Remote Sens. **2014**, *xx*, 1-x; doi:10.3390/——

OPEN ACCESS

remote sensing

ISSN 2072-4292

www.mdpi.com/journal/remotesensing

Article

An Empirical Study of Atmospheric Corrections Impacts on Hyperspectral Classification of Savanah Tree Species Using k-fold Cross-Validated Non-linear SVM

Morteza Shahriari Nia [1,*], Daisy Zhe Wang [1], Milenko Petrovic [2], Stephanie Ann Bohlman [3] and Paul Gader [1]

[1] Department of Computer and Information Science and Engineering, University of Florida, 432 Newell Dr., Gainesville, Florida 32611, USA

[2] Institute for Human and Machine Cognition, 15 SE Osceola Ave, Ocala, Florida 34471, USA

[3] School of Forest Resources and Conservation, 349 Newins Ziegler Hall, Gainesville, Florida 32611, USA

* Author to whom correspondence should be addressed; msnia@cise.ufl.edu

Received: xx / Accepted: xx / Published: xx

Abstract: Identifying savannah species at ecological scale is a key step in measuring biomass, carbon reserves, drought and invasive species spread predictions. In this paper we perform classification and geo-mapping of tree species from hyperspectral imagery collected using AVIRIS airborne sensors at pixel level. We provide a thorough comparison of the effects of ATCOR and FLAASH atmospheric corrections in prediction accuracy. We exploit Gaussian Filters to eliminate sensor measurements and calibration errors, to the best of our knowledge we are the first in employing Gaussian Filters for hyperspectral species classification. This is a pilot study for NEON in Ordway-Swisher Biological Station in north-central Florida. Among indexes we found that applying NDVI and NIR wavelength ($734.1nm$) threshold filters do not play constructive roles in SVM model performance for species classification. We observed that removal of water absorption bands to be most helpful. Species classification was performed using variety of Support Vector Machine kernels where Radial Basis outperformed others. Our classification produces accurate predictions of about 75%.

Keywords: Species classification; Hyperspectral; High spatial and spectral resolution; Pixel-level classification; Support Vector Machines; Ordway-Swisher Biological Station; National Ecological Observatory Network; Airborne Observation Platform protocols;

1. Introduction

Mapping tree species by remote sensing techniques is an essential step in understanding how planetrary species play roles at ecological scale. This will enable us to study land covers, climate change, invasive species, plant competitions, predict fire potentials and spreading routes, soil characteristics and etc [1,2]. This kind of research has only been possible via the technological advancements in remote sensing facilities such as hyperspectral imagery or Light Detection and Ranging (LiDAR). Various studies have dealt with identifying tree species at both pixel level and crown level and as technology becomes available and economically feasible, studies tend to cover larger areas. Carnegie Airborne Observatory[1] (CAO) is a major pioneer in employing airborne technology for remote sensing at ecology scale where they study large areas in Amazonians, Kruger National Park in South Africa, Madagascar among others. Colgan et al. [2] uses a two stage Support Vector Machines (SVM) at pixel level and at crown level for tree species classification where LiDAR measurements were used for crown segmentation. Féret and Asner [3] study the accuracy of various parametric/non-parametric supervised classification techniques and observed that there is a clear advantage in using Regularized Discriminant Analysis, Linear Discriminant Analysis, and Support Vector Machines among others.

Cho et al. [4] compares accuracies when different hyperspectral sensors of CAO, WorldView2 and QuickBird are utilized by convolving the 72 bands of CAO to eight and four multispectral channels available in the WorldView-2 and Quickbird satellite sensors, respectively. Interestingly enough they observed that WorldView-2 produced more accurate classification results then QuickBird and finally CAO. Clark et al. takes on another perspective and compares lab measurements to pixel and to crown level and try to identify important wavelength regions for species discrimination [5]. They observed that optimal regions of the spectrum for species discrimination varied with scale. However, near-infrared (700-1327nm) bands were consistently important regions across all scales. Bands in the visible region (437-700nm) and shortwave infrared (1994-2435nm) were more important at pixel and crown scales [5]. Clark et al. in another work evaluates the effects of different metrics used for classification (indexes, derivatives, signals themselves and all together) [6]. There are other tree species classification works such as [3,7–12] that share the same approach with minor variations.

There are other schools of thought that try to identify more context specific features of remote rensing. For example Baldeck and Asner [13] try to measure how similar beta diversity of regions are; they use distance measures such as Euclidean distance and K-means clustering in unsupervised models. Using these clustering techniques one can provide a quick understanding of beta diversities and avoid costly and time consuming field data collections. However, this line of research needs more work as about 50% of pixels are classified as *other*, therefore any conclusion at this scale of uncertainty is not necessarily helpful, the same holds in Baldeck et al.'s latter work in [14]. Sometimes specially tailored tools and methodologies in this field are necessary. As an example, one should note that different bands in a hyperspectral image have different signal to noise ratios, and Principal Components (PC) transform will not always result in components with a steadily increasing noise level. This makes setting a cut-off point

[1] http://cao.stanford.edu/

difficult. Minimum Noise Fraction (MNF) [15] is a modified PC transform which produces a set of principal component images ordered in terms of decreasing signal quality.

In this paper we perform species classification and geo-mapping at large scale ($37km^2$). This is part of the pilot study for National Ecological Observatory Network (NEON). In our approach we take on a novel low-level perspective at hyperspectral species classification by taking into account the employed atmospheric correction and Gaussian Filtering of reflectance values. This setup is unique to this paper and to the best of our knowledge this is the first paper that considers using Gaussian Filters on hyperspectral data for species classification with an emphasis on impact of atmospheric corrections on hyperspectral signals. This asserts how a proper preprocessing on data can enhance classification accuracy. Beyond that we characterize effects of various predictors such as NDVI, NIR and water absorption bands. Below we descibe the scope and goals of the NEON project and how this research paper fits in the big picture.

NEON is a long term ecology monitoring project for discovering, understanding and forecasting impacts of climate change, land use change, and invasive species at continental-scale. National Science Foundation (NSF) provides funding for NEON as a 30-year project starting 2016. Local ecological measurements at sites distributed within 20 ecoclimatic domains across the contiguous United States, Alaska, Hawaii, and Puerto Rico will be coordinated with high resolution, regional airborne remote sensing observations [16]. This statistically represent unmanaged wildland conditions across the US. NEON moves an additional 40 relocatable terrestrial sites every five to seven years to locations where they may best capture specific ecological phenomena such as land use change and regional nitrogen deposition. There are also 36 aquatic sites collect representative aquatic data, where many aquatic sites are located near core and relocatable terrestrial sites[2]. Airborne Observation Platform (AOP) would be the remote sensing platform with equipments of meter/sub-meter resolution for hyperspectral and LiDAR measurements. This paper is a pilot study on the pre-mission airborne hyperspectral data collected. No operation at the scale and time span of NEON has been ever carried out before and a thorough study of the opportunities and challenges ahead is necessary specifically in the *remote sensing* perspective where the volume of data can quickly become overwhelming [17].

2. Data Collection

The NEON Southeast Domain 3 contains the southern portions of the Gulf Coast states, half of South Carolina, and all of Florida except for the southern tip. The candidate core site for Domain 3 is located at the Ordway-Swisher Biological Station (OSBS)[3] which is a 37-square kilometer area in Putnam County in north-central Florida and is managed jointly by the University of Florida and the Nature Conservancy located at UTM coordinates: 3,285,000 Northing, 405,000 Easting and UTM Zone 17. OSBS features diverse natural forests, small pine plantations nearby, a range of wildlife species that reflects the area's ecological communities, and a 75-year history of low human impact. Nine major plant communities exist within the region as defined by the Florida Natural Areas Inventory (FNAI) and these diverse targets

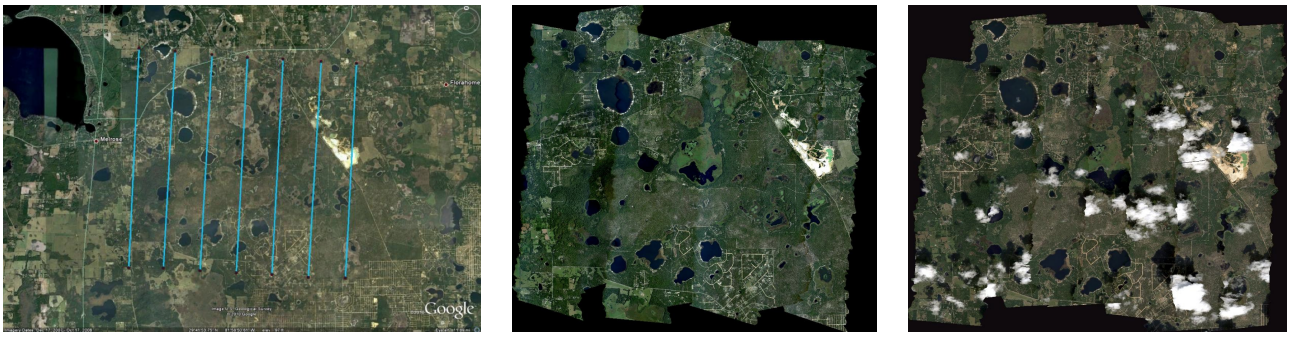
[2] http://www.neoninc.org/science/domains#sthash.aW1THj1N.dpuf
[3] http://ordway-swisher.ufl.edu/index.htm

Figure 1. JPL AVIRIS flights over OSBS [17]

(a) Flights ground tracks

(b) Hyperspectral true-color mosaic, morning 09/04/2010

(c) Hyperspectral true-color mosaic, mid-day 09/10/2010



are populated by sandhill, xeric hammock, upland mixed forest, baygalls, basin swamp, basin marsh, marsh lake, clastic upland lake and sandhill upland lakes. The sandhills community is managed using prescribed burning on a scheduled 3-year rotation. The ground sampling part of this campaign focused on a sandhill ecosystem dominated by Long-Leaf Pine (*Pinus Palustris*) and Turkey Oak (*Quercus Laevis*). The sandhill ecosystem at OSBS was selected for concurrent ground measurements because a NEON instrumented tower will be located within this ecosystem type [17,18].

Since the instrumentation slated for deployment on the eventual AOP remote sensing payloads were not yet available, airborne spectroscopic and LiDAR measurements were made during this campaign using existing systems that exhibit similar performance characteristics as the instrumentation under development [18]. It is important to note that the actual system for NEON will have better conformance of hyperspectral/LiDAR integrations with better spatial resolution.

Hypercpectral: AVIRIS (Airborne Visible/Infrared Imaging Spectrometer) operated by personnel from the Jet Propulsion Laboratory (JPL) deployed on a Twin Otter DeHavilland DHC-6-300 aircraft in partnership with the National Aeronautics and Space Administration Terrestrial Ecology Program was used to collect data. JPL has flown on two separate days over OSBS: morning of September 4, 2010 (between 9:30 and 10:30 am) and mid-day of September 10, 2010. Both of the flights were flown at approximately $4000m$ AGL at approximately 90 knots with zenith angle of 180.0 and azimuth angle of 0.0 with Altitude of $13kft$ (SOG $65 - 91kts$)[4] at mostly clear with some haze for Sept. 4th and some puffies for Sept. 10th. Details of these flights can be seen in Figure 1. Dependent on flight line, pixel size ranges from $3.3m$ to $3.6m$. Hyperspectral data was atmospherically corrected using FLAASH [19] and ATCOR [20] algorithms. There are eight flight lines and the total size of which is 20.5GB for FLAASH and 10.2GB for ATCOR data. There are 224 bands recorded with wavelengths from $365.93nm$ to $2496.24nm$.

Atmospheric characterization relied on measurements of a CIMEL sun photometer in coordination with the NASA Aerosol Robotic Network[5]. Measurements were collected on September 4, 2010 and

[4] NASA JPL AVIRIS Flight f100904t01: http://aviris.jpl.nasa.gov/cgi/flights_10.cgi?step=view_flightlog&flight_id=f100904t01 and http://aviris.jpl.nasa.gov/cgi/flights_10.cgi?step=view_flightlog&flight_id=f100910t01

[5] NASA AERONET: http://aeronet.gsfc.nasa.gov/cgi-bin/bamgomas_interactive

the derived atmospheric information was used to improve the atmospheric correction of the AVIRIS spectrometer data. Detailed measurements such as aerosol optical thickness, water vapor, and etc are available online[6]. FLAASH atmospheric correction is approximated as below [21]:

$$L_e \approx (\frac{(A+B)\rho_e}{1-\rho_e S} + L_a) \tag{1}$$

where ρ_e is an average surface reflectance for the pixel and a surrounding region, S is the spherical albedo of the atmosphere, L_a is the radiance back scattered by the atmosphere, and A and B are coefficients that depend on atmospheric and geometric conditions but not on the surface. On the other hand, ATCOR performs atmorpheric correction as follows [22]:

$$\rho = \frac{\pi\{d^2(c_0 + c_1 DN) - L_{path}\}}{\tau E_g} \tag{2}$$

where τ is the atmospheric (direct or beam) transmittance for a vertical path through the atmosphere, d is the earth-sun distance in astronomical units, c_0, c_1 and DN are the radiometric calibration offset, gain, and digital number, respectively. ρ is the surface reflectance and E_g is the global flux on the ground.

LiDAR: The National Center for Airborne Laser Mapping (NCALM)[7] coordinated the waveform-LiDAR flights in the same study areas and as with the AVIRIS flights as displayed in Figure 3 [17]. NCALM flew an Optech Gemini[8] waveform-LiDAR with a nominal altitude flight covering the entire OSBS on September 1, 2010. In total there were 33 flight lines collected over OSBS both in the morning and in the afternoon, 7 flight lines were re-flown due to clouds. Range scale of LiDAR data was at $1.0m$ and up to five returns were recorded. The parameters for the data collected are $70kHzPRF$, wide beam divergence of $0.8mrad$, $20deg$ half scan angle, and $40Hz$ scan frequency. The GPS/IMU navigation data are processed using the Applanix POSPac MMS software to determine the position, orientation, and trajectory of the aircraft. The discrete LiDAR return data are processed using the Optech DASHMap software to create point cloud data files in ASPRS LAS 1.2 format [23]. Each point data record contains information such as the intensity for the laser pulse and the X, Y, and Z geolocation of the point return. DASHMap reports up to four discrete returns: first, second, third, last. Using the geolocation information, the point clouds can be visualized 3-dimensionally.

3. species Classification

Upon the collection, orthorectification and atmospheric correction of hyperspectral and LiDAR data we work on identifying tree species. With the resolution of images being about 3 meters we are not getting pure pine or oak signals and there is lots of mixing pixels. A pixel's hyperspectral values can be a linear/nonlinear mixing of leaf, branch, soil, shade, and other signals. The timing of fights (September) adds to the challenge: leafs might not be as green or some trees might have already started to lose leaves

---

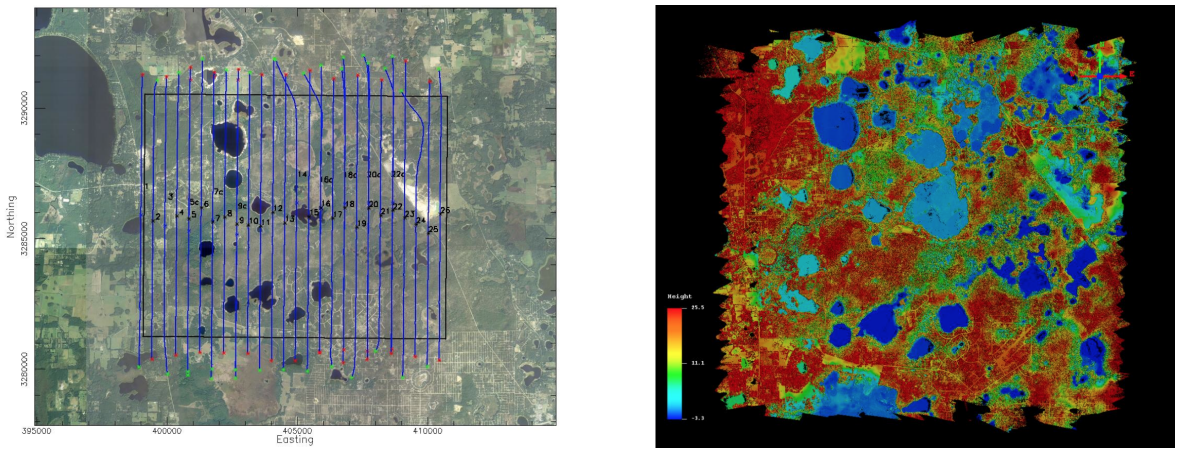
6 OSBS Aero Measurements: http://aeronet.gsfc.nasa.gov/cgi-bin/type_one_station_opera_v2_new?site=Ordway-Swisher

7 http://www.ncalm.cive.uh.edu/

8 Optech Gemini LiDAR sensor: http://www.optech.ca/gemini.htm

Figure 2. NCALM Optech Gemini LiDAR flights over OSBS 09/01/2010 [17]

(a) Flights ground tracks **(b)** LiDAR color height mosaic



and this leads to us getting more branch signals. Furthermore Longleaf Pine is a conifer (needlelaf), unlike broadleaf trees where signal return can be more accurate.

3.1. Field Data

For our classification task, we collected ground data of identifying tree species on February 28th, 2014. We drove to OSBS with a laptop that has ArcMap installed on and the ENVI image loaded on, we connected a professional grade GPS to the laptop. ArcMap reads GPS coordinates and mapped the polygons in the ENVI image. In this way, we marked several geo-polygons that had similar plant species in the ENVI image. Later on we overlayed the identified polygons with proper JPL AVIRIS flight which had the least amount of clouds. This approach works pretty accurate if you are not in a dense forest such as tropical forests where GPS signal under the tree canopy has high deviations due to NLOS (no-line of sight) of GPS signals. Even with these considerations, commercial GPS have sub-meter accuracy and when combined with error accumulated in orthorectification of flight images we still need to mark several land marks such as roads, big trees and etc to be able to re-verify marked points in the map and avoid shifts in coordinates.

Altogether we identified species for 1269 pixels. In Table 1 you can find the details of identified Regions of Interest (ROIs) along with their details.

The closer ROI Ids are means that they are collected at a closer vicinity (geographically/temporally), and the more distant ROI ids means that canopies are more apart. You can see that some species have lots of abundance (334 for Turkey Oak) and some have few (81 for Laurel Oak) this bias in population size inadverently affect classification accuracy.

The Oak (other) species represent generic oak specie (*specie details unknown*). Live Oak is specifically Sand Live Oak (*Quercus Geminata*), Laurel Oak represents *Quercus Hemisphaerica*, Turkey Oak is *Quercus Laevis*, Longleaf Pine is *Pinus Palustris*, and Pine (other) represents a mixture of different varieties of pines: Longleaf Pine (*Pinus palustris*), Loblolly Pine(*Pinus Taeda*) or Slash Pine (*Pinus Elliottii*).

We use a combination of grid search and heuristics to adjust SVM parameters for better performance.

Table 1. Field Data Specifications

	Broad Leaf								Conifer			
Specie	Oak (other)		Laurel Oak		Live Oak		Turkey Oak		Pine (other)		Longleaf Pine	
	ROI Id	# of Pixels	ROI Id	# of Pixels	ROI Id	# of Pixels	ROI Id	# of Pixels	ROI Id	# of Pixels	ROI Id	# of Pixels
	23	62	52	29	4	48	10	73	20	68	12	107
	19	19	50	20	13	37	15	34	14	67	9	30
	17	15	49	17	28	22	21	33	2	30	45	23
	35	15	48	9	29	22	11	26	5	24	18	22
	36	10	53	6	30	22	6	23	3	20	44	21
							34	23	42	17	37	17
							31	18	43	16	16	16
							25	17	1	13	40	15
							26	16	22	10	47	14
							32	16	38	10	39	12
							27	15			41	12
							7	14			46	11
							24	13			8	7
							33	13				
Total	5 ROIs	121 Pixels	5 ROIs	81 Pixels	5 ROIs	151 Pixels	14 ROIs	334 Pixels	10 ROIs	275 Pixels	13 ROIs	307 Pixels
Partial Total	29 ROIs - 687 Pixels								23 ROIs - 582 Pixels			
Grand Total	52 ROIs - 1269 Pixels											

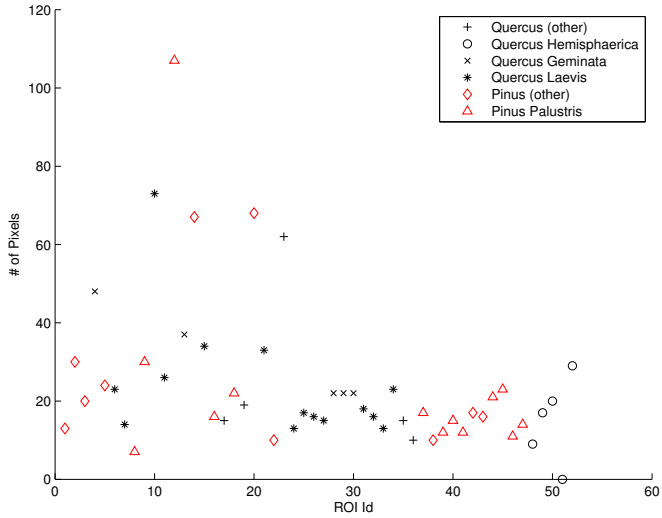
3.2. Preprocessing

We load hyperspectral images in Matlab using an in-house upgraded version of `enviread`, initially developed by Dr. Ian Howat at Ohio State University [24]. We check for the consistency of calibration and uniformness of pixel sizes. As different flights have different altitudes and hence pixel resolutions, this is an essential step to account for. Defining a hyperspectral image I with dimensionality (x, y, w, z), where $x \in X = [167000, 833000]$ meters represents the range of UTM Easting values, $y \in Y = [0, 9400000]$ meters represents UTM Northing values, $w \in W = \{1, \ldots, 224\}$ is the index of the reflectance wavelengths, and $z \in Z = \{1, \ldots, 60\}$ is the UTM zone of the image. Based on our observations, we take constant $\xi = 10000$ as a cut-off point to avoid erroneous sensor readings. There are some noises in JPL AVIRIS measurements such as negative reflectance values; the range of hyperspectral reflectance values is in range $[-32762, 32724]$: one should note that reflectance is the proportion of sun radiance signals which should be a positive value, but in normalized form reflectance is between zero and one. Below you can see the normalization process:

$$I_{xywz} = \begin{cases} 0 & \text{for} \quad I_{xywz} < 0 \\ 1 & \text{for} \quad I_{xywz} > \xi \\ \sqrt{\frac{I_{xywz}}{\xi}} & \text{otherwise} \end{cases} \tag{3}$$

To normalize we set negative reflectance values to zero and values greater than 10,000 to 10,000. To enhance the intensity of readings we take the square-root of signal returns. This procedure is due to the following facts: a) There is no standard output of reflectance data and even reflectance of a single crown at different pixels can be quiet different. Unlike minerals that have fixed and known reflectance values, trees can different signal returns based on generation, condition of growth (water quality, climate, soil, and etc) which can make this task challenging. b) We do not have ground reflectance values produced by NEON available to us. c) Due to low resolution of images signals are all mixed. Due to these reasons,

Figure 3. Field Data Distribution of ROIs



empirically obtained thresholds and ranges are inevitable. Regarding square root, one should note that without taking the square root the images lack proper day-light intensity and appear dark.

3.2.1. Filter Water Absorption Bands

In our calculations wavelengths corresponding to strong water vapor absorption bands in the atmosphere are excluded. This includes 1333.2nm to 1482.7nm, 1791.6nm to 1967.4nm, and 2406.9nm to 2496.2nm. Due to strong absorption at those wavelengths, a small radiance signal is measured by the instrument. This leads to errors in the reflectance calculation (as several highly random upward spikes in reflectance plots).

3.2.2. Filter Non-vegetated/Shaded Pixels

A filter of $NIR < 0.33$ excludes heavily shaded samples which usually have distorted reflectance signals [2]. Normalized Difference Vegetation Index (NDVI) is an index which shows how green a pixel is and is usually used to remove material that does not belong to planetary material such as roads, clouds and any not-vegetated area, even grass and so on. By properly applying an NDVI filter you will end up in major tree crowns. As you can see in 1, we have tens of pixels per ROI and considering that each pixel is about $3m$ wide and the fact that the species look into do not have wide crowns; this means that ROIs are generously marked and span several tree crowns. This includes the empty spaces between crowns, some shady pixels, understory grass, and etc. From within such generous canopy the initial belief is that to properly classify tree species we should get rid of low NDVI pixels from such large canopies. Contrary to this belief due to the mixing nature of hyperspectral pixels and aggregate operation of classificaion techniques in multi-dimensional space, removing low NDVI pixels from tree canopies degraded classification performance by about 10%. The take away of this experiment was that for specifically marked canopies it is better to preserve even low NDVI pixel as they might have some added value in the aggregate operation of the classification algorithm. NDVI is defined as below:

$$NDVI = \frac{NIR - VIS}{NIR + VIS} \tag{4}$$

where is NIR the reflectance in the reflective near-infrared wavelengths (725-1100 nm) and VIS is the reflectance in the visible (red) wavelengths (580-680 nm). The principle behind this is that VIS is in a part of the spectrum where chlorophyll causes considerable absorption of incoming radiation, and the NIR is in a spectral region where spongy mesophyll leaf structure leads to considerable reflectance [25,26]. For this purpose we chose the band at 665.6 nm for red and 734.1 nm for near-infrared. By filtering out pixels with $NDVI < 0.4$ we are essentially removing pixels that are not green.

3.2.3. Gaussian Filter

Real-life sensor measurements are far from perfect and there are many nosy readings along different bands. We take advantage of abundance of bands (224 bands) and take use of their aggregated information by applying a Gaussian filter. We take a Gaussian window w of size $N > 0$, the coefficients of a Gaussian window are computed as below:

$$w(n) = e^{-\frac{1}{2}(\alpha \frac{n}{N/2})^2} \tag{5}$$

where $-\frac{(N-1)}{2} \leq n \leq \frac{(N-1)}{2}$, and α is inversely proportional to the standard deviation (σ) of a Gaussian random variable ($\sigma = \frac{N}{2\alpha}$). Once we have Gaussian parameters we perform convolution to apply the smoothing factor. By convolving vectors $u \in \mathbb{R}^m$ and $v \in \mathbb{R}^n$, we will have vector $w \in \mathbb{R}^{m+n-1}$ such that:

$$w(k) = \sum_j u(j)v(k - j + 1) \tag{6}$$

3.3. Support Vector Machines

It is well known in literature that Support Vector Machines (SVM) outperforms other algorithms on species classification [2,4,14]. Our focus is on the impact of atmospheric corrections (FLAASH vs ATCOR) at pixel level classification using *non-linear multi-class SVM*. We parametrize SVM with k-fold cross validation where $k = 5$, which means for each model we perform classification 5 times where at each time a separate non-overlapping portion of ground data is specified for train and test purposes. The ratio of train vs test is 4 to 1 from total samples. We

Classifier non-linearity comes from taking the following non-linear functions as kernel for SVM:

- Polynomial function kernel.
- Radial Basis Function (RBF) kernel

Regarding multi-class classification, we create $\binom{c}{2}$ classifiers where c is the number of classes. In this case $c = 6$, hence we train $\binom{6}{2} = 15$ classifiers at each iteration of k-fold. We train all the classifiers and to decide on the class a given test case belongs to, we perform majority voting among classifiers to decide the class of each pixel. We avoid aggregating canopy pixels classes to determine the class of

the canopy as this process can be performed using heuristics such as majority of pixels or etc, which is irrelevant to the purpose of pixel-level classification. Furthermore, it might not be always feasible to assume that we always know the correct boundaries of a canopy in new data sets.

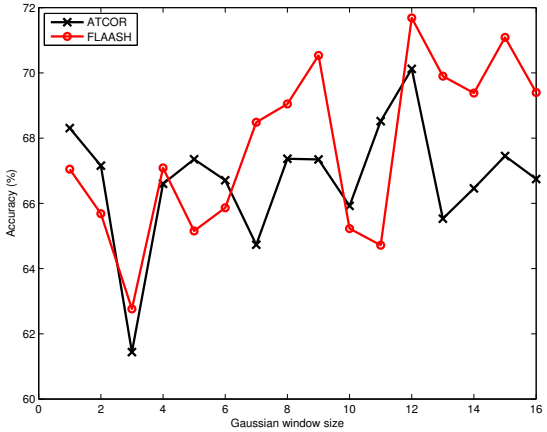
4. Results and Discussion

We study the impact of optimzing classifier parameters on FLAASH and ATCOR atmospheric corrections. Our focus is on the impact of data pre-processing filters on the performance of the species classificaion while we compare the two. An implementation of a one-vs-one multi-class k-fold cross-validation setup of SVM using non-linear kernels (polynomial and radial basis) functions is used for our model. In our scenario $k = 5$ as for some of our species we only have 5 canopies and we canot $k > 5$ for a valid implementation of k-fold cross validation. Majority vote determines the specie of a pixel. Test and train sets are canopy-aware, meaning that pixels of a single canopy is either used for training or test sets and not both. A mixture of grid search and heuristic optimization determines various parameters for SVM kernels, but for brevity we only present 2D diagrams here as other parameters settings are out of scope of this project. Specifically we set a cost of $C = \infty$ for misclassification, meaning that we have little to zero tolerance for improperly classified samples.

In Figure 4 we show how removing water absorption bands helps us improve prediction accuracy. Before removing these bands (Figure 4a) prediction accuracy starts at about 66% and rises to about 71% as we increase Gaussian window size (about 5% performance improvement). There is a positive trend across different window sizes which means we can expect better accuracy the bigger the window size gets. This stems from the fact that water absorption bands add to randomness and as we increase window size this randomness is dissolved to the majority of neighbors and would have less and less impact as window size increases. After removing water absorption bands (Figure 4b) prediction accuracy starts at 74% and rises to about 76.5% at a window size of 8. Beyond that there is no significant change in accuracy. As window size increases you can see that there is no predictable impact of Gaussian smoothing in accuracy; This is due to the fact that all the currently available data signals has soem level

Figure 4. Impact of Gaussian Window on Prediction Accuracy

(a) Before removing water absorption bands



(b) After removing water absorption bands

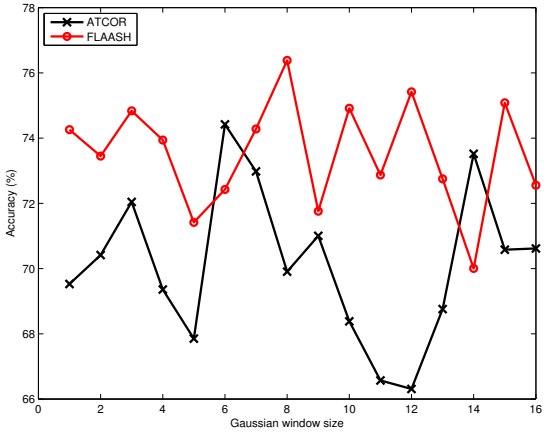
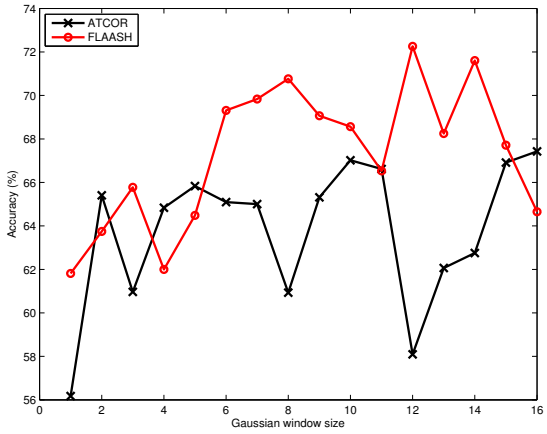


Figure 5. Impact of Removing low NDVI and NIR pixels



of inherent entropy (less noisy) and increasing window size beyond certain point just takes data away from their original meaning which can either help or harm accuracy in unpredictable ways.

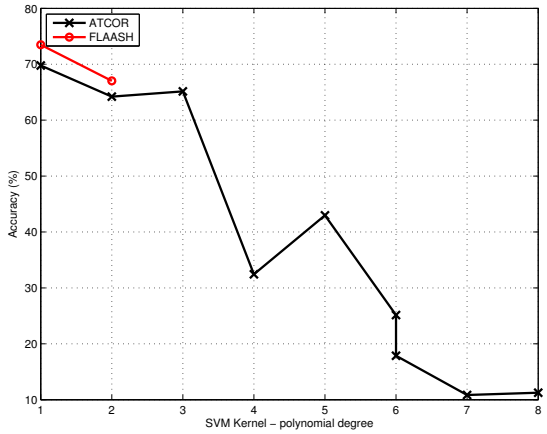
Next, we look at the impact of removing low NDVI-NIR pixels from dataset. As you can see in Figure 5 there is a general degradation of performance from the previous scenairo that we used the whole dataset. In FLASSH dataset we are at about 70% and ATCOR yields 66% accuracy. This is due to the fact that pixel size is large ($3m$) and canopies are large as well. According to Table 1, canopy sizes are in the range of $[6, 107]$ pixels. This generous marking of canopies includes areas with little green-ness, shaddows, branches and etc (low NDVI and low NIR values). But you should note that due to the mixing nature of reflectance values, even low NDVI/NIR pixels of a continous canopy still contains signals from the underlying species, Which might not be as green. Here we can see that including low NDVI/NIR pixels of a continous canopy actually helps the performance of the classification model with an impact of about 4%.

Finally, we look at the effect of atmospheric correction on prediction accuracy vs classification model parameters. $C, \sigma, P, MaxIter$ and *optimization method* are the knobs of SVM that we tune in a mixture of grid and heuristic search. C stands for cost or penalty of misclassification against simplicity of the decision surface, σ (or also known as γ in literature) defines how far the influence of a single training example reaches, with low values meaning 'far' and high values meaning 'close' in RBF function. P is the polinomial degree for polynomial function as kernel. $MaxIter$ is the maximum number of iterations the optimization function is supposed to run. and *optimization method* defines what optimization method we select. For parameters except σ and P we set $C = +\infty$, $MaxIter = 10,000$ and *optimization method* to quadratic programming. In what follows we evaluate the impact of P and σ respectively as demonstrated in Figure 6.

Figure 6a demonstrates the impact of polynomial degree P on prediction accuracy using a polynomial kernel in SVM. FLAASH atmospherically corrected data yields 73.5% of accuracy while ATCOR results in 69.8%. The simpler the polinomial, the better the performance; with more complicated polynomial we get a high bias classification model which performs poorly when evaluated on test data. For FLAASH atmospheric correction with $P \geq 3$ the optimization function does not reach cnvergence in the number of $MaxIter$. On the other hand, ATCOR data performs as predicted, the data seems to be showing a

Figure 6. Parameter tuning for classification algorithms

(a) Tuning polynomial order for SVM with polynomial kernel function



(b) Tuning σ in SVM with RBF kernel function

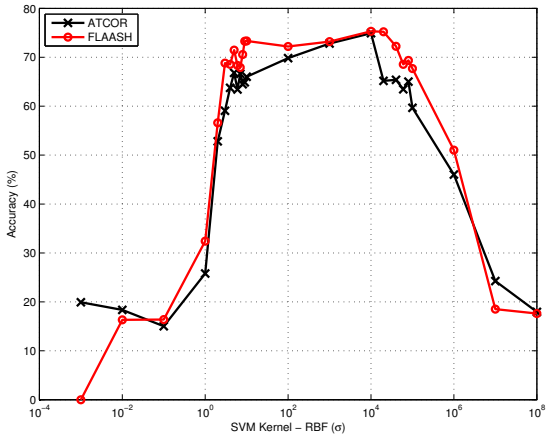


Table 2. Demo of confusion matrix using RBF kernel function with accuracy 75.2% near peak

		Predicted Class						
		Pine (other)	Longleaf Pine	Turkey Oak	Live Oak	Oak (other)	Laurel Oak	$\sum$
Known Class	Pine (other)	10	10	3	0	0	0	23
	Longleaf Pine	20	36	0	0	0	4	60
	Turkey Oak	1	0	62	0	0	0	63
	Live Oak	0	0	0	44	4	0	48
	Oak (other)	0	0	0	0	1	9	10
	Laurel Oak	1	0	0	0	0	5	6
	$\sum$	32	46	65	44	5	18	210

linear separability which as the kernel becomes more complex shows signs of overfitting as FLAASH. Accuracy gets as low as 10.8% and 11.3% with polinomila degrees of 7 and 8.

The Radial Basis kernel has better performance yields than polynomial. As shown in Figure 6b, with a peak at 75.3% we achieve the best results using FLAASH data while ATCOR comes close at 74.9%. RBF does not show good performance at either too low or too high σ values. σ is the inverse of the width of the RBF kernel (roughly defining the area of influence of a support vector); in other terms, it defines how much influence a single training example has. The larger σ is, the closer other examples must be to be affected. As RBF takes data to a higher dimentionality, on what σ provides: A small σ gives you a pointed bump in the higher dimensions, a large gamma gives you a softer, broader bump. So neither extreme shows a good fit and a middle point of $\sigma = 10,000$ provides best results. On the negative side, FLAASH begins with an accuracy of 0% and ATCOR at 19.9%, but they quickly get to a stable region close by to each other while FLAASH demonstrates superior performance in most of the cases. There is a good range of σ values ($[10, 10000]$) where we reach a somehow plain in prediction accuracy which shows a more robust performance than polynomia.

Table 2 demonstrates the confusion matrix of the proposed clasification model at a near peak setup (75.2%). You can see that the majority of missclassifications are between Pine (other) class and Longleaf Pine. Similar misclassification can be found between Oak (other) and other types of Oak. This is due ot the fact that this class of Oak or Pines is a mixture of different types of Oak or Pine repectively which might include Longleaf Pine or Laurel Oak and such conformance migt be unavoidable. The main advantage of this approach is that there is mere misclassification between oak vs Pine category. The oak category is rarely misclassified as pine, but pines have been mistaken with Turkey Oak and Laurel Oak. One should also note that different species of oak are not misclassified to each other. Laurel Oak, Turkey Oak and Live Oak are not mistaken for each other, this shows a good intra broad-leaf species classification.

5. Conclusions

Identifying species using remote sensing technologies such as hyperspectral and LiDAR sensors has a critical utility in studying global warming, bio-mass estimation, carbon preserves, invasive species identification and etc. In this paper we perform species classification using SVM over AVIRIS hyperspectral data available via NEON from Ordway Swisher Biological Station in north-central Florida. Our focus is on comparing FLAASH and ATCOR atmospheric corrections while we analize pre-processing techniques both at signal level (Gaussian Filter, Water Absorption Bands Filter) and pixel level (NDVI and NIR Filters). We found Gaussian Filter and Water Absorption Bands Filter to be helpful in prediction model performance but NDVI and NIR degraded the performance when applied to continous canopies of similar species. In all these scenarios we compared FLAASH and ATCOR atmospheric corrections and noted that FLAASH data supersedes ATCOR in majority of the scenarios. Our model performed robust in intra broad-leaf classification with minor inter conifer/broad-leaf misclassifications.

References

1. Scholes, R.; Archer, S. Tree-grass interactions in savannas. *Annual review of Ecology and Systematics* **1997**, *28*, 517–544.
2. Colgan, M.S.; Baldeck, C.A.; Féret, J.B.; Asner, G.P. Mapping savanna tree species at ecosystem scales using support vector machine classification and BRDF correction on airborne hyperspectral and LiDAR data. *Remote Sensing* **2012**, *4*, 3462–3480.
3. Féret, J.; Asner, G.P. Tree species discrimination in tropical forests using airborne imaging spectroscopy. *Geoscience and Remote Sensing, IEEE Transactions on* **2013**, *51*, 73–84.
4. Cho, M.A.; Mathieu, R.; Asner, G.P.; Naidoo, L.; van Aardt, J.; Ramoelo, A.; Debba, P.; Wessels, K.; Main, R.; Smit, I.P.; others. Mapping tree species composition in South African savannas using an integrated airborne spectral and LiDAR system. *Remote Sensing of Environment* **2012**, *125*, 214–226.
5. Clark, M.L.; Roberts, D.A.; Clark, D.B. Hyperspectral discrimination of tropical rain forest tree species at leaf to crown scales. *Remote sensing of environment* **2005**, *96*, 375–398.

6. Clark, M.L.; Roberts, D.A. Species-level differences in hyperspectral metrics among tropical rainforest trees as determined by a tree-based classifier. *Remote Sensing* **2012**, *4*, 1820–1855.
7. Dalponte, M.; Ørka, H.O.; Ene, L.T.; Gobakken, T.; Næsset, E. Tree crown delineation and tree species classification in boreal forests using hyperspectral and ALS data. *Remote sensing of environment* **2014**, *140*, 306–317.
8. Féret, J.B.; Asner, G.P. Semi-supervised methods to identify individual crowns of lowland tropical canopy species using imaging spectroscopy and LiDAR. *Remote Sensing* **2012**, *4*, 2457–2476.
9. Ghosh, A.; Fassnacht, F.E.; Joshi, P.; Koch, B. A framework for mapping tree species combining hyperspectral and LiDAR data: Role of selected classifiers and sensor across three spatial scales. *International Journal of Applied Earth Observation and Geoinformation* **2014**, *26*, 49–63.
10. Immitzer, M.; Atzberger, C.; Koukal, T. Tree species classification with random forest using very high spatial resolution 8-band WorldView-2 satellite data. *Remote Sensing* **2012**, *4*, 2661–2693.
11. Naidoo, L.; Cho, M.; Mathieu, R.; Asner, G. Classification of savanna tree species, in the Greater Kruger National Park region, by integrating hyperspectral and LiDAR data in a Random Forest data mining environment. *ISPRS Journal of Photogrammetry and Remote Sensing* **2012**, *69*, 167–179.
12. Ustin, S.L.; Gitelson, A.A.; Jacquemoud, S.; Schaepman, M.; Asner, G.P.; Gamon, J.A.; Zarco-Tejada, P. Retrieval of foliar information about plant pigment systems from high resolution spectroscopy. *Remote Sensing of Environment* **2009**, *113*, S67–S77.
13. Baldeck, C.A.; Asner, G.P. Estimating Vegetation Beta Diversity from Airborne Imaging Spectroscopy and Unsupervised Clustering. *Remote Sensing* **2013**, *5*, 2057–2071.
14. Baldeck, C.; Colgan, M.; Féret, J.B.; Levick, S.; Martin, R.; Asner, G. Landscape-scale variation in plant community composition of an African savanna from airborne species mapping. *Ecological Applications* **2014**, *24*, 84–93.
15. Green, A.A.; Berman, M.; Switzer, P.; Craig, M.D. A transformation for ordering multispectral data in terms of image quality with implications for noise removal. *Geoscience and Remote Sensing, IEEE Transactions on* **1988**, *26*, 65–74.
16. Kampe, T.U.; Johnson, B.R.; Kuester, M.; Keller, M. NEON: the first continental-scale ecological observatory with airborne remote sensing of vegetation canopy biochemistry and structure. *Journal of Applied Remote Sensing* **2010**, *4*, 043510–043510.
17. Krause, K.; Kuester, M. Airborne Observation Platform (AOP) Pathfinder 2010 Data Release. http://neoninc.org/pds/files/NEON.AOP.015068.pdf. The National Ecological Observatory Network is a project sponsored by the National Science Foundation and managed under cooperative agreement by NEON, Inc. The NEON 2010 Pathfinder data set is based on work supported by the National Science Foundation under Grant DBI-0752017.
18. Kampea, T.; Krausea, K.; Meiera, C.; Barnetta, D.; McCorkela, J. The NEON 2010 Airborne Pathfinder Campaign in Florida, 2010. NEON Technical Memo 002.
19. Adler-Golden, S.; Berk, A.; Bernstein, L.; Richtsmeier, S.; Acharya, P.; Matthew, M.; Anderson, G.; Allred, C.; Jeong, L.; Chetwynd, J. FLAASH, a MODTRAN4 atmospheric correction

package for hyperspectral data retrievals and simulations. Proc. 7th Ann. JPL Airborne Earth Science Workshop, 1998, pp. 97–21.

20. Richter, R.; Schläpfer, D. Atmospheric/topographic correction for satellite imagery. *DLR report DLR-IB* **2005**, pp. 565–01.
21. Adler-Golden, S.M.; Matthew, M.W.; Bernstein, L.S.; Levine, R.Y.; Berk, A.; Richtsmeier, S.C.; Acharya, P.K.; Anderson, G.P.; Felde, J.W.; Gardner, J.; others. Atmospheric correction for shortwave spectral imagery based on MODTRAN4. SPIE's International Symposium on Optical Science, Engineering, and Instrumentation. International Society for Optics and Photonics, 1999, pp. 61–69.
22. Richter, J.; Schläpfer, D. Atmospheric / Topographic Correction for Satellite Imagery. ATCOR-2/3 User Guide, Version 8.3.1, February 2014. ATCOR, 2014.
23. LAS Specification Version 1.2. http://www.asprs.org/a/society/committees/standards/asprs_las_format_v12.pdf. Approved by ASPRS Board 09/02/2008.
24. Howat, I. ENVI file reader, updated 2/9/2010. http://www.mathworks.com/matlabcentral/fileexchange/15629-envi-file-reader--updated-2-9-2010. This code initially developed by Ian Howat, and updated by Yushin Ahn, Ray Jung and CSI lab University of Florida.
25. Tucker, C.J. Red and photographic infrared linear combinations for monitoring vegetation. *Remote sensing of Environment* **1979**, *8*, 127–150.
26. Jackson, R.; Slater, P.; Pinter Jr, P. Discrimination of growth and water stress in wheat by various vegetation indices through clear and turbid atmospheres. *Remote sensing of environment* **1983**, *13*, 187–208.