were resampled to 1 km × 1 km spatial resolution to match the vegetation data. The feature processing procedures are specified in Figure 2.

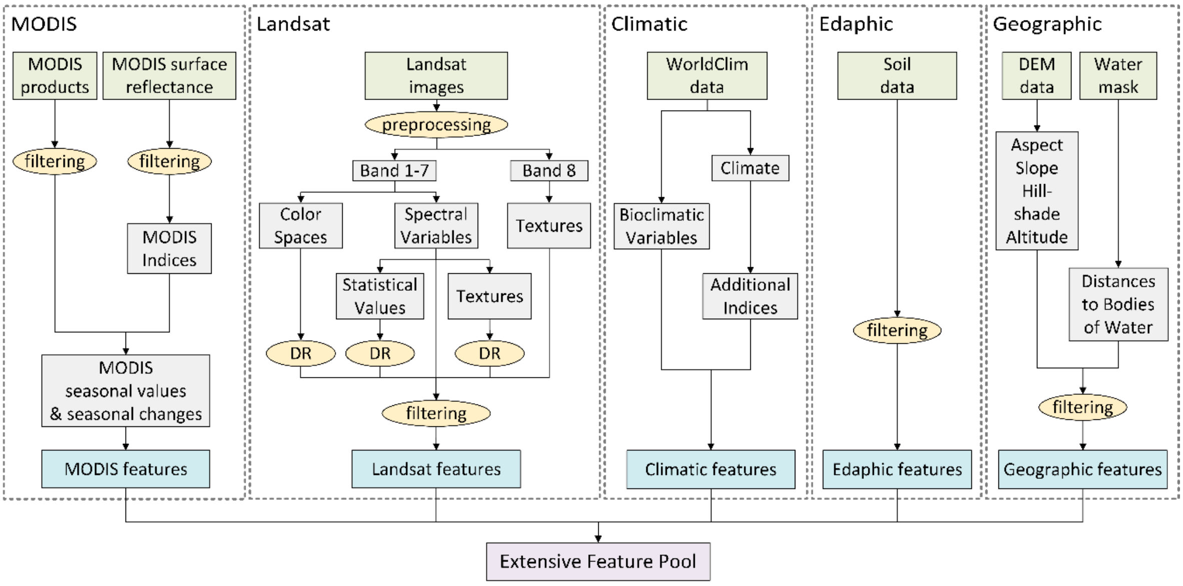

Figure 2. Extensive feature pool processing and computing procedures.

2.2. Modeling and Mapping

Developed from gradient boosting machine (GBM), an ensemble learning model adding new decision trees to complement those already built for better outcomes, XGBoost inherits the merits of GBM and provides state-of-the-art results with higher computing speed under many circumstances. Compared with GBM, XGBoost ameliorates the iterative optimization procedure [21,74]. In addition, several programming skills, such as pre-sorting feature values combined with histogram splitting and parallel and heterogeneous programming also promote classification performance and computing efficiency. The XGBoost model is not only suitable for data mining proposes, but it is also appropriate for image classification tasks such as land cover mapping in RS [21–23,75,76].

Bootstrap aggregating (bagging) ensemble learning, an effective ML framework known as subsampling samples (rows) or features (columns) with replacement as training sets and aggregating the trained base models, could avoid the negative effect of interdependency among variables, control overfitting, and increase robustness to noise. The framework was also proven to be efficient in enhancing model performances for various classifiers [77–81], which was also supported by our comparison of evaluations of mapping results via bagging vs. single model (Table S3). Besides, it may reduce the noise and increase the connectivity between patches of mapping results so that prediction would be more representative of natural vegetation patches (Box S1) [27]. Given these advantages, we used XGBoost as the base model of the bagging ensemble framework, then built the XGBoost bagging model as the classifier for vegetation mapping.

To evaluate the vegetation map, we randomly chose 497 or 2045 plots out of the total 3006 or 12269 plots (approximately 1/6) for DzB or NZ, respectively, as the test set, and the remaining, training set, for cross-validation and further evaluations. Distribution maps of train and test points are available in Figure 3. In addition, to prevent loss of information, the points in the test set from those classes with fewer than eight presences in association (DzB) were put back to the training set for class balance (in Section 2.2.2) and then removed before model training.

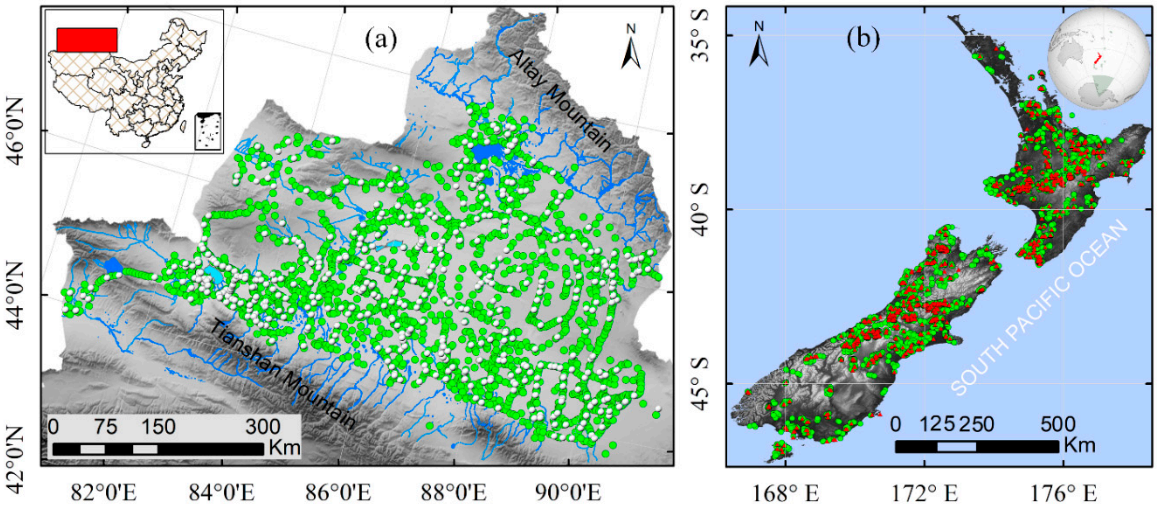

Figure 3. Distribution of sampling points of (**a**) Dzungarian Basin (DzB) and (**b**) New Zealand (NZ). Green circles represent train set while white circles or red triangles represent test sets.

2.2.1. Feature Selection

Finding a more discriminating and informative subset from the extensive feature pool is essential for increasing computing efficiency. Generally, the feature selection methods could be categorized as the filter-based method and the wrapper-based method [82]. However, without considering the combined effects of features, the filter-based method could not work quite well if the dataset contains multiple classes (Table S4). Therefore, we adopted a widely used wrapper-based feature selection method—the sequential forward floating selection (SFFS)—as the basic framework for feature selection [83–86]. SFFS is a bottom up search procedure that adds a remaining feature to the original empty set on every loop, which gives the highest evaluations under the objective function, followed by a backtracking step that excludes the worst features under certain conditions. However, even though it is improved from its predecessors, SFFS still suffers from the nesting effect, because forward inclusion is always unconditional [83,84]. A new improvement on SFFS, the improved forward floating selection (IFFS) method was reported for better solutions, because it incorporates a replacing step into the basic framework after the backtracking step [87]. However, our test on IFFS based on the extensive feature pool showed that when collinearity among features were very high, IFFS still suffered from the nesting effect.

Therefore, in this study, we made a minor improvement on IFFS; namely, revised IFFS, by adding a plus-L-minus-R search into the streamline for noting that non-useful features may be included into the selected set when highly correlated features exist. The plus-L-minus-R search excluded R features after incorporating L features from the feature pool. Its parameters R and L were set to 1 and 5 respectively, because the computing time would increase a lot if L is small, and non-useful features would be included if L is large. Besides, the replacing step from basic IFFS was optional for substituting either the most or the least important feature. The feature evaluations for plus-L-minus-R search were gathered from XGBoost inner feature importance scores, though feature evaluations of basic IFFS are kappa metrics computed by a stratified 10-fold cross-validation on the training set.

Additionally, several steps were conducted to improve the efficiency of the feature selection process. First, before computing, identical features with Pearson's correlations > 0.98 were removed. Second, after the loop completed, the set with the best evaluation (denoted as bestSET) and its location in the IFFS set sequence were recorded. The sets ahead of the bestSET with local maximum evaluations in the IFFS set sequence were then collected (denoted as maxSETs). Lastly, bestSET and the sets in maxSETs with evaluations >0.95*(evaluation of bestSET) were aggregated as the final feature set. The detailed feature selection procedures are specified in Figure 4.

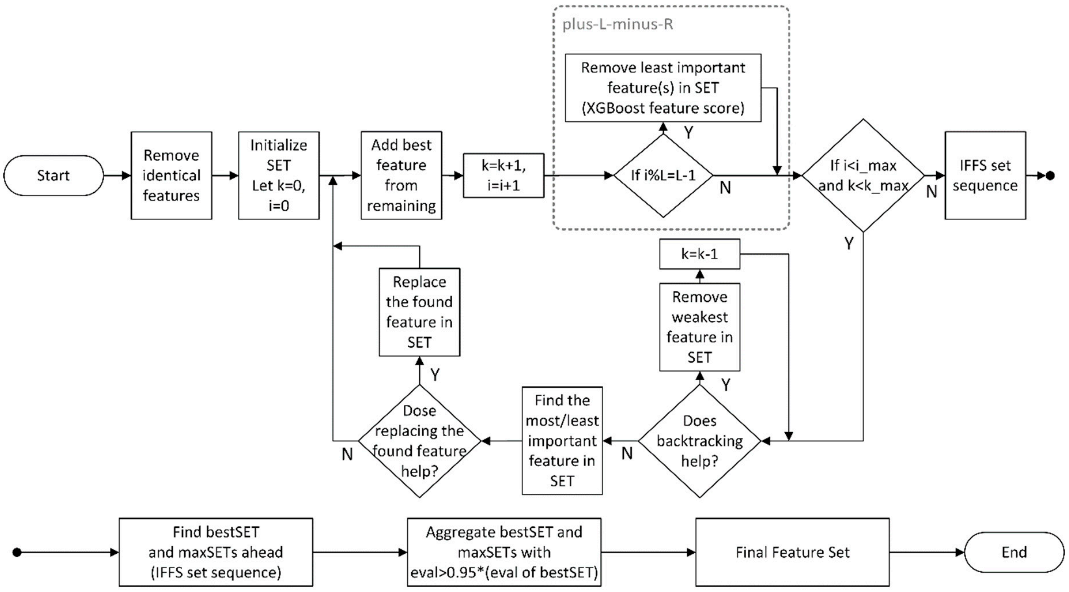
plus-L-minus-R
Start
Remove identical features
Initialize SET Let k=0, i=0
Add best feature from remaining
k=k+1, i=i+1
Remove least important feature(s) in SET (XGBoost feature score)
If i%L=L-1
Y
N
If i<i_max and k<k_max
IFFS set sequence
k=k-1
Replace the found feature in SET
Remove weakest feature in SET
Dose replacing the found feature help?
Find the most/least important feature in SET
Does backtracking help?
Find bestSET and maxSETs ahead (IFFS set sequence)
Aggregate bestSET and maxSETs with eval>0.95*(eval of bestSET)
Final Feature Set
End

Figure 4. Flow chart of feature selection, the revised improved forward floating selection (IFFS).

As a result, 49 or 45 of 456 features for vegetation subtypes or associations (DzB), and 41 or 40 of 356 features for habitats and vegetation classes (NZ) were selected, respectively. The texture variables played an important role in vegetation mapping, especially for cases with numerous classes. For example, the feature set of associations (DzB) comprised 34 textures, accounting for 76% in total. Also, the distances to water were incorporated in all the cases, indicating their importance in vegetation mapping. In addition, LST and its seasonal changes, which could reflect the energy balance of the land surface, were also included in almost all the cases.

To validate the contribution of the feature selection process to the mapping results, we compared the evaluations of results based on the selected set (by the revised IFFS, and the original IFFS) vs. the original feature pool (Table S4). The performances of mapping using the selected variables were not inferior to mapping using the original feature pool, indicating that the revised IFFS successfully selected a subset (~10% of total), useful in discriminating vegetation classes, from the extensive feature pool.

2.2.2. Vegetation Mapping

For vegetation data with HVCS, generally two ways for vegetation mapping are used: first, mapping the lower level classes directly and merging the result with higher level classes and second, mapping higher level classes as base map layer and then splitting the result with lower level classes. Deciding which one to choose always depends on expert knowledge and overall evaluations of maps. In DzB we found that mapping associations directly is less convincing, because some associations from the desert vegetation type with few observations had low accuracies. Thus, for DzB, we first computed a vegetation subtype map as the base map layer, and then split each subtype into association(s) according to HVCS; however, directly mapping was the only choice for NZ.

Before model training, we adopted synthetic minority over-sampling technique (SMOTE) to balance the training set after noting that an extremely imbalanced dataset could cause prediction errors for classes with few presences. The parameters of SMOTE were set to default, because it had optimal evaluations in this study. Then, a multiclass XGBoost bagging model was built on bootstrapped datasets (subsampling both columns and rows) from the training set, with the parameters of the XGBoost (the base model) set to default. The predicted results of each base model were then aggregated by weighted voting using kappa as the metric (Equations (2) and (3)). The probability maps of each vegetation class derived from the bagging framework were reserved for the uncertainty calculation (in Section 2.3). As for the parameters of the bagging framework, the optimal rates of 0.7 for both column and row sampling were found by grid search via stratified 10-fold cross-validation on the training set; the runtime (total base model number) was set to 610 because no further improvements on the evaluations or modelling stability were observed as the runtime increased (Figure S3).

The subtype map of DzB was used as the base map layer for association mapping in which each subtype was divided to corresponding association(s) according to HVCS with small multiclass XGBoost bagging models. SMOTE was also implemented for each subtype before model training. The vegetation type map was obtained from the vegetation subtype map by merging the corresponding classes based on HVCS (Figure 1). Because the observation data of DzB and NZ did not include any urban areas or bodies of water, these land use types were directly extracted from the resampled land use map of China, 2010, for DzB [88] and the land cover database classes version 4.1, for NZ (download website: https://lris.scinfo.org.nz/layer/48423-lcdb-v41-land-cover-database-version-41-mainland-new-zealand/). These land use types were not included in the evaluations.

Finally, the results were filtered with a majority filter of 3 × 3 window size to increase the connectivity of each class so that the distribution of vegetation patches would be more lifelike.

The XGBoost bagging model was implemented with XGBoost (version 0.81), scikit-learn (version 0.19.1), and Geospatial Data Abstraction Library (GDAL, version 2.3.2) in the Python 3 programming language [21,89,90]. Source codes, including feature computing, feature selection, and vegetation mapping, are available at https://github.com/ZH-pku/xgb_vegetation_mapping. All our mapping processes were executed on two NVIDIA TITAN X (Pascal) GPUs.

2.3. Evaluations

To evaluate the vegetation mapping results, metrics including overall accuracy (OA, Equation (2)), Cohen's kappa (kappa, Equations (3) and (4)), and the original confusion matrix were computed based on the predicted and real class of the reserved test points. In addition, to further analyze the mapping uncertainty, we computed the uncertainty of the base map layer via the confusion index (CI, Equation (5)) (Burrough, et al. 1997).

$$\mathrm{OA} = \frac{e}{N}. \tag{2}$$

$$p_e = \frac{1}{N^2}\sum_k n_k^r n_k^p. \tag{3}$$

$$\mathrm{kappa} = \frac{\mathrm{OA} - p_e}{1 - p_e}. \tag{4}$$

$$\mathrm{CI} = 1 - (\mu_{max} - \mu_{max-1}), \tag{5}$$

where e, and N in Equations (2) and (3) are number of correct predictions, and number of total observations; p_e, k, n_k^r, and n_k^p in Equations (3) and (4) are the hypothetical probability of chance agreement, the class predicted, number of observations of class k, and number of predictions of class k, respectively; μ_{max} and μ_{max-1} in Equation (5) are the largest and the second largest probability values of each pixel of all possible classes generated from the bagging framework.

3. Results

Following the above workflow, we produced the map of vegetation types (seven classes), subtypes (17 classes), and associations (43 classes) of DzB and the maps of 10 habitats and 41 vegetation classes of NZ, as shown in Figures 5 and 6.

For DzB, the metrics of the final mapping result were calculated on the test points within the DzB boundary (441 out of the chosen 497 points). The OA (kappa) of vegetation types, subtypes, and associations were 0.907 (0.821), 0.801 (0.739), and 0.748 (0.685), respectively. For NZ, the established models produced the habitat and vegetation class maps with evaluations of 0.946 (0.830) and 0.703 (0.666), respectively. We found that accuracy decreased with increments of class number, and among all levels of classification (vegetation types, subtypes, and associations).

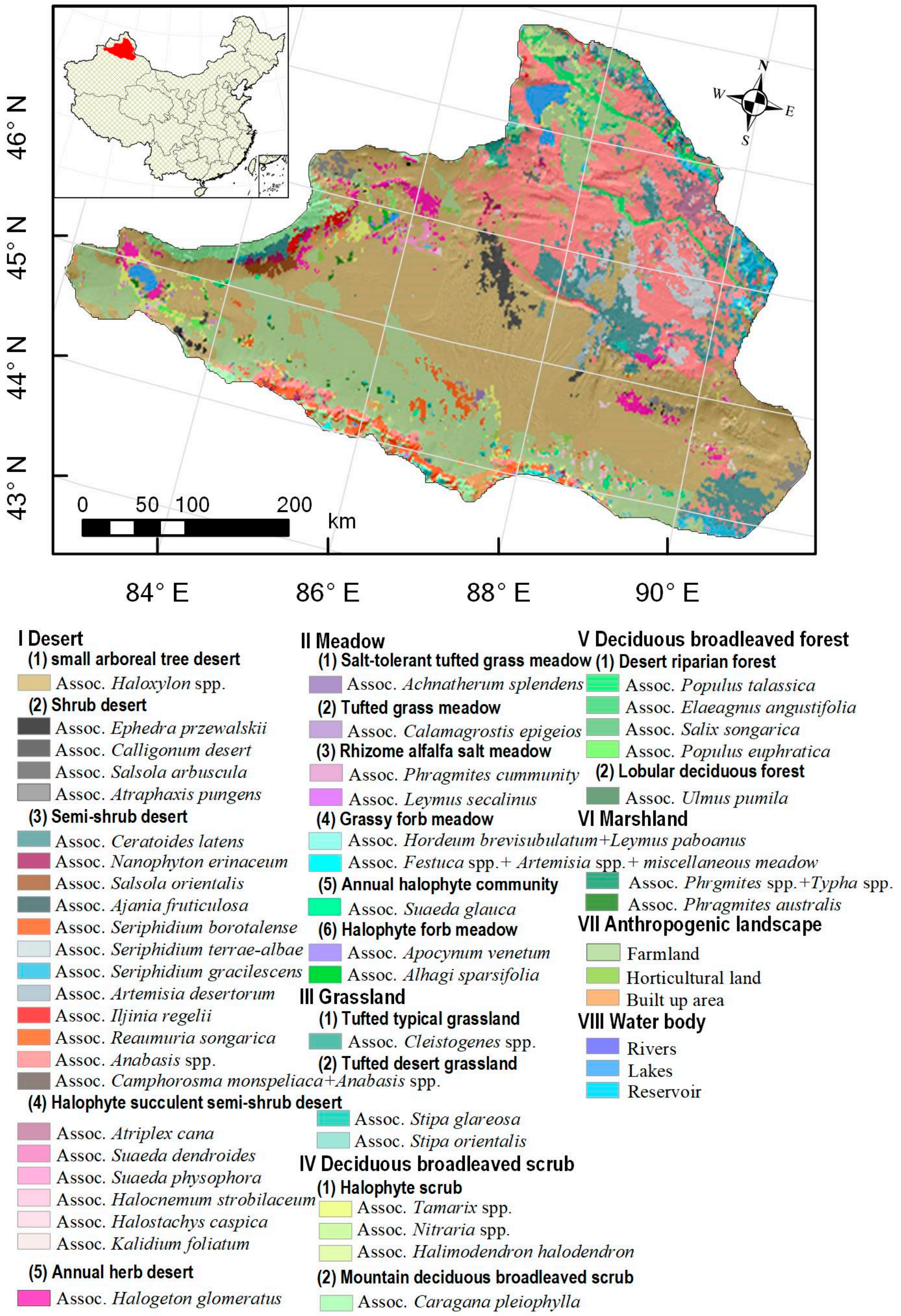
46° N
45° N
44° N
43° N
84° E
86° E
88° E
90° E
0
50
100
200
km
N
W
E
S
I Desert
(1) small arboreal tree desert
Assoc. *Haloxylon* spp.
(2) Shrub desert
Assoc. *Ephedra przewalskii*
Assoc. *Calligonum desert*
Assoc. *Salsola arbuscula*
Assoc. *Atraphaxis pungens*
(3) Semi-shrub desert
Assoc. *Ceratoides latens*
Assoc. *Nanophyton erinaceum*
Assoc. *Salsola orientalis*
Assoc. *Ajania fruticulosa*
Assoc. *Seriphidium borotalense*
Assoc. *Seriphidium terrae-albae*
Assoc. *Seriphidium gracilescens*
Assoc. *Artemisia desertorum*
Assoc. *Iljinia regelii*
Assoc. *Reaumuria songarica*
Assoc. *Anabasis* spp.
Assoc. *Camphorosma monspeliaca*+*Anabasis* spp.
(4) Halophyte succulent semi-shrub desert
Assoc. *Atriplex cana*
Assoc. *Suaeda dendroides*
Assoc. *Suaeda physophora*
Assoc. *Halocnemum strobilaceum*
Assoc. *Halostachys caspica*
Assoc. *Kalidium foliatum*
(5) Annual herb desert
Assoc. *Halogeton glomeratus*
II Meadow
(1) Salt-tolerant tufted grass meadow
Assoc. *Achnatherum splendens*
(2) Tufted grass meadow
Assoc. *Calamagrostis epigeios*
(3) Rhizome alfalfa salt meadow
Assoc. *Phragmites cummunity*
Assoc. *Leymus secalinus*
(4) Grassy forb meadow
Assoc. *Hordeum brevisubulatum*+*Leymus paboanus*
Assoc. *Festuca* spp.+ *Artemisia* spp.+ *miscellaneous meadow*
(5) Annual halophyte community
Assoc. *Suaeda glauca*
(6) Halophyte forb meadow
Assoc. *Apocynum venetum*
Assoc. *Alhagi sparsifolia*
III Grassland
(1) Tufted typical grassland
Assoc. *Cleistogenes* spp.
(2) Tufted desert grassland
Assoc. *Stipa glareosa*
Assoc. *Stipa orientalis*
IV Deciduous broadleaved scrub
(1) Halophyte scrub
Assoc. *Tamarix* spp.
Assoc. *Nitraria* spp.
Assoc. *Halimodendron halodendron*
(2) Mountain deciduous broadleaved scrub
Assoc. *Caragana pleiophylla*
V Deciduous broadleaved forest
(1) Desert riparian forest
Assoc. *Populus talassica*
Assoc. *Elaeagnus angustifolia*
Assoc. *Salix songarica*
Assoc. *Populus euphratica*
(2) Lobular deciduous forest
Assoc. *Ulmus pumila*
VI Marshland
Assoc. *Phrgmites* spp.+*Typha* spp.
Assoc. *Phragmites australis*
VII Anthropogenic landscape
Farmland
Horticultural land
Built up area
VIII Water body
Rivers
Lakes
Reservoir

Figure 5. Vegetation map of DzB.

An uncertainty analysis showed that 83.54% of the mapping area in DzB (according to the subtype map) was in complete agreement (CI ≤ 0.66), and increased to 94.96% if partial agreement (0.66 < CI ≤ 0.90) was included. For NZ, 89.71% and 71.20% of the mapping area showed complete agreement, which increased to 96.88% and 89.70% when partial agreement was included (Figure A1). We found no significant relationships between distance to sampling points and uncertainty, indicating that uncertainty variations may largely stem from the predictive power of features and the number of observations in each class.

We also recorded the time cost of computation of model training, testing, and prediction, as presented in Table S5. The results showed that XGBoost almost cost less time than any other ML algorithm in prediction. However, in the process of model training and testing, XGBoost did not necessarily perform better than the others.

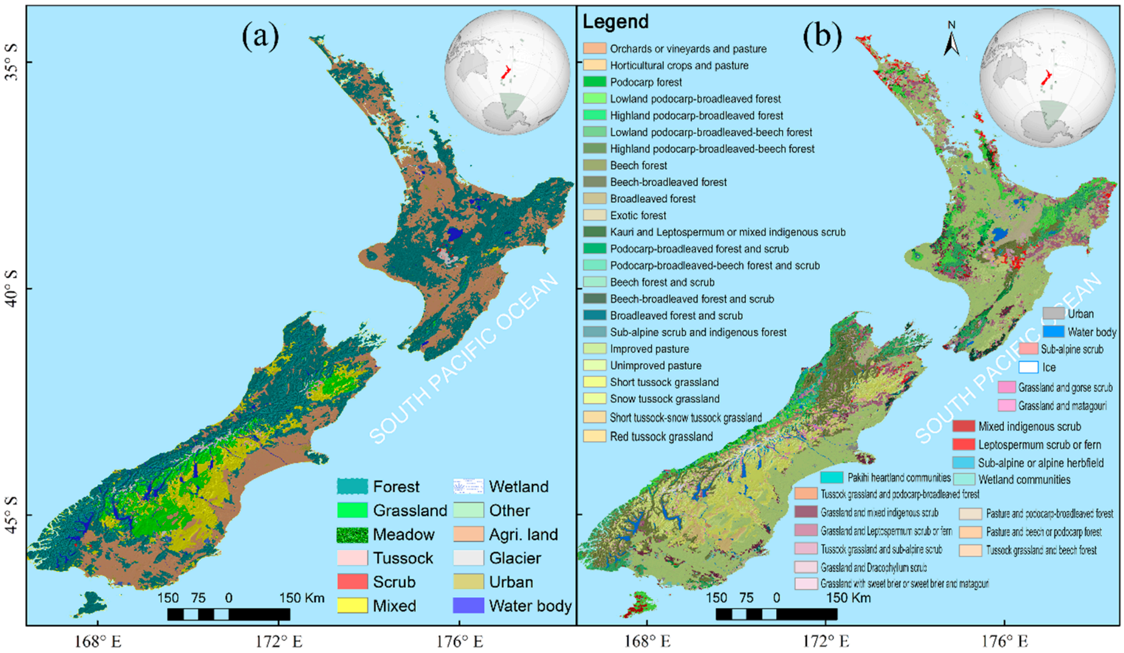
(a)
(b)
35° S
40° S
45° S
168° E
172° E
176° E
SOUTH PACIFIC OCEAN
N
Forest
Grassland
Meadow
Tussock
Scrub
Mixed
Wetland
Other
Agri. land
Glacier
Urban
Water body
150 75 0 150 Km
Legend
Orchards or vineyards and pasture
Horticultural crops and pasture
Podocarp forest
Lowland podocarp-broadleaved forest
Highland podocarp-broadleaved forest
Lowland podocarp-broadleaved-beech forest
Highland podocarp-broadleaved-beech forest
Beech forest
Beech-broadleaved forest
Broadleaved forest
Exotic forest
Kauri and Leptospermum or mixed indigenous scrub
Podocarp-broadleaved forest and scrub
Podocarp-broadleaved-beech forest and scrub
Beech forest and scrub
Beech-broadleaved forest and scrub
Broadleaved forest and scrub
Sub-alpine scrub and indigenous forest
Improved pasture
Unimproved pasture
Short tussock grassland
Snow tussock grassland
Short tussock-snow tussock grassland
Red tussock grassland
Urban
Water body
Sub-alpine scrub
Ice
Grassland and gorse scrub
Grassland and matagouri
Mixed indigenous scrub
Leptospermum scrub or fern
Sub-alpine or alpine herbfield
Pakihi heartland communities
Wetland communities
Tussock grassland and podocarp-broadleaved forest
Grassland and mixed indigenous scrub
Grassland and Leptospermum scrub or fern
Tussock grassland and sub-alpine scrub
Grassland and Dracophylum scrub
Grassland with sweet brier or sweet brier and matagouri
Pasture and podocarp-broadleaved forest
Pasture and beech or podocarp forest
Tussock grassland and beech forest

Figure 6. Vegetation maps of (**a**) habitats, and (**b**) vegetation classes of NZ.

4. Discussion

In this study, we produced high-resolution (~1 km^2) multiclass (>40 classes) vegetation maps in two regions (DzB and NZ) with relatively high accuracy based on a combination of extensive features and field survey data, using XGBoost as the ML classifier. As the confusion matrices of the two cases illustrated (Figure A2), the vegetation classes of DzB and NZ were effectively separated by our method. Because DzB and NZ consist of most of the terrestrial habitat types worldwide, it could be reasonable to say that the method used in this study has potential to be used in vegetation mapping in various terrestrial ecosystems.

To further assess our mapping results, we evaluated the climate change initiative (CCI) land cover map produced by the European Space Agency (ESA) [10], MODIS vegetation continuous fields (VCF) produced by the National Aeronautics and Space Administration (NASA) [11], and remap (RS online land cover mapping pipeline) [91], based on the test set. The general mapping procedures for these projects mainly depend on the surface spectral reflectance in RS imagery and limited environmental data (remap) [9,92]. We found that our approach was more accurate than the projects above (DzB vegetation types: $\mathrm{OA}_{\mathrm{CCI}} = 0.760$, $\mathrm{kappa}_{\mathrm{CCI}} = 0.334$; $\mathrm{OA}_{\mathrm{VCF}} = 0.609$, $\mathrm{kappa}_{\mathrm{VCF}} = 0.334$ and $\mathrm{OA}_{\mathrm{remap}} = 0.156$, $\mathrm{kappa}_{\mathrm{remap}} = 0.191$; $\mathrm{OA}_{\mathrm{this\ study}} = 0.907$, $\mathrm{kappa}_{\mathrm{this\ study}} = 0.821$; and for NZ habitats: $\mathrm{OA}_{\mathrm{CCI}} = 0.381$, $\mathrm{OA}_{\mathrm{VCF}} = 0.771$, $\mathrm{kappa}_{\mathrm{VCF}} = 0.473$, $\mathrm{OA}_{\mathrm{remap}} = 0.905$, $\mathrm{kappa}_{\mathrm{remap}} = 0.817$, $\mathrm{OA}_{\mathrm{this\ study}} = 0.946$, $\mathrm{kappa}_{\mathrm{this\ study}} = 0.830$). The evaluations of the compared projects were even lower when the number of classes increased (Tables S6–S8), indicating that our method outperformed them although they are effective in large scale mapping with a limited number of classes (< 20 classes) [9,92].

As the growth of plants is affected by both biotic and abiotic factors, a combination of environmental data and RS features could contribute to vegetation mapping, especially where the spectral differences of some classes are subtle or the effect of the microenvironment is significant. Here, we compared the mapping results with and without environmental data. For mapping without environmental data, the climatic, edaphic, and geographic data, as well as several MODIS products or indices (LST, Snow Index, TVDI, BCI, SMI, and SRSI) were removed from the feature pool. The feature selection was re-conducted, followed by training multiclass XGBoost bagging models directly on vegetation subtypes

(DzB), associations (DzB), habitats (NZ), and vegetation classes (NZ), with evaluations calculated on the test set. The calculations were repeated 16 times, by setting the base model number to 72 in bagging framework. The results stressed the importance of considering both RS features and environmental data in vegetation mapping (Table 1). The differences in evaluation were subtle (~0.05 OA) when the number of classes was small, but the gaps were large (~0.10 OA) as the number of classes increased. Meanwhile, the differences in sensitivity (a metric, true positive/condition positive) showed that natural vegetation types, e.g., deserts (−0.08 on average, associations of DzB), grasslands or meadows (−0.134 on average, associations of DzB), and forests and scrubs (−0.170 on average, vegetation classes of NZ), presented an obvious decrease in classification accuracy, while the sensitivities of artificial types like agricultural land or glaciers were almost the same or even increased. Also, without the information on the bodies of water, the sensitivities of wetlands suffered great losses of −0.333 in DzB and −0.666 in NZ. Thus, for cases with numerous vegetation classes, the mapping results could be more accurate if incorporating environmental data.

Table 1. Vegetation mapping evaluations with and without environmental data. With Env. represents mapping with environmental data; Without Env. represents mapping without environmental data. SD is standard deviation.

	Case	Num. of Classes (in Test Set)	OA ± SD	Kappa ± SD
With Env.	vege subtype (DzB)	20	0.798 ± 0.006	0.744 ± 0.007
	association (DzB)	50	0.764 ± 0.009	0.704 ± 0.011
	habitat (NZ)	10	0.944 ± 0.001	0.824 ± 0.002
	vege class (NZ)	41	0.701 ± 0.004	0.665 ± 0.004
Without Env.	vege subtype (DzB)	20	0.758 ± 0.009	0.694 ± 0.011
	association (DzB)	50	0.599 ± 0.012	0.534 ± 0.014
	habitat (NZ)	10	0.888 ± 0.002	0.679 ± 0.004
	vege class (NZ)	41	0.600 ± 0.006	0.550 ± 0.006

To compare the performance of XGBoost with other prevailing benchmark ML classifiers, we selected extremely randomized trees (extra-trees, ET), GBM, random forest (RF), quadratic discriminative analysis (QDA), SVM, and kNN as comparing models in this study. All the classifiers were programmed with scikit-learn (version 0.19.1), by setting all parameters to default [89]. The feature selection was conducted individually for each model on the extensive feature pool, with the plus-L-minus-R search removed if the inner feature importance score was unavailable (kNN, SVM, QDA). Then, we built multiclass predicting models using a bagging framework, the same as XGBoost, with the training set, and tested with the test set directly on vegetation types (DzB), vegetation subtypes (DzB), associations (DzB), habitats (NZ), and vegetation classes (NZ). The calculations were repeated 16 times, by setting the base model number to 72 in the bagging framework, considering the computing limitations. We found that the DT-based methods, i.e., XGBoost, ET, RF, and GBM could provide promising results (~10 classes, ~0.9 OA; ~40 classes, ~0.7 OA), but kNN, SVM, and QDA could only perform fairly well when the number of classes was small (Figure 7). These results evinced the extendibility of DT-based algorithms in vegetation mapping. To further compare the DT-based methods, we re-conducted the mapping processes using the bagging framework, and setting the base model number to 610. Then, we computed the uncertainty according to Equation (5). The comparison of mean uncertainty values indicated that XGBoost performed the best in all the cases (Table 2). In addition, the percentages of area of complete agreement or little agreement were the highest or lowest when using XGBoost as the base model (Figure S4), highlighting the accuracy and robustness of the mapping results of XGBoost among the DT-based methods. Furthermore, a comparison of the computing time showed that the predicting time of XGBoost (using GPUs) was less than other compared methods in most cases. Although during model training and testing XGBoost was not the most time-saving ML algorithm, as prediction is crucial and the most time-consuming part in the mapping processes when the target resolution is high, it is reasonable to say that XGBoost is efficient in vegetation mapping. Compared with its predecessor GBM, XGBoost improved a lot in model performances, especially under conditions with numerous classes. Further evidence could be found in the realm of ecology. For example, Sandino, Pegg, Gonzalez and Smith [24] found that with reliable field data, XGBoost can distinguish forests infected with pathogens as high as OA = 97.32%. Dong, Xu, Wang and Pu [23] also successfully mapped submerged plants using XGBoost. Therefore, considering the advantages in programming and computing, XGBoost is a suitable method for vegetation mapping.

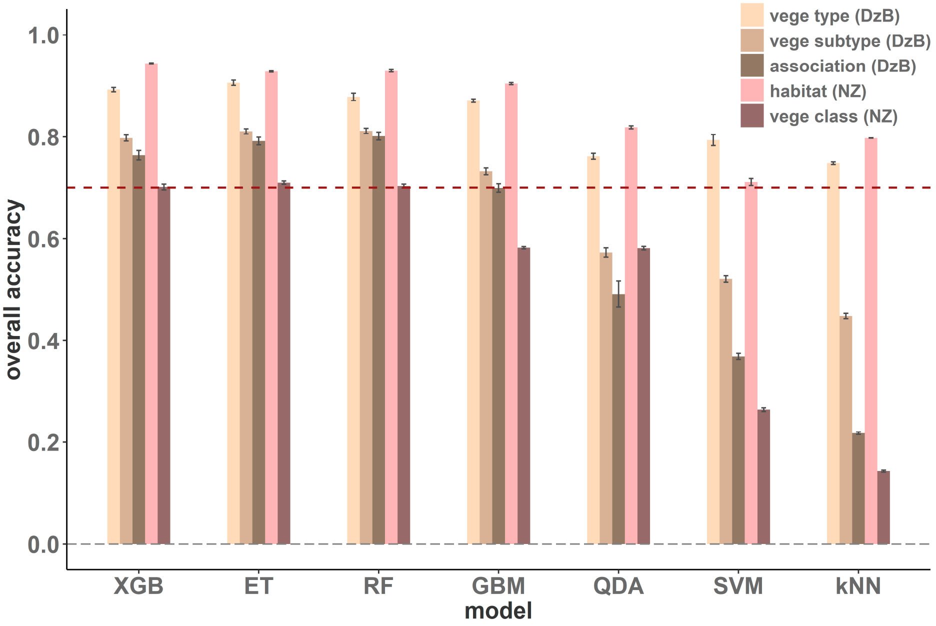

Figure 7. Comparison of OA (± standard deviation (SD)) of different machine learning (ML) models. The selected models include extreme gradient boosting (XGB), extra trees (ET), random forest (RF), gradient boosting machine (GBM), quadratic discriminative analysis (QDA), support vector machine (SVM), and k-nearest neighbors (kNN).

Table 2. Mean uncertainty values of vegetation mapping using extreme gradient boosting (XGB), extra trees (ET), random forest (RF), and gradient boosting machine (GBM).

Case	Num. of Classes (in Test Set)	XGB	ET	RF	GBM
vege type (DzB)	7	0.117	0.146	0.208	0.142
vege subtype (DzB)	20	0.285	0.378	0.450	0.300
association (DzB)	50	0.411	0.491	0.467	0.568
habitat (NZ)	10	0.182	0.239	0.248	0.235
vege class (NZ)	41	0.401	0.490	0.490	0.448

As the plant community is characterized by its dominant species, the simplified field survey procedures we advocated could be an alternative solution to the second problem: the paucity of vegetation field survey data due to the costly, labor-intensive and time-consuming fieldwork. With targeted information on the main species aided by carefully taken ground photographs, the simplified field survey could successfully collect the information needed for vegetation mapping with reduced labor intensity, which was proven in our experiment where RLCs from the simplified method were almost the same as those from a standard investigation ($R^2 = 0.97$, $p < 0.01$) (Figure S5).

By far, the studies on vegetation mapping showed no highly accurate results with numerous classes like ours [93]. By comparison, our method performed better, partly because of the extensive features we computed. Interdependency between features may be of potential concern for some mapping scientists. However, first, with an effective feature selection method, highly redundant or irrelevant features could be removed; second, due to improvements on the algorithm level (including the bootstrapping design of the bagging framework to control overfitting), our method could eliminate the possible negative effect of collinearity [21,77,80]. Some features may be pointless to ecological explanation, but the interpretability of variables should be prioritized in explanation-oriented research [94], and for a mapping perspective, more attention should be paid to the predictive power and the generality of the proposed methods instead; in other words, evaluation standards should be question-oriented.

Although outstanding in performance, our workflow still has room for further improvements. First, the simplified field survey might not be feasible in tropical and subtropical regions where the communities are more diverse and complex with fewer dominant species, defining many of which is quite hard if not impossible. Therefore, the accuracy of species information and RLCs derived from ground photos could not be guaranteed. In light of this, we suggest to use the simplified field survey procedures in temperate and polar zones. Also, the RLCs derived from ground photos might contain bias without careful treatment of plant overlaps. Second, the feature pool could be more extensive. By incorporating data from various satellites or ground monitoring sites, more details about the spectral characteristics of vegetation or environmental conditions could be unveiled, that is beneficial to enhancing performance. A change of phenology patterns or temporal features could contribute to discriminating vegetation classes [26,95]. To detect accurately, RS reflectance data with a shorter cadence (~8 day or less) should be employed, on which frequency analysis methods (e.g., Fourier transform) could reveal the patterns. Third, determining the scale of a vegetation map is still problematic. Since the time of using RS techniques in vegetation mapping, the mismatch of scales between field survey data and RS data is still an unsolved problem. Generally, the scale of a large part of RS or environmental data (>500 m) is far beyond the scale of field survey (1–100 m), because one can only accurately identify plant species within ~10 m boundary. The basic logic of field surveying in vegetation mapping is to choose one point within a certain region as representative. However, the process could not be bias-free in the sight from an adult height. In ML fashion vegetation mapping, the scale should be accordant with RS data. With the development of RS technology, ever-increasing satellite data with higher resolution (<20 m), indeed close to the scale of field survey, have been generated, like reflectance data from Sentinels, GaoFens, and WorldViews. Thus, a robust fusion method of low-resolution and high-resolution data should be considered in feature computing. More object-based methods and convolutional neural networks (CNN)-based feature extraction methods, as approaches to combining different data, should be addressed [96,97]. Also, using an unmanned aerial vehicle (UAV) may be an ideal solution to this problem [76]. With the sight from a higher elevation, photos from UAVs could provide information with the scale between traditional field surveys and RS, which may be helpful to ensure the field survey data being more accurate and representative. In this sense, the necessity of data fusion methods in feature computing should be more emphasized. Lastly, although the workflow was tested on two different cases, the extendibility may still need further assessment through more cases. However, due to the lack of sound vegetation inventory data, testing the method with more cases is quite hard at present. Besides, as our cases included most of the terrestrial habitat types, it is safe to say that our method shed some light on future vegetation monitoring.

As supported by various studies in monitoring land use, habitat, or vegetation cover changes, etc., the ground photography-aided methods have broader application prospects [98–100]. In recent years, citizen science (or crowd sourcing) is becoming an alternative data source for plant phenology, as well as for wild life observation and disease monitoring [101]. Numerous projects or monitoring platforms have been launched or built where volunteers all over the world contribute their ground photos useful to researchers, such as PhenoCam [102], Penguin Watch [102], eBird [103], Season Spotter [104], and Field Photo Library by SciStarter [9]. With the workflow in this study, it is possible to launch and build timely vegetation monitoring programs in which volunteers upload ground photographs with geographical references and identify the main species with the guidance of experts. In addition, with an efficient ML algorithm, the platform could automatically map the vegetation in high resolution with relatively high accuracy based on extensive features. Indeed, increasing demands on computing power, especially when dealing with raster imagery of a large spatial extent, may be a potential concern. Fortunately, with emerging advancements in computational technology (e.g., super computers and heterogeneous computing devices) and revolutionary cloud computing platforms like GEE, ML model training and predicting could be accelerated [14,105]. Thus, with the advanced technologies, relatively high accuracy, and potential extendibility, building high-resolution vegetation monitoring systems using the workflow we proposed in various spatial scales (even worldwide) matching various needs could be a possible and important path we may consider in the next step.

5. Conclusions

In this study, we produced high-resolution multiclass vegetation maps using extensive features and vegetation data from simplified field surveys or simulation through XGBoost models in the cases containing most of the habitats of terrestrial ecosystems. The overall accuracies of DzB (0.907, for vegetation types; 0.801, for vegetation subtypes; 0.748, for associations) and NZ (0.946, for habitats; 0.703, for vegetation classes) were higher than many of the proposed vegetation mapping projects.

The findings indicated that our approach has potential in producing high accuracy vegetation mapping in various terrestrial ecosystems. In addition, it opened a possibility of building a timely vegetation monitoring platform with the aid of citizen science programs, Google Earth Engine, and advanced computing technologies.

Supplementary Materials: The following are available online at http://www.mdpi.com/2072-4292/11/12/1505/s1, Figures S1–S5, Tables S1–S8, Box S1.

Author Contributions: Conceptualization, H.Z. and A.E.; methodology, H.Z. and A.E.; software, H.Z.; validation, A.E. and H.Z.; formal analysis, H.Z., A.E. and J.X.; investigation, A.E.; resources, J.F. and J.Z.; data curation, H.Z. and A.E.; Writing—Original Draft preparation, A.E., H.Z. and S.T.; Writing—Review and Editing, J.F., S.T., S.W., Z.T. and J.Z.; visualization, H.Z. and A.E.; supervision, J.F.; project administration, J.F.; funding acquisition, J.F.

Funding: This research was funded by (1) the National Natural Science Foundation of China, and (2) the National Key Research and Development Program of China, grant number (1) 31621091, (2) 2017YFC0503906.

Acknowledgments: We thank M.S. Bin Yuan of the School of Electronics Engineering and Computer Science, Peking University and M.S. Weijia Wang of Tencent Holdings Limited, Shenzhen, China for machine learning methodology supports. We thank Yu Liu and Wenjie Fan of the School of Earth and Space Sciences, Peking University, Yuan Lou of Ohio State University, and David A. Yuen of Columbia University for their precious advices on the discussion part, and materials and methods part. We also thank Ke Guo of the Institute of Botany, the Chinese Academy of Sciences for his valuable advices on the HVCS of DzB.

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

Appendix A

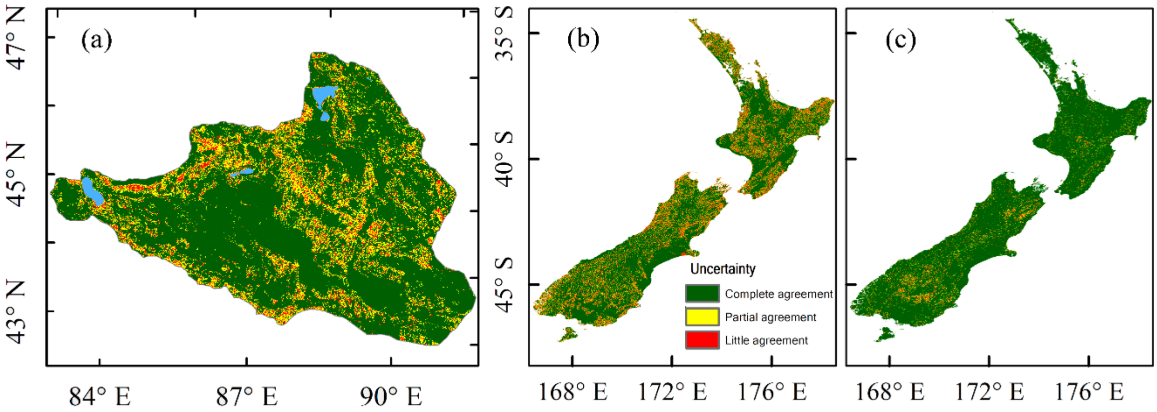

Figure A1. Uncertainty maps of (**a**) vegetation subtypes (DzB), (**b**) vegetation classes (NZ), and (**c**) habitats (NZ). Little agreement represents 0 ≤ CI ≤ 0.66; partial agreement represents 0.66 < CI ≤ 0.90; and complete agreement represents 0.90 < CI ≤ 1.

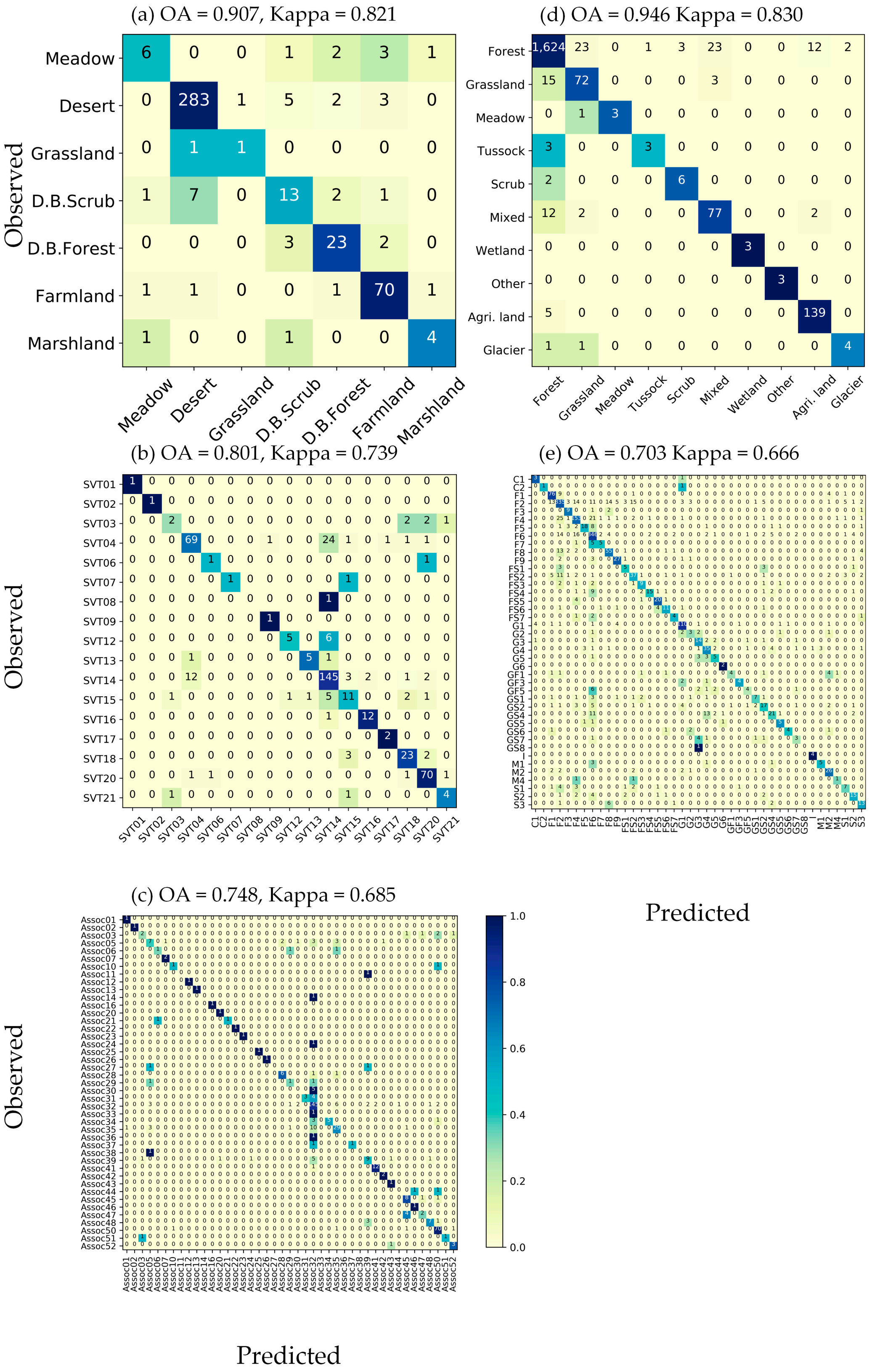

Figure A2. Confusion matrices of (**a**) vegetation types (DzB), (**b**) vegetation subtypes (DzB), (**c**) associations (DzB), (**d**) habitats (NZ), and (**e**) vegetation classes (NZ) based on the test sets. The meanings of vegetation class abbreviations were explained in Table S1 (DzB) and S2 (NZ) in Supplementary Materials.

References

1. Assessment, M.E. Millennium ecosystem assessment. In *Ecosystems and Human Well-Being: Biodiversity Synthesis*; World Resources Institute: Washington, DC, USA, 2005.
2. Diaz, S.; Pascual, U.; Stenseke, M.; Martin-Lopez, B.; Watson, R.T.; Molnar, Z.; Hill, R.; Chan, K.M.A.; Baste, I.A.; Brauman, K.A.; et al. Assessing nature's contributions to people. *Science* **2018**, *359*, 270–272. [CrossRef] [PubMed]
3. Pereira, H.M.; Ferrier, S.; Walters, M.; Geller, G.N.; Jongman, R.H.G.; Scholes, R.J.; Bruford, M.W.; Brummitt, N.; Butchart, S.H.M.; Cardoso, A.C.; et al. Essential Biodiversity Variables. *Science* **2013**, *339*, 277–278. [CrossRef] [PubMed]
4. Wang, C.Y.; Guo, Z.Z.; Wang, S.T.; Wang, L.P.; Ma, C. Improving Hyperspectral Image Classification Method for Fine Land Use Assessment Application Using Semisupervised Machine Learning. *J. Spectrosc.* **2015**, *2015*, 969185. [CrossRef]
5. Forzieri, G.; Alkama, R.; Miralles, D.G.; Cescatti, A. Satellites reveal contrasting responses of regional climate to the widespread greening of Earth. *Science* **2017**, *356*, 1140–1144. [CrossRef] [PubMed]
6. Staver, A.C.; Archibald, S.; Levin, S.A. The Global Extent and Determinants of Savanna and Forest as Alternative Biome States. *Science* **2011**, *334*, 230–232. [CrossRef] [PubMed]
7. Küchler, A.W. *Vegetation Mapping*; Ronald Press Co.: New York, NY, USA, 1967; pp. 853–855.
8. Malatesta, L.; Attorre, F.; Altobelli, A.; Adeeb, A.; De Sanctis, M.; Taleb, N.M.; Scholte, P.T.; Vitale, M. Vegetation mapping from high-resolution satellite images in the heterogeneous arid environments of Socotra Island (Yemen). *J. Appl. Remote Sens.* **2013**, *7*, 073527. [CrossRef]
9. Pettorelli, N.; Laurance, W.F.; O'Brien, T.G.; Wegmann, M.; Nagendra, H.; Turner, W. Satellite remote sensing for applied ecologists: Opportunities and challenges. *J. Appl. Ecol.* **2014**, *51*, 839–848. [CrossRef]
10. Defourny, P.; Kirches, G.; Brockmann, C.; Boettcher, M.; Peters, M.; Bontemps, S.; Lamarche, C.; Schlerf, M.; Santoro, M. Land Cover CCI. Product User Guide Version 2. Available online: http://maps.elie.ucl.ac.be/CCI/viewer/download.php (accessed on 25 June 2019).
11. Giri, C.; Zhu, Z.; Reed, B. A comparative analysis of the Global Land Cover 2000 and MODIS land cover data sets. *Remote Sens. Environ.* **2005**, *94*, 123–132. [CrossRef]
12. Naghibi, S.A.; Pourghasemi, H.R.; Dixon, B. GIS-based groundwater potential mapping using boosted regression tree, classification and regression tree, and random forest machine learning models in Iran. *Environ. Monit. Assess.* **2016**, *188*, 44. [CrossRef]
13. Franklin, J. Predictive vegetation mapping: Geographic modelling of biospatial patterns in relation to environmental gradients. *Prog. Phys. Geogr.* **1995**, *19*, 474–499. [CrossRef]
14. Gorelick, N.; Hancher, M.; Dixon, M.; Ilyushchenko, S.; Thau, D.; Moore, R. Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sens. Environ.* **2017**, *202*, 18–27. [CrossRef]
15. Jordan, M.I.; Mitchell, T.M. Machine learning: Trends, perspectives, and prospects. *Science* **2015**, *349*, 255–260. [CrossRef] [PubMed]
16. Breckling, B.; Dong, Q. Uncertainty in Ecology and Ecological Modelling. In *Handbook of Ecosystem Theories and Management*; CRC Press: Boca Raton, FL, USA, 2000; p. 51.
17. Zhang, C.Y.; Selch, D.; Cooper, H. A Framework to Combine Three Remotely Sensed Data Sources for Vegetation Mapping in the Central Florida Everglades. *Wetlands* **2016**, *36*, 201–213. [CrossRef]
18. Su, L.H. Optimizing support vector machine learning for semi-arid vegetation mapping by using clustering analysis. *ISPRS J. Photogramm. Remote Sens.* **2009**, *64*, 407–413. [CrossRef]
19. Zhang, C.Y.; Xie, Z.X. Object-based Vegetation Mapping in the Kissimmee River Watershed Using HyMap Data and Machine Learning Techniques. *Wetlands* **2013**, *33*, 233–244. [CrossRef]
20. De Colstoun, E.C.B.; Story, M.H.; Thompson, C.; Commisso, K.; Smith, T.G.; Irons, J.R. National Park vegetation mapping using multitemporal Landsat 7 data and a decision tree classifier. *Remote Sens. Environ.* **2003**, *85*, 316–327. [CrossRef]
21. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.
22. Mitchell, R.; Frank, E. Accelerating the XGBoost algorithm using GPU computing. *PeerJ Comput. Sci.* **2017**, *3*, e127. [CrossRef]

23. Dong, H.; Xu, X.; Wang, L.; Pu, F. Gaofen-3 PolSAR Image Classification via XGBoost and Polarimetric Spatial Information. *Sensors* **2018**, *18*, 611. [CrossRef]
24. Sandino, J.; Pegg, G.; Gonzalez, F.; Smith, G. Aerial Mapping of Forests Affected by Pathogens Using UAVs, Hyperspectral Sensors, and Artificial Intelligence. *Sensors* **2018**, *18*, 944. [CrossRef]
25. Man, C.D.; Nguyen, T.T.; Bui, H.Q.; Lasko, K.; Nguyen, T.N.T. Improvement of land-cover classification over frequently cloud-covered areas using Landsat 8 time-series composites and an ensemble of supervised classifiers. *Int. J. Remote Sens.* **2018**, *39*, 1243–1255. [CrossRef]
26. Zhong, L.H.; Hu, L.N.; Zhou, H. Deep learning based multi-temporal crop classification. *Remote Sens. Environ.* **2019**, *221*, 430–443. [CrossRef]
27. Hirayama, H.; Sharma, R.C.; Tomita, M.; Hara, K. Evaluating multiple classifier system for the reduction of salt-and-pepper noise in the classification of very-high-resolution satellite images. *Int. J. Remote Sens.* **2019**, *40*, 2542–2557. [CrossRef]
28. Jiang, H.; Li, D.; Jing, W.L.; Xu, J.H.; Huang, J.X.; Yang, J.; Chen, S.S. Early Season Mapping of Sugarcane by Applying Machine Learning Algorithms to Sentinel-1A/2 Time Series Data: A Case Study in Zhanjiang City, China. *Remote Sens.* **2019**, *11*, 861. [CrossRef]
29. Liu, Y.; Guo, Q.; Tian, Y. A software framework for classification models of geographical data. *Comput. Geosci.* **2012**, *42*, 47–56. [CrossRef]
30. Ferrier, S.; Guisan, A. Spatial modelling of biodiversity at the community level. *J. Appl. Ecol.* **2006**, *43*, 393–404. [CrossRef]
31. Ferrier, S. Mapping spatial pattern in biodiversity for regional conservation planning: Where to from here? *Syst. Biol.* **2002**, *51*, 331–363. [CrossRef] [PubMed]
32. Whittaker, R.H. Classification of natural communities. *Bot. Rev.* **1962**, *28*, 1–239. [CrossRef]
33. Whittaker, R.H. *Ordination and Classification of Communities*; Junk: The Hague, The Netherlands, 1973; Volume 5.
34. Somerfield, P.J. Identification of the Bray-Curtis similarity index: Comment on Yoshioka. *Mar. Ecol. Prog. Ser.* **2008**, *372*, 303–306. [CrossRef]
35. Zhengyi, W. *Chinese Vegetation*; Science Press: Beijing, China, 1980.
36. Myers, N.; Mittermeier, R.A.; Mittermeier, C.G.; da Fonseca, G.A.B.; Kent, J. Biodiversity hotspots for conservation priorities. *Nature* **2000**, *403*, 853–858. [CrossRef]
37. Wardle, P. *Vegetation of New Zealand*; CUP Archive: Cambridge, UK, 1991.
38. Wiser, S.K.; Thomson, F.J.; De Caceres, M. Expanding an existing classification of New Zealand vegetation to include non-forested vegetation. *N. Z. J. Ecol.* **2016**, *40*, 160–178. [CrossRef]
39. GBIF.org. Global Biodiversity Information Facility. 2018. Available online: https://www.gbif.org/ (accessed on 25 June 2019).
40. Newsome, P.F.J. *Vegetative Cover Map of New Zealand*, 2nd ed.; National Water and Soil Conservation Authority by the Water and Soil Directorate: Wellington, New Zealand, 1987.
41. Hall, D.K.; Riggs, G.A.; Salomonson, V.V.; Digirolamo, N.E.; Bayr, K.J. MODIS snow-cover products. *Remote Sens. Environ.* **2002**, *83*, 181–194. [CrossRef]
42. Hansen, M.C.; Defries, R.S.; Townshend, J.R.G.; Sohlberg, R.; Dimiceli, C.; Carroll, M. Towards an operational MODIS continuous field of percent tree cover algorithm: Examples using AVHRR and MODIS data. *Remote Sens. Environ.* **2002**, *83*, 303–319. [CrossRef]
43. Nicholson, S.E.; Davenport, M.L.; Malo, A.R. A comparison of the vegetation response to rainfall in the Sahel and East Africa, using normalized difference vegetation index from NOAA AVHRR. *Clim. Chang.* **1990**, *17*, 209–241. [CrossRef]
44. Myneni, R.; Knyazikhin, Y.; Park, T. MOD15A2H MODIS/Terra Leaf Area Index/FPAR 8-Day L4 Global 500 m SIN Grid V006. NASA EOSDIS Land Processes DAAC. 2015. Available online: https://lpdaac.usgs.gov/dataset_discovery/modis/modis_products_table/mod15a2h_v006 (accessed on 16 October 2016).
45. Zhengming, W. *MODIS Land Surface Temperature Products Users' Guide*; University of California: Santa Barbara, CA, USA, 2013.
46. Qi, J.; Chehbouni, A.; Huete, A.R.; Kerr, Y.H.; Sorooshian, S. A Modified Soil Adjusted Vegetation Index. *Remote Sens. Environ.* **1994**, *48*, 119–126. [CrossRef]

47. Rahimzadeh-Bajgiran, P.; Omasa, K.; Shimizu, Y. Comparative evaluation of the Vegetation Dryness Index (VDI), the Temperature Vegetation Dryness Index (TVDI) and the improved TVDI (iTVDI) for water stress detection in semi-arid regions of Iran. *ISPRS J. Photogramm. Remote Sens.* **2012**, *68*, 1–12. [CrossRef]
48. 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]
49. Huo, A.D.; Chen, X.H.; Li, H.K.; Hou, M.; Hou, X.J. Development and testing of a remote sensing-based model for estimating groundwater levels in aeolian desert areas of China. *Can. J. Soil Sci.* **2011**, *91*, 29–37. [CrossRef]
50. Rao, B.R.M.; Sankar, T.R.; Dwivedi, R.S.; Thammappa, S.S.; Venkataratnam, L.; Sharma, R.C.; Das, S.N. Spectral Behavior of Salt-Affected Soils. *Int. J. Remote Sens.* **1995**, *16*, 2125–2136. [CrossRef]
51. Collado, A.D.; Chuvieco, E.; Camarasa, A. Satellite remote sensing analysis to monitor desertification processes in the crop-rangeland boundary of Argentina. *J. Arid Environ.* **2002**, *52*, 121–133. [CrossRef]
52. Chander, G.; Markham, B.L.; Helder, D.L. Summary of current radiometric calibration coefficients for Landsat MSS, TM, ETM+, and EO-1 ALI sensors. *Remote Sens. Environ.* **2009**, *113*, 893–903. [CrossRef]
53. Massetti, A.; Sequeira, M.M.; Pupo, A.; Figueiredo, A.; Guiomar, N.; Gil, A. Assessing the effectiveness of RapidEye multispectral imagery for vegetation mapping in Madeira Island (Portugal). *Eur. J. Remote Sens.* **2016**, *49*, 643–672. [CrossRef]
54. Beckschäfer, P.; Fehrmann, L.; Harrison, R.D.; Xu, J.; Kleinn, C. Mapping Leaf Area Index in subtropical upland ecosystems using RapidEye imagery and the randomForest algorithm. *IFor.-Biogeosci. For.* **2014**, *7*, 1–11. [CrossRef]
55. Huete, A.R. A Soil-Adjusted Vegetation Index (Savi). *Remote Sens. Environ.* **1988**, *25*, 295–309. [CrossRef]
56. Hall, D.K.; Riggs, G.A.; Salomonson, V.V. Development of Methods for Mapping Global Snow Cover Using Moderate Resolution Imaging Spectroradiometer Data. *Remote Sens. Environ.* **1995**, *54*, 127–140. [CrossRef]
57. Haralick, R.M.; Shanmugam, K.; Dinstein, I.H. Textural Features for Image Classification. *IEEE Trans. Syst. Man Cybern.* **1973**, *SMC-3*, 610–621. [CrossRef]
58. Zhu, C.; Yang, X. Study of remote sensing image texture analysis and classification using wavelet. *Int. J. Remote Sens.* **1998**, *19*, 3197–3203. [CrossRef]
59. Nickolls, J.; Buck, I.; Garland, M.; Skadron, K. Scalable parallel programming with CUDA. In Proceedings of the ACM SIGGRAPH 2008, Los Angeles, CA, USA, 11–15 August 2008; p. 16.
60. Pohl, C.; Van Genderen, J.L. Review article multisensor image fusion in remote sensing: Concepts, methods and applications. *Int. J. Remote Sens.* **1998**, *19*, 823–854. [CrossRef]
61. Chen, D.Y.; Brutsaert, W. Satellite-sensed distribution and spatial patterns of vegetation parameters over a tallgrass prairie. *J. Atmos. Sci.* **1998**, *55*, 1225–1238. [CrossRef]
62. Podest, E.; Saatchi, S. Application of multiscale texture in classifying JERS-1 radar data over tropical vegetation. *Int. J. Remote Sens.* **2002**, *23*, 1487–1506. [CrossRef]
63. Hutchinson, M.; Xu, T.; Houlder, D.; Nix, H.; McMahon, J. *ANUCLIM 6.0 User's Guide*; Australian National University: Canberra, Australia, 2009.
64. Kriticos, D.J.; Jarošik, V.; Ota, N. Extending the suite of bioclim variables: A proposed registry system and case study using principal components analysis. *Methods Ecol. Evol.* **2014**, *5*, 956–960. [CrossRef]
65. Fick, S.E.; Hijmans, R.J. WorldClim 2: New 1-km spatial resolution climate surfaces for global land areas. *Int. J. Climatol.* **2017**, *37*, 4302–4315. [CrossRef]
66. Thornthwaite, C.W. An approach toward a rational classification of climate. *Geogr. Rev.* **1948**, *38*, 55–94. [CrossRef]
67. Thornthwaite, C.W. The water balance. *Publ. Clim.* **1957**, *8*, 1–104.
68. Fang, J.; Yoda, K. Climate and vegetation in China II. Distribution of main vegetation types and thermal climate. *Ecol. Res.* **1989**, *4*, 71–83. [CrossRef]
69. Kira, T. *A New Classification of Climate in Eastern Asia as the Basis for Agricultural Geography*; Horticultural Institute Kyoto University: Kyoto, Japan, 1945.
70. Hengl, T.; de Jesus, J.M.; MacMillan, R.A.; Batjes, N.H.; Heuvelink, G.B.M.; Ribeiro, E.; Samuel-Rosa, A.; Kempen, B.; Leenaars, J.G.B.; Walsh, M.G.; et al. SoilGrids1km-Global Soil Information Based on Automated Mapping. *PLoS ONE* **2014**, *9*, e105992. [CrossRef] [PubMed]

71. Sethian, J.A. A fast marching level set method for monotonically advancing fronts. *Proc. Natl. Acad. Sci. USA* **1996**, *93*, 1591–1595. [CrossRef] [PubMed]
72. Pekel, J.-F.; Cottam, A.; Gorelick, N.; Belward, A.S. High-resolution mapping of global surface water and its long-term changes. *Nature* **2016**, *540*, 418. [CrossRef]
73. USGS. *Landsat 8 (L8) Data Users Handbook*; USGS: Reston, VA, USA, 2015; Volume 1.
74. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
75. Georganos, S.; Grippa, T.; Vanhuysse, S.; Lennert, M.; Shimoni, M.; Kalogirou, S.; Wolff, E. Less is more: Optimizing classification performance through feature selection in a very-high-resolution remote sensing object-based urban application. *Gisci. Remote Sens.* **2018**, *55*, 221–242. [CrossRef]
76. Sandino, J.; Gonzalez, F.; Mengersen, K.; Gaston, K.J. UAVs and Machine Learning Revolutionising Invasive Grass and Vegetation Surveys in Remote Arid Lands. *Sensors* **2018**, *18*, 605. [CrossRef]
77. Breiman, L. Bagging predictors. *Mach. Learn.* **1996**, *24*, 123–140. [CrossRef]
78. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
79. Dietterich, T.G. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Mach. Learn.* **2000**, *40*, 139–157. [CrossRef]
80. Bryll, R.; Gutierrez-Osuna, R.; Quek, F. Attribute bagging: Improving accuracy of classifier ensembles by using random feature subsets. *Pattern Recognit.* **2003**, *36*, 1291–1302. [CrossRef]
81. Rokach, L. Ensemble-based classifiers. *Artif. Intell. Rev.* **2010**, *33*, 1–39. [CrossRef]
82. Kohavi, R.; John, G.H. Wrappers for feature subset selection. *Artif. Intell.* **1997**, *97*, 273–324. [CrossRef]
83. Pudil, P.; Novovicova, J.; Kittler, J. Floating Search Methods in Feature-Selection. *Pattern Recognit. Lett.* **1994**, *15*, 1119–1125. [CrossRef]
84. Somol, P.; Pudil, P.; Novovicova, J.; Paclik, P. Adaptive floating search methods in feature selection. *Pattern Recognit. Lett.* **1999**, *20*, 1157–1163. [CrossRef]
85. Guyon, I.; Elisseeff, A. An introduction to variable and feature selection. *J. Mach. Learn. Res.* **2003**, *3*, 1157–1182.
86. Chandrashekar, G.; Sahin, F. A survey on feature selection methods. *Comput. Electr. Eng.* **2014**, *40*, 16–28. [CrossRef]
87. Nakariyakul, S.; Casasent, D.P. An improvement on floating search algorithms for feature subset selection. *Pattern Recognit.* **2009**, *42*, 1932–1940. [CrossRef]
88. Wu, B. *Land Cover of China*; Science Press: Beijing, China, 2017.
89. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
90. Warmerdam, F. The geospatial data abstraction library. In *Open Source Approaches in Spatial Data Handling*; Springer: New York, NY, USA, 2008; pp. 87–104.
91. Murray, N.J.; Keith, D.A.; Simpson, D.; Wilshire, J.H.; Lucas, R.M. REMAP: An online remote sensing application for land cover classification and monitoring. *Methods Ecol. Evol.* **2018**, *9*, 2019–2027. [CrossRef]
92. Álvarez-Martínez, J.M.; Jiménez-Alfaro, B.; Barquín, J.; Ondiviela, B.; Recio, M.; Silió-Calzada, A.; Juanes, J.A. Modelling the area of occupancy of habitat types with remote sensing. *Methods Ecol. Evol.* **2017**, *9*, 580–593. [CrossRef]
93. Hengl, T.; Walsh, M.G.; Sanderman, J.; Wheeler, I.; Harrison, S.P.; Prentice, I.C. Global mapping of potential natural vegetation: An assessment of Machine Learning algorithms for estimating land potential. *PeerJ* **2018**, *6*, e5457. [CrossRef] [PubMed]
94. Elith, J.; Leathwick, J.R. Species Distribution Models: Ecological Explanation and Prediction across Space and Time. *Annu. Rev. Ecol. Evol. Syst.* **2009**, *40*, 677–697. [CrossRef]
95. Lloyd, D. A phenological classification of terrestrial vegetation cover using shortwave vegetation index imagery. *Int. J. Remote Sens.* **1990**, *11*, 2269–2279. [CrossRef]
96. Blaschke, T. Object based image analysis for remote sensing. *ISPRS J. Photogramm. Remote Sens.* **2010**, *65*, 2–16. [CrossRef]
97. Scarpa, G.; Gargiulo, M.; Mazza, A.; Gaetano, R. A CNN-based fusion method for feature extraction from sentinel data. *Remote Sens.* **2018**, *10*, 236. [CrossRef]

98. Clark, P.E.; Hardegree, S.P. Quantifying vegetation change by point sampling landscape photography time series. *Rangel. Ecol. Manag.* **2005**, *58*, 588–597. [CrossRef]
99. Michel, P.; Mathieu, R.; Mark, A.F. Spatial analysis of oblique photo-point images for quantifying spatio-temporal changes in plant communities. *Appl. Veg. Sci.* **2010**, *13*, 173–182. [CrossRef]
100. Roush, W.; Munroe, J.S.; Fagre, D.B. Development of a spatial analysis method using ground-based repeat photography to detect changes in the alpine treeline ecotone, Glacier National Park, Montana, USA. *Arct. Antarct. Alp. Res.* **2007**, *39*, 297–308. [CrossRef]
101. Dickinson, J.L.; Zuckerberg, B.; Bonter, D.N. Citizen Science as an Ecological Research Tool: Challenges and Benefits. *Annu. Rev. Ecol. Evol. Syst.* **2010**, *41*, 149–172. [CrossRef]
102. Brown, T.B.; Hultine, K.R.; Steltzer, H.; Denny, E.G.; Denslow, M.W.; Granados, J.; Henderson, S.; Moore, D.; Nagai, S.; SanClements, M.; et al. Using phenocams to monitor our changing Earth: Toward a global phenocam network. *Front. Ecol. Environ.* **2016**, *14*, 84–93. [CrossRef]
103. Sullivan, B.L.; Aycrigg, J.L.; Barry, J.H.; Bonney, R.E.; Bruns, N.; Cooper, C.B.; Damoulas, T.; Dhondt, A.A.; Dietterich, T.; Farnsworth, A. The eBird enterprise: An integrated approach to development and application of citizen science. *Biol. Conserv.* **2014**, *169*, 31–40. [CrossRef]
104. Kosmala, M.; Crall, A.; Cheng, R.; Hufkens, K.; Henderson, S.; Richardson, A.D. Season Spotter: Using Citizen Science to Validate and Scale Plant Phenology from Near-Surface Remote Sensing. *Remote Sens.* **2016**, *8*, 726. [CrossRef]
105. Keckler, S.W.; Dally, W.J.; Khailany, B.; Garland, M.; Glasco, D. GPUs and the future of parallel computing. *IEEE Micro* **2011**, *31*, 7–17. [CrossRef]