Importance of Vegetation Type in Forest Cover Estimation

Anuj Karpatne, Mace Blank, Michael Lau, Shyam Boriah, Karsten Steinhaeuser, Michael Steinbach and Vipin Kumar
Department of Computer Science & Engineering, University of Minnesota
{anuj,blank,mwlau,sboriah,ksteinha,steinbac,kumar}@cs.umn.edu

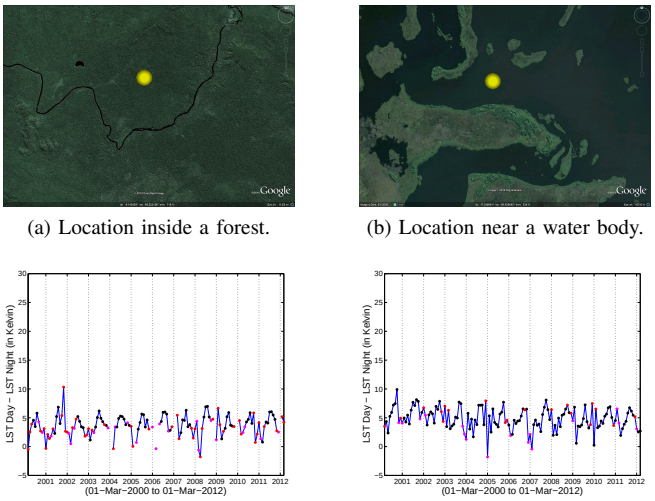
***Abstract*—Forests are an important natural resource that play a major role in sustaining a number of vital geochemical and bioclimatic processes. Since damage to forests due to natural and anthropogenic factors can have long-lasting impacts on the ecosystem of the planet, monitoring and estimating forest cover and its losses at global, regional and local scales is of primary concern. Developing forest cover estimation techniques that utilize remote sensing datasets offers global applicability at high temporal frequencies. However, estimating forest cover using satellite observations is challenging in the presence of heterogeneous vegetation types, each having its unique data characteristics. In this paper, we explore techniques for incorporating information about the vegetation type in forest cover estimation algorithms. We show that utilizing the vegetation type improves performance regardless of the choice of input data or forest cover learning algorithm. We also provide a mechanism to automatically extract information about the vegetation type by partitioning the input data using clustering.**

I. Introduction

Forests are a crucial ecological resource; they are responsible for preserving biodiversity, maintaining soil fertility, regulating water resources, sustaining economic activities and supporting geochemical and climatological cycles. Thus, estimating the amount of forest cover over large areas is important for many reasons. Specifically, quantifying forest cover is a necessary input for studies linking changes in forests with natural and anthropogenic events [4, 10, 17].

Field-based techniques for monitoring forest cover provide the most detailed and accurate information about the health of forests. However, these studies are expensive and are thus limited to smaller regions of interest, or are undertaken at irregular time intervals. Remote sensing-based techniques, however, provide a cost-effective solution for monitoring forest cover. Recently, a number of methods have been developed for rapidly estimating forest cover at coarse spatial resolution using remote sensing data such as vegetation index and land surface temperature as input [9, 19, 20].

Recent well-known forest cover estimation techniques are based on a variety of learning methods, including logistic regression [20] and regression trees [19]. These approaches learn a single global model that is applied for all vegetation types. However, as we observe in Figure 1, the input data can be very similar for two completely different land cover types. In this example, the mean land surface temperature (used as input in [20]) for a location inside a forest is almost identical to that of a water body. Thus, it is important to devise forest cover estimation techniques that behave differently in different vegetation domains, thus preserving vegetation heterogeneity.



(a) Location inside a forest. (b) Location near a water body.

(c) LST time series for the example location inside a forest. (d) LST time series for the example location near a water body.

Figure 1. An example of a location inside a forest marked by a yellow dot in (a), and a location near a water body marked by a yellow dot in (b), which show similar LST characteristics in (c) and (d) respectively. *Note:* Most figures in this paper are best seen in color.

In this paper, we investigate the role of vegetation type in estimating forest cover and explore techniques for identifying and exploiting characteristics of the data which are indicative of vegetation type. We focus this study on the region of Mato Grosso, Brazil ($7°$–$19°S$, $62°$–$50°W$) and evaluate the performance of the proposed techniques on a variety of learning algorithms, using vegetation index and land surface temperature observations as input. We demonstrate that the enriched regression models provide *substantially* better estimates of forest cover than a recently developed logistic regression approach [20].

The key contributions of the paper are as follows:

- We systematically investigate the role of vegetation type in estimating forest cover and explore techniques that model heterogeneous relationships of different vegetation types with the forest cover.

- We comparatively evaluate the ability of vegetation indices and land surface temperature datasets in discriminating between different vegetation types, and we demonstrate their effectiveness in relation to standard land cover labels.
- We comprehensively show that utilizing information about the vegetation type improves performance regardless of the choice of (1) input data, or (2) learning algorithm.

II. Related Work

Predictive vegetation modeling using environmental variables has been explored in a number of scenarios for estimating response variables such as soil properties [14], species distribution [2], disease count [21] and the presence or absence of a vegetation type [12]. Such techniques have been devised for varying kinds of environmental data types, which can be boolean, discrete or continuous. Further, predictive vegetation models often incorporate information about the spatial properties of the domain [16] using various approaches such as auto-regressive methods [11], geostatistical methods [3], parameter estimation methods [15] and geographically weighted regression [22].

van Leeuwen et al. [20] presented an approach for estimating forest cover using remote sensing datasets obtained through MODIS (Moderate Resolution Imaging Spectroradiometer). They explored the use of land surface temperature (LST) for forest cover estimation and highlighted the importance of using the difference between the day and night land surface temperatures for estimating forest cover, which best captures the influences of multiple geophysical pathways on forest cover. They also found that LST observations during the dry season provide the maximum predictive power about the forest cover as compared to other seasonal periods. Furthermore, van Leeuwen et al. suggested a technique for detecting the months corresponding to the dry season. The month with the maximum average difference between LST Day and LST Night over a period of 6 years (2000 to 2005) was identified for each location, and a 3-month window centered around the detected month was used to mark the dry season at every location. Logistic regression models were tested using LST in combination with the normalized vegetation index (NDVI) as input, using correlation as performance measure. In this paper, the van Leeuwen et al. [20] approach is presented in the quantitative evaluation (Section VI) as the baseline logistic regression method.

III. Data

In this study, we use the land surface temperature (LST), normalized difference vegetation index (NDVI), enhanced vegetation index (EVI) and land cover type (LC) datasets obtained from MODIS. The datasets are at a spatial resolution of 0.05° on the geographic Climate Modeling Grid (CMG), and distributed through the Land Processes Distributed Active Archive Center [1]. We provide a description of each of the datasets below:

Land Surface Temperature (LST): The Collection 5 monthly composited and averaged level-3 MOD11C3 product was used for extracting LST which is available at 0.05° geographic CMG. LST is derived from thermal infrared bands and measures the land surface temperature during the day as well as the night. We only consider cloud-free observations of LST for evaluation.

Normalized Difference Vegetation Index (NDVI) and *Enhanced Vegetation Index* (EVI): NDVI and EVI are obtained from the monthly Level-3 MOD13C2 product at 0.05° geographic CMG. The MODIS NDVI product contains atmospherically corrected bi-directional surface reflectances masked for water, clouds and cloud shadows. On the other hand, EVI uses the blue band to remove residual atmospheric contamination caused by smoke and sub-pixel thin clouds. EVI also uses feedback adjustment to minimize canopy background variations and to enhance vegetation sensitivity from sparse to dense vegetation conditions.

Land Cover Type (LC): The Land Cover Type is an yearly MCD12C1 product, available at 0.05° geographic CMG, which provides the dominant land cover type using a supervised classification approach. The primary land cover scheme identifies 17 land cover classes defined by the International Geosphere Biosphere Programme (IGBP), which includes 11 natural vegetation classes, 3 developed and mosaicked land classes, and 3 non-vegetated land classes.

IV. Methods

The task of estimating the forest cover at a particular location in a given year can be realized as a regression problem in a supervised setting, where the forest cover is the dependent variable and the predictor variables consist of input data from remote sensing datasets. More formally, we can express the estimation problem as $Y = f(X)$, where $Y \in [0, 1]$ is the forest cover at a given location in a year, interpreted as the proportion of pixel area (where a pixel corresponds to a $0.05^\circ \times 0.05^\circ$ MODIS grid) covered by forests, X is the observed input data obtained from remote sensing datasets, such as LST, NDVI and EVI at the same location and year, and f is the objective function which has to be learned.

We note that LST, NDVI and EVI contain 12 monthly observations at a given location in a year and several techniques can be devised for extracting annual features from each of them, which best model forest cover when used as predictor variables. Also, the land surface temperature (measured using LST) and vegetation index (measured using NDVI and EVI) represent unique characteristics about the domain and hence offer complementarities that can be harnessed by employing multivariate analysis techniques. Furthermore,

spatial and temporal dependencies among the observations can be leveraged by modeling structural inter-relationships inherently present in the data. However, since the primary focus of this paper is to understand the effects of incorporating vegetation type in the learning algorithm, we perform an in-depth exploration of simple univariate predictor variables and basic learning algorithms, and present techniques for enhancing the performance of a learning algorithm by making it aware of the vegetation type. Specifically, we explore the following combinations of predictor variables and learning algorithms:

Predictor Variables: We construct predictor variables from the 12 monthly observations of LST, NDVI and EVI in a particular year at a given location. For NDVI and EVI, we consider the mean of the 12 values in a year as the predictor variable. For LST, we consider the mean difference between LST Day and LST Night during the months corresponding to the dry season in that year, using the same approach as mentioned in [20].

Learning Algorithms:

- Logistic Regression: Y is expressed as the following function of X:

$$\ln\left(\frac{Y}{1-Y}\right) = \beta^T X \tag{1}$$

  where β is the set of model parameters which must be learned using the Iteratively Reweighted Least Squares (IRLS) algorithm [13]. Logistic Regression is traditionally used for classification problems where the likelihood of Y is assumed to share a linear relationship with X. It is also commonly used in regression scenarios when the dependent variable has been extracted from binary data, such as the proportion of forest cover in the pixel area at a coarse resolution.
- Linear Regression: A linear relationship between Y and X is assumed, which can be described as

$$Y = \beta^T X \tag{2}$$

  where β represents the set of parameters which are estimated using the Ordinary Least Squares (OLS) method [7, 13].
- Regression Trees: Similar to the use of decision trees in classification, regression trees [5] are used for estimating the value of a dependent variable at the leaf node, by performing splitting decisions at internal nodes, as described in Figure 2. This involves choosing the attribute for splitting at each internal node, which provides the maximum reduction in the Mean Squared Error (MSE) of the subsequent subtrees. To prevent over-fitting, we perform pruning by restricting the number of internal nodes in the learned regression tree to be less than or equal to 30.

Instead of developing a single learning algorithm which estimates the forest cover for every vegetation type, independent regression algorithms can be designed for each vegetation type, incorporating vegetation heterogeneity in the regression scheme. Since traditional learning algorithms attempt to learn the heterogeneous dependencies of multiple vegetation types and forest cover in a single framework, they suffer in performance by being too restrictive in nature. In contrast, a scheme utilizing multiple independent learning algorithms, each corresponding to a different vegetation type, offers more flexibility and thus can take advantage of the additional degrees of freedom for improving its estimation power.

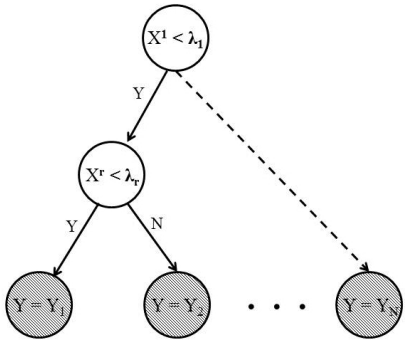
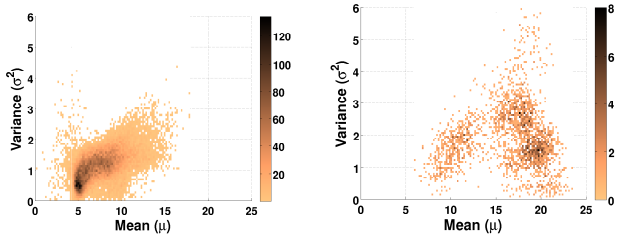


Figure 2. An example of a regression tree, where the leaf nodes (shaded) estimate the value of the dependent variable, while the internal nodes perform splitting decisions.

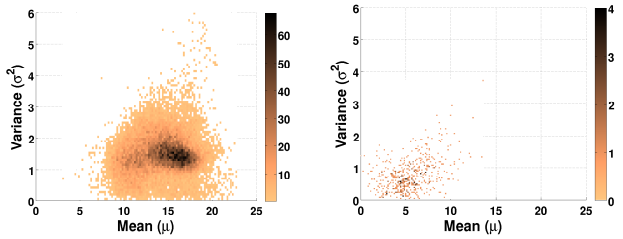
Learning multiple regression algorithms requires segmentation of the observation space into multiple regions, so as to demarcate the domain of application of different learning algorithms. This can be achieved by characterizing each input data instance with a vegetation type label, both during the training and testing phases. We describe the following two simple techniques for achieving the desired segmentation of the input data depending upon the vegetation type at a location: (a) Partitioning the Feature Space, and (b) Time Series Clustering. They are described in detail as follows:

A. Partitioning the Feature Space

The observations are first projected to a lower-dimensional feature space, where the features are selected based on their ability in discriminating between different vegetation types. As an example, forests exhibit high inter-annual NDVI and EVI mean, and low inter-annual LST mean, while the intra-annual variance of LST, NDVI and EVI is low. On the other hand, farms show a characteristically higher intra-annual NDVI, EVI and LST variance due to the presence of cultivation cycles. Further, wetlands and locations near the water bodies show similar behavior in NDVI and EVI mean as that of a forested location. We thus explore the benefit in using inter-annual mean (μ) and intra-annual variance (σ^2) of LST, NDVI and EVI, computed over a period of 5 years (2000 to 2004), for representing different vegetation types.

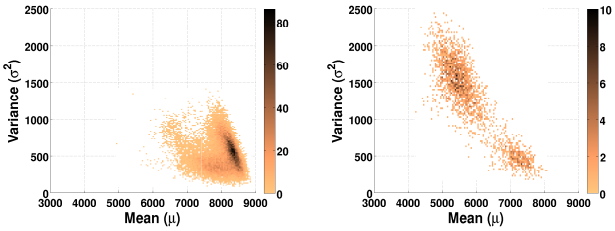


(a) LST in Forests (b) LST in Farms

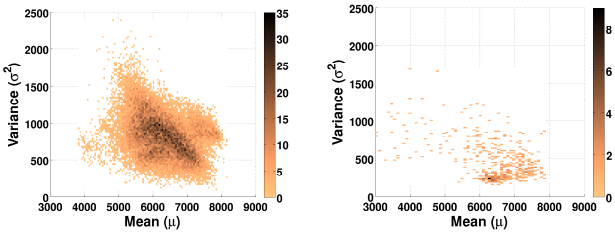


(c) LST in Grasslands (d) LST near Water Locations

Figure 3. Distribution of locations in the (μ, σ^2) feature space of LST for different vegetation types. The intensity of color at a μ, σ^2 pair in each of the subfigures denotes the number of samples observed with that pair of values in the two-dimensional histogram using μ and σ^2.



(a) NDVI in Forests (b) NDVI in Farms



(c) NDVI in Grasslands (d) NDVI near Water Locations

Figure 4. Distribution of locations in the (μ, σ^2) feature space of NDVI for different vegetation types. The intensity of color at a μ, σ^2 pair in each of the subfigures denotes the number of samples observed with that pair of values in the two-dimensional histogram using μ and σ^2.

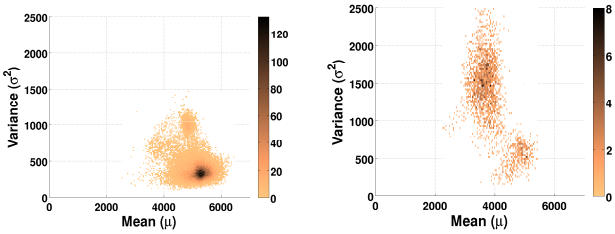
Figures 3, 4 and 5 present the distribution of locations inside the following four vegetation types (identified using MODIS LC labels): Forests (LC $\in \{1-5\}$), Farms (LC $\in \{12, 14\}$), Grasslands (LC $\in \{8-10\}$) and Water Locations (LC $\in \{0, 11\}$) in the (μ, σ^2) feature space of LST, NDVI and EVI respectively. They illustrate the differences in their behavior in the two-dimensional (μ, σ^2) feature space. A two-dimensional grid partitioning is then performed on the (μ, σ^2) feature space and the multiple regions constructed in this fashion correspond to unique vegetation types.

B. Time Series Clustering

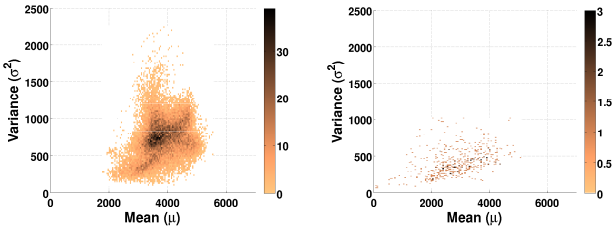
We employ k-means clustering, with varying values of k, for clustering LST, NDVI and EVI time series observations over a period of 5 years (2000 to 2004). The identified clusters can then be considered to represent distinct vegetation types of the domain, and unique regression algorithms for each cluster can then be learned and evaluated.

V. Evaluation Setup

We evaluate the performance of different forest cover learning algorithms using observations from years 2000 to 2004 at each location for training, and observations from years 2005 to 2009 for testing. This ensures temporal separation between the training and testing data for correct cross-validation of the models. Experiments using other spatial and temporal partitioning into training and testing subsets, such as training using observations in 2005 to 2009 and testing over observations in 2000 to 2004 were also carried out, which showed similar trends in the evaluation results.



(a) EVI in Forests (b) EVI in Farms



(c) EVI in Grasslands (d) EVI near Water Locations

Figure 5. Distribution of locations in the (μ, σ^2) feature space of EVI for different vegetation types. The intensity of color at a μ, σ^2 pair in each of the subfigures denotes the number of samples observed with that pair of values in the two-dimensional histogram using μ and σ^2.

A. Validation Data

The Program for the Estimation of Deforestation in the Brazilian Amazon (PRODES) [6], provides an annual deforestation product for each state in the Brazilian Amazon, which is generated from the analysis of high-resolution Landsat Thematic Mapper (TM) images by the Brazilian Instituto Nacional de Pesquisas Espaciais (INPE). PRODES data was downloaded at 90 m resolution in GeoTiff format, which was then converted to a vector (polygon) format.

The polygons were then reprojected from Datum SAD69 to WGS84 projectioning scheme. Using the base map of forest cover in 2000 and the increments of annual deforestation available every year, we calculated the forest cover for years in 2000 and 2009. The PRODES forest cover was used as the true forest cover for the purpose of evaluation.

B. Evaluation Metrics

Let Y_i denote the true observation of the dependent variable, and let $\widehat{Y}_i$ be its estimated value. The following evaluation metrics can then be defined for analyzing the performance of a regression model, which are commonly used in statistics, machine learning and data mining [8, 18]:

- *Pearson's Correlation (Corr):* The correlation between Y and $\widehat{Y}$ is a measure of their linear relationship and can be defined as

$$\text{Corr}(Y, \widehat{Y}) = \frac{\sum (Y_i - \bar{Y})(\widehat{Y}_i - \widehat{\bar{Y}})}{\sqrt{\sum (Y_i - \bar{Y})^2 \sum (\widehat{Y}_i - \widehat{\bar{Y}})^2}} \quad (3)$$

  where $\bar{Y}$ and $\widehat{\bar{Y}}$ denote the sample means of Y and $\widehat{Y}$ respectively.
- *Coefficient of Determination (R^2):* R^2 measures the proportion of variability in the dependent variable explained by the regression model and thus is a measure for the goodness of fit of the model, formally defined as

$$R^2 = 1 - \frac{\sum (Y_i - \widehat{Y}_i)^2}{\sum (Y_i - \bar{Y})^2} \quad (4)$$

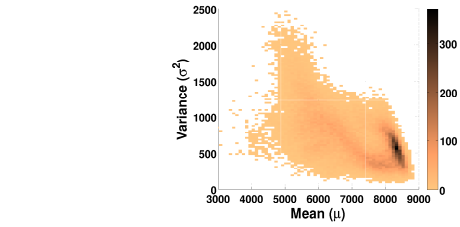
We study the effects of vegetation-specific modeling on several learning algorithms and examine their performance during the testing phase using correlation and $1 - R^2$ as our evaluation measures.

VI. Experimental Results and Discussion

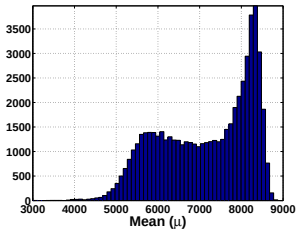
A. Partitioning the Feature Space

We first consider partitioning the input data space using land cover type labels available from LC data in year 2000, and denote this partitioning scheme as *Use*LC. We specifically construct the following 4 partitions: Forests (LC $\in \{1-5\}$), Farms (LC $\in \{12, 14\}$), Grasslands (LC $\in \{8-10\}$) and Water Bodies (LC $\in \{0, 11\}$), each of which is learned and evaluated using a separate regression algorithm. Evaluation results of *Use*LC experiment are illustrated in Table I.

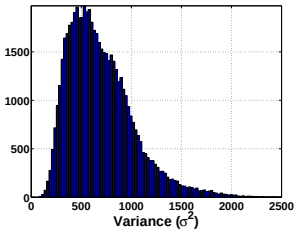
We next consider projecting each data instance to a two-dimensional feature space comprising of the inter-annual mean (μ), and the intra-annual variance (σ^2) of NDVI over a period of 5 years (2000 to 2004). Let L be the set of all locations. Figure 6 illustrates the distribution of locations in the (μ,σ^2) space of NDVI, which can be partitioned using the following schemes :



(a) (μ, σ^2) of NDVI



(b) μ of NDVI



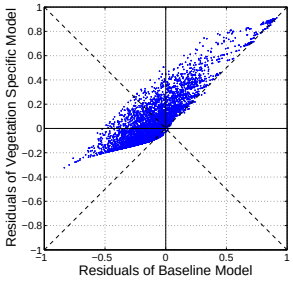
(c) σ^2 of NDVI

Figure 6. Distribution of all locations in the (μ, σ^2) space of NDVI, represented as histograms in (b) and (c), and as a two-dimensional histogram in (a). The intensity of color at a μ, σ^2 pair in (a) denotes the number of samples observed with that pair of values in the two-dimensional histogram using μ and σ^2.

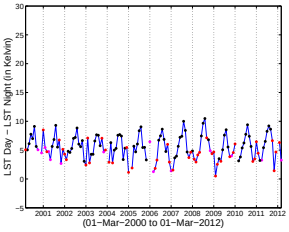
- $Part(3 \times 2)$: 6 regions are created by performing a two-dimensional grid partitioning, where μ is split in 3 divisions: $\mu \in [0, 5000]$, $\mu \in (5000, 8000]$ and $\mu \in (8000, 1000]$, and σ^2 is split in 2 divisions: $\sigma^2 \in [0, 750]$, and $\sigma^2 > 750$.
- $Part(4 \times 2)$: 8 regions are created by performing a two-dimensional grid partitioning, where μ is split in 4 divisions: $\mu \in [0, 5000]$, $\mu \in (5000, 7000]$, $\mu \in (7000, 8000]$ and $\mu \in (8000, 1000]$, and σ^2 is split in 2 divisions: $\sigma^2 \in [0, 750]$, and $\sigma^2 > 750$.
- $Part(4 \times 3)$: 12 regions are created by performing a two-dimensional grid partitioning, where μ is split in 4 divisions: $\mu \in [0, 5000]$, $\mu \in (5000, 7000]$, $\mu \in (7000, 8000]$ and $\mu \in (8000, 1000]$, and σ^2 is split in 3 divisions: $\sigma^2 \in [0, 400]$, $\sigma^2 \in [400, 750]$ and $\sigma^2 > 750$.

We now examine the locations in a specific partition in order to understand their physical correspondence with a specific vegetation type. Furthermore, we analyze the behavior of residuals in a given partition and investigate the benefits of vegetation specific modeling in the context of that partition.

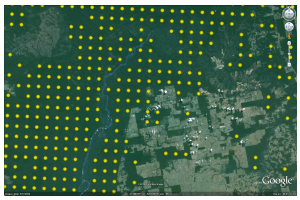
Figure 7 shows the results of a detected partition that corresponds to the forest vegetation type, as shown as a sample image in Figure 7(c). Sample LST and NDVI time series are shown in Figures 7(b) and 7(d) respectively, which illustrate low intra-annual variance of LST and NDVI, low inter-annual LST mean and high inter-annual NDVI mean for a location inside the forest vegetation type. Figure 7(a) provides a scatter plot of the residuals of the baseline approach (along the x-axis), and the residuals of the vegetation



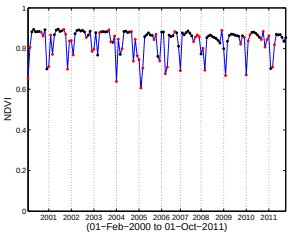
(a) Scatter Plot of Residuals



(b) Sample LST time series

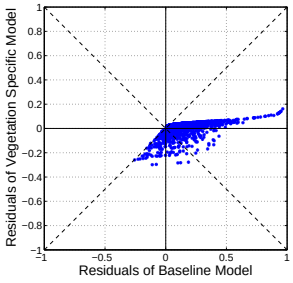


(c) Sample image of the detected region

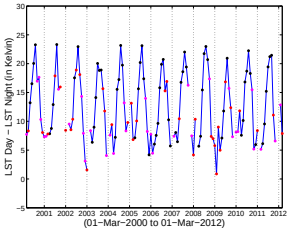


(d) Sample NDVI time series

Figure 7. Analysis of a detected partition corresponding to a forest vegetation type. The detected locations in the partition are shown as yellow dots in the sample image shown in (c). Example LST and NDVI time series of a sample location in the partition are shown in (b) and (d) respectively. A scatter plot for the residuals of the baseline model along with the residuals of vegetation specific modeling is shown in (a).

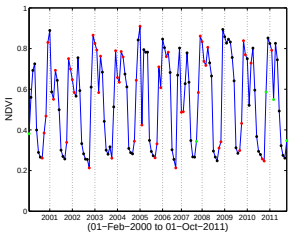


(a) Scatter Plot of Residuals



(b) Sample LST time series

(c) Sample image of the detected region



(d) Sample NDVI time series

Figure 8. Analysis of a detected partition corresponding to a cropland vegetation type. The detected locations in the partition are shown as orange dots in the sample image shown in (c). Example LST and NDVI time series of a sample location in the partition are shown in (b) and (d) respectively. A scatter plot for the residuals of the baseline model along with the residuals of vegetation specific modeling is shown in (a).

specific approach (along the y axis) for each observation inside the detected partition during the testing phase. It can be observed that most of the observations in the forest vegetation type have a negative baseline residual, in which case the residual of the vegetation specific modeling has a lower magnitude than the baseline model, indicating a better goodness of fit.

Similarly, Figure 8 presents the case of a partition that corresponds to farms, as shown in Figure 8(c). The LST and EVI time series, as shown in Figures 8(b) and 8(d) respectively, show high inter-annual variance that is characteristic of farms. The analysis of the residuals in Figure 8(a) suggests that most of the observations in the partition correspond to a high residual value for the baseline model, whereas the residual of the vegetation specific model in such cases is low in magnitude. These examples clearly illustrate the physical interpretation of the detected partitions and the performance improvement obtained by partitioning the input space based on vegetation type.

Table I presents evaluation results for the above-mentioned partitioning schemes on multiple combinations of predictor variable (LST, NDVI, and EVI) and learning algorithms (logistic, linear and regression tree). The baseline method refers to the case where no information about the vegetation type is used. Hence, the logistic regression models for the baseline approach correspond to the models proposed by [20] using either LST, NDVI or EVI. It can be observed that partitioning schemes utilizing the μ and σ^2 feature space of NDVI provide significant improvement in the correlation and $1 - R^2$ values as compared to the baseline approach. This illustrates the power of incorporating vegetation type in forest cover estimation. Further, partitioning experiments using μ and σ^2 feature space of NDVI perform at par with experiments involving LC labels, showing promise in devising better partitioning schemes using LST, NDVI and EVI data, as opposed to using LC labels which are noisy and inaccurate.

B. Time Series Clustering

We implement k-means clustering on NDVI time series using observations at each location over a period of 5 years (2000 to 2004). For each of the clusters obtained corresponding to different vegetation types, an independent regression algorithm is then learned and evaluated. Table II provides performance results of schemes utilizing k-means clustering. It can be observed that k-means generally performs better than feature space partitioning experiments using the same number of clusters or partitions. Moreover, the improvement in the correlation and $1 - R^2$ measures as compared to the baseline approach is significant.

For each cluster obtained using the k-means clustering with $k = 6$, we examine its centroid LST time series by considering the LST observations for locations inside the

Table I
EVALUATION RESULTS FOR SCHEMES WHICH INVOLVE PARTITIONING THE FEATURE SPACE INTO MULTIPLE REGIONS, EACH CORRESPONDING TO A DIFFERENT VEGETATION TYPE

Using LST	Logistic		Linear		Reg. Tree	
Model	Corr	$1-R^2$	Corr	$1-R^2$	Corr	$1-R^2$
Baseline	*0.7394*	*0.4553*	*0.7197*	*0.4871*	*0.7414*	*0.4528*
*Use*LC	0.8100	0.3442	0.8103	0.3441	0.8113	0.3425
$Part(3\times 2)$	0.8255	0.3229	0.8218	0.3297	0.8238	0.3261
$Part(4\times 2)$	0.8476	0.2892	0.8465	0.2914	**0.8425**	**0.2975**
$Part(4\times 3)$	**0.8495**	**0.2881**	**0.8485**	**0.2898**	0.8422	0.3003
Using NDVI	Logistic		Linear		Reg. Tree	
Model	Corr	$1-R^2$	Corr	$1-R^2$	Corr	$1-R^2$
Baseline	*0.8416*	*0.2959*	*0.8179*	*0.3383*	*0.8434*	*0.2972*
*Use*LC	0.8540	0.2738	0.8528	0.2784	0.8535	0.2787
$Part(3\times 2)$	0.8649	0.2530	0.8596	0.2628	0.8595	**0.2616**
$Part(4\times 2)$	**0.8683**	**0.2472**	0.8648	0.2530	**0.8596**	0.2624
$Part(4\times 3)$	0.8669	0.2491	**0.8668**	**0.2494**	0.8551	0.2698
Using EVI	Logistic		Linear		Reg. Tree	
Model	Corr	$1-R^2$	Corr	$1-R^2$	Corr	$1-R^2$
Baseline	*0.7428*	*0.4562*	*0.7126*	*0.5036*	*0.7527*	*0.4383*
*Use*LC	0.8118	0.3481	0.8109	0.3495	0.8143	0.3432
$Part(3\times 2)$	0.8399	0.3083	0.8379	0.3128	0.8294	0.3251
$Part(4\times 2)$	0.8516	0.2880	0.8507	0.2895	**0.8439**	**0.2993**
$Part(4\times 3)$	**0.8553**	**0.2824**	**0.8549**	**0.2833**	0.8423	0.3033

cluster. As shown in Figure 9, the six obtained centroid time series show significant separability among each other. Further, it can be observed that the different centroid time series correspond to unique characteristics of different vegetation types. For example, the centroid time series named T1 in Figure 9 shows high intra-annual variability which is characteristic of farms. On the other hand, the centroid time series named T6 in Figure 9 shows low inter-annual mean indicative of a forest vegetation type. Hence, k-means presents a technique for automatically learning multiple regression algorithms using LST, NDVI and EVI observations, where each detected cluster corresponds to a different vegetation type.

We further analyze the performance of the logistic regression scheme using LST as the predictor variable, by varying the number of clusters used during k-means clustering from $k = 1$ to $k = 15$. Figure 10 presents the behavior of correlation and $1 - R^2$ on varying the number of clusters for the aforementioned regression scheme. It can be seen that the gain in performance by increasing the number of clusters is initially larger for smaller values of k, and slowly converges to a constant value after some k, when the number of clusters is sufficient for modeling the heterogeneity of the vegetation types.

VII. CONCLUSIONS AND FUTURE WORK

There exists a heterogeneity in the relationships of remote sensing data in different vegetation types with the forest

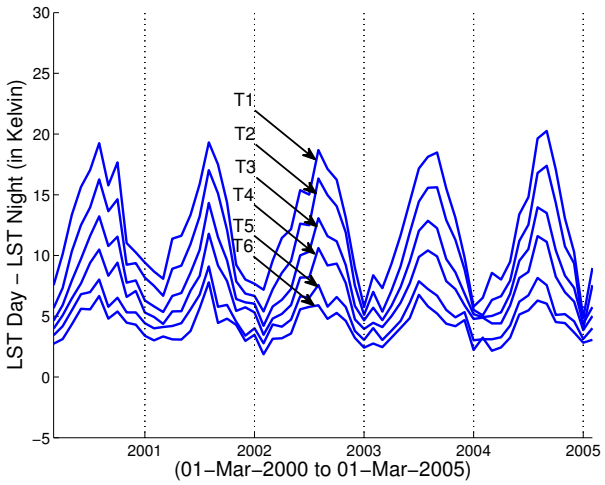


Figure 9. Centroid LST time series obtained by k-means clustering with $k = 6$.

Table II
EVALUATION RESULTS FOR SCHEMES INVOLVING k-MEANS CLUSTERING OF TIME SERIES, WITH VARYING VALUES OF k

Using LST	Logistic		Linear		Reg. Tree	
Model	Corr	$1-R^2$	Corr	$1-R^2$	Corr	$1-R^2$
Baseline	*0.7394*	*0.4553*	*0.7197*	*0.4871*	*0.7414*	*0.4528*
$k=6$	0.8432	0.2956	0.8431	0.2969	0.8364	0.3071
$k=8$	0.8496	0.2878	0.8492	0.2894	0.8399	0.3041
$k=12$	**0.8566**	**0.2787**	**0.8560**	**0.2804**	**0.8459**	**0.2966**
Using NDVI	Logistic		Linear		Reg. Tree	
Model	Corr	$1-R^2$	Corr	$1-R^2$	Corr	$1-R^2$
Baseline	*0.8416*	*0.2959*	*0.8179*	*0.3383*	*0.8434*	*0.2972*
$k=6$	0.8548	0.2731	0.8599	0.2652	0.8521	0.2785
$k=8$	0.8629	0.2582	0.8648	0.2551	0.8523	0.2767
$k=12$	**0.8691**	**0.2469**	**0.8685**	**0.2481**	**0.8531**	**0.2761**
Using EVI	Logistic		Linear		Reg. Tree	
Model	Corr	$1-R^2$	Corr	$1-R^2$	Corr	$1-R^2$
Baseline	*0.7428*	*0.4562*	*0.7126*	*0.5036*	*0.7527*	*0.4383*
$k=6$	0.8464	0.2943	0.8450	0.2969	0.8364	0.3096
$k=8$	0.8547	0.2818	0.8534	0.2840	0.8430	**0.3001**
$k=12$	**0.8610**	**0.2726**	**0.8602**	**0.2739**	**0.8439**	0.3003

cover at a particular location in a given year. Developing a single regression framework that models the behavior of every vegetation type with the forest cover thus increases the complexity of the problem, adversely affecting the performance of the algorithm. In this paper, we explored techniques that incorporate information about the vegetation type at a particular location for modeling its distinctive behavioral relationship with forest cover. The following two approaches were proposed and evaluated for segmenting the data space: (a) partitioning the feature space, and (b) clustering time series. A distinct regression algorithm was then developed for each detected segment. Experimental results

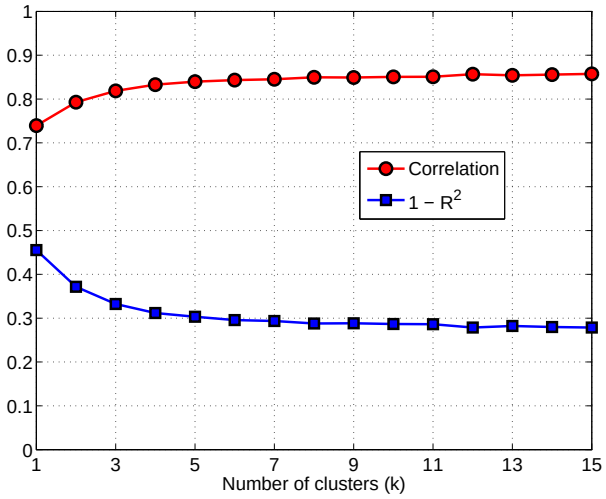


Figure 10. Variation in the performance of logistic regression model using LST as the predictor variable when the number of clusters is varied in k-means clustering

of the proposed techniques show a significant enhancement in the performance of a number of learning algorithms: logistic regression, linear regression, and regression tree, using either LST, NDVI, or EVI as the predictor variable.

The proposed techniques provide a preliminary insight into the role of vegetation type in modeling forest cover showing significant scope for future work in this direction. By analyzing the residuals in different vegetation types, it can be observed that the models can be further improved by using feature extraction techniques that best capture the information in multiple datasets. Specifically, the complementarities of multiple remote sensing datasets can be harnessed for constructing more robust representations of the information about vegetation type. The proposed techniques for incorporating information about vegetation type can be further improved by exploring ensemble methods using feature selection and feature space partitioning. The spatial and temporal structures in the data can be leveraged to construct more accurate regression models. Furthermore, regularization techniques can be explored for preventing local models from overfitting in the presence of training datasets with small sample size. Finally, the proposed approaches can be applied and evaluated in multiple regions of study, and in other vegetation-based regression domains involving interesting and representable heterogeneities in the data.

VIII. Acknowledgements

This research was supported in part by the National Science Foundation under Grants IIS-1029711 and IIS-0905581, as well as the Planetary Skin Institute. Access to computing facilities was provided by the University of Minnesota Supercomputing Institute.

References

[1] Land Processes Distributed Active Archive Center. http://lpdaac.usgs.gov.

[2] M. Austin. Spatial prediction of species distribution: an interface between ecological theory and statistical modelling. *Ecological modelling*, 157(2):101–118, 2002.

[3] T. Bishop and A. McBratney. A comparison of prediction methods for the creation of field-extent soil property maps. *Geoderma*, 103(1):149–160, 2001.

[4] G. B. Bonan. Forests and climate change: Forcings, feedbacks, and the climate benefits of forests. *Science*, 320(5882): 1444–1449, 2008.

[5] L. Breiman, J. Friedman, R. Ohlsen, et al. *Classification and regression trees*. CRC Press, Boca Raton, FL, 1984.

[6] G. Câmara, D. d. M. Valeriano, and J. V. Soares. Metodologia para o cálculo da taxa anual de desmatamento na Amazônia legal. São José dos Campos, INPE, 2006.

[7] A. Dobson. *An introduction to generalized linear models*. CRC press, 2002.

[8] N. Draper and H. Smith. Applied regression analysis (wiley series in probability and statistics). 1998.

[9] H. Eva, S. Carboni, F. Achard, et al. Monitoring forest areas from continental to territorial levels using a sample of medium spatial resolution satellite imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 65(2):191–197, 2010.

[10] J. Gillis. With deaths of forests, a loss of key climate protectors. *The New York Times*, October 1 2011.

[11] M. Gumpertz, J. Graham, and J. Ristaino. Autologistic model of spatial pattern of phytophthora epidemic in bell pepper: effects of soil variables on disease presence. *Journal of Agricultural, Biological, and Environmental Statistics*, pages 131–156, 1997.

[12] F. He, J. Zhou, and H. Zhu. Autologistic regression model for the distribution of vegetation. *Journal of agricultural, biological, and environmental statistics*, 8(2):205–222, 2003.

[13] P. McCullagh and J. Nelder. *Generalized linear models*. Chapman & Hall/CRC, 1989.

[14] N. McKenzie and P. Ryan. Spatial prediction of soil properties using environmental correlation. *Geoderma*, 89(1):67–94, 1999.

[15] J. Miller and J. Franklin. Modeling the distribution of four vegetation alliances using generalized linear models and classification trees with spatial dependence. *Ecological Modelling*, 157(2):227–247, 2002.

[16] J. Miller, J. Franklin, and R. Aspinall. Incorporating spatial dependence in predictive vegetation models. *Ecological Modelling*, 202(3):225–242, 2007.

[17] P. Moutinho and S. Schwartzman. *Tropical Deforestation and Climate Change*. Amazon Institute for Environmental Research, 2005.

[18] P. Tan, M. Steinbach, V. Kumar, et al. *Introduction to data mining*. Pearson Addison Wesley Boston, 2006.

[19] J. R. G. Townshend, M. Carroll, C. DiMiceli, et al. Vegetation Continuous Fields MOD44B, Collection 5. University of Maryland, College Park, MD, 2011.

[20] T. T. van Leeuwen, A. J. Frank, Y. Jin, et al. Optimal use of land surface temperature data to detect changes in tropical forest cover. *Journal of Geophysical Research*, 116, 2011.

[21] J. Wakefield. Disease mapping and spatial regression with count data. *Biostatistics*, 8(2):158–183, 2007.

[22] Q. Wang, J. Ni, and J. Tenhunen. Application of a geographically-weighted regression analysis to estimate net primary production of chinese forest ecosystems. *Global Ecology and Biogeography*, 14(4):379–393, 2005.