CLASSIFICATION OF CHEATGRASS INVASION USING BIOPHYSICAL AND REMOTE SENSING DATA

Aaron Tuor & Kyle Larson
Pacific Northwest National Laboratory
{aaron.tuor, kyle.larson}@pnnl.gov

ABSTRACT

In this study we explore machine learning approaches which incorporate geophysical data and readings from multiple remote sensing platforms for predictive mapping of vegetation over very large, ecologically diverse regions. We focus specifically on mapping an invasive exotic annual grass, cheatgrass (*Bromus tectorum*), that has become a major land management issue in the Western United States. We explore the advantages of integrating imagery from both the LandSat7 and MODIS platforms which have complementary spectral bandwidth, spatial resolution, and temporal frequency. Our experiments demonstrate that machine learning methods integrating biophysical variables and imagery from multiple platforms can be beneficial to the predictive mapping of invasive vegetation. Our best performing model is used to create a map of coverage predictions to aid in land management over 288 million acres of the Western United States.

1 INTRODUCTION

Accurate and up to date maps of vegetation coverage are critical assets for land management agencies tasked with mitigating damage wrought by invasive species. Current methods for mapping invasive annual grasses from satellite and aerial imagery usually involve choosing one platform (Rogan et al., 2008). However multiple platforms may potentially be useful, e.g., using finer spatial resolution imagery improves prediction Jeevalakshmi et al. (2016), but platforms with more frequent temporal resolution are also important for detecting invasive species which maintain growth cycles distinct from native vegetation (West et al., 2017). In this work we demonstrate the effectiveness of machine learning models which integrate biophysical data and multiple satellite platfroms.

Cheatgrass is an invasive grass, prevalent in the western United States (USDA, 2015) which threatens ecosystem function, rangeland health, and human safety (Knapp, 1996). The basis to many of these threats is an increase in fine fuels, which can lead to increased fire frequency and consequent loss of native vegetation and wildlife habitat (Bradley & Mustard, 2005). Compounding this, cheatgrass is more resilient than native vegetation to returning growth following a fire (Pellant et al., 1990; Melgoza et al., 1990). Knowledge about significant contributors to fire risk is increasingly critical in light of recent record breaking wildfire seasons. However, while cheatgrass abounds throughout much of the western U.S., detailed spatial information about its presence and abundance are still lacking across this extent (Bradley & Mustard, 2005). Our study addresses this gap by producing a map of coverage estimates across approximately 288 million acres of suspected cheatgrass habitat.

2 DATASET

In this section we explain the data used in this study: spatial data of various biophysical parameters, time series of MODIS and Landsat-7 data, and field observations of cheatgrass cover. The study area spans 23 EPA Level III ecoregions Omernik & Griffith (2014) which we use as a categorical variable. Two other categorical biophysical variables, soil moisture-temperature regimes and existing vegetation type, were selected because they are useful indicators of conditions that affect plant community dynamics. We use an additional 50 continuous variables related to climate, elevation, and location. Table 1 contains brief descriptions of the continuous input variables.

Variable	Description
$X_{1:2}$	latitude & longitude
X_3	Elevation.
X_4	Potential Relative Radiation.
X_5	Median winter precipitation.
X_6	Median growing degree days.
$X_{7:50}$	30-year climatic conditions.
$X_{51:323}$	Landsat-7 bands 1-10.
$X_{324:664}$	MODIS bands 1-8.

Table 1: Continuous variables (Left)

Subset	Values	Description
D_1	$\mathbf{x}_{1:50}$	Biophysical
D_2	$\mathbf{x}_{1:50}$ and $\mathbf{x}_{324:664}$	Biophysical + MODIS
D_3	$\mathbf{x}_{1:323}$	Biophysical + LandSat
D_4	$\mathbf{x}$	All values

Table 1a: Variable Subsets (Above)

Time series of both annual- and seasonal-composite MODIS (Terra, 2001-2016) and Landsat-7 (2003-2012) data were used in this study. Both sets of imagery were composited to reduce residual cloud and aerosol contamination. We used both annual and seasonal composite images for spring and summer periods which correspond to distinguishing periods in the life cycle of cheatgrass. To accomodate varying spatial granularity (30m^2-250m^2) across satellite platforms, we aligned the data using bi-linear interpolation to the finest spatial resolution available. The aligned data set over the full study region comprises ≈3 billion 30m^2 contiguous geographic locations in the western U.S.[1].

We obtained 6650 field observations from multiple studies that collected vegetation measurements (on transects ranging from 25-m to 100-m) between 2001 and 2014. Among these locations is a strong break in the distribution of cheatgrass canopy cover values at approximately 2 percent canopy cover which represents approximately equal proportions of field measurements (48 % and 52%). We train models on field measurement locations to predict between these two canopy cover classes above and below this natural break. With $\mathbf{x}$ as the vector of values associated with a location, Table 1a defines data subsets we use for evaluating potential gains and losses in classification accuracy by the addition of spectral data to a baseline set of biophysical variables.

3 METHODS

In this section we outline four machine learning methods for coverage classification: Random Forest (RF), Logistic Regression (LR), Multi-Layer Perceptron (MLP), and a joint Recurrent Neural Network model (JRNN). While we explore model performance over each of the 4 input subsets defined in Table 1a for all models (excluding D_1 for JRNN), there a few differences in the treatment of the inputs. For satellite data, the JRNN model uses a time series format, whereas the MLP, LR, and RF models treat this data as a high-dimensional flat vector. Categorical values are mapped to vectors in two ways: one-hot vectors for RF, and embedding vectors for LR, MLP, and JRNN. For all models, continuous variables were standardized to have zero mean and unit variance.

We chose RF for comparison as it has been shown to perform well for predictive vegetation mapping Cutler et al. (2007), and typically provides competitive results compared to deep learning models with limited training data Kussul et al. (2017). Using sci-kit learn's (Pedregosa et al., 2011) RF implementation, we experimented with the full range of available hyperparameters to find the best model configuration for each data subset.

3.1 LOGISTIC REGRESSION

To provide perspective on the value of deep learning for deriving predictive feature representations for ecological modeling, we explore a baseline linear LR model based directly on the standardized predictor variables. Let $D_p \in \mathbb{R}^n$ be a vector of input values as defined in Table 1a, and q be the size of an embedding vector. The input for a given location to our LR model is a vector $\hat{\mathbf{x}} \in \mathbb{R}^{3q+n}$ composed of standardized real values of continuous predictor variables and categorical embeddings. The output, $\hat{\mathbf{y}} \in \mathbb{R}^2$, is a predicted distribution over classes of cheatgrass canopy cover:

$$\hat{\mathbf{x}} = \left[\mathbf{w}^\intercal_{\texttt{eco}} \quad \mathbf{w}^\intercal_{\texttt{cov}} \quad \mathbf{w}^\intercal_{\texttt{soil}} \quad D_p^\intercal\right]^\intercal, \qquad \hat{\mathbf{y}} = \texttt{softmax}(\mathbf{W}\hat{\mathbf{x}} + \mathbf{b}) \tag{1}$$

[1]Currently pursuing permissions to release our aligned dataset collected from many sources.

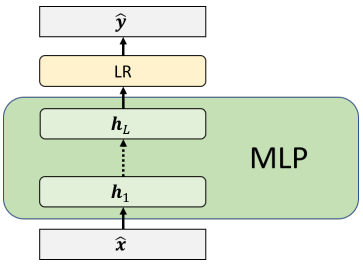

Figure 1: **M**ulti-**L**ayer **P**erceptron

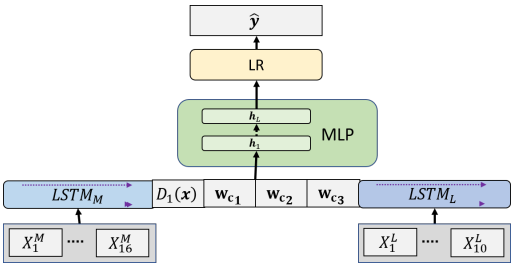

Figure 2: **J**oint **R**ecurrent **N**eural **N**etwork

The embedding vectors $\mathbf{w}_{\texttt{cov}}$, $\mathbf{w}_{\texttt{eco}}$, and $\mathbf{w}_{\texttt{soil}} \in \mathbb{R}^q$, correspond to the general cover class, ecoregion, and soil moisture temperature regime class values for a location. The matrix $\mathbf{W} \in \mathbb{R}^{2\times 3q+n}$, vector $\mathbf{b} \in \mathbb{R}^2$, and embedding matrices are model parameters.

3.2 Deep Learning Models

Deep neural networks were chosen as prospective models as past studies have demonstrated their effectiveness for remote vegetation mapping (Kussul et al., 2017). The deep learning (DL) models used in this study can be viewed as a composition of transformations, mapping the input to a discriminative latent representation. Figures 1, 2 show computational graphs of the DL models.

The input to the MLP model is the same as defined for the LR model. The MLP consists of L hidden layers $\mathbf{h}$ which are defined as:

$$\mathbf{h}_l = \texttt{RELU}(\mathbf{W}_l\mathbf{h}_{l-1} + \mathbf{b}_l), \quad \text{where} \quad \mathbf{h}_0 = \widehat{\mathbf{x}}, \quad \text{and} \quad \texttt{RELU}(\mathbf{z})_i = \texttt{max}(\mathbf{z}_i, 0)\,. \tag{2}$$

The output of the MLP is then used as input to a LR classifier (Equation 1) with hidden weights, and biases $(\mathbf{W}_l, \mathbf{b}_l))$ optimized jointly with LR parameters.

Our final model (JRNN) employs a joint composition of bidirectional RNNs (Schuster & Paliwal, 1997) with Long Short Term Memory (Hochreiter & Schmidhuber, 1997), which preprocess sequences of satellite data for input to an MLP. Let $\mathcal{X}^{\texttt{M}}$, $\mathcal{X}^{\texttt{L}}$ be two time series where $\mathcal{X}_t^{\texttt{M}}$, $\mathcal{X}_t^{\texttt{L}}$ are vectors of annual, spring, and summer composite pixel values of the t-th year for the spectral bands of the MODIS and LandSat imagery respectively. We define two bidirectional LSTM networks, $\texttt{LSTM}_{\texttt{M}}$ and $\texttt{LSTM}_{\texttt{L}}$, for the respective satellite time-series . The LSTMs provide condensed vector representations of the platform time series for a given location. These vectors are concatenated with the categorical embeddings and the continuous biophysical variables for the input to an MLP (Equation 2) but with:

$$\widehat{\mathbf{x}} = \left[\mathbf{w}_{\texttt{soil}}^{\intercal} \quad \mathbf{w}_{\texttt{cover}}^{\intercal} \quad \mathbf{w}_{\texttt{eco}}^{\intercal} \quad D_1^{\intercal} \quad \texttt{LSTM}_{\texttt{M}}\left(\mathcal{X}^{\texttt{M}}\right)^{\intercal} \quad \texttt{LSTM}_{\texttt{L}}\left(\mathcal{X}^{\texttt{L}}\right)^{\intercal}\right]^{\intercal} \tag{3}$$

for JRNN models incorporating all satellite data (D_4). Only one LSTM network is used for models with data from a single satellite platform (D_2, D_3). LSTM parameters are optimized jointly, end-to-end, with the successive MLP and LR classifier parameters of the full JRNN model.

For both the MLP and JRNN models, we use Dropout (Srivastava et al., 2014) and Batch Normalization (Ioffe & Szegedy, 2015) to avoid overfitting and stabilize gradient updates respectively. We fit the parameters of the JRNN, MLP, and LR models[2] by minimizing the cross-entropy between the ground truth and predictions with the ADAM (Kingma & Ba, 2014) optimization algorithm.

4 Results and Analysis

For each data subset, we trained model classes (RF, LR, MLP, JRNN) across 200 random configurations of their respective hyperparameters. Due to the limited amount of samples and wide ecological and environmental diversity across the study area, we used random 5-fold cross-validation splits (with equivalent joint distributions of ecoregion and coverage class) for each experimental run.

[2]Implemented in Tensorflow (Abadi et al., 2016). Code to be open sourced pre-publication.

	Mean Accuracy over Folds				Standard Deviation of Accuracy			
	D_1	D_2	D_3	D_4	D_1	D_2	D_3	D_4
LR	76.712	79.611	81.215	**81.441**	0.021	0.019	0.019	0.012
RF	80.149	81.667	82.484	**82.794**	0.024	0.024	0.022	0.025
MLP	79.604	82.010	82.449	**83.192**	0.025	0.014	0.015	0.012
JRNN	—	82.050	82.580	**83.184**	—	0.011	0.012	0.012

Table 2: Classification accuracy of models for subsets of data

Model	*D1<D2*	D1<D3	*D1<D4*	*D2<D3*	*D2<D4*	*D3<D4*
LR	0.002	0.001	0.003	0.028	0.008	0.335
RF	0.033	0.006	0.003	0.073	0.021	0.277
MLP	0.011	0.004	0.002	0.293	0.102	0.193
JRNN				0.115	0.012	0.096

Table 3: p-values from standard one-tailed t-test for improvement in accuracy. Colored cells denote passing a test for a particular α value. $\alpha = 0.01$, $\alpha = 0.05$, $\alpha = 0.1$, Not significant.

Table 2 shows the average accuracy across cross-validation folds for the most accurate models trained on the 4 subsets of input variables. Table 3 shows the p-values from one-tailed t-tests of accuracy improvement for models that incorporate satellite imagery. For all models, incorporating satellite data from either LandSat or MODIS provides significant gains in accuracy. LR and RF show more gains for LandSat than MODIS but no significant gains from using both platforms. While the deep learning models are ambivalent to finer time granularity (MODIS) or finer spatial granularity (LandSat) imagery, only the JRNN model gains from multiple platforms.

5 Conclusion

The main goal of this study is a reliable map of cheatgrass invasion estimates over critical areas for use by ecology scientists and land management agencies. Our final predictive coverage map is composed of ensemble predictions (the average probability of cheatgrass occurence across cross-validation fold models) from the most accurate model (MLP using all variables)[3]. The final map was preliminarily validated from inspection by an ecologist who found that sampled final map predictions are justified as credible by these location's biophysical factors. We are openly distributing the final mapping product to interested ecologists and land management agencies. Our hope is that this mapping product can aid in mitigating detrimental effects of cheatgrass invasion, and promote new field studies for critical locations that may have heightened fire risk.

The present study covers the widest geographic extent for estimates of cheatgrass invasion to date. While our performance assesments of prospective models are specific to cheatgrass, they may lend insight and guidance to future researchers creating predictive models from a fusion of biophysical measurements and/or multiple satellite platform imagery. We chose an MLP model for a final map due to higher accuracy and lower computational cost. However, we suspect that predictive vegetation mapping can highly benefit from a greater amount of field data and creative deep learning architectures like our JRNN model that show improvement from incorporating multiple remote sensing platforms.

Acknowledgements

This work was funded by the Deep Learning for Scientific Discovery Laboratory Directed Research and Development investment at the Pacific Northwest National Laboratory. Pacific Northwest National Laboratory is operated for the U.S. Department of Energy by Battelle under Contract DE-AC05-76RL01830.

[3]Contact kyle.larson@pnnl.gov for access to the final mapping product.

REFERENCES

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, pp. 265–283, 2016.

Bethany A Bradley and John F Mustard. Identifying land cover variability distinct from land cover change: cheatgrass in the great basin. *Remote Sensing of Environment*, 94(2):204–213, 2005.

D Richard Cutler, Thomas C Edwards Jr, Karen H Beard, Adele Cutler, Kyle T Hess, Jacob Gibson, and Joshua J Lawler. Random forests for classification in ecology. *Ecology*, 88(11):2783–2792, 2007.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

D Jeevalakshmi, S Narayana Reddy, and B Manikiam. Land cover classification based on NDVI using LANDSAT8 time series: A case study Tirupati region. In *Communication and Signal Processing (ICCSP), 2016 International Conference on*, pp. 1332–1335. IEEE, 2016.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Paul A Knapp. Cheatgrass (bromus tectorum l) dominance in the great basin desert: history, persistence, and influences to human activities. *Global environmental change*, 6(1):37–52, 1996.

Nataliia Kussul, Mykola Lavreniuk, Sergii Skakun, and Andrii Shelestov. Deep learning classification of land cover and crop types using remote sensing data. *IEEE Geoscience and Remote Sensing Letters*, 14(5):778–782, 2017.

Graciela Melgoza, Robert S Nowak, and Robin J Tausch. Soil water exploitation after fire: competition between bromus tectorum (cheatgrass) and two native species. *Oecologia*, 83(1):7–13, 1990.

James M Omernik and Glenn E Griffith. Ecoregions of the conterminous united states: evolution of a hierarchical spatial framework. *Environmental management*, 54(6):1249–1266, 2014.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

Mike Pellant et al. The cheatgrass-wildfire cycle: are there any solutions. In *McArthur et al.(eds.) Proceedings of a Symposium on cheatgrass invasion, shrub die-off, and other aspects of shrub biology and management. US For. Serv., Int. Res. Sta., Gen. Tech. Rep. INT-276. Ogden, UT*, pp. 11–18, 1990.

John Rogan, Janet Franklin, Doug Stow, Jennifer Miller, Curtis Woodcock, and Dar Roberts. Mapping land-cover modifications over large areas: A comparison of machine learning algorithms. *Remote Sensing of Environment*, 112(5):2272–2283, 2008.

Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, 1997.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

NRCS USDA. The plants database (http://plants. usda. gov, 6 may 2015). National Plant Data Team, Greensboro, 2015.

Amanda M West, Paul H Evangelista, Catherine S Jarnevich, Sunil Kumar, Aaron Swallow, Matthew W Luizza, and Stephen M Chignell. Using multi-date satellite imagery to monitor invasive grass species distribution in post-wildfire landscapes: An iterative, adaptable approach that employs open-source data and software. *International Journal of Applied Earth Observation and Geoinformation*, 59:135–146, 2017.