Remote Sensing of Environment 233 (2019) 111382

Contents lists available at ScienceDirect

Remote Sensing of Environment

journal homepage: www.elsevier.com/locate/rse

Assimilating optical satellite remote sensing images and field data to predict surface indicators in the Western U.S.: Assessing error in satellite predictions based on large geographical datasets with the use of machine learning

Junzhe Zhang, Gregory S. Okin*, Bo Zhou

Department of Geography, University of California, Los Angeles, CA 90095, USA

ARTICLE INFO

Edited by Jing M. Chen

Keywords:
Random forest
BLM
AIM and LMF
Drylands
Data assimilation
Remote sensing
k-Fold cross-validation

ABSTRACT

Indicators of vegetation composition, vegetation structure, bare ground cover, and gap size in drylands potentially gives information about the condition of ecosystems, in part because they are strongly related to factors such as erosion, wildlife habitat characteristics, and the suitability for some land uses. Field data collection based on points does not produce spatially continuous information about surface indicators and cannot cover vast geographic areas. Remote sensing is possibly a labor- and time-saving method to estimate important biophysical indicators of vegetation and surface condition at both temporal and spatial scales impossible with field methods. Regression models based on machine learning algorithms, such as random forest (RF), can build relationships between field and remotely sensed data, while also providing error estimates. In this study, field data including over 15,000 points from the Assessment, Inventory, and Monitoring (AIM) and Landscape Monitoring Framework (LMF) programs on Bureau of Land Management (BLM) lands throughout the Western U.S., Moderate Resolution Imaging Spectroradiometer (MODIS) bidirectional reflectance distribution function (BRDF) parameters, MODIS nadir BRDF-adjusted reflectance (NBAR), and Landsat 8 Operational Land Imager (OLI) surface reflectance products with ancillary data were used as predictor variables in a k-fold cross-validation approach to RF modeling. RF regression models were built to predict fourteen indicators of vegetation cover and height, as well as bare gap parameters. The RF model estimates exhibited good correlations with independent samples, with a low bias and a low RMSE. External cross-validation showed good agreement with out-of-bag (OOB) errors produced by RF and also allowed mapping prediction uncertainty. Predicted distribution maps of the surface indicators were produced by using these relationships across the arid and semiarid Western U.S. The bias and RMSE distribution maps show that the sample insufficiency and unevenly pattern of sample strongly impact the accuracy of the RF regression and prediction. The results from this study clearly show the utility of RF as a means to estimate multiple dryland surface indicators from remotely sensed data, and the reliability of the OOB errors in assessing the accuracy of the predictions.

1. Introduction

The Western U.S. is largely composed of arid and semiarid lands that provide a variety of important ecosystem goods and services, but land degradation in these areas, a critical global issue in the 21st century (Bestelmeyer et al., 2015), can be severe. In the mostly dry Western U.S. vegetation can be sparse and composed of a mix of life forms (i.e., woody and herbaceous plants), often with a considerable amount of non-photosynthetic vegetative material. This complex vegetation structure and bare soil cover are important in regard to the functioning of these lands. For instance, large amounts of bare connected soil can make these environments susceptible to erosion by wind and water (Ludwig et al., 2007; Okin et al., 2009). As a result of the need for monitoring, specialized biophysical surface indicators of vegetation and surface condition have been developed for dryland ecosystems (e.g., Herrick et al., 2015). The large variations in bare soil cover, vegetation cover, and vegetation structure at different landscape levels are strongly related to erosion, determine wildlife habitat characteristics, and control the suitability for some land uses, making the use of multiple indicators critical in the monitoring and management of lands in the Western U.S. and elsewhere (Herrick et al., 2010; Knippertz and Stuut, 2014). As the largest manager of land in the arid and semiarid Western U.S., the Bureau of Land Management (BLM) has developed the Assessment, Inventory, and Monitoring (AIM) and Landscape

* Corresponding author at: 315 Portola Plaza, Department of Geography, UCLA, CA 90095, USA.
E-mail address: okin@geog.ucla.edu (G.S. Okin).

https://doi.org/10.1016/j.rse.2019.111382
Received 2 July 2019; Received in revised form 11 August 2019; Accepted 20 August 2019
Available online 09 September 2019

Monitoring Framework (LMF) programs to systematically collect information on lands it manages throughout the western states (MacKinnon et al., 2011).

In situ observation is a commonly used method to measure surface conditions. Field data collection based on points does not provide spatially or temporally continuous information about the surface and is susceptible to under-sampling, even in a relatively small area (Karl et al., 2014). Moreover, measuring surface indicators in situ is time-consuming and laborious, especially in harsh or remote areas (Elzinga et al., 1998; Holthausen et al., 2005). Remote sensing is a practical method for detecting surface indicators at different temporal and spatial scales within a short time frame (Sun et al., 2008). Several studies have shown that assimilating satellite remote sensing images and field data can generate surface indicators at relatively large landscape scales (Booth and Cox, 2009; Jones et al., 2018; Karl et al., 2012; Laliberte et al., 2004; Luscier et al., 2006; McCord et al., 2017). However, remote sensing techniques may have difficulty measuring all surface indicators with the required accuracy and precision (Marsett et al., 2006). In addition, the relationship between surface reflectance and surface indicators is usually nonlinear (Duniway et al., 2012) and the spatial cover and temporal density of field data collection can be limited. The small training sample size and nonlinearity make the use of traditional regression approaches problematic for assimilating remote sensing images and field data (Duniway et al., 2012; Liang et al., 2012). Machine learning algorithms, on the other hand, which were developed first by artificial intelligence scientists, excel at solving nonlinear problems and can overcome the issue caused by small sample size (Lary et al., 2016). As examples, a Bayesian additive regression tree (BART) model has been applied to estimate six surface indicators in Northern California and Nevada based on AIM data and high spatial resolution satellite images (McCord et al., 2017); a 30-m annual vegetation map of the Western U.S. was created based on remotely sensed and field data by using the random forest (RF) regression approach (Jones et al., 2018); and a 100-m soil property and class maps of the U.S. was generated based on land cover and gSSURGO polygon data by using a tree-based regression model (Ramcharan et al., 2018).

In our study, a RF regression model based on the Frequentist framework (Breiman, 2001; Herrick et al., 2010; Jones et al., 2018; Leenaars et al., 2017; McCord et al., 2017; Ramcharan et al., 2018) was employed to derive the relationships between AIM and LMF field data and remotely sensed data, combined with ancillary data. We added bidirectional reflectance distribution function (BRDF) parameter products to our machine learning-based regression method to help retrieve the structural indicators (i.e., plant height and bare soil gap size), as BRDF is sensitive not just to surface brightness, but surface architecture as well, and therefore potentially correlates better with structural indicators (Li and Strahler, 1986; Jones et al., 2018). The objectives of this study were to: (1) build RF regression models for fourteen biophysical surface indicators of vegetation and surface condition, (2) apply the resulting RF regression models to generate predicted distribution maps for these indicators across the arid and semiarid Western U.S., and (3) provide external k-fold cross-validation estimates of error and map the error distribution in the study area. The addition of external cross-validation, as opposed to the internal cross-validation that inherent to the RF approach, provides the opportunity to understand the limitations on RF predictions in conditions more closely approximating what a land manager might experience. They also provide the opportunity to produce geographical estimates of error to better represent geographical variability in the quality of estimates which may be used to better contextualize predictions or to prioritize the location of new measurements.

2. Random forest algorithm

2.1. Decision tree-based modeling

Decision trees are commonly used classification and regression methods in remote sensing analysis (Friedl and Brodley, 1997). Decision trees create tree-like models in which each internal node represents a test on an independent variable, each branch represents a criterion of the test, and each 'leaf' represents a result of the test. The tests measure the homogeneity (i.e., Gini impurity or mean squared error, MSE) between a descendant and its parent variable. If the root (i.e., dependent variable) is categorical, this approach yields a "classification tree". If the root (i.e., dependent variable) is continuous, this approach yields a "regression tree". Because the AIM and LMF surface indicators are continuous, we only concern ourselves here with the regression tree approach. The most well-known regression tree approach is the classification and regression tree (CART) algorithm (Breiman et al., 1984; Zhang et al., 2013; Zhang et al., 2012), which is a nonparametric algorithm that recursively partitions the dataset through simple regression models into increasingly smaller subsets by the same splitting decision (i.e., core function). Each simple regression model (i.e., regression plane) only has one dependent variable and the relationship between the regression planes is nonlinear.

Although CART shows good performance in regression, another approach has been to employ an ensemble regression tree model (i.e., additive trees or 'forests'), which is an algorithm that synthesizes multiple related but different models, to improve the accuracy and precision of predictive analytics (Lary, 2010). Specifically, for the regression tree model, two or more regression trees are built based on different subsets of training samples (Dietterich, 2000). The final result of the ensemble regression tree model is the weighted result based on the outcome of each tree. RF (Breiman, 2001) and BART (Chipman et al., 2010) are the two most commonly used ensemble regression tree approaches.

2.2. Random forest regression model

RF has been successfully used for regression in many disciplines (Pal, 2005) and is characterized by the bagging (i.e., bootstrap aggregating) approach (Breiman, 1996). RF has three qualities that recommend its use here. First, RF builds multiple regression trees independently by using different bootstrapped sample subsets of training samples (Steinberg and Colla, 2009). There may be some outliers in one of the bootstrapped sample subsets, but each tree relies on its own subset, so the sensitivity of RF to outliers is reduced. Second, each node of a tree is split by using a randomly chosen independent variable among the entire set of independent variable (Liaw and Wiener, 2002), and RF chooses the subset of trees with the least error as the final output, making RF robust against overfitting (Rodriguez-Galiano et al., 2012). Third, in the bagging approach, RF randomly chooses sample subsets from the training samples with replacement (i.e., bootstrap), which means that even a small dataset can be sampled multiple times making RF resilient to sample insufficiency (Breiman, 2001). Although RF has some advantages compared to other machine learning regression algorithms, it is difficult to use it in datasets with missing data (Pal, 2005). One advantage of the AIM/LMF sampling approach is that there are very few missing data. Nonetheless, the standard approach of RF in these cases is to separate a dependent variable with missing data into two dependent variables: one continuous variable consisting of the present data and one categorical variable that labels missing data (present data are marked as 1 and missing data are marked as 0). However, adding the categorical dependent variable reduces the importance of present data and impacts the measurement of the homogeneity (i.e., Gini impurity or MSE) of the original dependent variable (Murphy, 2012).

RF has two hyperparameters (parameters that control

implementation of the algorithm, in contrast to parameters that are determined through running the algorithm) to control tree growth: the maximum number of independent variables for each tree and the number of trees used to produce the forest. The maximum number of independent variables controls the depth of the tree and is tuned to generate the most efficient expression of the model. The number of trees controls the size of the forest and is tuned to find enough trees to improve the accuracy of the model without overly increasing computational costs. Here, MSE was used as the criterion to split the tree. If the MSE of the descendant node is smaller than a threshold, then the branch stops growing and the leaf of this branch is a possible outcome of the tree. If the MSE of the descendant node is greater than a threshold, then the node is the parent node for the next descendant node. Inherent to the RF approach is out-of-bag (OOB) cross-validation in which samples are divided into different training (usually 70–80%) and validation (usually 20%–30%) sets in the production of each tree, from which an overall OOB error can be estimated (Breiman, 2001). In this approach, all data are eventually used to produce the final tree, and thus OOB errors tend to underestimate the true error in the predictions.

2.3. k-Fold cross-validation

A complementary approach to cross-validation is external cross-validation, in which small, random samples (usually 5%–10%) are withheld from the predictions entirely, and then error is estimated based on the final model's ability to predict these withheld points. This typically done some number, *k*, of times (i.e., k-fold cross-validation) and can be done with or without replacement. k-Fold cross-validation is a commonly used cross-validation method for a machine learning algorithm. The aim of k-fold cross-validation is to employ unseen data to estimate the performance of an algorithm (James et al., 2013; Russell and Norvig, 2011). Because the omitted data are not included in the production of the model, these external estimates of error are higher than the OOB errors, but better reflect the error that might be expected by a user of the model (Roberts et al., 2017; Segal, 2004; Svetnik et al., 2003). Thus, the advantage of k-fold cross-validation is that it can utilize all samples as training and testing samples, which leads a less biased or less optimistic estimate for the performance of the machine learning algorithm (Kuhn and Johnson, 2013). In addition, this approach to cross-validation means that *k* different predictions are made, allowing the production of distribution of estimates that can be used as an indicator of the precision of the estimates. In our study, because each data point is associated with a geographical location, we can make geographically explicit estimates of error which potentially has utility in prioritizing the location of new measurements or spatially contextualizing the quality of a prediction in a certain area.

3. Data and methodology

3.1. Study area

AIM and LMF measurements taken from 2013 to 2017 in eleven states in the Western U.S. were used in this study (Fig. 1). Generally, the Western U.S. has an arid and semiarid climate; however, the west coast of California has a Mediterranean climate (Westerling et al., 2006). Deserts, semiarid and arid areas, and mountains make up most of the land cover in the area. The main types of vegetation are grass and shrub with a small fraction of the forest (Loveland et al., 1991).

The study area covers about four hundred level IV ecoregions based on the National Gap Analysis Project (GAP) dataset according to McMahon et al. (2001). Within each ecoregion, the biotic (e.g., vegetation) and abiotic (e.g., climate) phenomena are similar (McMahon et al., 2001). In our study, we included every ecoregion that contains more than two AIM or LMF sites (Fig. 1). The urban areas, dry lakes, and lakes in those ecoregions were removed from the study area by using the GAP dataset (McMahon et al., 2001).

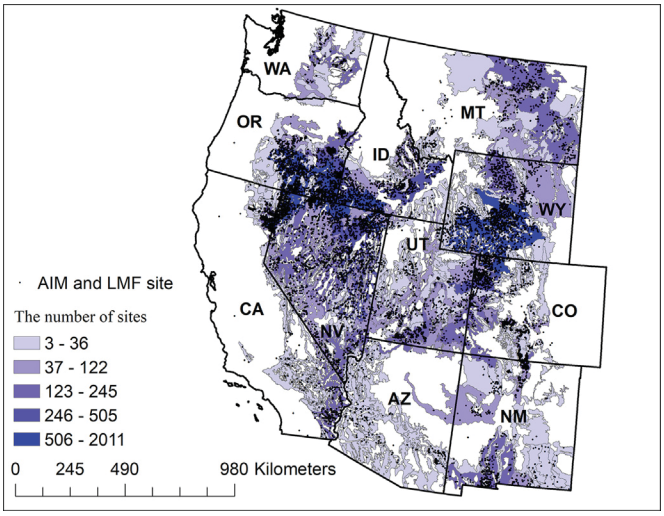

Fig. 1. The pattern of Assessment, Inventory, and Monitoring (AIM) and Landscape Monitoring Framework (LMF) sites in the study area (about 400 level IV ecoregions). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.2. Field data collection

The Landscape Approach Data Portal (https://landscape.blm.gov/geoportal) contains field measurements from BLM lands (MacKinnon et al., 2011). The AIM and LMF dataset has a total of 20 common surface indicators including bare ground cover, native vegetation cover, invasive vegetation cover, plant height, and soil stability, measured using consistent methods, such as line-point-intercept (LPI), gap intercept, and belt transect (Herrick et al., 2017; MacKinnon et al., 2011). AIM and LMF data were collected by different BLM surveying and mapping teams (MacKinnon et al., 2011). These teams contributed to approximately 50 projects in different states across the Western U.S. Fourteen surface indicators were selected for this study (Table 1).

3.3. Remote sensing products and ancillary data

Three types of remote sensing products and three types of ancillary data were used as independent variables in RF regression models (Table 2). These remote sensing products are Moderate Resolution Imaging Spectrometer (MODIS) BRDF parameters, MODIS nadir BRDF-adjusted reflectance (NBAR), and Landsat 8 Operational Land Imager (OLI) surface reflectance. MODIS data have 500-m resolution and OLI data have 30-m resolution. Ancillary data including climate variables, topographic variables, soil texture variables were also included. Each dataset covers the whole study area.

The MODIS BRDF parameters product (MCD43A1) contains the model kernels for each MODIS band obtained from the Ross Thick-Li Sparse BRDF model used by MODIS to characterize the angular distribution of reflected light (Schaaf et al., 2002). For each of the seven MODIS bands (Table 2), there are three kernel weights (i.e., isotropic, geometric, and volumetric). The isotropic kernel weight (k_{iso}) represents the bidirectional reflectance of a simple, flat isotropic scatter, the geometric kernel weight (k_{geo}) represents the bidirectional reflectance of a surface containing a large number of objects (plants), and the volumetric kernel weight (k_{vol}) represents the bidirectional reflectance of a homogeneous thick medium consisting of randomly located scattering plane facets with a particular volume density (Roujean et al., 1992). The MODIS NBAR product (MCD43A4) constitutes a prediction of reflectance viewed from the nadir in each MODIS band (Strahler et al., 1999). The Landsat 8 OLI surface reflectance product contains seven bands and has a much higher spatial resolution and superior noise characteristics compared to MODIS (Roy et al., 2014).

Table 1
The list of all surface indicators in this study.

Surface indicators	Description
Gap 25–50	The fraction of the transect comprised of bare soil gaps between 25 cm and 50 cm.
Gap 51–100	The fraction of the transect comprised of bare soil gaps between 51 cm and 100 cm.
Gap 101–200	The fraction of the transect comprised of bare soil gaps between 101 cm and 200 cm.
Gap 201–250	The fraction of the transect comprised of bare soil gaps between 201 cm and 250 cm
Gap > 250	The fraction of the transect comprised of bare soil gaps > 250 cm.
Bare soil cover	The percent bare ground cover.
Total vegetation cover	The percent canopy cover of herbaceous and woody plants (both invasive and non-invasive).
Sagebrush cover	The percent canopy cover of sagebrush.
Sagebrush height	The mean value of the heights of living or dead sagebrush.
Herbaceous height	The mean value of the heights of living or dead herbaceous plants.
NInvPerennial Grass	The percent canopy cover of non-invasive perennial grasses.
NInv Shrub	The percent canopy cover of non-invasive shrubs.
NInvPerennial Forb	The percent canopy cover of non-invasive perennial forbs.
InvAnnual Grass	The percent canopy cover of invasive annual grasses.

Table 2
Spatial resolution and number of predictors for remote sensing products and ancillary data.

Variable	Original spatial resolution (m)	Number of bands
MODIS BRDF parameters	500	30
MODIS NBAR	500	7
OLI surface reflectance	30	7
Vegetation indices	500 or 30	4
Climate	1000	3
Topography	90	3
Soil texture	1000	3

For each AIM or LMF site, the nearest neighbor, closest-in-time cloud-free MODIS (daily) and OLI (every 16 days) pixel value (surface reflectance of MODIS and Landsat 8 and the kernel weights of BRDF parameters) were extracted from images downloaded from Google Earth Engine.

Climate variables used in this study were monthly mean (30 days prior to the collection date of AIM or LMF samples) precipitation, monthly mean temperature, and monthly mean solar radiation. Monthly mean precipitation, temperature, and solar radiation were taken from the daily surface weather and climatological summaries (https://daymet.ornl.gov/), which have 1000-m resolution. To keep all climate variables dimensionless, all of them were converted to normalized value (Maclaurin et al., 2016) by using the following equation:

$$\hat{x} = \frac{x - \min(X)}{\max(X) - \min(X)}, \quad (1)$$

where $\hat{x}$ represents the normalized value at a certain pixel, x represents the original value at a certain pixel, X represents all the values in the entire study area. The range of all three climate variables used in the RF regression is therefore [0,1], which is also the range of the remote sensing variables.

Topographic variables used in this study were elevation, slope, and aspect derived from Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) Global Digital Elevation Model (GDEM) data with 90-m resolution (Fujisada et al., 2005), downloaded from Natural Resources Conservation Service (NRCS, https://datagateway.nrcs.usda.gov/). To keep all topographic variables dimensionless, elevation and slope were converted to normalized value using Eq. 1. Since the impact of aspect is different on the east side and west side, a -cosθ function was used at a pixel if the pixel had an eastern orientation (0°–179°), while a *-cos(360°-θ)* function was used if the pixel had a western orientation (180° to 359°) (Hafez et al., 2017; Smith, 1977). The range of all three topographic variables used in the RF regression is therefore [0,1].

Soil texture variables used in this study were the fraction of clay, silt, and sand in the topsoil layer, which were derived from USDA's State Soil Geographic Database (STATSGO) with 1000-m resolution (Miller and White, 1998), downloaded from NRCS (https://datagateway.nrcs.usda.gov/). The fractions of clay, silt, and sand are percentages and are therefore unitless. The range of all three soil texture variables used in the RF regression is therefore [0,1].

3.4. Random forest implementation and mapping

In RF, the remote sensing and ancillary variables were treated as independent variables ($X_1 \ldots X_n$) and each surface indicator was treated as a dependent variable (Y). Fourteen RF models based on the relationships between the fourteen surface indicators and remote sensing and ancillary variables were used to generate the predicted distribution maps. Python 3.6 with Scikit-learn 0.18 package (Pedregosa et al., 2011) was used to create the RF regression models and output of the resulting of cross-validation. ArcGIS 10.4 (ESRI, 380 New York Street, Redlands, CA 92373, USA) was used to extract the pixel value of remotely sensed and ancillary data and generate the predicted distribution maps.

After extraction of the values for remotely sensed and ancillary data (as discussed above), we conducted initial testing of RF models, and finally set 8 as the maximum number of independent variables in each tree and 100 as the number of trees to produce in the forest to provide a good balance between error reduction and computation time (Fig. 2).

In our implementation of k-fold cross-validation, we set k to 20 without replacement. In this approach, each sample was omitted exactly one time (i.e., 5% omission each time), in random order, and each surface indicator was predicted 20 times. This led to the production of 20 separate RF models for each of the fourteen surface indicators. RF models were produced with an 80%–20% split for training and validation to make out-of-bag (OOB) testing data, from which OOB error for each model could be calculated. Predictions of the fully omitted 5% of samples were then made. Thus, for each indicator, each sample point is associated with a single in situ value and a single prediction derived from a model in which the sample was not included in training data. Each sample point also has 19 (20 minus 1) predictions from models in which the point was included as training data. And ensemble mean of these 19 points was calculated as the average of these 19 values.

The contribution of each variable to the overall regression, as the total decrease in MSE from splitting on the variable based on the method of Scikit-learn 0.18 package (Pedregosa et al., 2011). The total decrease in MSE provides estimates of the importance of the variables in the RF model results. The contributions reported here are the average of each of the 20 RF models for each indicator.

To produce continuous prediction maps of each indicator, 20 individual maps were produced for each indicator, and the final prediction map was calculated as the mean of these 20 maps. Because AIM and LMF data were usually collected from June to September, remote

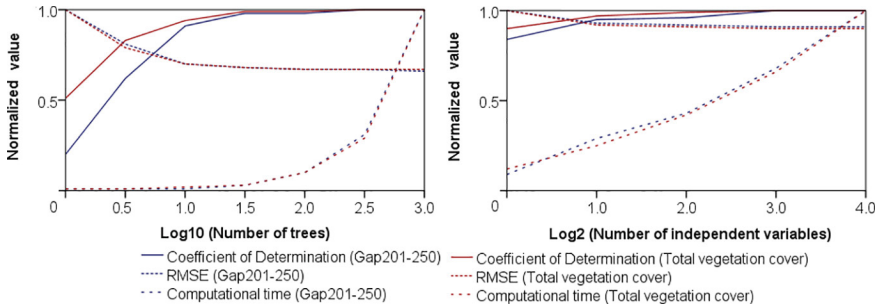

Fig. 2. RF regression model performance as a function of number of trees and the maximum number of independent variables (using the indicator of Gap 201–250 and total vegetation cover as examples). Y axis represents the value in the normalized scale, which is the ratio of the present value to the maximum value.

sensing images during the mid-summer were selected to predict the distribution of surface indicators. For the MODIS products, we chose MODIS BRDF parameters and NBAR products collected on July 20th, 2016 that covered the Western U.S. A sequence of images (June 13–August 6, 2016) was chosen to create a cloud-free OLI surface reflectance product to cover the Western U.S. We chose the summer of 2016 to make the prediction maps of surface indicators because this is the summer with the largest number of AIM and LMF points.

3.5. Error estimation and representation

Three statistics were employed to evaluate the relationship between model predictions and in situ measurements. The coefficient of determination (R^2) represents how well the RF model predictions correlated with in situ measurements:

$$R^2 = 1 - \frac{\sum_{t=1}^{n} (y_t - \hat{y}_t)^2}{\sum_{t=1}^{n} (y_t - \overline{y}_t)^2}, \quad (2)$$

where y_t represents a value of in situ measurement, $\overline{y}_t$ represents the mean value of in situ measurements, and $\hat{y}_t$ represents a value of the RF estimate. R^2 does not, however, provide an estimate for how well the RF model predicts the correct values. For that, additional error metrics are required. Mean error (ME) provides an estimate of the bias of the RF estimates:

$$ME = \frac{1}{n} \sum_{t=1}^{n} (\hat{y}_t - y_t). \quad (3)$$

The root-mean-square error (RMSE) provides an estimate of the overall error of the RF estimates:

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n} (\hat{y}_t - y_t)^2}{n}}. \quad (4)$$

These three statistics (i.e., R^2, ME, and RMSE) were calculated using both OOB samples and external (from the 5% of samples left out of each RF model) samples to produce estimates of OOB and external error, respectively.

3.6. Spatial characteristics of error

Because each in situ data point was geographically located and had associated with it a prediction where it was not included in the training data, our approach allows characterization of the spatial distribution predictions errors. Error (ME and RMSE) distributions for each indicator were estimated and mapped as the mean value of all MEs or RMSEs in each ecoregion in the study area.

4. Results

For the purposes of discussion, we only show five surface indicators, a mixture of structural and non-structural (i.e., Gap > 250, Bare soil cover, Total vegetation cover, Herbaceous height, and NInvPerennial Grass cover), representing indicators that were both well-predicted and poorly-predicted. Additional results for the other surface indicators may be found in the supplementary data.

4.1. Model evaluation

Our results indicate strong positive relationships exist between predicted values and in situ values, with external R^2 values ranging from 0.21 for the poorly-predicted variable to 0.70 for the well-predicted variables (Table 3 and Fig. 3). Most of the MEs of surface indicators are positive but only the ME of Gap > 250 is negative with absolute values lower than 0.57 (Fig. 3). Most indicators show the same pattern of over-prediction at low values and under-prediction at high

Table 3
Error metrics for individual surface indicators.

Surface indicator	Coefficient of determination			Mean error			RMSE		
	Internal	OOB	External	Internal	OOB	External	Internal	OOB	External
Gap 25–50	0.88	0.23	0.22	0.06	0.05	0.16	1.97	5.08	5.01
Gap 51–100	0.89	0.28	0.26	0.05	0.01	0.13	2.51	6.38	6.37
Gap 101–200	0.89	0.32	0.32	0.08	0.20	0.22	2.89	7.25	7.28
Gap 201–250	0.92	0.49	0.51	0.24	0.47	0.57	5.92	15.34	14.90
Gap > 250	0.92	0.45	0.46	−0.07	−0.53	−0.20	8.30	21.15	21.03
Bare soil cover	0.94	0.61	0.60	0.09	0.16	0.25	4.51	11.18	11.37
Total vegetation cover	0.95	0.70	0.71	0.02	0.10	0.09	4.82	12.22	12.16
Sagebrush cover	0.89	0.32	0.31	0.18	0.52	0.48	3.42	8.57	8.69
Sagebrush height	0.88	0.22	0.24	0.32	1.59	0.97	8.41	20.93	21.18
Herbaceous height	0.93	0.56	0.58	0.11	0.70	0.42	3.88	9.78	9.78
NInvPerennial Grass	0.92	0.47	0.46	0.19	0.67	0.57	5.29	13.26	13.38
NInv Shrub	0.89	0.25	0.27	0.17	0.52	0.37	4.18	10.66	10.55
NInvPerennial Forb	0.88	0.24	0.21	0.11	0.30	0.26	1.54	3.86	3.88
InvAnnual Grass	0.90	0.37	0.37	0.25	0.41	0.57	4.46	11.50	11.23

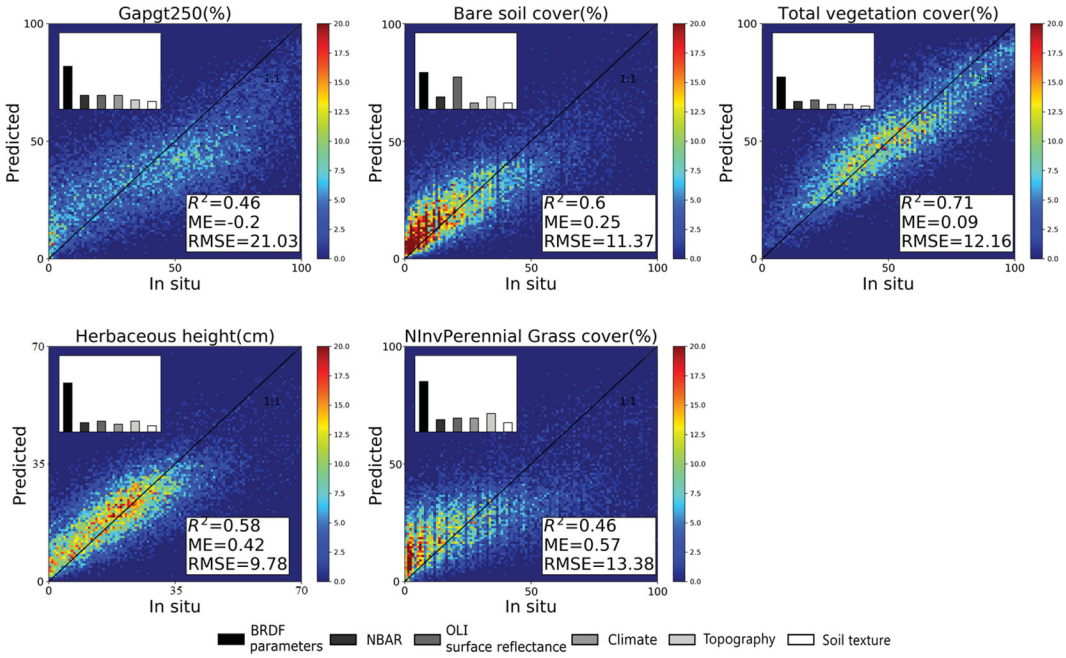

Fig. 3. Correlations between model-predicted external values, calculated using the external k-fold cross-validation, and in situ values of five surface indicators and the relative contributions of remote sensing and ancillary variables to the regressions (inset). The diagonal represents the 1:1 line. The color bare shows the density of points. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

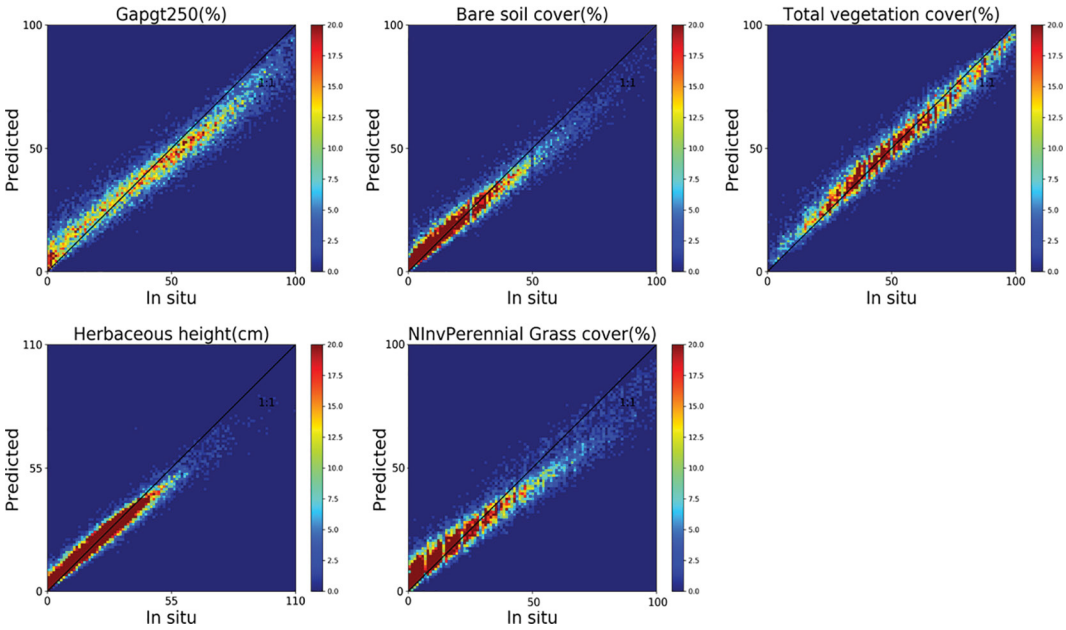

Fig. 4. Correlations between ensemble mean predicted and in situ values used for internal error calculations. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

values (Fig. 3). MODIS BRDF parameters contribute more to the regressions than any other variables. Climate and topographic variables have the second greatest contributions to the regressions. In contrast, MODIS NBAR, OLI surface reflectance, and soil texture contribute relatively little to the regressions (Fig. 3). The two vegetation indices for both satellite surface reflectance products contribute very little, so their contributions are not shown. The differences in R^2, ME, and RMSE between OOB samples and external samples are very small (Table 3). For internal predictions (i.e., predictions of points used in the training), correlations are much tighter with R^2 is considerably higher, and |ME| and RMSE considerably lower, than either external or OOB estimates (Table 3 and Fig. 4). Internal RMSE for all variables is generally 2.5 times lower than the OOB or external error estimates.

Our results indicate that there are some variances in indicator estimates during the k-fold cross-validation, though the magnitude of the variance is small compared to the range of the predictions (Fig. 5). For some indicators (e.g., Herbaceous height and NInvPerennial Grass cover), there appeared to be significant correlations between ensemble variance and mean. For other indicators (e.g., Gap ≥ 250 and Total vegetation cover), there was no clear correlation between ensemble variance and mean. This suggests that it is not necessary to take great care when conducting the cross-validation to sub-select points that span the range of indicator values.

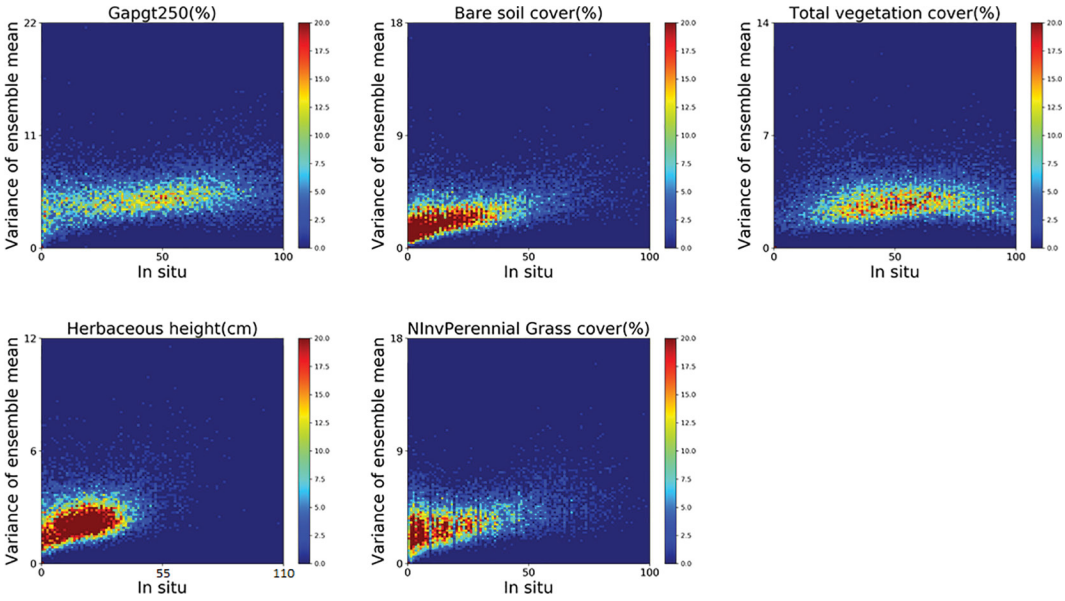

Fig. 5. Correlations between ensemble variance (calculated as the difference between the 90th and 10th percentiles of the ensemble of 20 k-fold cross-validation runs) and the ensemble means. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

4.2. Prediction map

The RF regression models were used to generate predicted distribution maps. The same indicators as shown in Fig. 3 were selected to create the predicted distribution maps in eleven selected ecoregions across the Western U.S. (Fig. 6). These maps show reasonable patterns of the indicators, showing the lower total vegetation cover, herbaceous height and non-invasive perennial grass cover in more arid regions such as Mojave Basin and Range and Sonoran Basin and Range, as well as the larger covers of Gap > 250 cm and bare soil cover in drier regions.

4.3. Error distribution maps

Error (ME and RMSE) distribution maps were produced based on the mean value of ME and RMSE of all AIM and LMF sites in each selected ecoregion using the external (cross-validation) errors. The same indicators as shown in Fig. 3 were selected to create the error distribution maps in eleven selected ecoregions across the Western U.S. (Fig. 7). Compared to the distribution of AIM and LMF sites (Fig. 1), ME and RMSE are closer to zero in the areas where have more sites, for example, in the Wyoming Basin and Northern Basin and Range. Despite this, there is no clear geographical pattern for ME or RMSE for any of

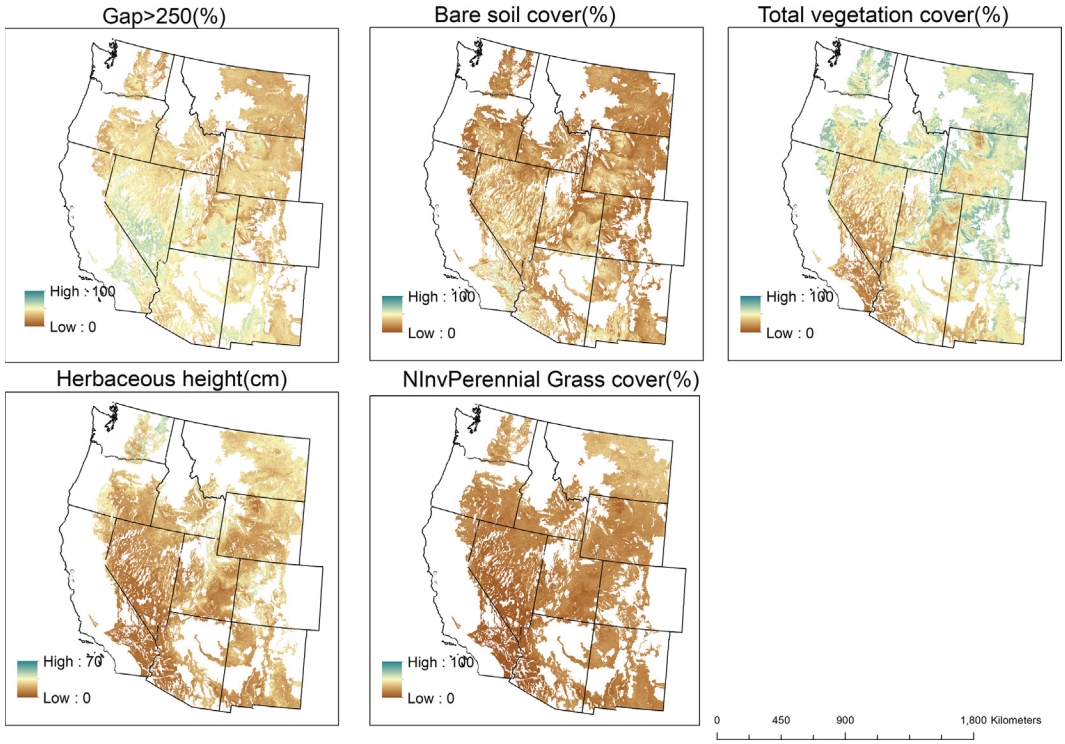

Fig. 6. Ensemble mean distribution maps of surface indicators in eleven selected ecoregions of the Western U.S. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

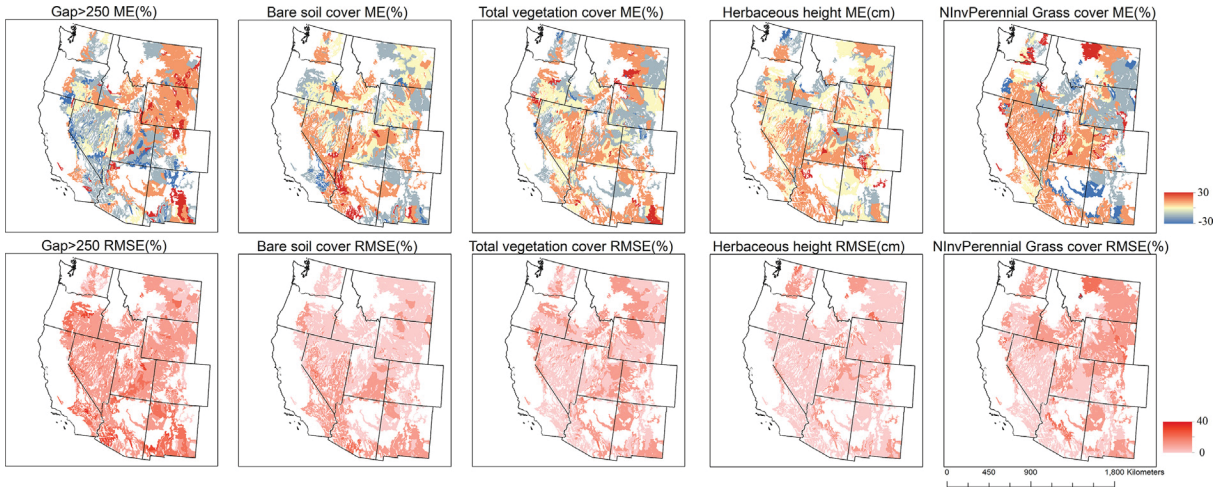

Fig. 7. Distribution maps of mean error (ME) and root mean square error (RMSE) of five surface indicators. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the variables. For Gap ≥ 250, for instance, positively- and negatively-biased ecoregions abut one another. Likewise, there are no clear relationships between RMSE (Fig. 7) and the mean value for individual ecoregions (Fig. 6).

Analysis of individual ecoregions, in fact, indicates that ecoregions with a larger number of points tend to have lower error (Fig. 8). However, this is a sufficient but not necessary condition. Ecoregions with more than about one hundred points tend to have RMSE below the median. However, there are ecoregions with many fewer points that have lower RMSE than ecoregions with many points.

5. Discussions

In general, our results indicate that there is potential for using RF to estimate AIM and LMF indicators based on optical remote sensing products combined with location-specific climate, topographic, and soil variables as predictors, though clearly some are more amenable to prediction than others. Based on the RF work so far, the correlations between model prediction and in situ measurement and the statistical evaluation of the regressions indicate that assimilating optical satellite remote sensing images and field data can provide good predictions of those indicators in arid and semiarid areas in the Western U.S. This is consistent with recent work by Jones et al. (2018).

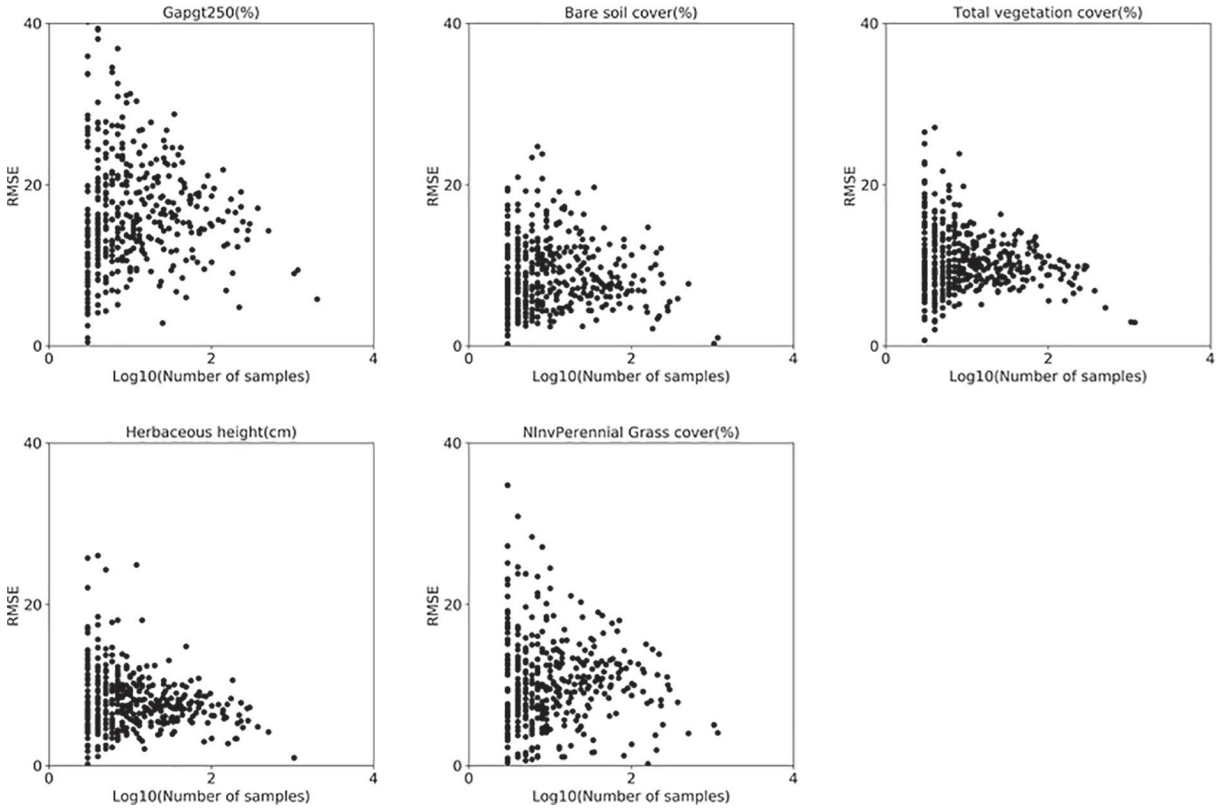

Fig. 8. RMSE of predictions in ecoregion polygons plotted against the number of points in each polygon.

5.1. Sample insufficiency and unevenly pattern of the sample

AIM and LMF projects that are on the BLM public lands are measured by different surveying and mapping teams, so there are high concentration AIM samples in specific areas that were targeted for monitoring. In addition, AIM and LMF samples are limited to BLM lands and are not evenly distributed within the study areas. For example, in the Sonoran Basin and Range ecoregion, which covers the part of California and Arizona, AIM and LMF samples are concentrated at Southern California, but there are fewer samples in Arizona (Fig. 1). This uneven spatial distribution of samples has the potential to result in incorrect predictions of surface indicators in those areas without any samples (Fig. 6). RF regression models can address the problem of sample insufficiency because it adopts the bagging approach, which can convert a small dataset to a large dataset by randomly sampling with the replacement from the original dataset (Breiman, 2001). However, RF regression model cannot correct the uneven spatial distribution of input sampling sites. That said, our analysis indicates that a small sample number is not sufficient to produce a high prediction error (Fig. 8). Sample insufficiency is not, alone, a predictor of prediction error at the ecoregion level.

5.2. Contribution of independent variables

Our results clearly indicate that BRDF parameter variables have a strong influence on the RF regression models (Fig. 3). This was true regardless of whether the variable quantified cover (i.e., Bare soil cover and Total vegetation cover) or structure (e.g. Herbaceous height and Gap > 250). Although the BRDF parameters do not, directly, include information on surface brightness in different bands, we suspect that the differences in BRDF parameters, arising from differences in reflectance nonetheless contain considerable spectral information. Indicating by the previous studies (Zhu et al., 2011; He et al., 2012; Pisek et al., 2012; Jiao et al., 2014; Gao et al., 2003), BRDF parameters is linked to both structural and surface cover indicators Thus, the BRDF parameters are doing double duty, providing information about surface structure through the parameters themselves, and providing information about reflectance through the differences in parameters for different bands. MODIS NBAR and Landsat 8 OLI surface reflectance have the lower contributions to the RF regression and prediction (Fig. 3). That is because NBAR is retrieved from BRDF parameters and OLI surface reflectance has a strong correlation with BRDF parameters. Therefore, those two datasets are redundant. Moreover, those two vegetation indices (NDVI and NDNVI) are all based on MODIS NBAR and OLI surface reflectance, so they are also redundant and have a very low contribution compared to other variables. However, because of the bagging algorithm, RF is able to preferentially select the most important independent variables to build the regression model. Therefore, aside from the complexity of the model, there really is no penalization for adding independent variables.

Although the MODIS pixels are considerably larger than the actual field measurements, which are better matched to the scale of OLI pixels, there is little indication that the finer-scale OLI reflectance contributed more than MODIS reflectance to the RF models, which might be caused by the errors in registration of either field data or Landsat remote sensing pixels. However, MODIS pixel has a lower spatial resolution but totally covers the AIM or LMF field site, so MODIS products can get rid of the registration error. Clearly, over smaller scales with a dense network of points, the finer information provided by higher-resolution satellites would be important in differentiating between values. However, for this continental-scale analysis, there does not appear to be a significant advantage of this finer-scale information.

5.3. Appropriate estimators of RF prediction errors

One of the advantages of RF over other methods is its use of bagging, in which, for each iteration, a certain percentage of the data is randomly chosen to be left out of the tree-making process and is held back for testing. In the iterative RF approach where many trees are made, every point is eventually used to produce the final RF model and predictions. Thus, the OOB error for an RF model is based on estimates in which every point is, eventually, used in training. The use of OOB error as a metric of prediction quality has, therefore, been criticized as a form of model overfitting (Bernard et al., 2012). Our results indicate that the OOB errors and the external error estimates produced through k-fold cross-validation in which there truly are independent samples are nearly identical. We conclude, therefore, that the use of OOB error as a metric of RF prediction quality is an acceptable metric, at least in the application here. Obviously, the use of internal error estimates (in which the final model is used to predict the value of each point which is then compared to the in situ value which was used in training) is inappropriate as a metric of error because it dramatically estimates error (in this case by a nearly constant factor of 2.5).

5.4. Spatial distributions of error

An intuitive assumption about any model prediction is that larger sample sizes produce better estimates (lower error) and that, conversely, smaller sample sizes produce worse estimates (higher error). The first aspect of that assumption appears to be true in the context of this study. Ecoregions with a high number of samples (one hundred samples in this study) are generally predicted better than the median (Fig. 8). However, there are many ecoregions with a small number of samples that are predicted as well as, or better than ecoregions with large sample sizes. Therefore, the concentration of samples is not, in and of itself, a predictor of accuracy. The sample insufficiency in some ecoregions (most of the ecoregions have less than ten samples) and the unevenly distributed samples prevent us from investigating the differences in errors among ecoregions. In fact, in the analysis presented here, there is little indication of what contributes to prediction error at the ecoregion level. Considerable additional work including adding more samples (around one hundred samples) evenly distributed in each ecoregion will be required to address this issue and it is possible that a relatively simple answer does not exist.

5.5. Limitation of RF

Despite performing well in many cases, RF regression has considerable difficulty in cases where a variable may have a valid value for some points but not others. For example, the indicator of sagebrush height cannot have a value at a site where sagebrush does not exist and using a value of zero for sagebrush height is not equivalent indicating that sagebrush isn't present. In a case such as this, to model the sagebrush height, two variables must be used: a dependent variable (i.e., sagebrush height) with all of the present values of sagebrush height and an additive dependent variable (i.e., sagebrush missing) marking areas where sagebrush is present as 1 and areas where sagebrush is not present as 0. The additive dependent variable becomes as important as the original dependent variable during tree growth because some additional variables may be employed to build the regression of the additive dependent variable. Moreover, because of the low cardinality (i.e., 1 and 0) of the additive dependent variable, there is only one option to split the additive dependent, thus impacting the gradient of homogeneity.

In the context of predictions of AIM/LMF indicators, this is a critical factor to keep in mind. For variables such as sagebrush height, we observe the lowest coefficients of variation and highest errors among the variables tested (Table 3). Some of the error in these predictions may be due to this problem, and methodological improvements are needed for managing this “null values” problem in the context of predicting certain landscape indicators.

6. Conclusions

In this study, we employed a machine learning-based regression model (i.e., random forest) to assimilate satellite remote sensing images (i.e., MODIS BRDF parameters, NBAR, and Landsat 8 OLI surface reflectance) and in situ measurements (i.e., AIM and LMF data collection) from four hundred selected ecoregions of the Western U.S. The field data collection used consistent methods and was performed by different teams or offices in different states, so the data are reproducible. Within these data, the predicted distribution maps of nineteen surface indicators, which are related to vegetation composition, vegetation structure, and bare ground cover, were created. The correlations between the model predicted values and in situ values of all surface indicators are strongly positive. The MODIS BRDF parameters product tends to contribute more to the regression than other predictors.

These results exhibit the potential for predicting, using optical imagery and ancillary data, the distribution of important dryland indicators using RF. However, there are caveats. First, predictions are strongest in areas that have in situ data. Care must be taken when extrapolating out of these areas. Second, this approach tends to underpredict at high values and over-predict at low values. The errors at high values generally contribute low relative error and may be within acceptable ranges for many applications. However, the high relative error is likely at low values, and care must be taken in cases with low values. Nonetheless, within these limits, this study shows how these relationships can be extended to produce spatially continuous datasets coupled with quantitative estimates of the error. Therefore, assimilating satellite remote sensing images and field data using machine learning methods can provide usable predictions of the surface indicators in drylands. There are many potential uses for such prediction maps that extend beyond the management mandate for which the original in situ data were commissioned. For instance, predicted distribution maps from our study could be employed as inputs for climate models (particularly bare soil cover, which is a common variable in global and regional models, e.g., Xue and Shukla, 1993) to forecast the potential for dust emission in the Western U.S.

Acknowledgment

This work was funded by the National Aeronautics and Space Administration (NASA, USA) grant NNX17AG50G and the National Science Foundation (NSF, USA)-funded Jornada Basin Long Term Ecological Research Network (LTER) program DEB-1235828. We thank Matt Zebrowski, the cartographer of Department of Geography, UCLA, for providing the assistance with the figures and tables in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.rse.2019.111382.

References

Bernard, S., Sébastien, A., Laurent, H., 2012. Dynamic random forests. Pattern Recogn. Lett. 33, 1580–1586.
Bestelmeyer, B.T., Okin, G.S., Duniway, M.C., Archer, S.R., Sayre, N.F., Williamson, J.C., Herrick, J.E., 2015. Desertification, land use, and the transformation of global drylands. Front. Ecol. Environ. 13, 28–36.
Booth, D.T., Cox, S.E., 2009. Dual-camera, high-resolution aerial assessment of pipeline revegetation. Environ. Monit. Assess. 158, 23–33.
Breiman, L., 1996. Bagging predictors. Mach. Learn. 24, 123–140.
Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.
Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A., 1984. Classification and Regression Trees. CRC press.
Chipman, H.A., George, E.I., McCulloch, R.E., 2010. BART: Bayesian additive regression trees. Ann. Appl. Stat. 4, 266–298.
Dietterich, T.G., 2000. An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization. Mach. Learn. 40, 139–157.
Duniway, M.C., Karl, J.W., Schrader, S., Baquera, N., Herrick, J.E., 2012. Rangeland and pasture monitoring: an approach to interpretation of high-resolution imagery focused on observer calibration for repeatability. Environ. Monit. Assess. 184, 3789–3804.
Elzinga, C.L., Salzer, D.W., Willoughby, J.W., 1998. Measuring & Monitering Plant Populations. CreateSpace Independent Publishing Platform.
Friedl, M.A., Brodley, C.E., 1997. Decision tree classification of land cover from remotely sensed data. Remote Sens. Environ. 61, 399–409.
Fujisada, H., Bailey, G.B., Kelly, G.G., Hara, S., Abrams, M.J., 2005. ASTER DEM performance. IEEE Trans. Geosci. Remote Sens. 43, 2707–2714.
Gao, F., Schaaf, C.B., Strahler, A.H., Jin, Y., Li, X., 2003. Detecting vegetation structure using a kernel-based BRDF model. Remote Sens. Environ. 86 (2), 198–205.
Hafez, A., Soliman, A., El-Metwally, K., Ismail, I., 2017. Tilt and azimuth angles in solar energy applications–a review. Renew. Sust. Energ. Rev. 77, 147–168.
He, L., Chen, J.M., Pisek, J., Schaaf, C.B., Strahler, A.H., 2012. Global clumping index map derived from the MODIS BRDF product. Remote Sens. Environ. 119, 118–130.
Herrick, J.E., Lessard, V.C., Spaeth, K.E., Shaver, P.L., Dayton, R.S., Pyke, D.A., Jolley, L., Goebel, J.J., 2010. National ecosystem assessments supported by scientific and local knowledge. Front. Ecol. Environ. 8, 403–408.
Herrick, J.E., et al., 2015. Monitoring Manual for Grassland, Shrubland, and Savanna Ecosystems Volume 1: Core Methods.
Herrick, J.E., Van Zee, J.W., Havstad, K.M., Burkett, L.M., Whitford, W.G., Bestelmeyer, B.T., Melgoza, A., Pellant, M., Pyke, D.A., Remmenga, M.D., 2017. Monitoring manual for grassland, shrubland and savanna ecosystems. Volume I: core methods. In: USDA ARS Las Cruces, New Mexico. The University of Arizona Press.
Holthausen, R., Czaplewski, R.L., DeLorenzo, D., Hayward, G., Kessler, W.B., Manley, P., McKelvey, K.S., Powell, D.S., Ruggiero, L.F., Schwartz, M.K., 2005. Strategies for Monitoring Terrestrial Animals and Habitats.
James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An Introduction to Statistical Learning. Springer.
Jiao, Z., Hill, M.J., Schaaf, C.B., Zhang, H., Wang, Z., Li, X., 2014. An anisotropic flat index (AFX) to derive BRDF archetypes from MODIS. Remote Sens. Environ. 141, 168–187.
Jones, M.O., Allred, B.W., Naugle, D.E., Maestas, J.D., Donnelly, P., Metz, L.J., Karl, J., Smith, R., Bestelmeyer, B., Boyd, C., 2018. Innovation in rangeland monitoring: annual, 30 m, plant functional type percent cover maps for US rangelands, 1984–2017. Ecosphere 9, e02430.
Karl, J.W., Duniway, M.C., Schrader, T.S., 2012. A technique for estimating rangeland canopy-gap size distributions from high-resolution digital imagery. Rangel. Ecol. Manag. 65, 196–207.
Karl, J.W., Taylor, J., Bobo, M., 2014. A double-sampling approach to deriving training and validation data for remotely-sensed vegetation products. Int. J. Remote Sens. 35, 1936–1955.
Knippertz, P., Stuut, J.-B.W., 2014. Mineral Dust : A Key Player in the Earth System. Springer.
Kuhn, M., Johnson, K., 2013. Applied Predictive Modeling. Springer.
Laliberte, A.S., Rango, A., Havstad, K.M., Paris, J.F., Beck, R.F., McNeely, R., Gonzalez, A.L., 2004. Object-oriented image analysis for mapping shrub encroachment from 1937 to 2003 in southern New Mexico. Remote Sens. Environ. 93, 198–210.
Lary, D.J., 2010. Artificial Intelligence in Geoscience and Remote Sensing. INTECH Open Access Publisher.
Lary, D.J., Alavi, A.H., Gandomi, A.H., Walker, A.L., 2016. Machine learning in geosciences and remote sensing. Geosci. Front. 7, 3–10.
Leenaars, J.G., Wheeler, I., Wright, M.N., Batjes, N.H., Bauer-Marschallinger, B., Blagotić, A., Mantel, S., Heuvelink, G., Mendes de Jesus, J., Guevara, M.A., 2017. SoilGrids250m: Global Gridded Soil Information Based on Machine Learning.
Li, X.W., Strahler, A.H., 1986. Geometric-optical bidirectional reflectance modeling of a conifer forest canopy. IEEE Trans. Geosci. Remote Sens. 24, 906–919.
Liang, S., Li, X., Wang, J., 2012. Advanced Remote Sensing: Terrestrial Information Extraction and Applications. Academic Press.
Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. R news 2, 18–22.
Loveland, T., Merchant, J., Brown, J., Ohlen, D., 1991. Development of a land-cover characteristics database for the conterminous U. S. Photogramm. Eng. Remote. Sens. 57, 1453–1463.
Ludwig, J.A., Bastin, G.N., Chewings, V.H., Eager, R.W., Liedloff, A.C., 2007. Leakiness: a new index for monitoring the health of arid and semiarid landscapes using remotely sensed vegetation cover and elevation data. Ecol. Indic. 7, 442–454.
Luscier, J.D., Thompson, W.L., Wilson, J.M., Gorham, B.E., Dragut, L.D., 2006. Using digital photographs and object-based image analysis to estimate percent ground cover in vegetation plots. Front. Ecol. Environ. 4, 408–413.
MacKinnon, W.C., Karl, J.W., Toevs, G.R., Taylor, J.J., Karl, S., Spurrier, C.S., Herrick, J.E., 2011. BLM Core Terrestrial Indicators and Methods. US Department of the Interior, Bureau of Land Management, National Operations Center Denver.
Maclaurin, G., Sengupta, M., Xie, Y., Gilroy, N., 2016. Development of a MODIS-derived Surface Albedo Data Set: An Improved Model Input for Processing the NSRDB. National Renewable Energy Lab (NREL).
Marsett, R.C., Qi, J., Heilman, P., Biedenbender, S.H., Watson, M.C., Amer, S., Weltz, M., Goodrich, D., Marsett, R., 2006. Remote sensing for grassland management in the arid southwest. Rangel. Ecol. Manag. 59, 530–540.
McCord, S.E., Buenemann, M., Karl, J.W., Browning, D.M., Hadley, B.C., 2017. Integrating remotely sensed imagery and existing multiscale field data to derive rangeland indicators: application of Bayesian additive regression trees. Rangel. Ecol. Manag. 70, 644–655.
McMahon, G., Gregonis, S.M., Waltman, S.W., Omernik, J.M., Thorson, T.D., Freeouf, J.A., Rorick, A.H., Keys, J.E., 2001. Developing a spatial framework of common ecological regions for the conterminous United States. Environ. Manag. 28, 293–316.

Miller, D.A., White, R.A., 1998. A conterminous United States multilayer soil characteristics dataset for regional climate and hydrology modeling. Earth Interact. 2, 1–26.
Murphy, K.P., 2012. Machine Learning: A Probabilistic Perspective. MIT press.
Okin, G.S., Parsons, A.J., Wainwright, J., Herrick, J.E., Bestelmeyer, B.T., Peters, D.C., Fredrickson, E.L., 2009. Do changes in connectivity explain desertification? Bioscience 59, 237–244.
Pal, M., 2005. Random forest classifier for remote sensing classification. Int. J. Remote Sens. 26, 217–222.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.
Pisek, J., Rautiainen, M., Heiskanen, J., Mõttus, M., 2012. Retrieval of seasonal dynamics of forest understory reflectance in a Northern European boreal forest from MODIS BRDF data. Remote Sens. Environ. 117, 464–468.
Ramcharan, A., Hengl, T., Nauman, T., Brungard, C., Waltman, S., Wills, S., Thompson, J., 2018. Soil property and class maps of the conterminous United States at 100-meter spatial resolution. Soil Sci. Soc. Am. J. 82, 186–201.
Roberts, D.R., Bahn, V., Ciuti, S., Boyce, M.S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J.J., Schröder, B., Thuiller, W., 2017. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography 40, 913–929.
Rodriguez-Galiano, V.F., Ghimire, B., Rogan, J., Chica-Olmo, M., Rigol-Sanchez, J.P., 2012. An assessment of the effectiveness of a random forest classifier for land-cover classification. ISPRS J. Photogramm. Remote Sens. 67, 93–104.
Roujean, J.L., Leroy, M., Deschamps, P.Y., 1992. A bidirectional reflectance model of the earths surface for the correction of remote-sensing data. J. Geophys. Res.-Atmos. 97, 20455–20468.
Roy, D.P., Wulder, M., Loveland, T.R., Woodcock, C., Allen, R., Anderson, M., Helder, D., Irons, J., Johnson, D., Kennedy, R., 2014. Landsat-8: science and product vision for terrestrial global change research. Remote Sens. Environ. 145, 154–172.
Russell, S., Norvig, P., 2011. Artificial Intelligence a Modern Approach 3rd Edition Pdf. Pearson Education Asia, Hong Kong.
Schaaf, C.B., Gao, F., Strahler, A.H., Lucht, W., Li, X.W., Tsang, T., Strugnell, N.C., Zhang, X.Y., Jin, Y.F., Muller, J.P., Lewis, P., Barnsley, M., Hobson, P., Disney, M., Roberts, G., Dunderdale, M., Doll, C., d'Entremont, R.P., Hu, B.X., Liang, S.L., Privette, J.L., Roy, D., 2002. First operational BRDF, albedo nadir reflectance products from MODIS. Remote Sens. Environ. 83, 135–148.
Segal, M.R., 2004. Machine Learning Benchmarks and Random Forest Regression.
Smith, J., 1977. Vegetation and microclimate of east-and west-facing slopes in the grasslands of Mt Wilhelm, Papua New Guinea. J. Ecol. 39–53.
Steinberg, D., Colla, P., 2009. CART: Classification and Regression Trees. Springer.
Strahler, A.H., Muller, J., Lucht, W., Schaaf, C., Tsang, T., Gao, F., Li, X., Lewis, P., Barnsley, M.J., 1999. MODIS BRDF/Albedo Product: Algorithm Theoretical Basis Document Version 5.0. NASA.
Sun, W., Liang, S., Xu, G., Fang, H., Dickinson, R., 2008. Mapping plant functional types from MODIS data using multisource evidential reasoning. Remote Sens. Environ. 112, 1010–1024.
Svetnik, V., Liaw, A., Tong, C., Culberson, J.C., Sheridan, R.P., Feuston, B.P., 2003. Random forest: a classification and regression tool for compound classification and QSAR modeling. J. Chem. Inf. Comput. Sci. 43, 1947–1958.
Westerling, A.L., Hidalgo, H.G., Cayan, D.R., Swetnam, T.W., 2006. Warming and earlier spring increase western US forest wildfire activity. Science 313, 940–943.
Xue, Y., Shukla, J., 1993. The influence of land surface properties on Sahel climate. Part 1: desertification. J. Clim. 6, 2232–2245.
Zhang, J., Zhu, W., Wang, L., Jiang, N., 2012. Evaluation of similarity measure methods for hyperspectral remote sensing data. In: Geoscience and Remote Sensing Symposium (IGARSS). IEEE, pp. 4138–4141 2012 IEEE International.
Zhang, J., Zhu, W., Dong, Y., Jiang, N., Pan, Y., 2013. A spectral similarity measure based on changing-weight combination method. Acta Geodaetica et Cartographica Sinica 42, 418–424.
Zhu, G., Ju, W., Chen, J.M., Gong, P., Xing, B., Zhu, J., 2011. Foliage clumping index over China's landmass retrieved from the MODIS BRDF parameters product. IEEE Trans. Geosci. Remote Sens. 50 (6), 2122–2137.