Integrating Machine Learning and Knowledge-Based Soil Inference Classification in the White Mountain National Forest
Jessica Philippe | April 16, 2020
What are we doing here?

Digital Soil Mapping in Practice
The White Mountain National Forest
– Initial Soil Survey
- Not our first DSM project
Who am I?

A Soil Scientist
And GIS specialist

– Working for NRCS for 16 years
– B.S. in Natural Resources Planning and M.S. in Soil and Water Science
– I started my soils career just as ArcSIE was coming into existence; I've never had to knock on any doors or use a stereoscope

But I promise, I’ve dug a LOT of holes.
Digital Soil Mapping

- The generation of geographically referenced soil databases based on quantitative relationships between spatially explicit environmental data and measurements made in the field and laboratory (McBratney et al., 2003)

- The spatial prediction of soil classes or properties from point data using a statistical algorithm
Soil-Landscape Models

SCORPAN (McBratney et al., 2003)

\[ S = f(S, C, O, R, P, A, N) + \varepsilon \]

- **Soil**, at a specific point in **space** and time
  - Soil classes, Sc
  - Soil attributes, Sa
- Empirical **quantitative** function of **environmental covariates**
  - Soil (class, or directly or remotely sensed property)
  - Climate
  - Organisms
  - Relief
  - Parent Material
  - Age
  - \( N = \text{Spatial Position} \)
- Plus an estimation of **error or uncertainty**
Digital Soil Mapping

**Conventional soil mapping**
- “Where is the boundary between two soils?”
- Focus on the marginal areas

**Digital soil mapping**
- Central concept is well defined
- Variation expressed across the landscape

![Digital soil mapping example](image-url)
Hybrid Approach

Machine learning + knowledge-based soil inference

Variable Importance and Initial Modeling
Random Forests
R

Final Modeling
Knowledge-based Inference
ArcSIE
Project Area

800,000 acres in New Hampshire and Maine

MLRA 143 – Northeastern Mountains
Challenges and Opportunities

Balancing operational and academic approaches to soil survey
Challenges and Opportunities

Continually changing resources and demands on our time
Challenges and Opportunities

Continually changing resources and demands on our time
Digital Soil Mapping

Basic Steps

- Data processing
- Landscape stratification
- Modeling
  - Knowledge extraction
  - Environmental layer selection
  - Prediction
- Field verification
- Post processing
- Correlation and publication

Iterative
WMNF Approach

Catena-by-catena modeling
BTI = Knowledge-based
ATI = hybrid approach
The rest = ???
Knowledge-based DSM

**Purpose:** extract the soil-landscape model that lives in an experienced soil scientist’s head and turn it into something consistent and reproducible.

Soil Scientists’ Knowledge → Soil Inference Engine → Fuzzy Soil Membership Map → Environmental Data
Elevation 200–600m is typical for soil A.

As elevation deviates from this range, the soil’s similarity to type A gradually decreases.
Knowledge-based DSM

Output from soil inference is a fuzzy membership map

Everywhere that is similar to the defined environment shows a high fuzzy membership
Digital Soil Mapping

Knowledge-based DSM

Expert Knowledge

Training Data + Covariates → Predictive model \( (f) \) → Soil Map
Field Data

849 Existing data points (series-level)
Random Forests

Decision tree classification

- CART models; classification and regression trees

Ensemble method

- Fit multiple CART models to independent “bootstrap” samples of the data and then combine the predictions
Random Forests

Variable Importance
- Shows how important each variable is in classifying the data
- Ordered top-to-bottom, most-to-least important
Data layers

Multi-resolution
valley bottom
flatness

Relative position

Terrain ruggedness index

Topographic wetness index
ATI points details

228 training points
31 cases
81 validation points
Results from Random Forests class prediction
Cases where RF is incorrect, used in global CBR
Fuzzy membership map from cases
First iteration of case-based reasoning combined with RF results.
“Final” map combining fuzzy memberships and random forests results
Accuracy Assessment

Random Forests: 58% overall accuracy

Hybrid: 74% overall accuracy
Hybrid Approach
Machine learning + knowledge-based soil inference
Hybrid Approach

Machine learning + knowledge-based soil inference
Hybrid Approach

Machine learning + knowledge-based soil inference
Hybrid Approach

Machine learning + knowledge-based soil inference
Hybrid Approach
Machine learning + knowledge-based soil inference
Thank You!

For more information:
Jessica.Philippe@usda.gov