From the ground up: Prototyping an improved LANDFIRE base product via state-of-the-art data preprocessing and modeling techniques

Introduction

-Lifeform, including tree, shrub, and herb classifications, is the base layer for all other LANDFIRE vegetation and fuels products
-Predictor variables for raster data are extracted from composited Landsat imagery (spectral bands and derivative indices), topographic, and climate datasets and are hosted within the LANDFIRE Reference Database (LFRDB)
-Dependent data sources, including field collected data, are also archived within the LFRDB
-Multiple machine learning and model optimization algorithms are available for comparison and the best results are selected
-The **overall goal** is to maximize model accuracy using the latest generation of machine learning tools (Fig. 1)

Methods And Flowchart

1. Examine the LFRDB to get per plot lifeform labels, predictor variables, and determine whether the plot has spatially intersected a disturbed area within the last five years within the study areas (Fig. 2)
2. Remove outliers using IsolationForests to assess each Landsat band. Perform lifeform class balancing by mirroring LANDFIRE Remap Existing Vegetation Cover (EVC) distribution. Separate data into 10% test and 90% train

3a. Determine which hyperparameter tuned models, including XGBoost, Random Forest (RF), XGBoost RF, and Light Gradient Boosting Model, provide the best overall and per class model by examining test data and apply the best and See5 models to input geospatial independent variable datasets

3b. Create unique image segments using Normalized Differenced Vegetation Index (NDVI) median, minimum, and maximum values

4. Merge most accurate tree and herb/shrub lifeform products by selecting the tree pixels in the best tree lifeform product and herb/shrub pixels from the best herb/shrub lifeform product
5. Use image segments as a mode filter on modeled lifeform products
6. Error assessment of withheld test to calculate overall accuracy and per class Mathew's correlation coefficients (MCC)
7. Revision of input data by adding additional expert opinion plots and subsequent rerunning of steps 2-6

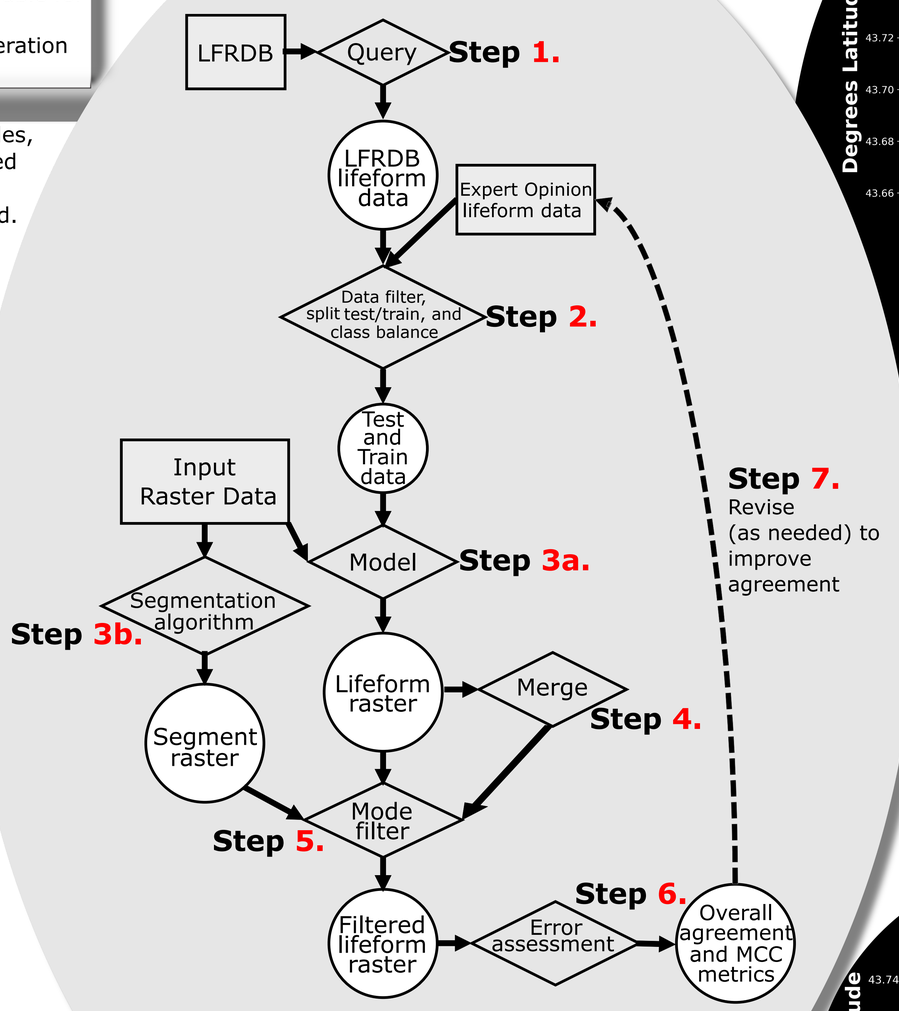

Figure 1: *Overview of the entire mapping process, starting with querying the LANDFIRE Reference Database (LFRDB) and ending with the production of output maps with accompanying error analysis*

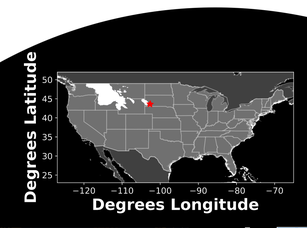

Figure 2: *Location of study test areas (white - overall, and red star - zoomed in Figures 3-5)*

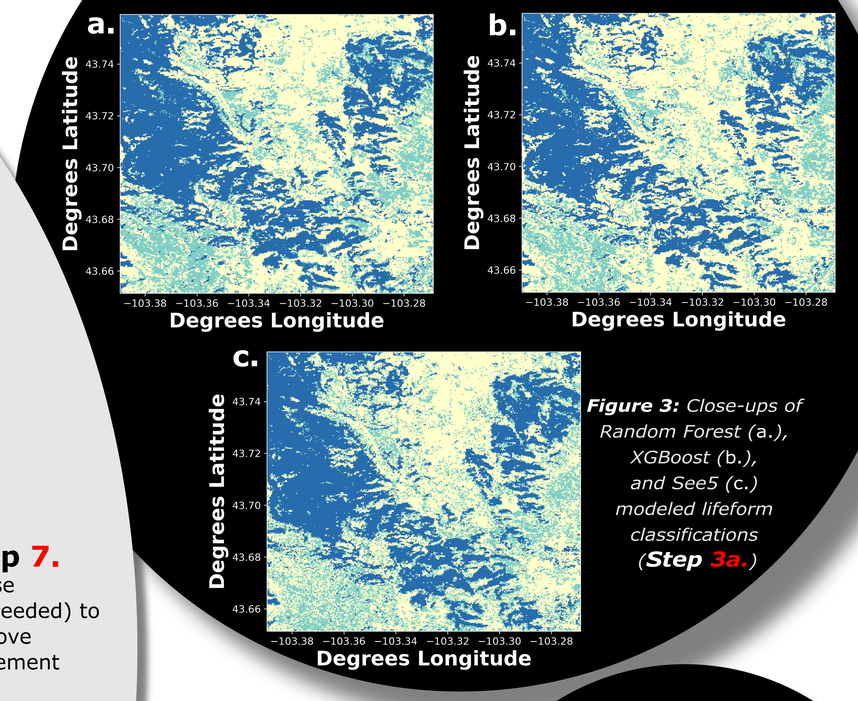

Figure 3: *Close-ups of Random Forest (a.), XGBoost (b.), and See5 (c.) modeled lifeform classifications (**Step 3a.**)*

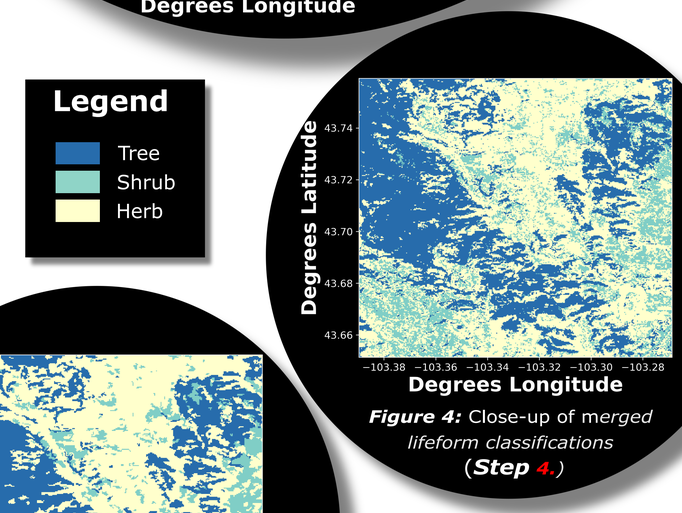

Figure 4: *Close-up of merged lifeform classifications (**Step 4.**)*

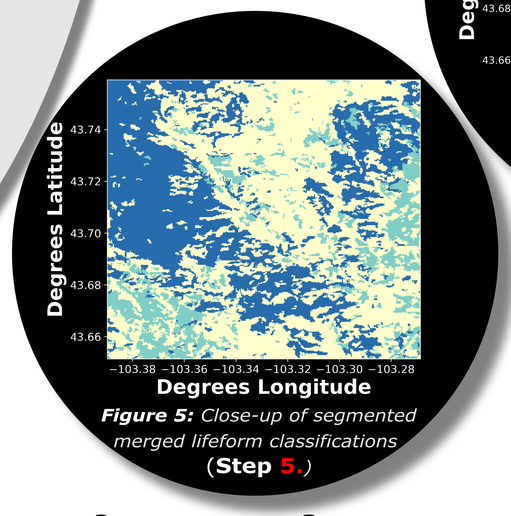

Figure 5: *Close-up of segmented merged lifeform classifications (**Step 5.**)*

Results and Discussion

Model/Filter	N	Model Error (%)	Map Error (%)	Tree MCC	Shrub MCC	Herb MCC
Step 3a. - RF	29207	20	22	75	62	36
Step 3a. - XGBoost	29207	19	23	71	61	39
Step 3a. - See5	29207	13	24	73	60	23
Step 4. - Merge	X	X	22	75	62	40
Step 5. - Segmentation Filter	X	X	21	76	64	33

Table 1: *Results from performing data query (Step 1.), data filter (Step 2.), Random Forest (RF), XGBoost, and See5 modeling (Step 3a.), lifeform product merge (Step 4.), and production of final lifeform product (Step 5.). Lifeform models and maps were assessed for percent error and lifeform tree, shrub, and herb, class accuracies were assessed by Mathew's Correlation Coeficients (MCC) scaled by 100*

-Hyperparameter tuned Random Forest and Xgboost maps outperformed See5 overall in this study (Table 1, Figs. 3-5)
-Merging modeled lifeform products, i.e., best output from tree, shrub and herb models, can improve overall accuracies and per-class Mathew's Correlation Coefficient (MCC) values (Table 1, Fig. 4)
-Using segmentation as a mode filter improved both overall map error and per class MCC for some classes (Table 1, Fig. 5). Accuracy metrics also illustrate that a simultaneous reduction of MCC may occur in at least one class (Table 1)
-These findings indicate that the revised procedures illustrated may improve the overall pre-revision lifeform accuracy
-All proceses have been scripted in Python to enable easier handoff to LANDFIRE production team and facilitate peer review of methods
-Overall and per-class MCC metrics error metrics enable LANDFIRE analysts to assess model performance

Joshua J. Picotte[1], Timothy D. Hatten[2], Daryn Dockter[3], Sanath Sathyachandran Kumar[1], Brian Tolk[3], and Inga La Puma[3]
[1]ASRC Federal Data Solutions, contractor to the United States Geological Survey(USGS), Earth Resources Observation and Science (EROS) Center, Sioux Falls, SD 57198, USA.
[2]USGS, EROS Center, Sioux Falls, SD 57198, USA.
[3]KBR, contractor to the USGS, EROS Center, Sioux Falls, SD 57198, USA.