New Project

Refining stand-level species distribution estimates using alternative small area estimation methods and high-resolution auxiliary information

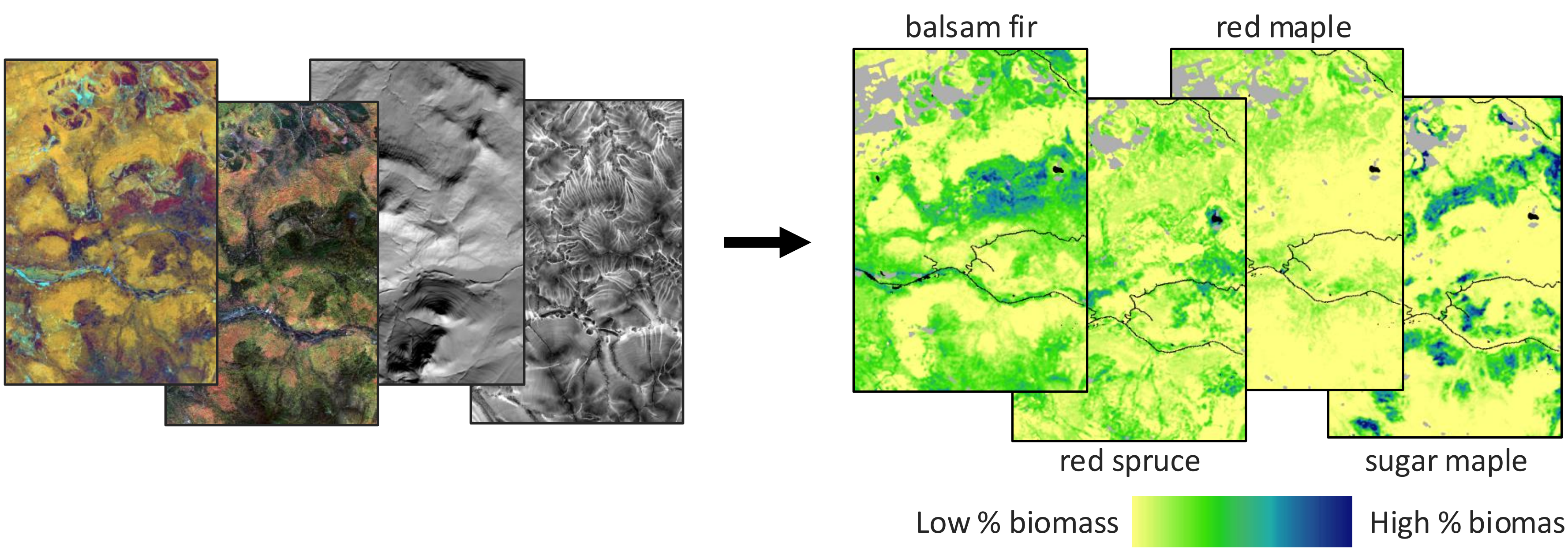
Kasey Legaard	UMaine
Ken Bundy	UMaine
Mike Premer	Weyerhaeuser
Ian Prior	Seven Islands Land Company
Ian Taviss	JD Irving, Ltd.
Nathaniel Naumann	PotlatchDeltic
Andy Lister	US Forest Service FIA
John Shaw	US Forest Service FIA
Karin Riley	US Forest Service FIA
Phil Radtke	Virginia Tech

Presenter: K. Legaard

Justification

Forest remote sensing applications have tended to support pixel-level inference, with estimates of uncertainty that apply to pixels

Species distributions from plot/pixel level machine learning:



Justification

To the extent that we wish to provide usable stand-level estimates, we need provide stand-level uncertainty

But we should do that while carrying forward relevant technological and methodological advances:

- Use of multi-spectral, multi-temporal, multi-source remote sensing data
- Statistical/machine learning to make sense of it all

Justification

Unit-level SAE by mixed effects Random Forest (MERF):

$$y = f(X) + Zv + \epsilon$$

Unit-level error

Nonparametric Random Forest modeling fixed effects (conditional mean of measurement y given auxiliary data X)

Random effects accounting for hierarchical dependencies of plot measurement data (plots within stands)

Estimation of areal means:

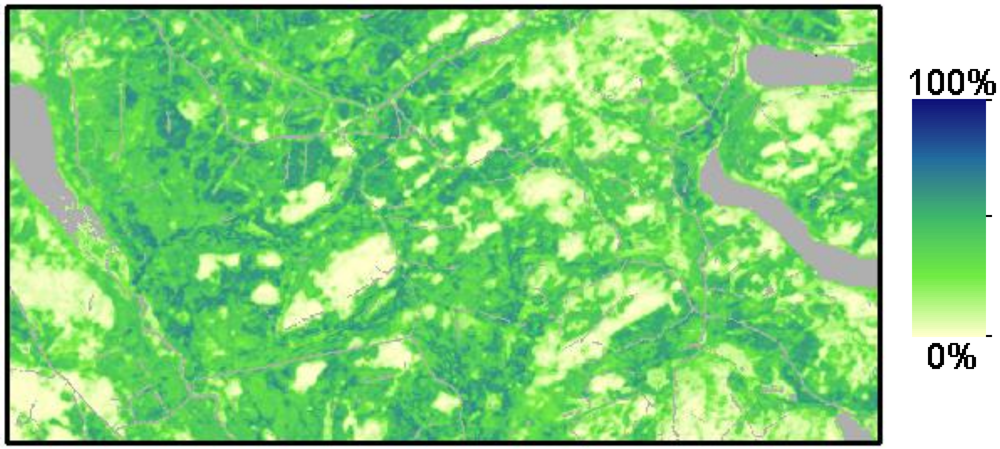
$$\hat{\mu}_i = \bar{\hat{f}}(X_i) + Z_i\hat{v}_i \quad \text{for} \quad i = 1, \ldots D.$$

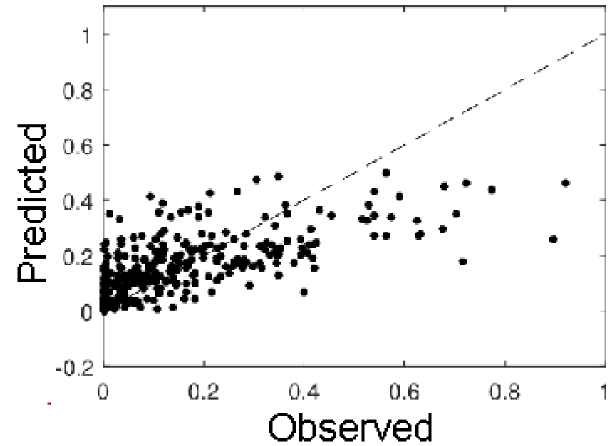
Uncertainty estimated by nonparametric bootstrap

Hajjem, A., Bellavance, F. & Larocque, D. (2014) Mixed-effects random forest for clustered data. Journal of Statistical Computation and Simulation, 84(6), 1313–1328.

Justification

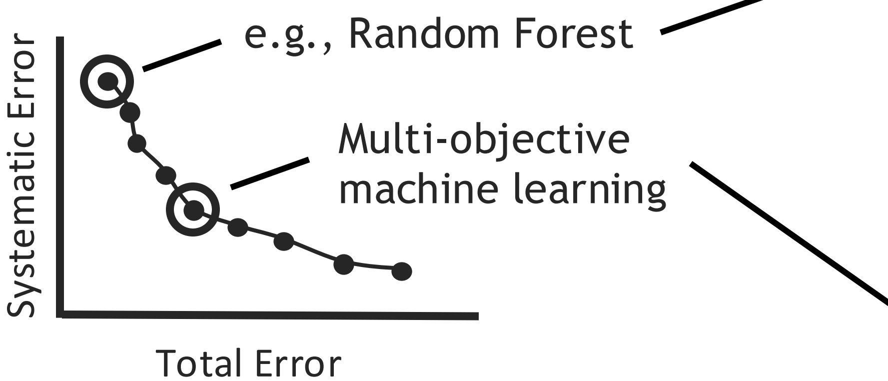
Balsam fir %AGLB, Random Forest:

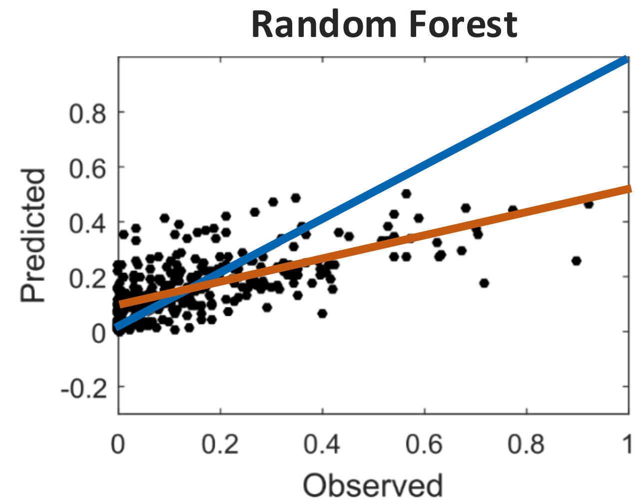


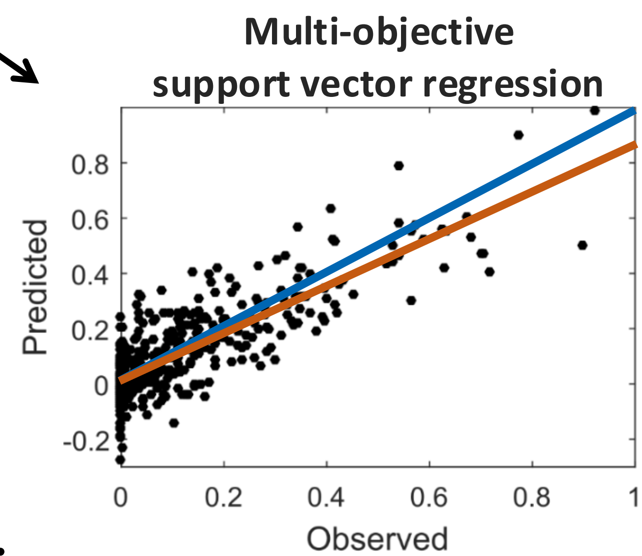


Justification

Location and measurement uncertainty introduces a tradeoff between total and systematic prediction error:





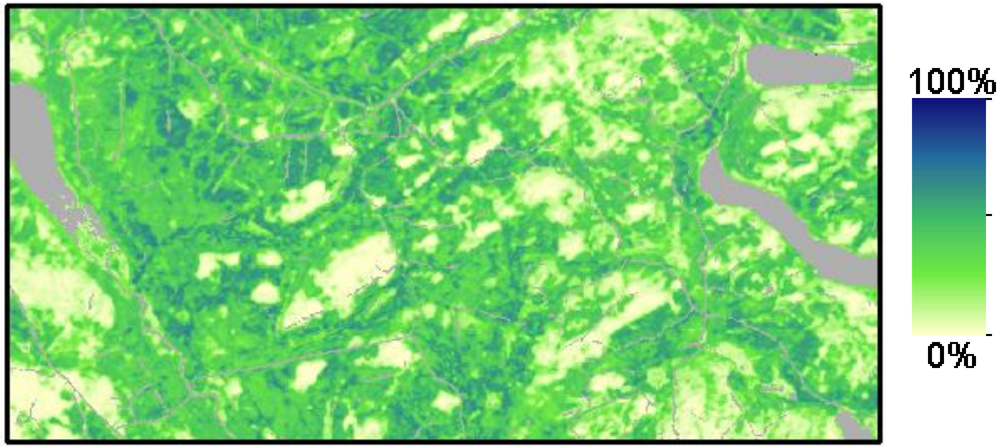


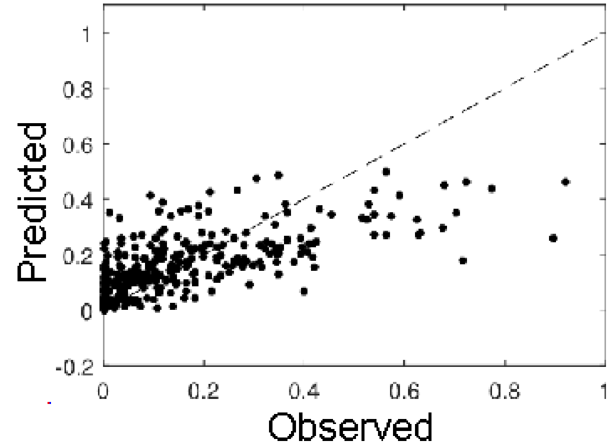
Minimization of total error causes attenuation bias in regression models and systematic error in maps.

Multi-objective machine learning minimizes both total and systematic error.

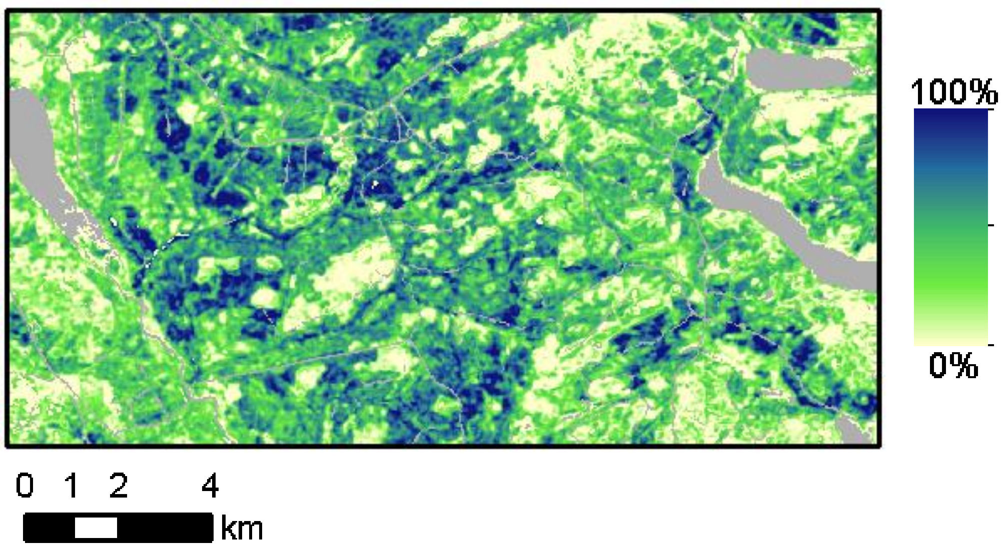
Justification

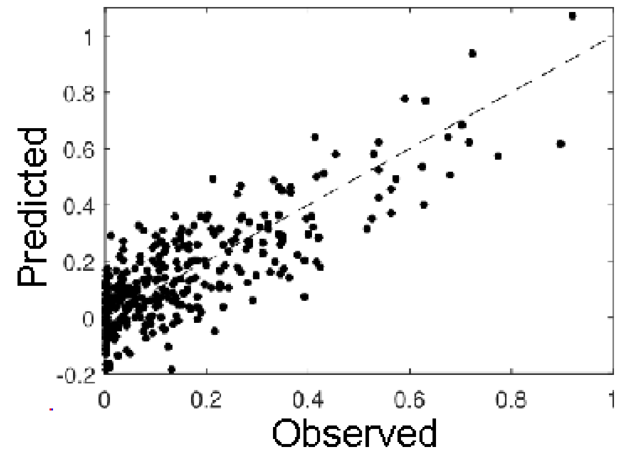
Balsam fir %AGLB, Random Forest:





Balsam fir %AGLB, multi-objective ML:

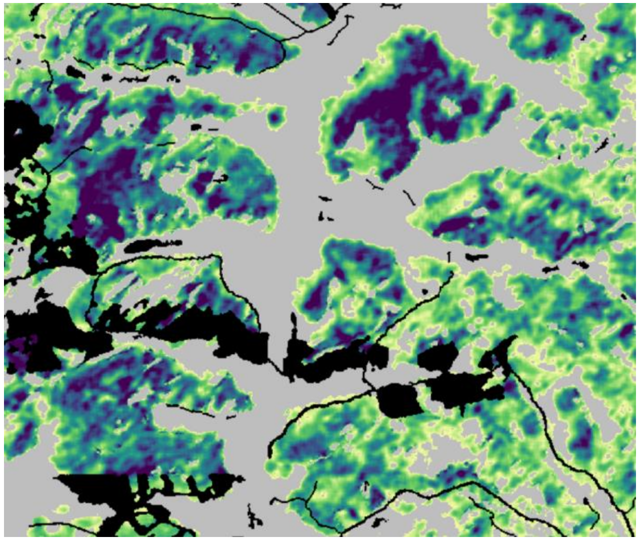




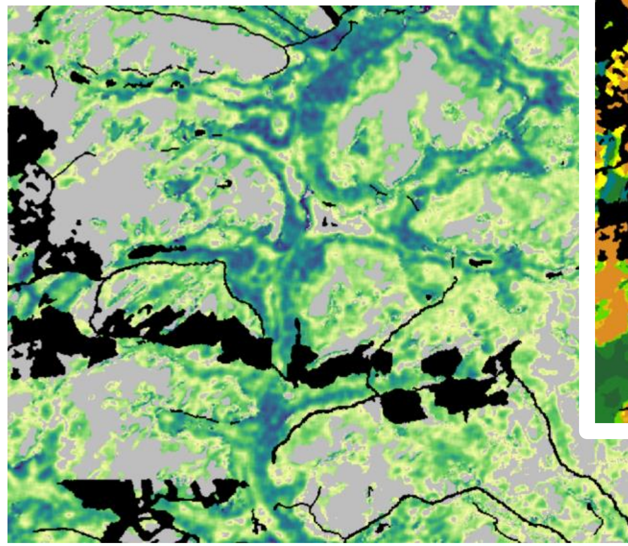
Multi-objective ML used for species (and forest type) mapping in Maine:

Justification

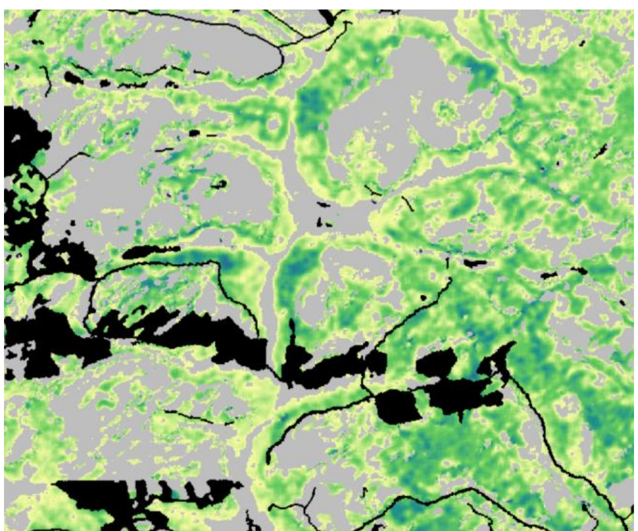
Sugar maple



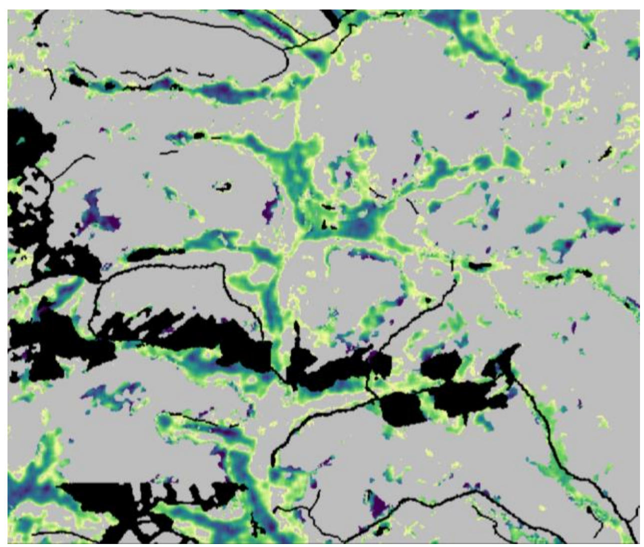
Balsam fir

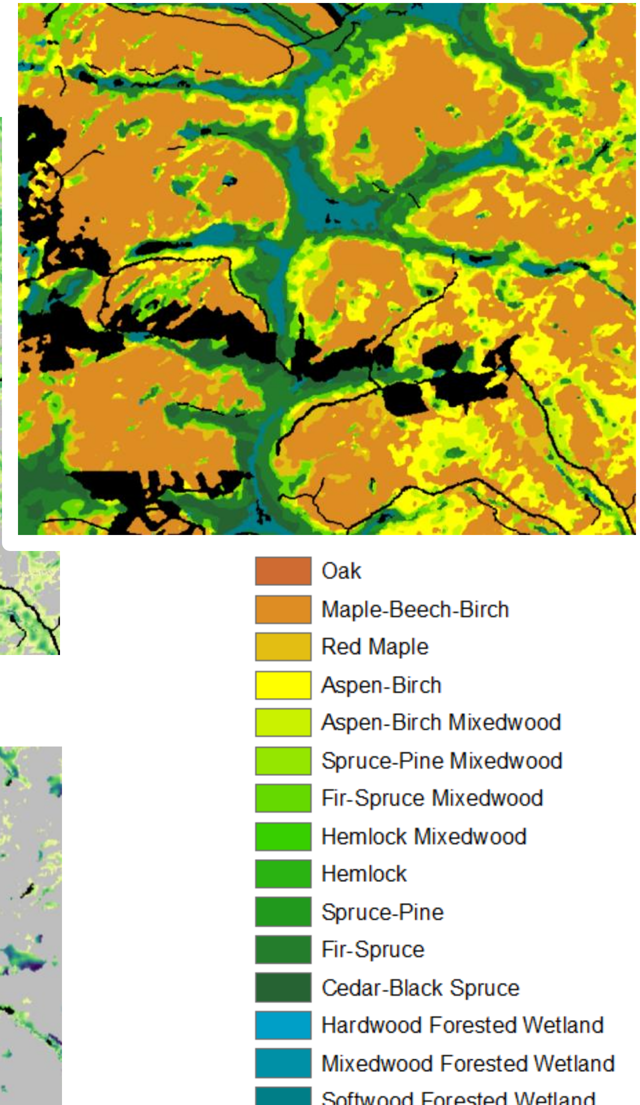


Paper birch



Northern white cedar





Objectives

Project goal:

Improve stand-level species distribution estimates using nonlinear, nonparametric small area estimation methods and high-resolution auxiliary information.

Supporting objectives:

(1) Generalize the mixed effects Random Forest framework to other ML model architectures

(2) Test for improvements to current ML-derived species distribution maps after employing additional information via stand-level SAE

(3) Evaluate SAE methods on independent stand-level species data provided by project partners

(4) Generalize SAE methods across different forest types and regions (mixed forests of Maine, mixed conifer forests of northern Idaho)

Methods

Mixed effects machine learning:

$$y = f(X) + Zv + \epsilon$$

Unit-level error

Machine learning model architecture of choice

Random effects

Objective is to generalize MERF by substituting alternative ML models for estimation of fixed effects

* Where *f(X)* is a support vector machine, we can insert our multi-objective ML into a mixed effects SAE framework

Methods

Mixed effects <u>multi-objective</u> machine learning:

Specification of a fixed effects model that reduces systematic error and better represents variability in species distributions

Especially important for estimation in stands with little or no plot measurement data

Stand-level uncertainty estimated using the same nonparametric bootstrap approach employed by MERF

Methods

kNN-based SAE for less prevalent/abundant species:

kNN methods associate one or more measurement plots to target pixel locations based on a measure of similarity between plots and pixels

All species are modeled simultaneously – predicted species mixtures match or closely match mixtures measured somewhere

Typically less accurate for an individual species when that species can be modeled well independently

Methods

Multi-model approach to stand-level species:

Use mixed effects multi-objective ML to predict a subset of species well; use it when it works

Use established kNN-based small area estimators for any remaining species of importance (typically species that are not well represented by measurement data)

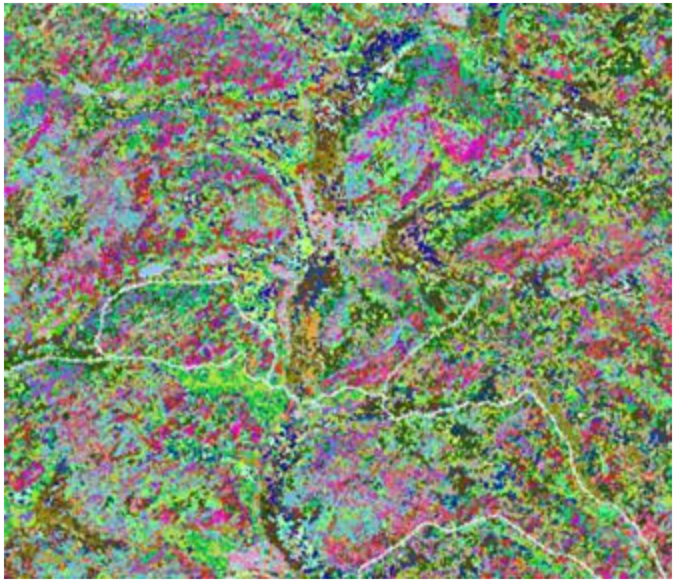
Incorporate multi-objective ML predictions into the similarity metric used for kNN imputation, so that kNN estimates are consistent with species predicted well by multi-objective ML

Methods

A broadly similar kNN imputation approach was adopted by TreeMap, developed at the RMRS:

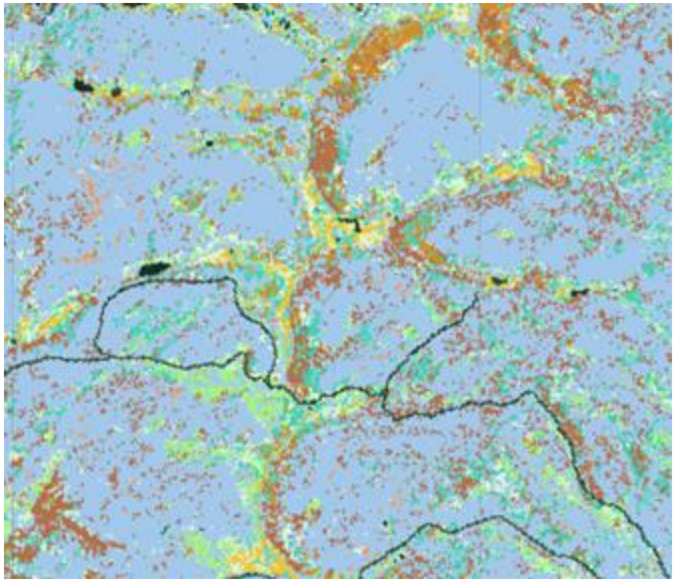
Used national LANDFIRE data products to associate FIA plots to target pixels

FIA plot identifier (k = 1)



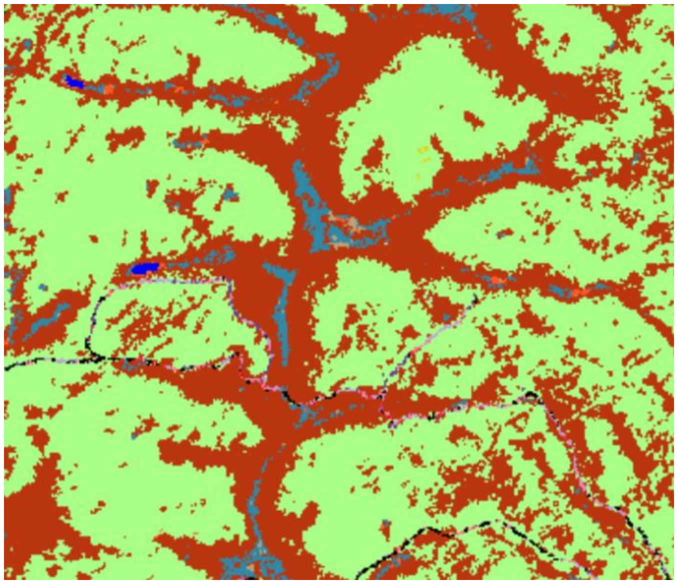
→

Forest type (FIA algorithm)

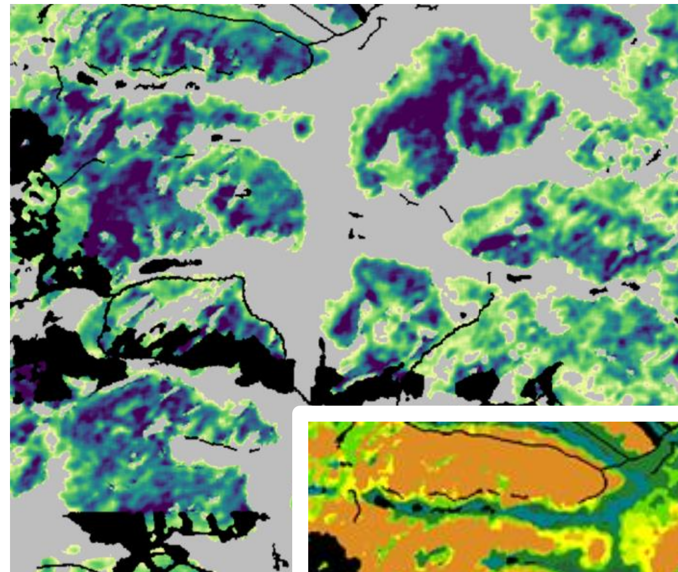


Methods

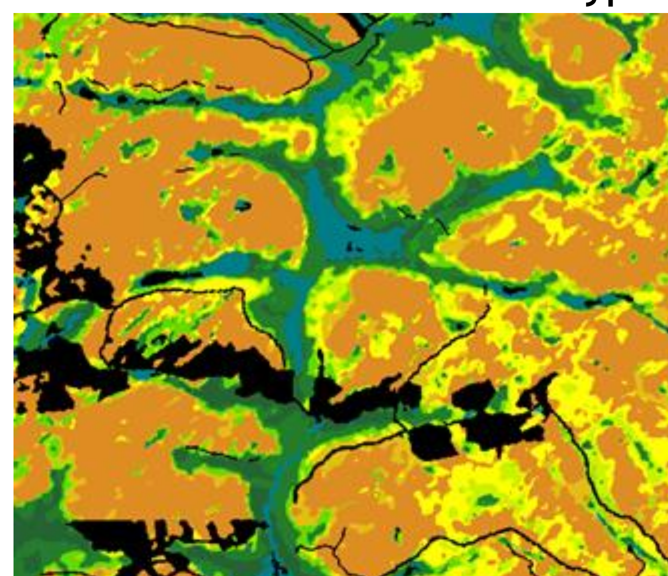
LANDFIRE existing veg type was the only composition information used in the TreeMap kNN imputation:



Can we get more out of kNN-based estimators by using more detailed composition information?



Individual species



Forest types

Opportunity to collaborate with TreeMap to evaluate costs/benefits of alternative workflows for operational use at scale

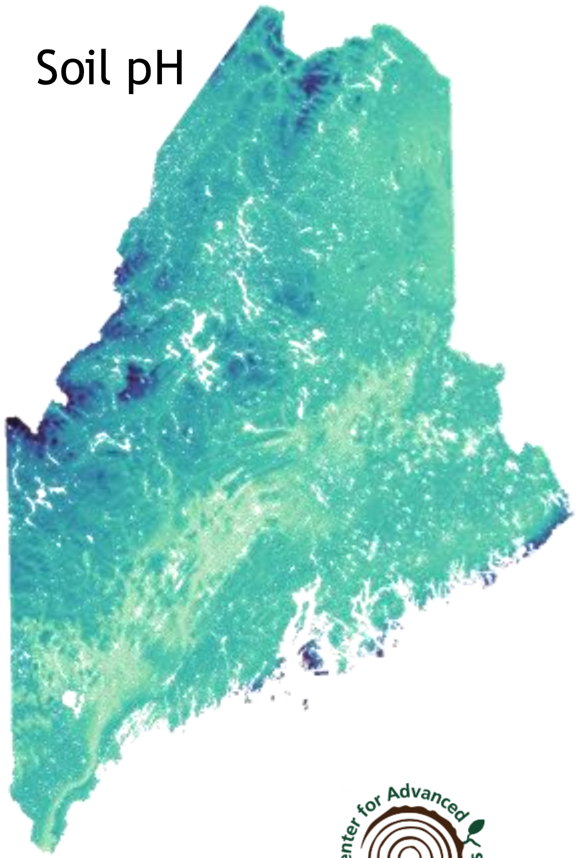
Methods

Testing alternative auxiliary data:

Currently use multi-temporal Sentinel-2 to map overstory species, and additional covariates haven't helped much at the pixel level

Can new auxiliary data improve stand-level species predictions?

- Digital soil data
  - 7 variables available for the state of Maine at ~20 m resolution



Methods

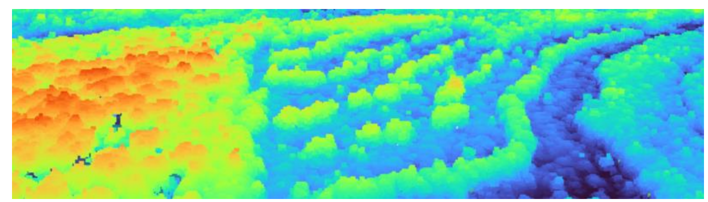
Testing alternative auxiliary data:

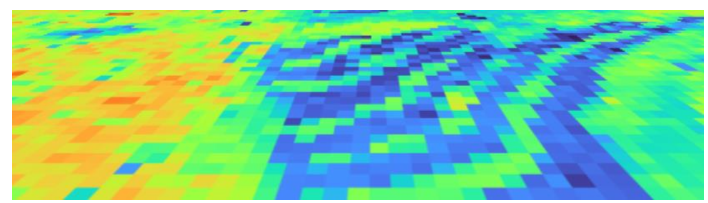
Currently use multi-temporal Sentinel-2 to map overstory species, and additional covariates haven't helped much at the pixel level

Can new auxiliary data improve stand-level species predictions?

- Digital soil data
  - 7 variables available for the state of Maine at ~20 m resolution
- Gridded vegetation height metrics from 3D NAIP, computed using fast and scalable computational methods
- 3D vegetation change from multi-temporal 3D NAIP
  - 2018, 2021, 2023-24 collections in Maine

2021 NAIP DSM





95th height percentile, 10 m grid

Deliverables

- Mixed effects ML software implementation
    - Python code integrating sklearn pipelines?
    - Eventual no-code access via a cloud-based SAE platform under development at UMaine (NCASI PSAE; PI A. Weiskittel)
- Demonstrated applications of mixed effects multi-objective ML plus kNN-based estimators for stand-level SAE, in multiple mixed forest types and regions
    - May expand application scope further, potentially through other funding sources
- Evaluation of novel auxiliary datasets for potential improvements in stand-level species predictions
    - Includes integration of fast, scalable 3D NAIP processing and investigation of multivariate change detection strategies applied to 3D NAIP

Company Benefits

- Development and testing of alternative SAE algorithms and workflows that leverage the benefits of machine learning, established remote sensing methods, and new auxiliary data
- Evaluation of new, scalable high-resolution auxiliary data sets
- Eventual integration of algorithms, workflows, and auxiliary data products into a cloud-based platform designed for accessible SAE
- Potential for refined management and planning using high-resolution species data and robust stand-level inferenc
- Potential to inform future refinements to national programs or standards through new collaborations with FIA
  - Knowledge exchange with programs using similar methods
  - Potential ancillary information for post-stratification, model-assisted regression, non-response bias correction