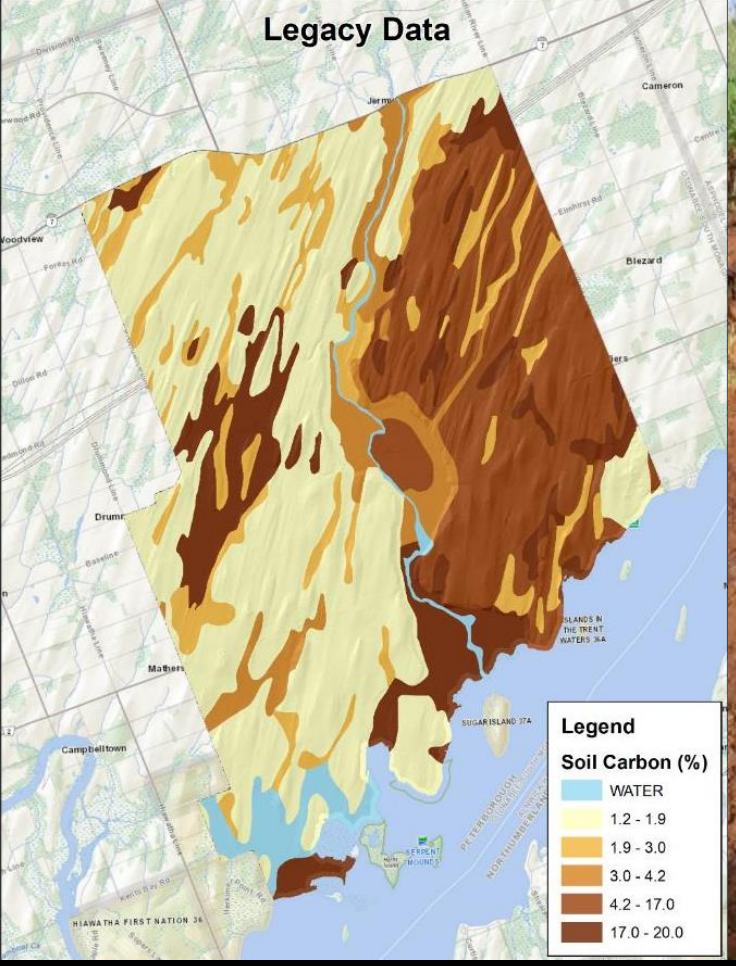
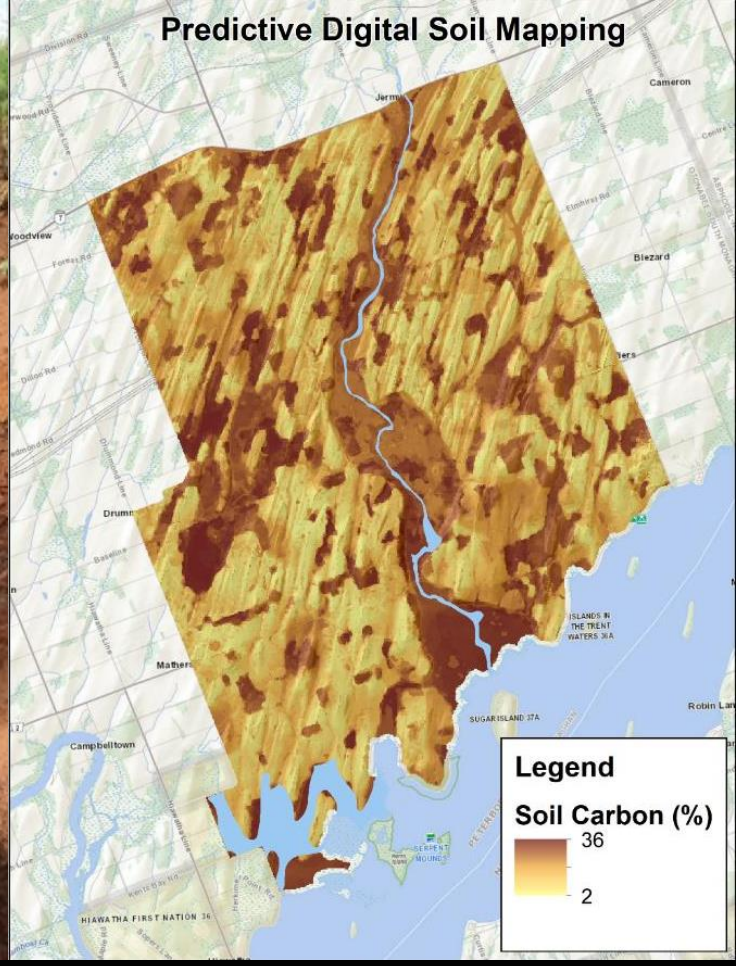




5. Basics of Machine-Learning with Applications for Digital Soil Mapping

Canadian Digital Soil Mapping Workshop, 2020

CDSM WG
CANADIAN SOCIETY OF SOIL SCIENCE
SOCIÉTÉ CANADIENNE DE LA SCIENCE DU SOL
Ontario
UNIVERSITY of GUELPH
DALHOUSIE UNIVERSITY

About Us

Brandon Heung

Dalhousie University

Brandon.Heung@dal.ca

Daniel Saurette

Ontario Ministry of Agri-Food and Rural Affairs

University of Guelph

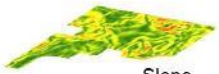
Daniel.Saurette@ontario.ca

Review of DSM Framework

STEP 1

Covariates

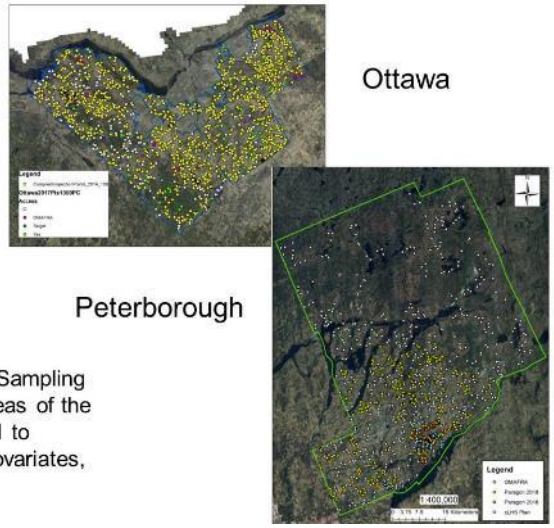
DEM



Slope

Wetness

Sample Plan



Conditioned Latin Hypercube Sampling (cLHS) in R, isolates which areas of the landscape need to be sampled to capture the variability of the covariates, optimized sample plan

STEP 2

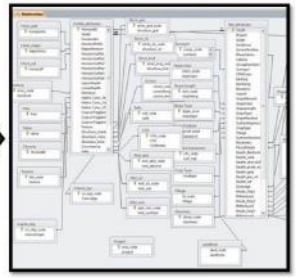
STEP 3

Lab Analysis

Test for:

- TOC
- TN
- pH
- OM
- Etc.

Database Management

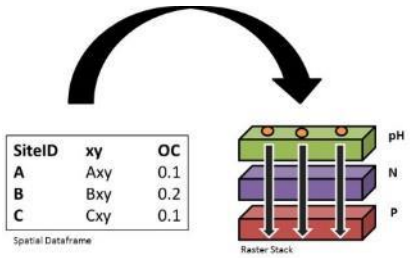


Link new data into database:

- Add in new field data
- Add in analytical data
- Add in bulk density data
- QA/QC

Review of DSM Framework

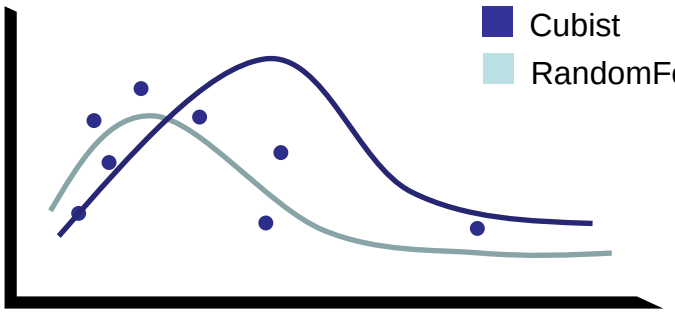
STEP 4a



SiteID	xy	OC	pH	N	P
A	Axy	0.1	6.5	0.1	0.2
B	Bxy	0.2	6.6	0.2	0.2
C	Cxy	0.1	6.8	0.3	0.1

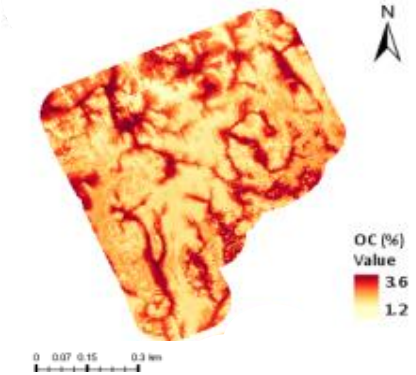
Add covariate data to training site data

STEP 4b

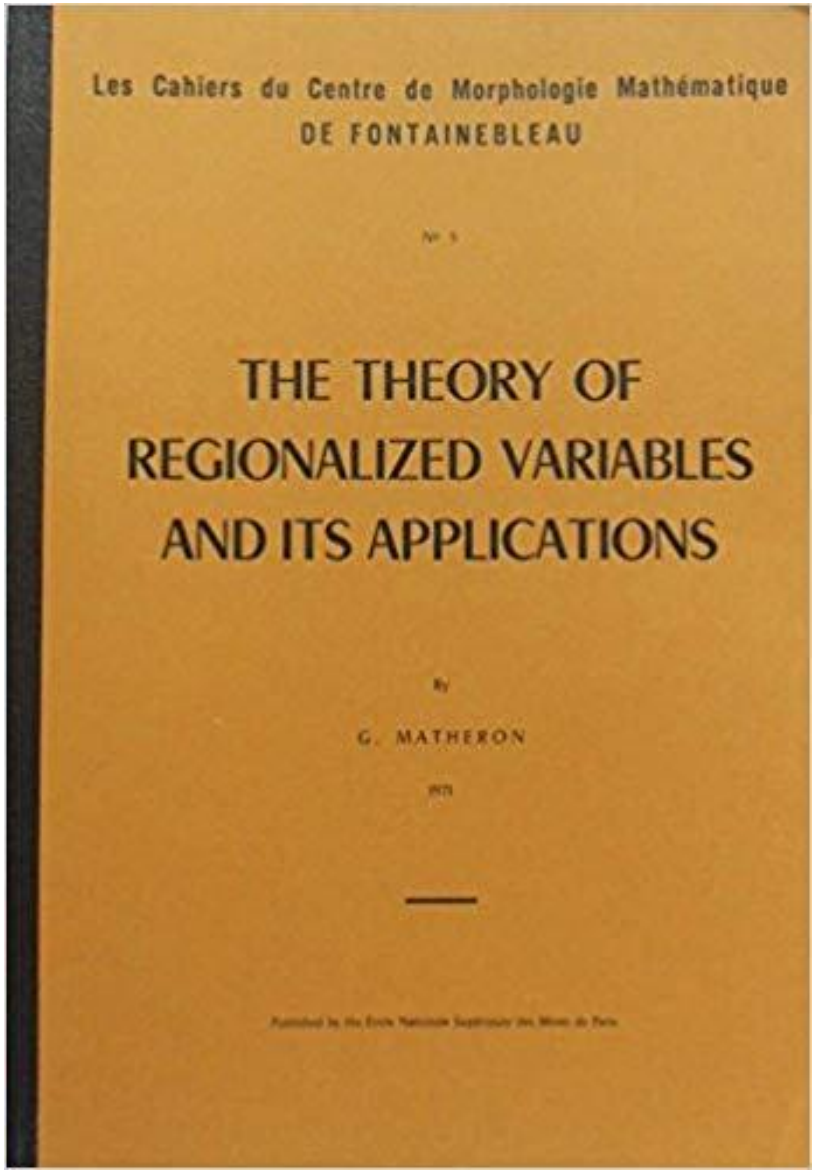


Find best fit model and validate it using GOOF statistics.

STEP 4c



Using covariate raster data, apply model across landscapes to predict soil characteristic throughout the area at different soil depths.



Universal Model of Soil Variation

$$Z(s) = Z^*(s) + \varepsilon(s) + \varepsilon$$

Z(s): Target soil variable (type or property)

Z*(s): Deterministic Component
Modelled using statistical soil-landscape model (e.g. regression or classification algorithm)

ε(s): Stochastic Component
Spatial structure that is modelled with a variogram (e.g. kriging)

ε: Spatially Uncorrelated Noise
Can't be modelled with the available data or models at a scale

Burrough and McDonnell, 1998

Available online at www.sciencedirect.com
SCIENCE DIRECT®
Geoderma 117 (2003) 3–52
GEODERMA
www.elsevier.com/locate/geoderma

On digital soil mapping

A.B. McBratney[a,*], M.L. Mendonça Santos[b], B. Minasny[a]

[a] Australian Centre for Precision Agriculture, Faculty of Agriculture, Food and Natural Resources, McMillan Building A05, The University of Sydney, Sydney, New South Wales 2006, Australia
[b] EMBRAPA-Centro Nacional de Pesquisa de Solos, Rua Jardim Botânico 1024, 22.460-000, Rio de Janeiro, RJ, Brazil

Received 19 November 2002; received in revised form 14 May 2003; accepted 5 June 2003

Abstract

We review various recent approaches to making digital soil maps based on geographic information systems (GIS) data layers, note some commonalities and propose a generic framework for the future. We discuss the various methods that have been, or could be, used for fitting quantitative relationships between soil properties or classes and their 'environment'. These include generalised linear models, classification and regression trees, neural networks, fuzzy systems and geostatistics. We also review the data layers that have been, or could be, used to describe the 'environment'. Terrain attributes derived from digital elevation models, and spectral reflectance bands from satellite imagery, have been the most commonly used, but there is a large potential for new data layers. The generic framework, which we call the scorpan-SSPFe (soil spatial prediction function with spatially autocorrelated errors) method, is particularly relevant for those places where soil resource information is limited. It is based on the seven predictive scorpan factors, a generalisation of Jenny's five factors, namely: (1) *s*: soil, other or previously measured attributes of the soil at a point; (2) *c*: climate, climatic properties of the environment at a point; (3) *o*: organisms, including land cover and natural vegetation; (4) *r*: topography, including terrain attributes and classes; (5) *p*: parent material, including lithology; (6) *a*: age, the time factor; (7) *n*: space, spatial or geographic position. Interactions (*) between these factors are also considered. The scorpan-SSPFe method essentially involves the following steps:

(i) Define soil attribute(s) of interest and decide resolution ρ and block size β.
(ii) Assemble data layers to represent Q.
(iii) Spatial decomposition or lagging of data layers.
(iv) Sampling of assembled data (Q) to obtain sampling sites.
(v) GPS field sampling and laboratory analysis to obtain soil class or property data.
(vi) Fit quantitative relationships (observing Ockham's razor) with autocorrelated errors.
(vii) Predict digital map.

* Corresponding author. Tel.: +61-2-9351-3214; fax: +61-2-9351-3706.
E-mail address: alex.mcbratney@acss.usyd.edu.au (A.B. McBratney).

doi:10.1016/S0016-7061(03)00223-4

Introducing the *SCORPAN* Model

$$S_{c,a} = f(s, c, o, r, p, a, n) + \varepsilon$$

$S_{c,a}$: Soil classes or attributes

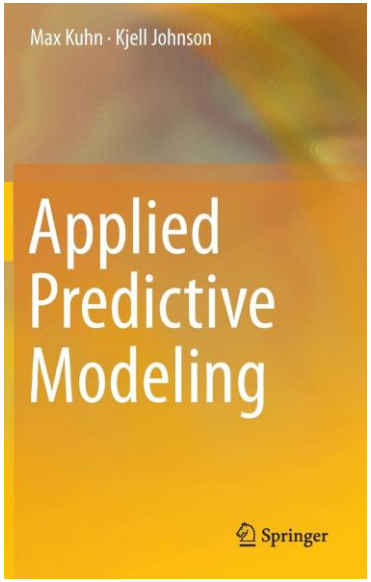
From Jenny's Equation:
c: Climate
o: Organisms, vegetation
r: Topography, landscape attributes
p: Parent material, lithology
a: Age or time factor

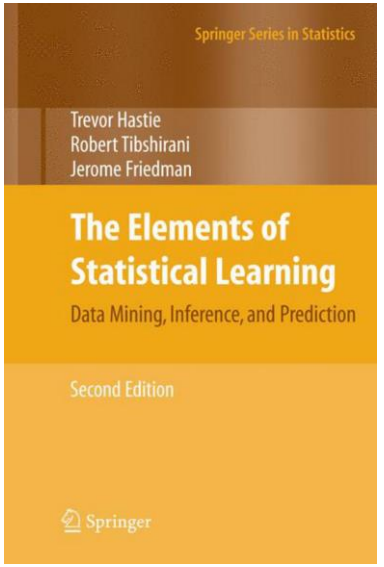
Additions:
s: soil, other properties of a soil at a point
n: space, spatial position – relative to sampling locations and environmental features
ε*:* spatially autocorrelated residuals

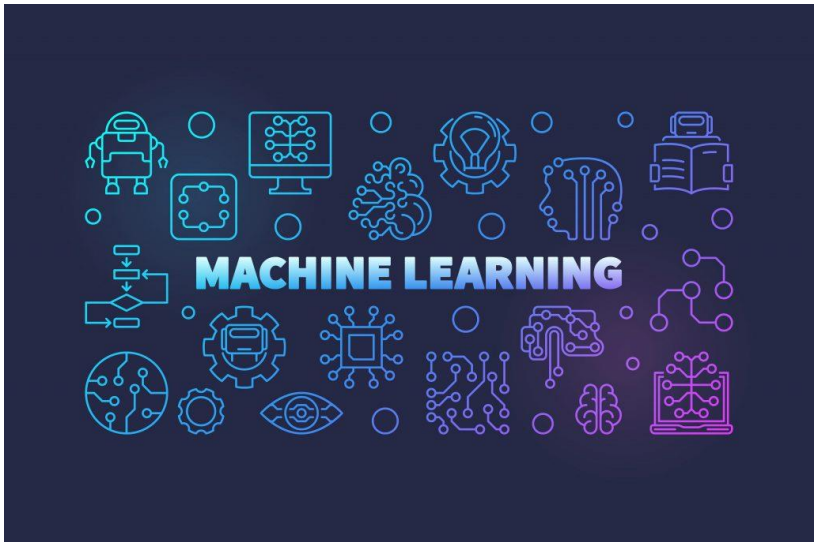
f(): **Quantitative function *f* linking *S* to *scorpan* factors**

Machine Learning

- We have a training dataset with a corresponding suite of environmental covariates
- (Semi-)Automated process of uncovering patterns from large datasets where patterns are used on new datasets for prediction purposes.
- Extraction of soil-environmental relationships.
  - Soils = f(climate, organisms, parent material, topography and time)







Types of Predictive Models

Tree-Based: CART (Boosting, Bagging), BART, MART, Random Forest, C5.0

Linear-Models: MLR, GLM, GAM, Logistic Regression

Model-Trees: Cubist, Logistic Model Tree

Expert-Knowledge: Fuzzy Inference Systems, Rule-Induction Algorithms

Neural Networks: Artificial Neural Networks, Convolutional Neural Networks

Support Vector Machines: Linear SVM, Polynomial SVM, Radial Basis Function SVM

Distance-Based Learners: k-Nearest Neighbours, Nearest Shrunken Centroid

Geostatistical: Ordinary Kriging, Co-Kriging, Regression Kriging

There are so many!!

Geoderma 265 (2016) 62–77

Contents lists available at ScienceDirect

Geoderma

journal homepage: www.elsevier.com/locate/geoderma

An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping

Brandon Heung [a,*], Hung Chak Ho [b], Jin Zhang [a], Anders Knudby [c], Chuck E. Bulmer [d], Margaret G. Schmidt [a]

[a] Soil Science Lab, Department of Geography, Simon Fraser University, 8888 University Drive, Burnaby, BC, V5A 1S6, Canada
[b] Remote Sensing and Spatial Predictive Modeling Lab, Department of Geography, Simon Fraser University, 8888 University Drive, Burnaby, BC, V5A 1S6, Canada
[c] Department of Geography, University of Ottawa, 60 University, Ottawa, ON, K1N 6N5, Canada
[d] British Columbia Ministry of Forests Lands and Natural Resources Operations, Natural Resource Sciences Section, Vernon, BC, V1B 2C7, Canada

Geoderma 239–240 (2015) 68–83

Contents lists available at ScienceDirect

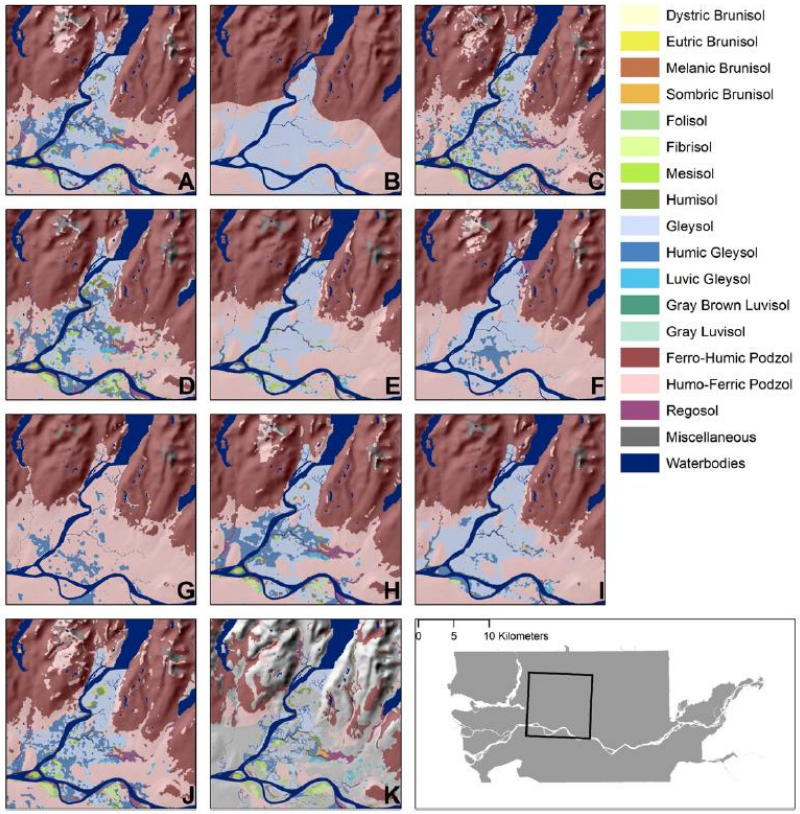
Geoderma

journal homepage: www.elsevier.com/locate/geoderma

Machine learning for predicting soil classes in three semi-arid landscapes

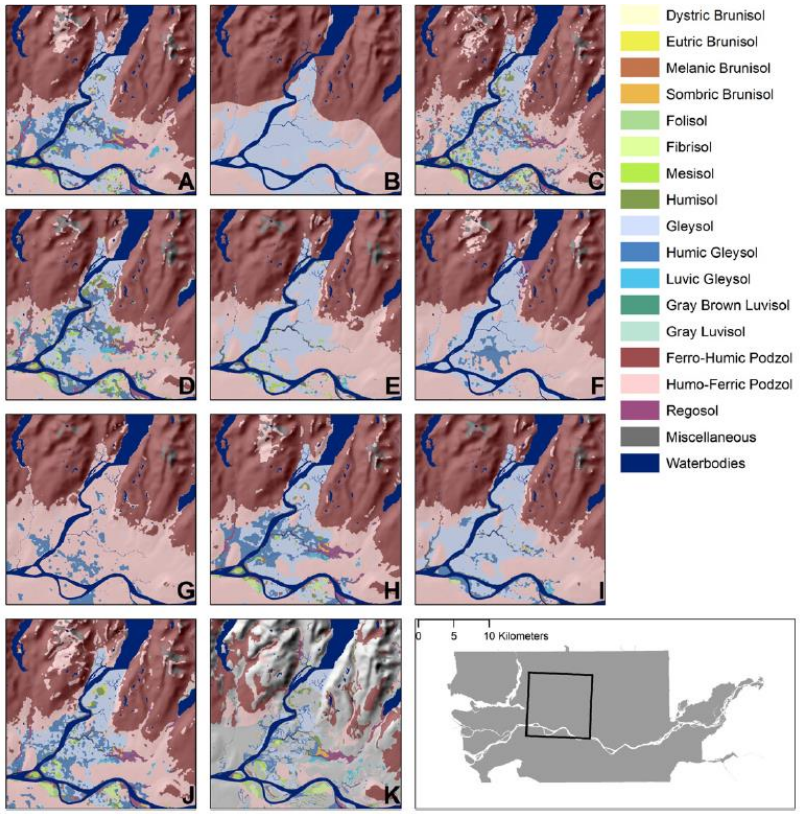
Colby W. Brungard [a,*], Janis L. Boettinger [a], Michael C. Duniway [b], Skye A. Wills [c], Thomas C. Edwards Jr. [d]

[a] Department of Plants, Soils and Climate, 4820 Old Main Hill, Utah State University, Logan, UT 84322, USA
[b] U.S. Geological Survey, Southwest Biological Science Center, 2290 SW Resource Blvd, Moab, UT 84532, USA
[c] National Soil Survey Center, Natural Resources Conservation Service, United States Department of Agriculture, 100 Centennial Mall North, Lincoln, NE 68508, USA
[d] U.S. Geological Survey, Utah Cooperative Fish and Wildlife Research Unit, Department of Wildland Resources, Utah State University, Logan, UT 84322, USA



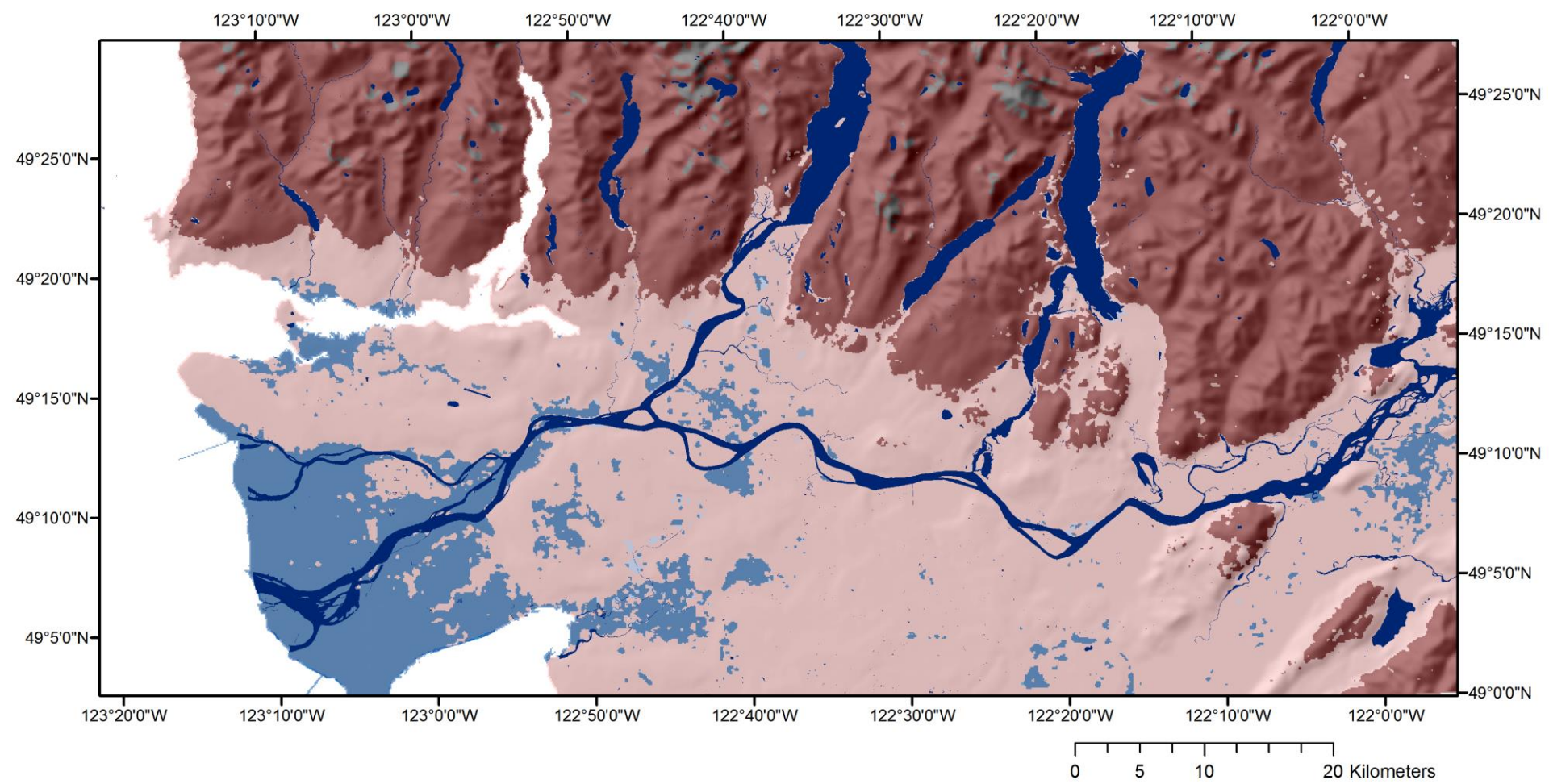
Things to Keep in Mind

- Each model predicts one realization of a DSM – there are an infinite number of realizations.
- Effectiveness of the model are dependent on:
  - Study Area (Extent and Resolution)
  - Target Variable
  - Environmental Covariates
  - Model Structure (e.g. Linear vs. Hierarchical)
- Each model has its own set of hyperparameters (i.e. model settings) that need to be optimized.



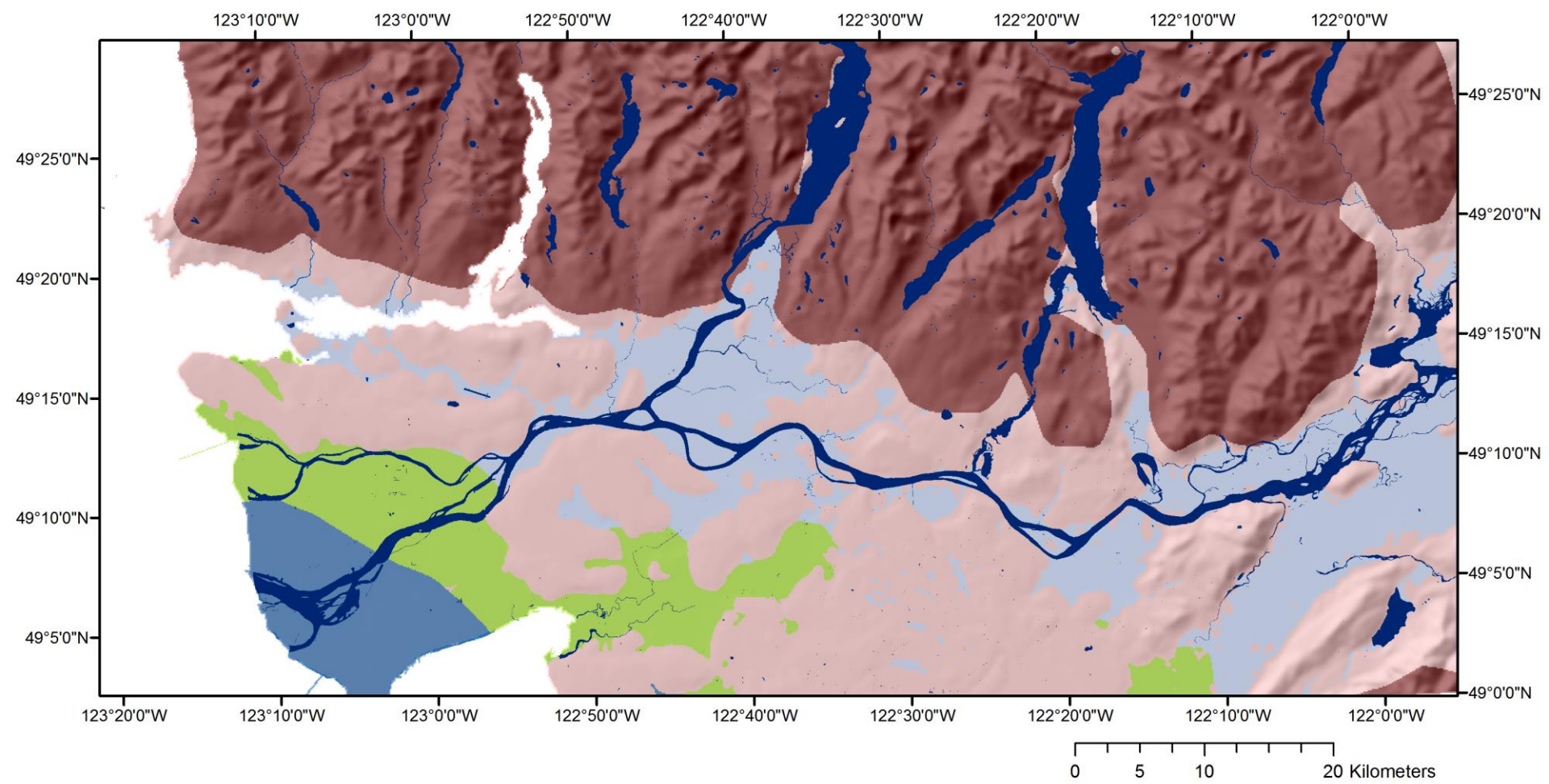
Things to Keep in Mind

- Some models are more computationally efficient than others.
- Some models tend to overfit while others tend to underfit.
- Each model may produce drastically different results given the same inputs.
- Model comparisons should be part of ‘Best Practices’.
- Always perform a visual assessment of the results to make sure that the DSMs are consistent with pedological knowledge.



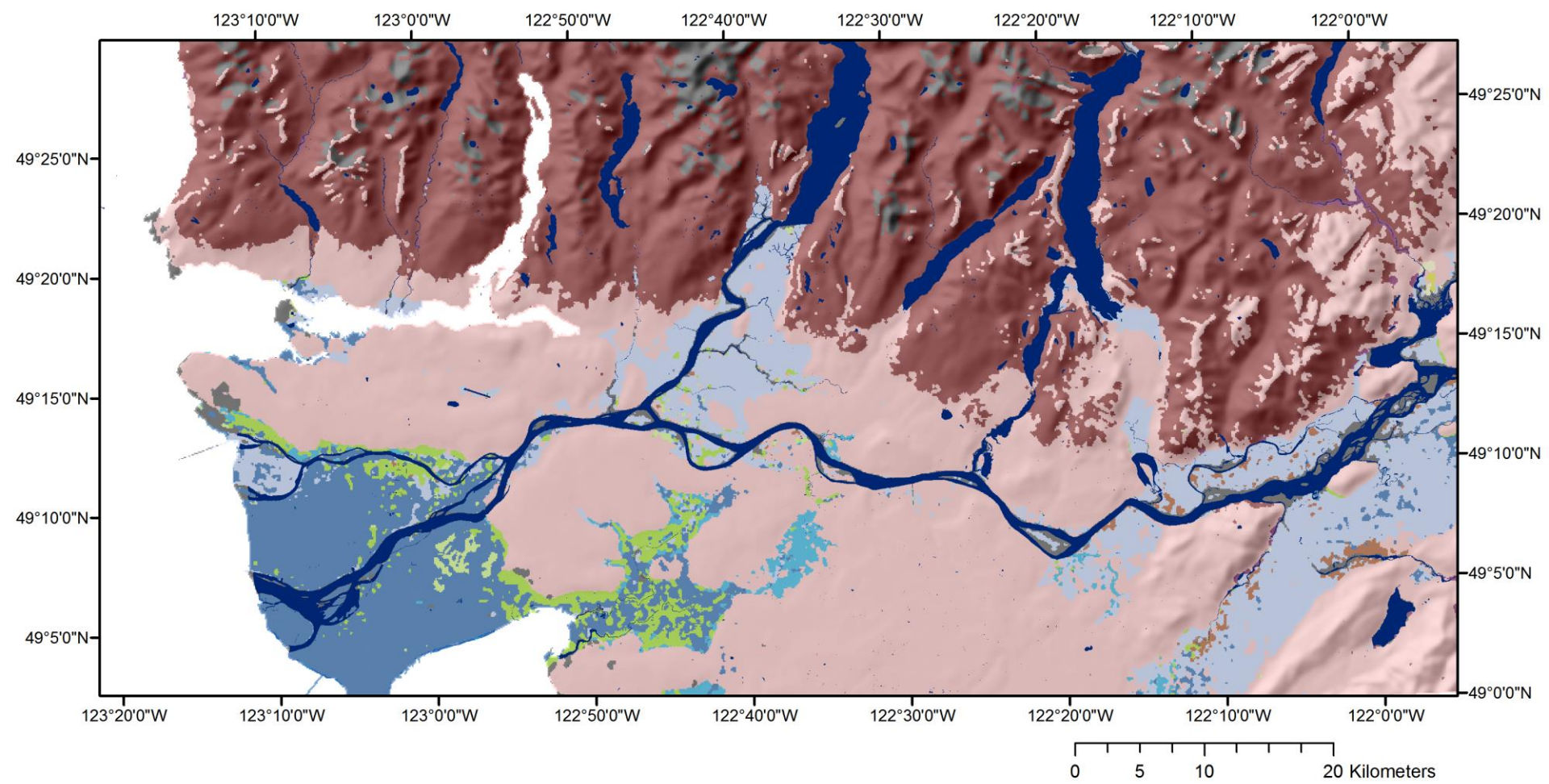
Area-Weighted: Nearest Shrunken Centroid

$C = 40\%$



$C = 42\%$

Area-Weighted: CART

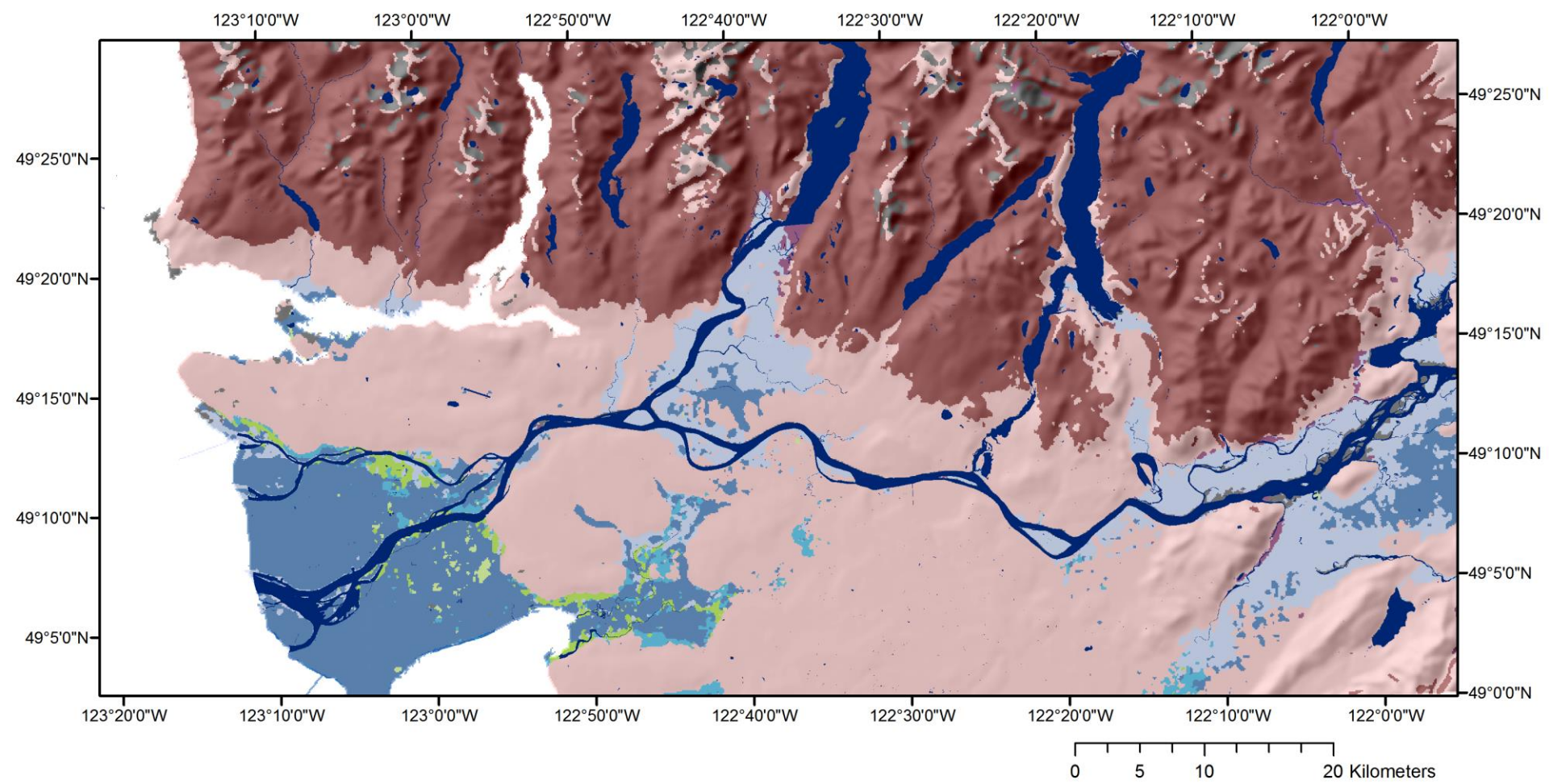


Soil Great Groups

- Dystric Brunisol
- Eutric Brunisol
- Melanic Brunisol
- Sombric Brunisol
- Ferro-Humic Podzol
- Humo-Ferric Podzol
- Folisol
- Fibrisol
- Mesisol
- Humisol
- Regosol
- Gray Brown Luvisol
- Gray Luvisol
- Gleysol
- Humic Gleysol
- Luvic Gleysol
- Bedrock, Rock Outcrop, Recent Alluvium, Talus
- Waterbodies

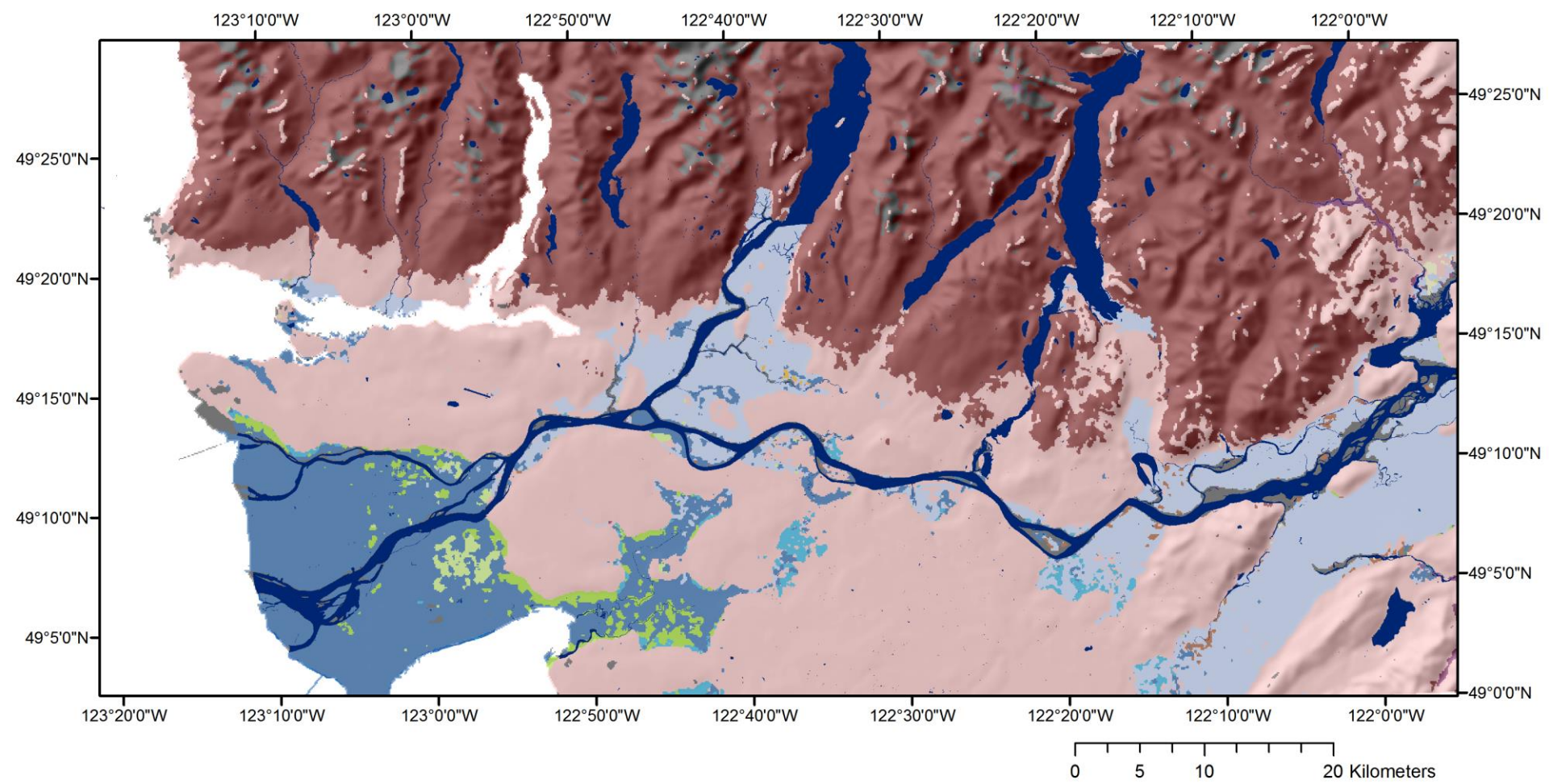
Area-Weighted: Multinomial Logistic Regression

$C = 48\%$



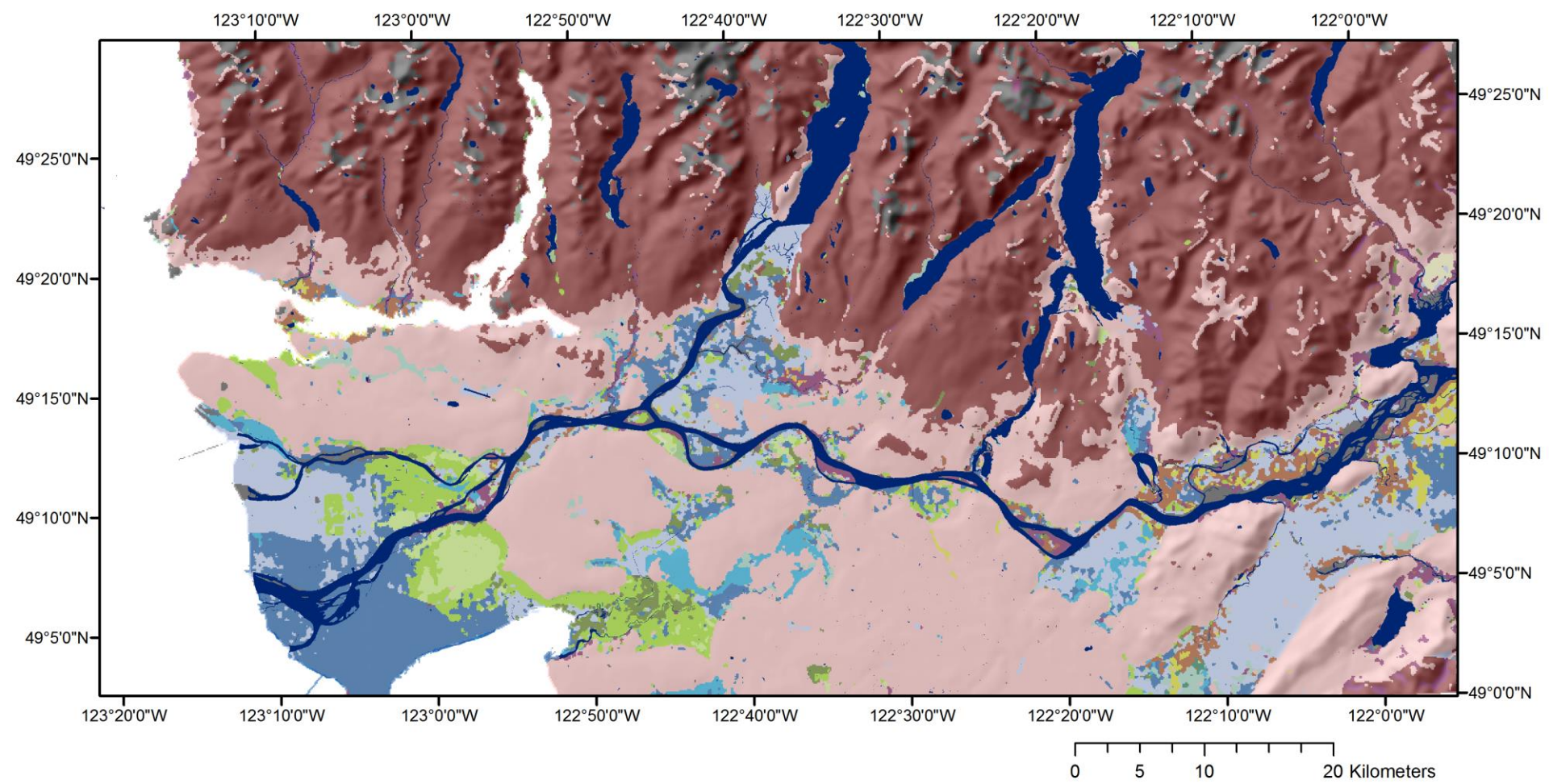
Area-Weighted: Artificial Neural Network

$C = 49\%$



Area-Weighted: Support Vector Machine - Linear

$C = 50\%$

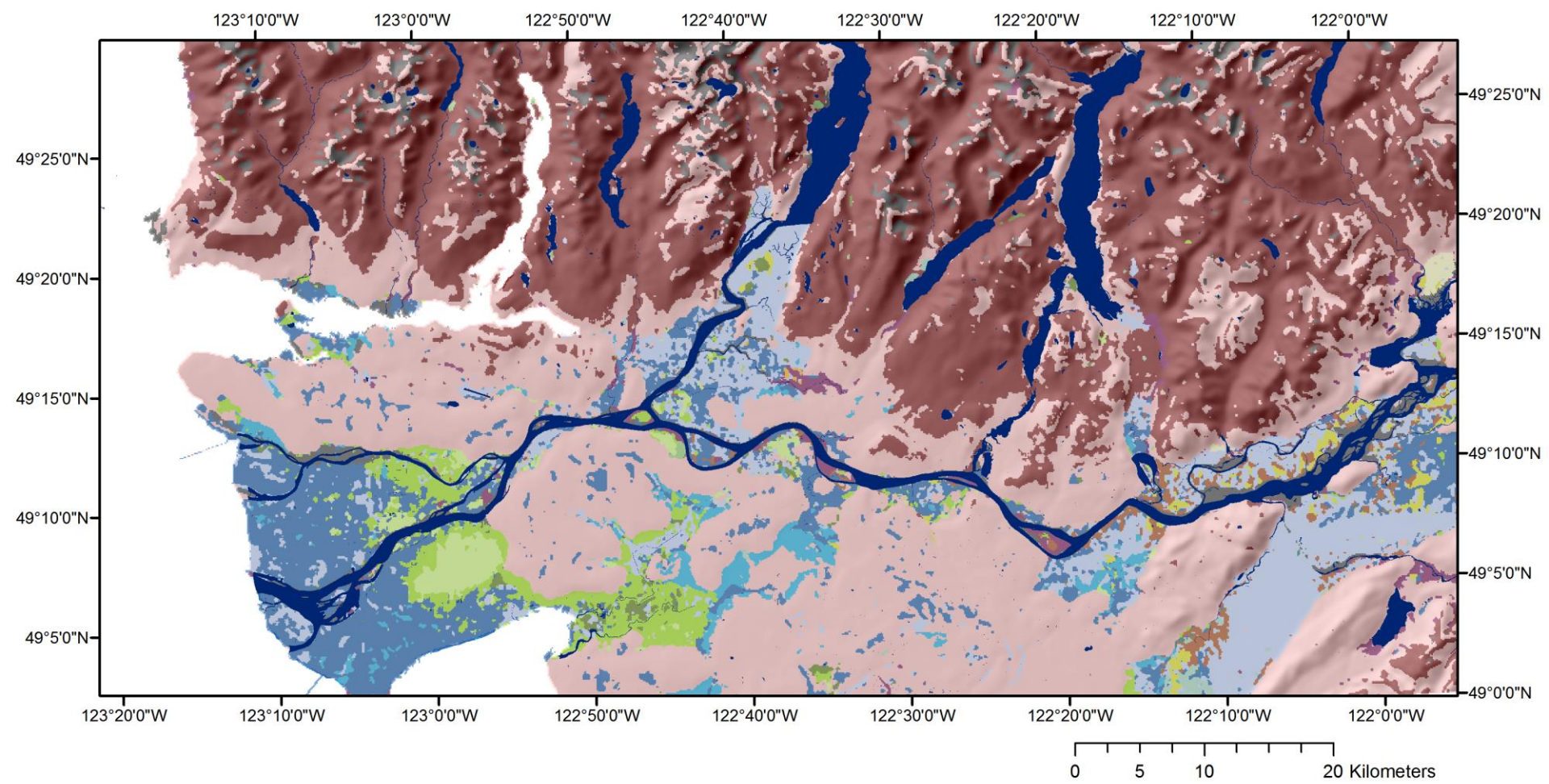


Soil Great Groups

- Dystric Brunisol
- Eutric Brunisol
- Melanic Brunisol
- Sombric Brunisol
- Ferro-Humic Podzol
- Humo-Ferric Podzol
- Folisol
- Fibrisol
- Mesisol
- Humisol
- Regosol
- Gray Brown Luvisol
- Gray Luvisol
- Gleysol
- Humic Gleysol
- Luvic Gleysol
- Bedrock, Rock Outcrop, Recent Alluvium, Talus
- Waterbodies

$C = 65\%$

Area-Weighted: Logistic Model Tree

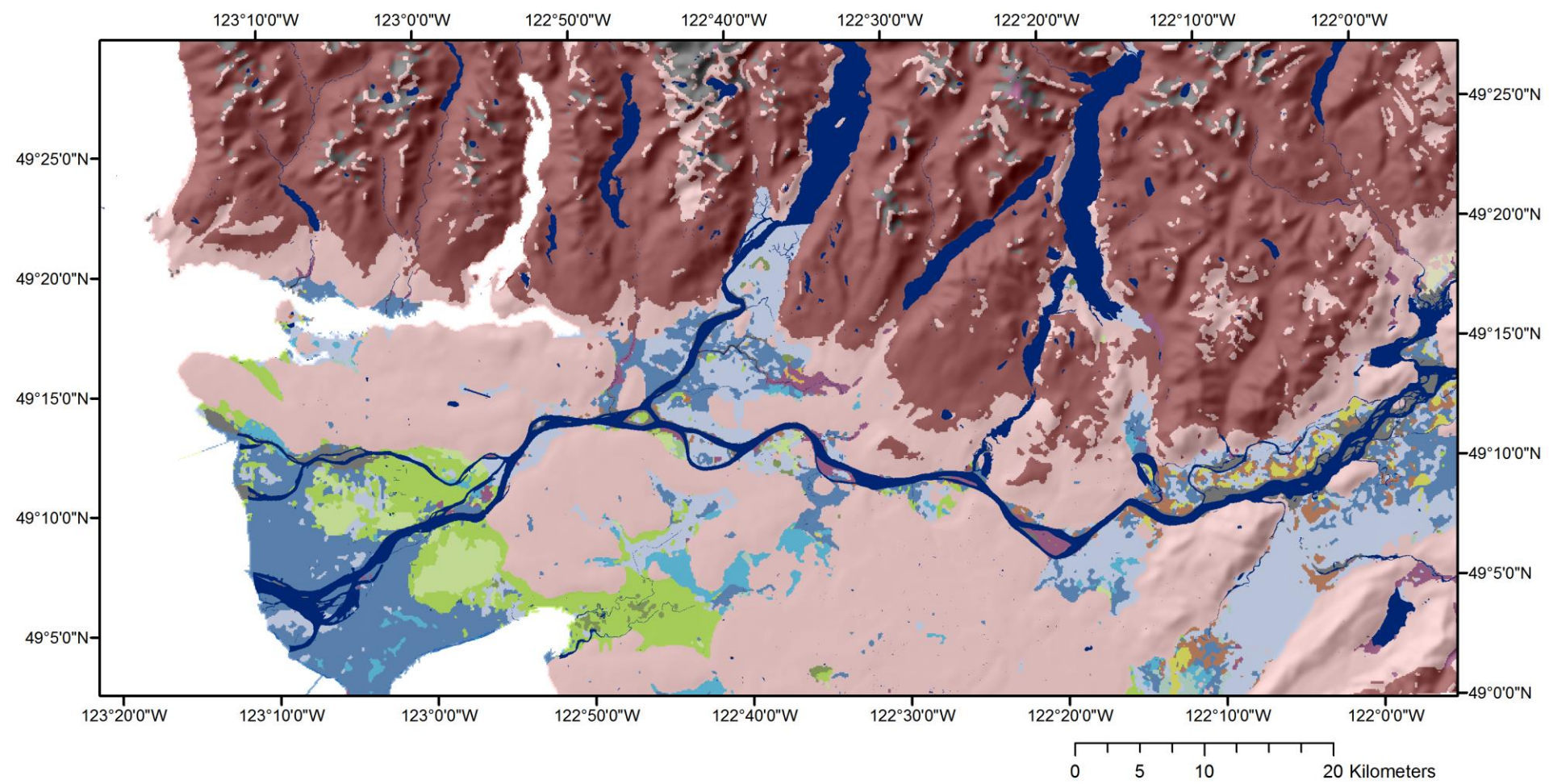


Soil Great Groups

- Dystric Brunisol
- Eutric Brunisol
- Melanic Brunisol
- Sombric Brunisol
- Ferro-Humic Podzol
- Humo-Ferric Podzol
- Folisol
- Fibrisol
- Mesisol
- Humisol
- Regosol
- Gray Brown Luvisol
- Gray Luvisol
- Gleysol
- Humic Gleysol
- Luvic Gleysol
- Bedrock, Rock Outcrop, Recent Alluvium, Talus
- Waterbodies

Area-Weighted: Support Vector Machine - Radial Basis Function

$C = 66\%$

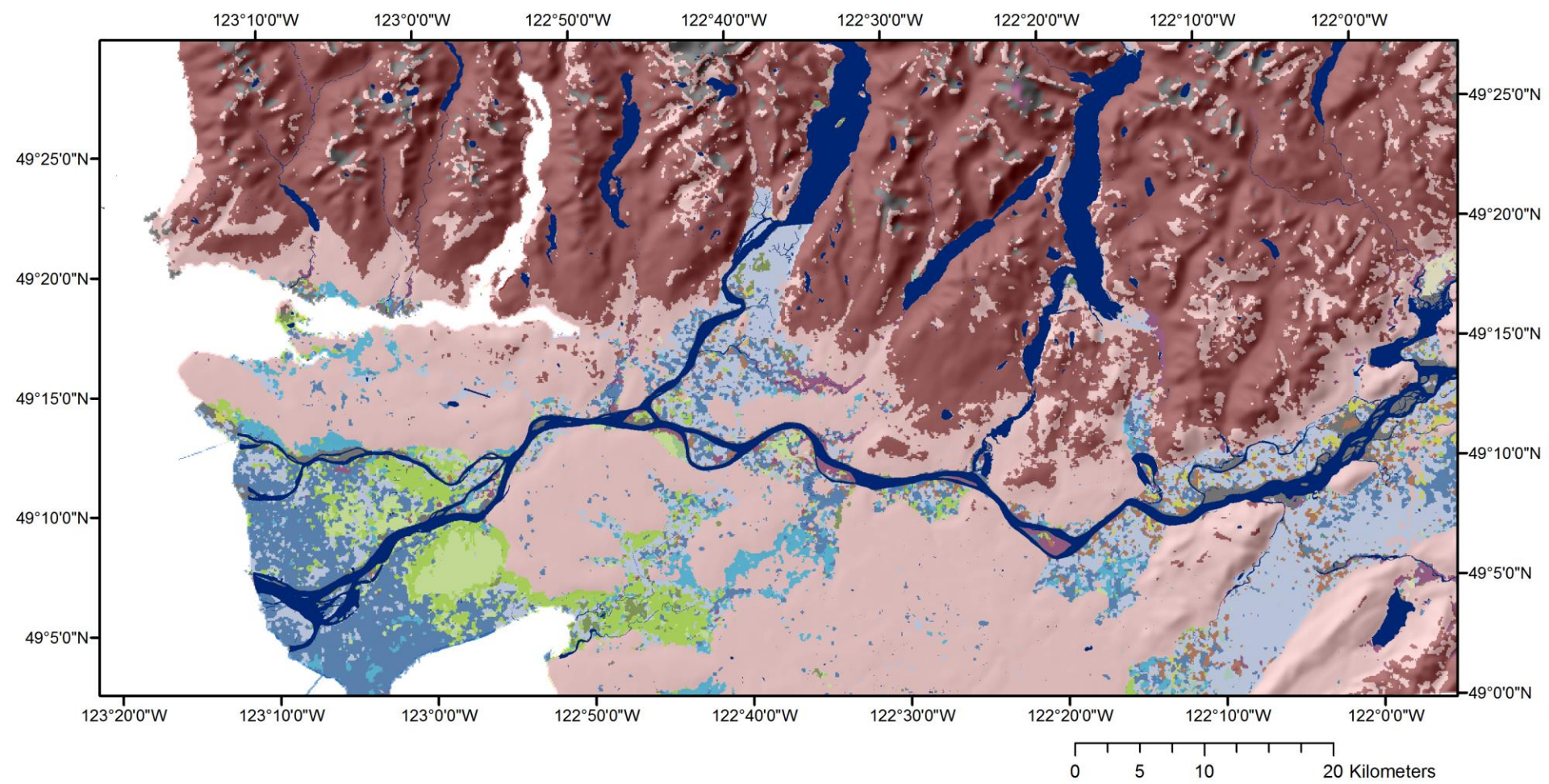


Soil Great Groups

- Dystric Brunisol
- Eutric Brunisol
- Melanic Brunisol
- Sombric Brunisol
- Ferro-Humic Podzol
- Humo-Ferric Podzol
- Folisol
- Fibrisol
- Mesisol
- Humisol
- Regosol
- Gray Brown Luvisol
- Gray Luvisol
- Gleysol
- Humic Gleysol
- Luvic Gleysol
- Bedrock, Rock Outcrop, Recent Alluvium, Talus
- Waterbodies

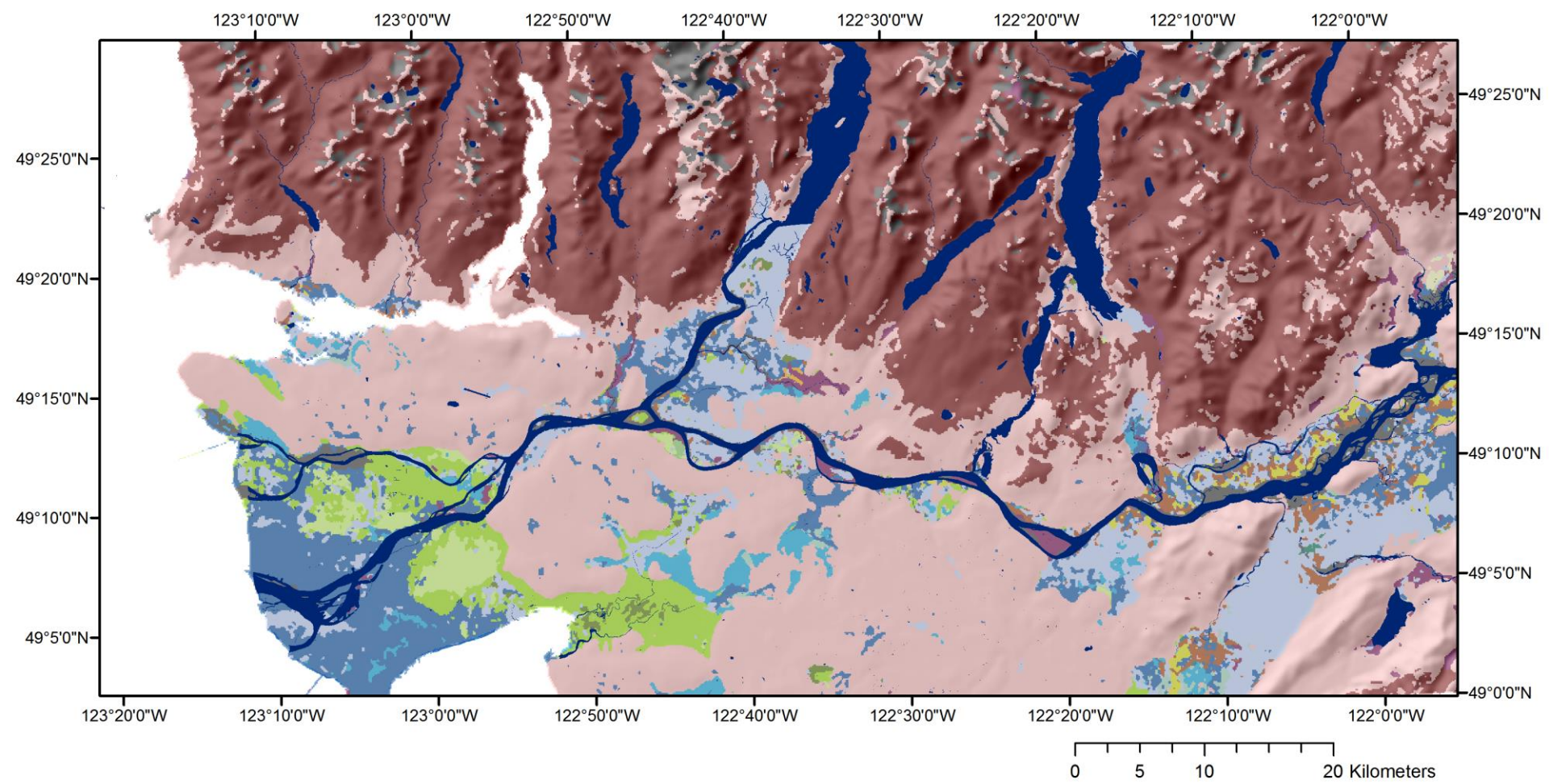
$C = 66\%$

Area-Weighted: Random Forest



Area-Weighted: k-Nearest Neighbours

$C = 70\%$

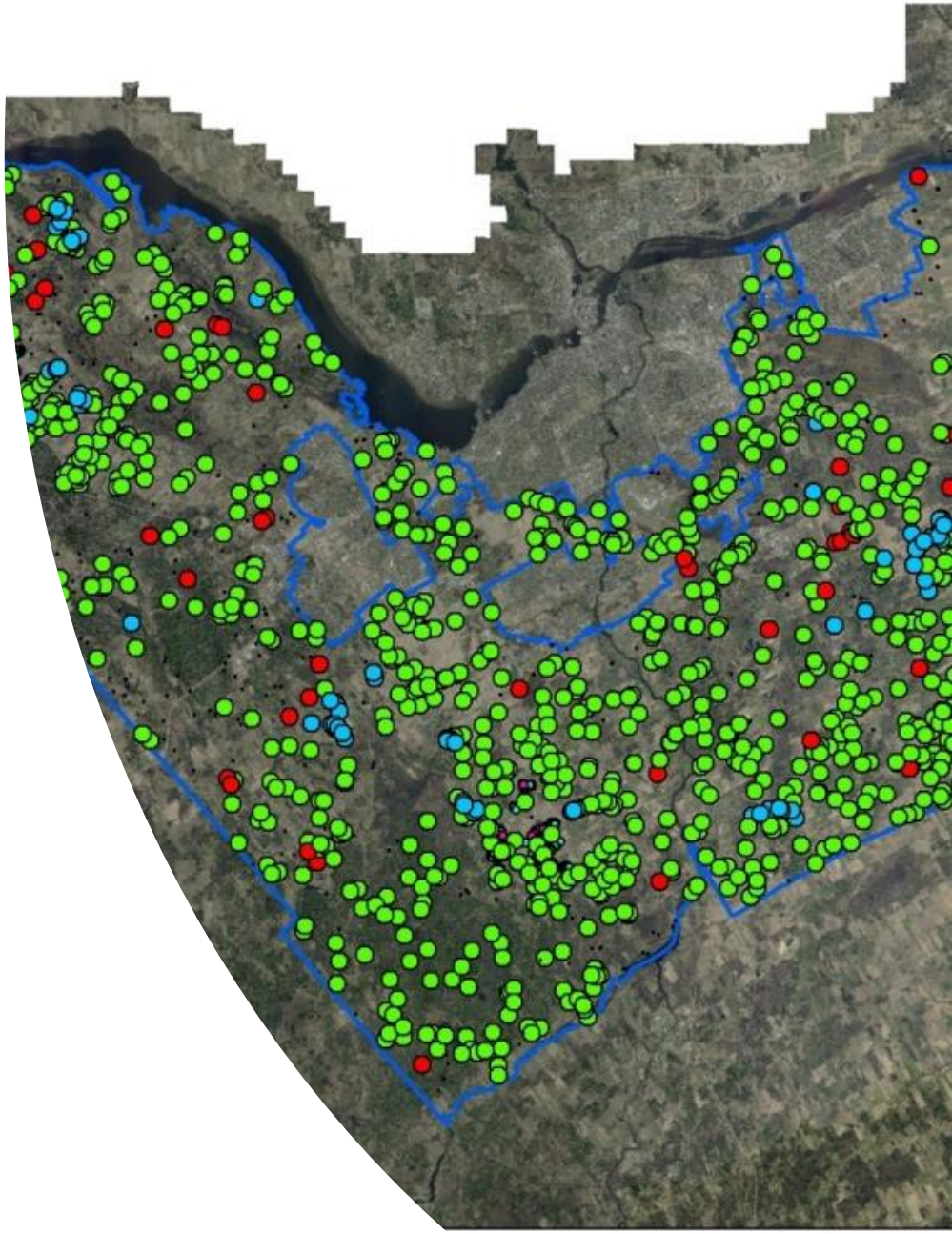


$C = 70\%$

Area-Weighted: CART with Bagging

Challenges with Validation

- Field sampling campaigns are extremely expensive – this is often the limiting factor for many studies.
- Sample sizes are often limited – but a decent number of sample points are required to train the machine-learning model.
- Need to maximize our datasets for both training and validation purposes in order to build good predictive models.



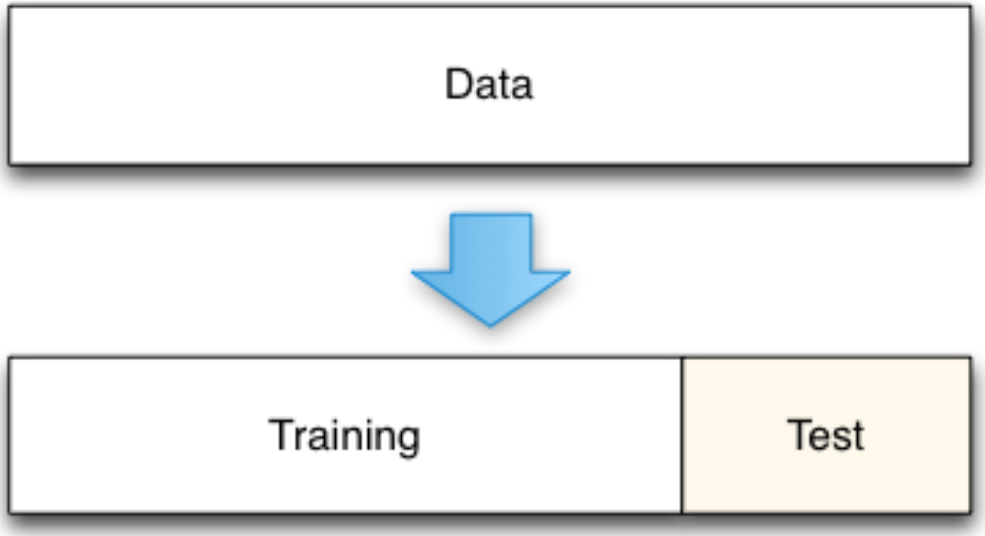
Cross-Validation

A form of validation where the full dataset is partitioned into training and validation datasets.

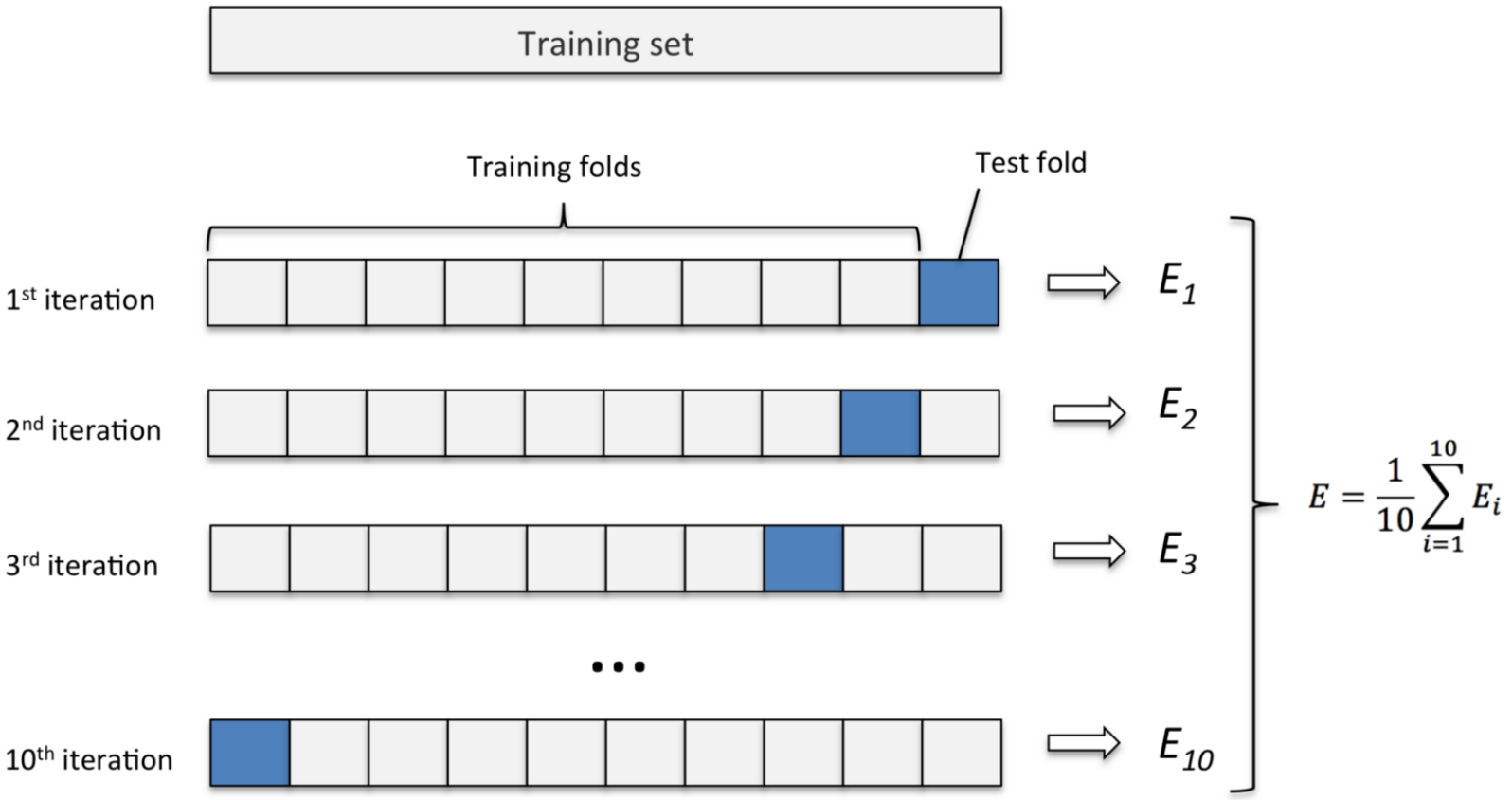
Some Common Options:

- (Repeated) Random Hold-Out Cross-Validation
- (Repeated) *k*-Fold Cross-validation
- Leave-One-Out Cross-Validation

Hold-Out Cross Validation

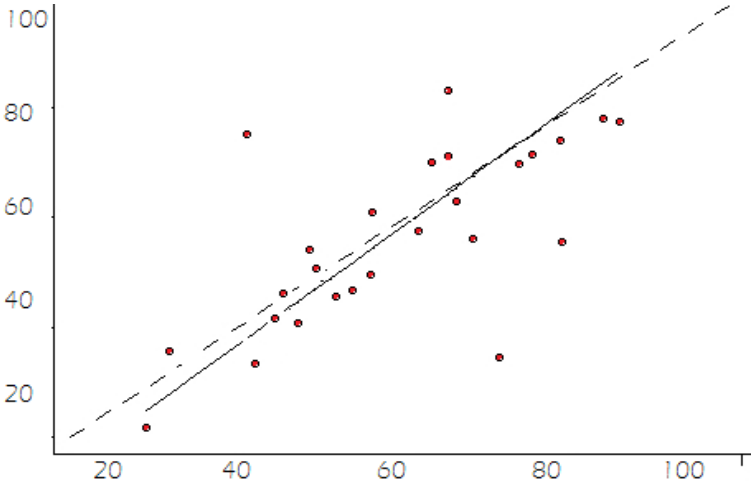


k- Fold Cross Validation



Accuracy Metrics – Continuous Properties

- R^2
  - measures the precision of the relationship
  - looks at observed vs predicted values
- **Concordance**
  - **Lin's concordance correlation coefficient**
  - **Evaluates accuracy and precision**
  - **Compares to a 45 degree (or 1:1) line**



- RMSE
  - Measures the difference between predicted values and observed values
  - In same units as the target variable

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(obs_i - pred_i)^2}{n}}$$

- Bias
  - Informs us if the model is consistently over-predicting or under-predicting
  - Even if the model has no bias (bias = 0), it can still be over- and under-predicting, just in equal amounts

$$bias = \frac{\sum_{i=1}^{n} pred_i - obs_i}{n}$$

Accuracy Metrics - Categorical

- Cohen's Kappa Coefficient (κ)
  - Measures agreement between predicted vs. actual classes
  - Accounts for by-chance agreement between predicted vs. actual classes

- Overall Correctness
  - The proportion of points that were correctly classified

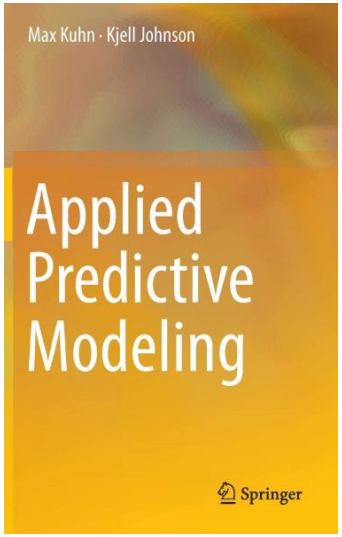
$$k = \frac{p_o - p_e}{1 - p_e}$$

Predicted	Actual																		
	Polygon-Derived Prediction																		
	BC	BLC	DBC	DGC	DB	EB	F	FHP	G	GL	H	HFP	HG	HR	LG	M	R	SB	Total
BC	**18**	0	2	0	0	4	0	0	0	0	1	0	0	0	0	1	0	0	26
BLC	1	**18**	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	21
DBC	0	2	**31**	2	0	0	0	0	0	2	0	0	2	0	0	0	0	0	39
DGC	0	0	0	**1**	0	0	0	0	0	1	0	0	0	0	0	1	0	0	3
DB	2	3	2	0	**88**	7	1	0	1	6	0	3	2	0	0	0	6	2	123
EB	2	3	3	8	4	**84**	0	0	0	8	0	0	2	2	2	0	9	0	127
F	0	0	0	0	1	0	**0**	0	0	0	0	0	0	0	0	0	0	0	1
FHP	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0	0
G	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0	0	0
GL	0	3	1	4	3	3	0	0	0	**33**	0	1	0	0	0	0	2	1	51
H	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0	0	0	0	0
HFP	0	0	0	0	6	1	0	1	0	5	0	**28**	1	1	0	2	1	4	50
HG	0	0	0	0	0	0	0	0	1	0	0	0	**7**	1	0	0	2	0	11
HR	0	0	0	1	0	0	0	0	0	0	0	0	0	**3**	0	0	0	0	4
LG	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	0	0	0	0
M	0	0	0	0	1	0	0	0	0	2	0	0	0	0	0	**0**	0	0	3
R	0	0	1	0	0	3	0	0	0	0	0	0	1	1	0	0	**12**	0	18
SB	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	**0**	1
Total	23	29	42	16	103	102	1	1	2	57	1	33	15	8	2	4	32	7	

The *caret* package

- Common framework for applying machine-learning algorithms for classification and regression purposes.
- Streamlines the optimization of hyper-parameters
- Streamlines the validation processes
- Allows for parallel processing to increase efficiency
- About 100 different algorithms

```r
### MODEL TRAINING & VALIDATION ###

fitControl <- trainControl(
  method = "repeatedcv", number=10, repeats=2, allowParallel = TRUE,
  returnResamp = "all", savePredictions = TRUE
)

### OPTIMIZATION OF HYPERPARAMETERS ###

mtry=c(2, 4, 6, 8, 10, 12, 14, 16, 17)
RF_tune <- data.frame(mtry=mtry)

### PARALLEL PROCESSING ###
library(doParallel)
library(foreach)
registerDoParallel(cores=5)

### MODEL COMPARISON ###

# RANDOM FOREST #
RF <- train(
  data = Training_Points,
  ML_Equation,
  method = "rf",
  tuneGrid = RF_tune,
  ntree = 500,
  trControl = fitControl)

# k NEAREST NEIGHBOURS #
knn <- train(
  data = Training_Points,
  ML_Equation,
  method = "kNN",
  tuneGrid = RF_tune,
  trControl = fitControl)
```



Least Angle Regression (`method = 'lars'`)
For regression using package **lars** with tuning parameters:

- Fraction (`fraction`, numeric)

Least Angle Regression (`method = 'lars2'`)
For regression using package **lars** with tuning parameters:

- Number of Steps (`step`, numeric)

Least Squares Support Vector Machine (`method = 'lssvmLinear'`)
For classification using package **kernlab** with tuning parameters:

- Regularization Parameter (`tau`, numeric)

Least Squares Support Vector Machine with Polynomial Kernel (`method = 'lssvmPoly'`)
For classification using package **kernlab** with tuning parameters:

- Polynomial Degree (`degree`, numeric)
- Scale (`scale`, numeric)
- Regularization Parameter (`tau`, numeric)

Least Squares Support Vector Machine with Radial Basis Function Kernel (`method = 'lssvmRadial'`)
For classification using package **kernlab** with tuning parameters:

- Sigma (`sigma`, numeric)
- Regularization Parameter (`tau`, numeric)

Linear Discriminant Analysis (`method = 'lda'`)
For classification using package **MASS** with no tuning parameters.
Linear Discriminant Analysis (`method = 'lda2'`)
For classification using package **MASS** with tuning parameters:

- Number of Discriminant Functions (`dimen`, numeric)

Linear Discriminant Analysis with Stepwise Feature Selection (`method = 'stepLDA'`)
For classification using packages **klaR** and **MASS** with tuning parameters:

- Maximum Number of Variables (`maxvar`, numeric)
- Search Direction (`direction`, character)

Linear Distance Weighted Discrimination (`method = 'dwdLinear'`)
For classification using package **kerndwd** with tuning parameters:

- Regularization Parameter (`lambda`, numeric)
- q (`qval`, numeric)

Linear Regression (`method = 'lm'`)
For regression with tuning parameters:

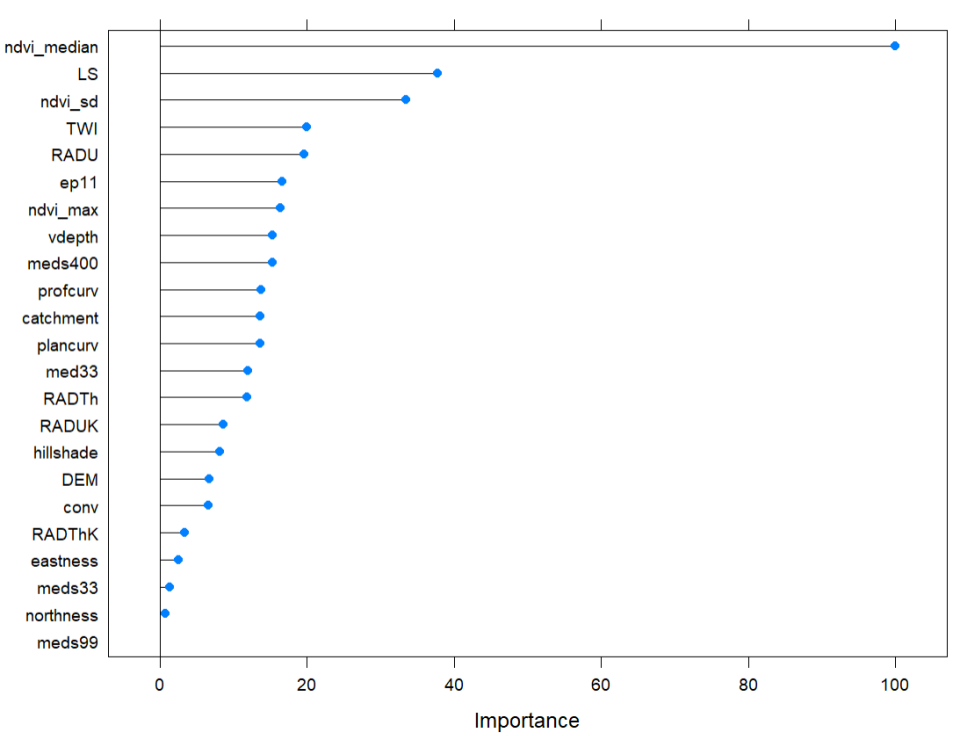
On Irrelevant Predictors

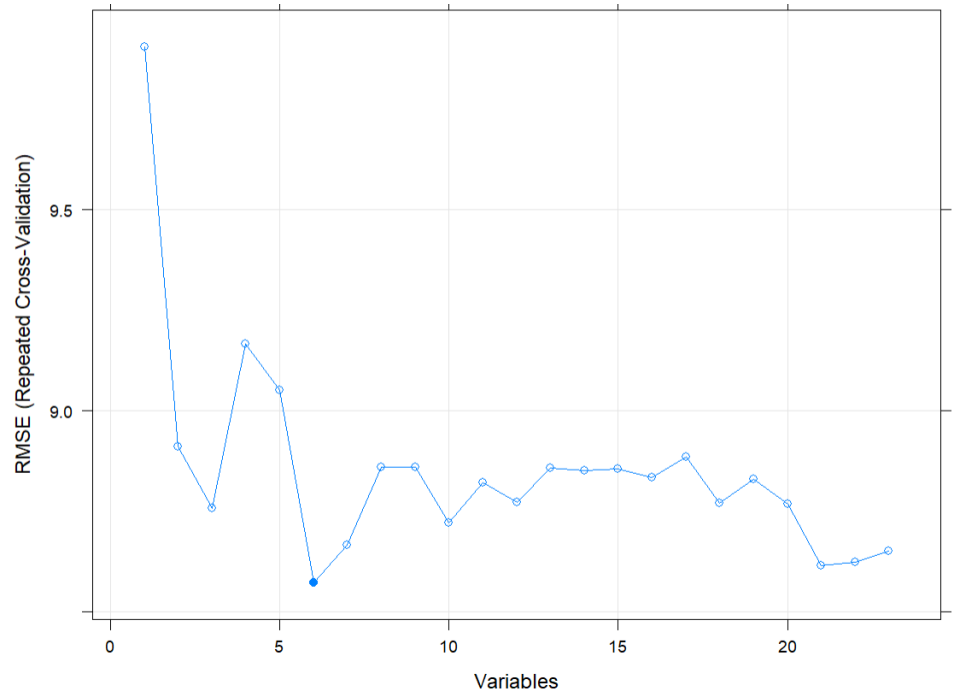
- Some models are sensitive to the number of predictors.
- Irrelevant predictors should be removed to:
  - Improve our ability to interpret soil-environmental relationships.
  - Potentially increase accuracy.
  - Make the model as small as possible.
  - Ensure generalizability.

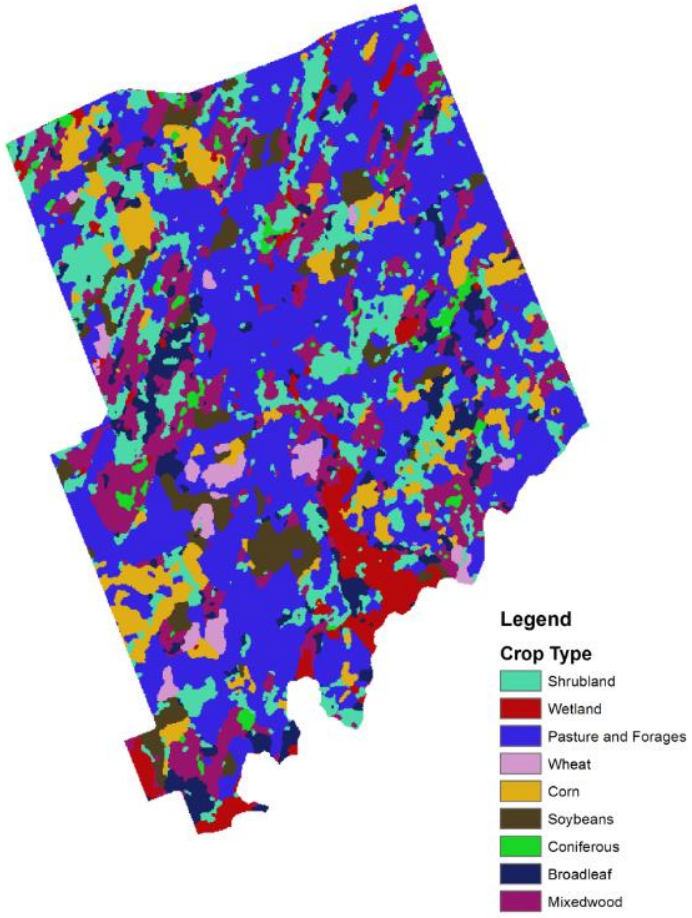
Recursive Feature Elimination

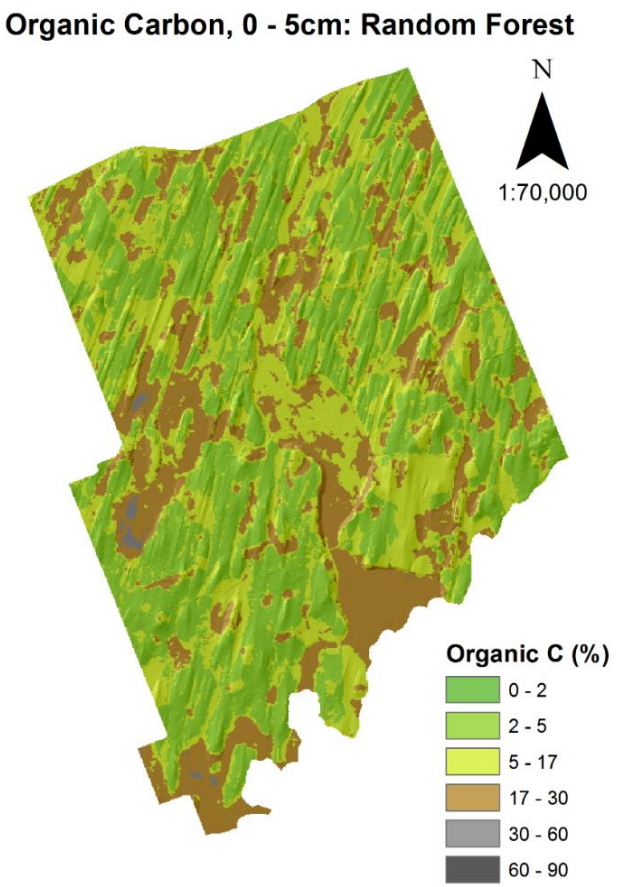
Variable importance analysis can be carried out for each model.

- **Step 1:** Fit model.
- **Step 2:** Generate variable importance metrics.
- **Step 3:** Calculate accuracy metrics.
- **Step 4:** Remove least important variable.
- **Step 5:** Repeat Steps 1-4 until only 1 variable remains.







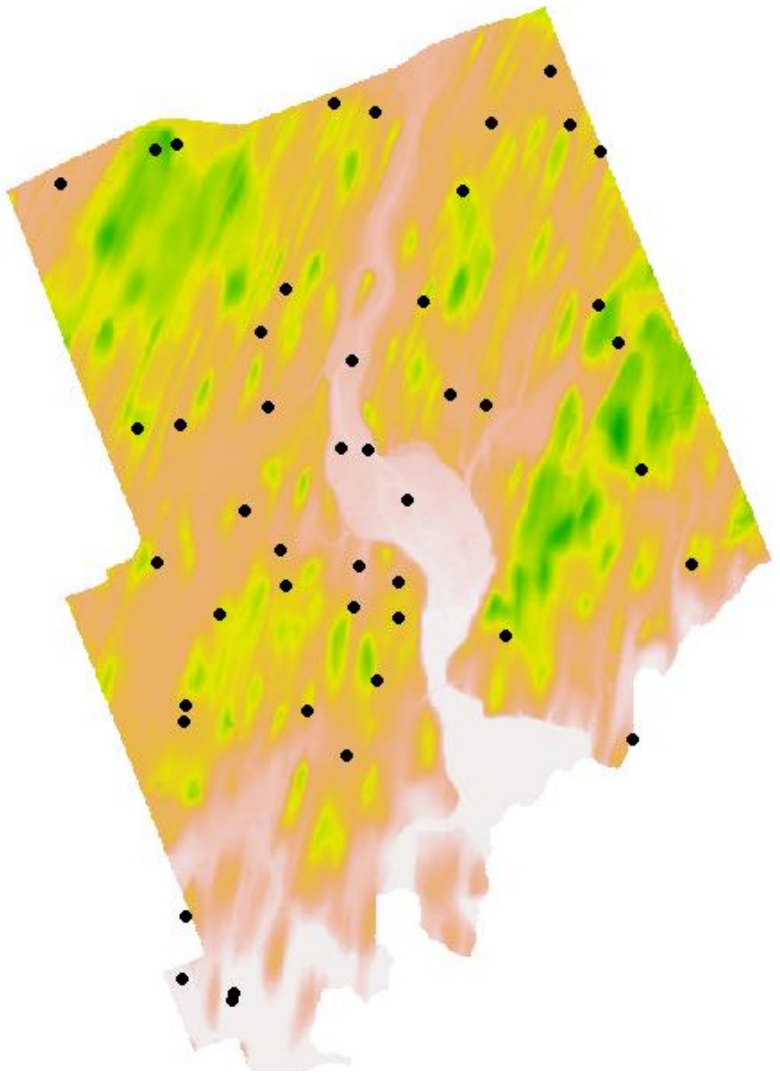


Case Study I: Modelling Continuous Data

Canadian Digital Soil Mapping Workshop, 2019

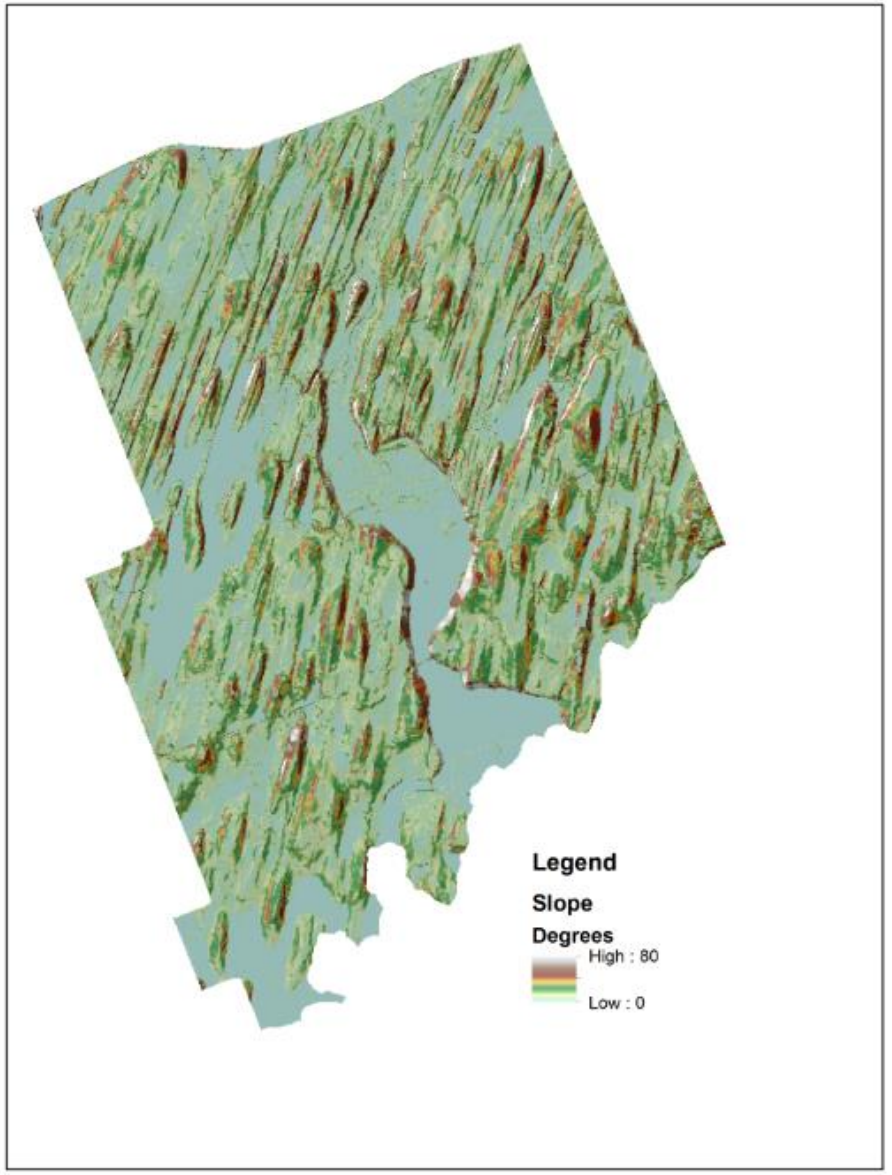
Soil Data

- 94 locations were sampled on a horizon-by-horizon basis using a Conditioned Latin Hypercube Sample Design
- 297 samples were analyzed for particle size fractions, organic carbon, and pH
- The upper and lower depth values were recorded for each horizon



Covariates (scorpan)

- Topographic indices derived from a digital elevation model (*r* factor)
  - Slope position
  - Elevation deviation
  - Wetness index
  - Profile curvature
  - Topographic ruggedness index
- Crop data (*o* factor)
- Gamma radiometric data (*s* factor)



Case Study I: Summary

1. **Harmonize Soil Data**: Everything is in a horizon basis. We will use an equal-area spline function to estimate soil attribute values at common depths (0, 5, 15, 30, 60, 100 cm).

2. **Develop Training Data:** Each soil observation will be spatially intersected with the suite of soil-environmental covariates.

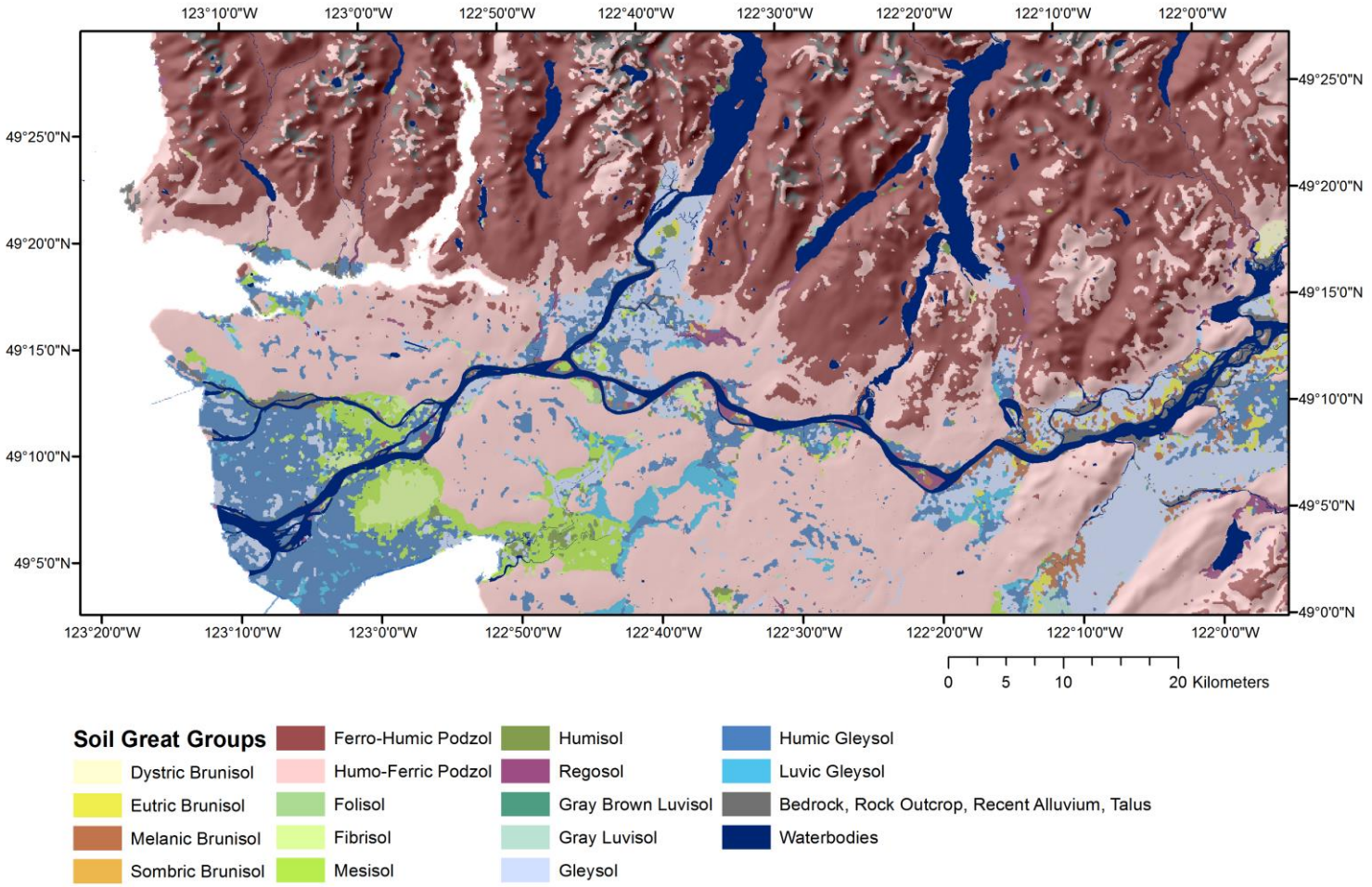
3. **Model Training & Validation:** We will train the Cubist and Random Forest models and optimize their hyperparameters. Validation will be performed using multiple replicates of 10-fold cross validation.

Case Study I: Summary

4. **Variable Importance Analysis:** Using caret package, we will identify which variables contribute the most to the prediction of the target variable.

5. **Map Prediction:** The final map will be predicted and visualized.

6. **Recursive Feature Elimination:** Irrelevant variables will be sequentially removed.



Case Study II: Modelling Categorical Data

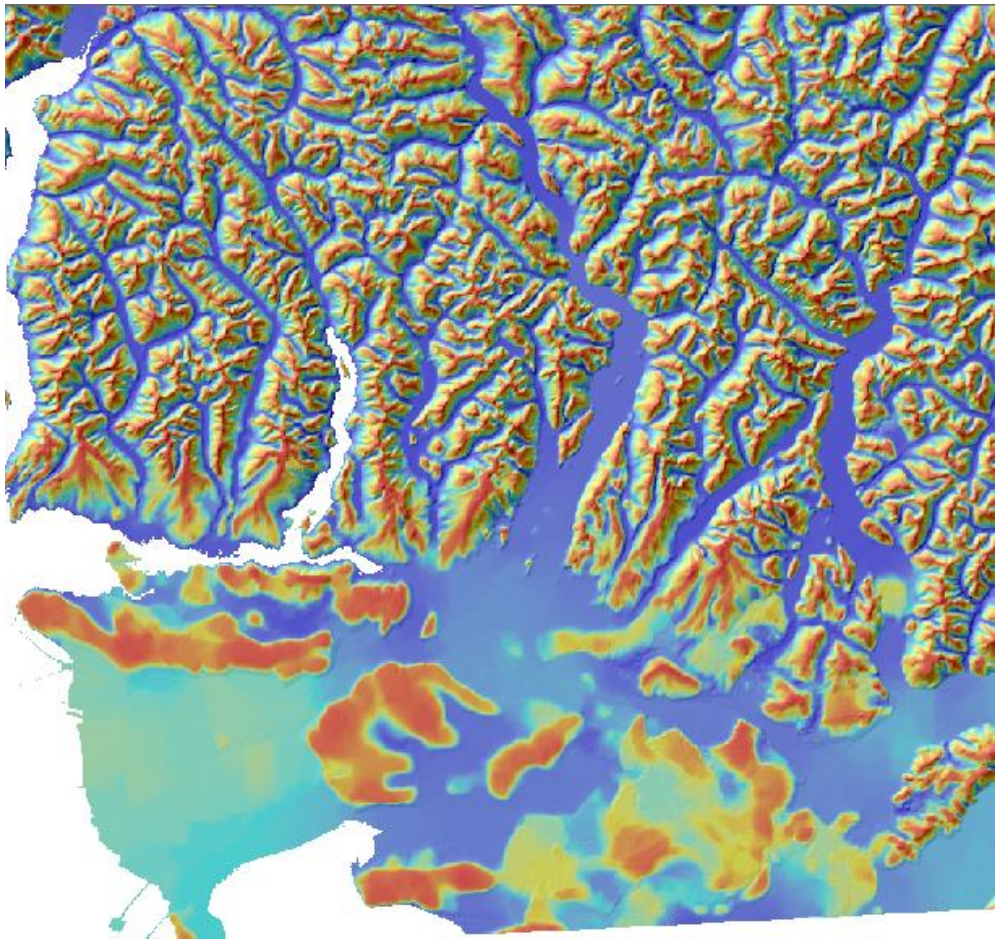
Canadian Digital Soil Mapping Workshop, 2019

Soil Data

- Using the Soils of the Langley-Vancouver Map Area soil survey, synthetic training data was extracted from single-component map units
- 3,757 were randomly generated based on 15 soil great groups

Covariates (scorpan)

- Temperature & Precipitation (*c* factor)
- Biogeoclimatic Ecosystem Classification (*o* factor)
- Topographic indices derived from a digital elevation model (*r* factor)
- Bedrock geology (*p* factor)
- Distance to river and stream (*n* factor)

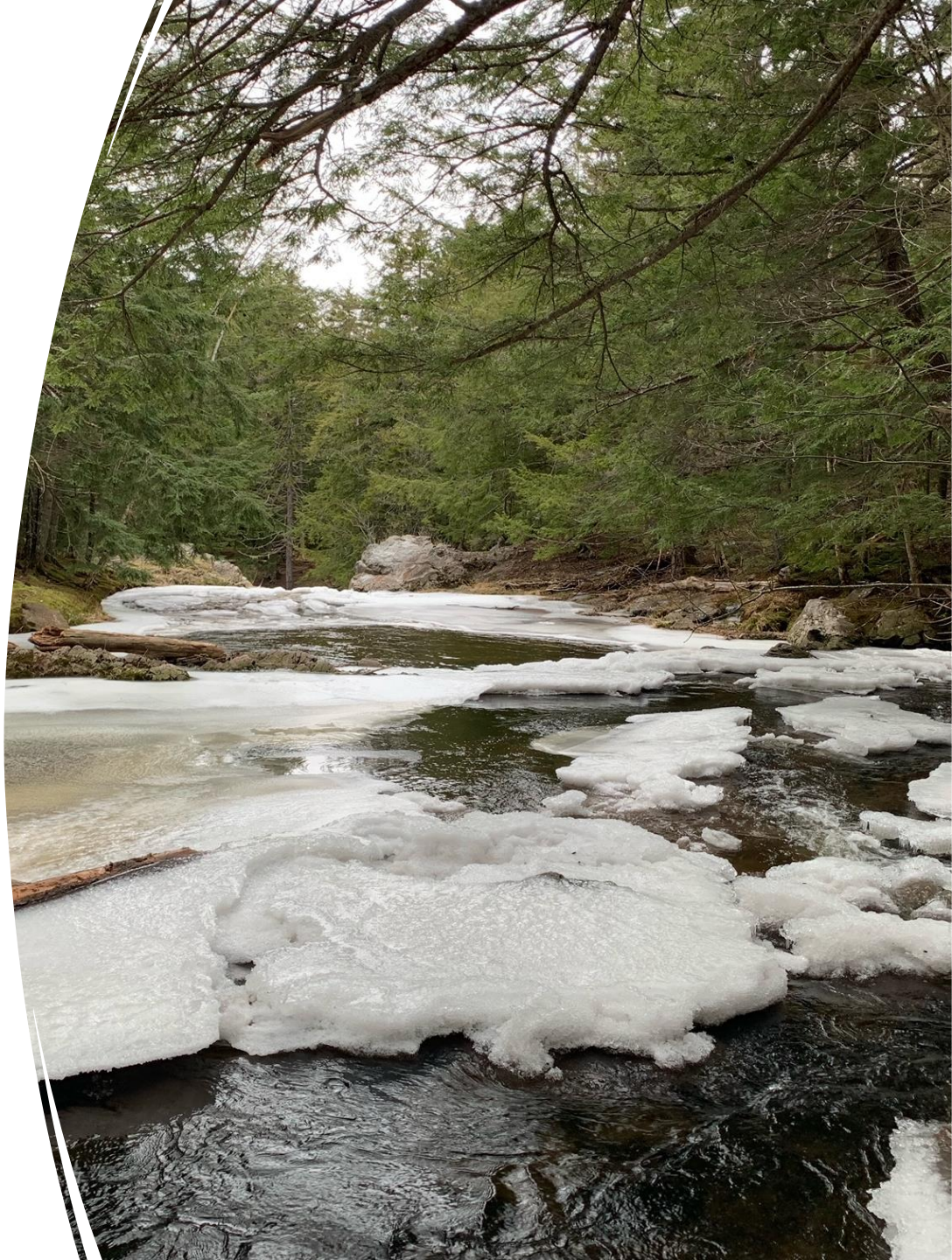


Case Study II: Summary

1. **Develop Training Data:** Each soil observation will be spatially intersected with the suite of soil-environmental covariates. This time we will use a parallel processing technique.

2. **Model Training & Validation:** We will train the Cubist and Random Forest models and optimize their hyperparameters. Validation will be performed using multiple replicates of 10-fold cross validation.

3. **Map Prediction:** The final map will be predicted and visualized.

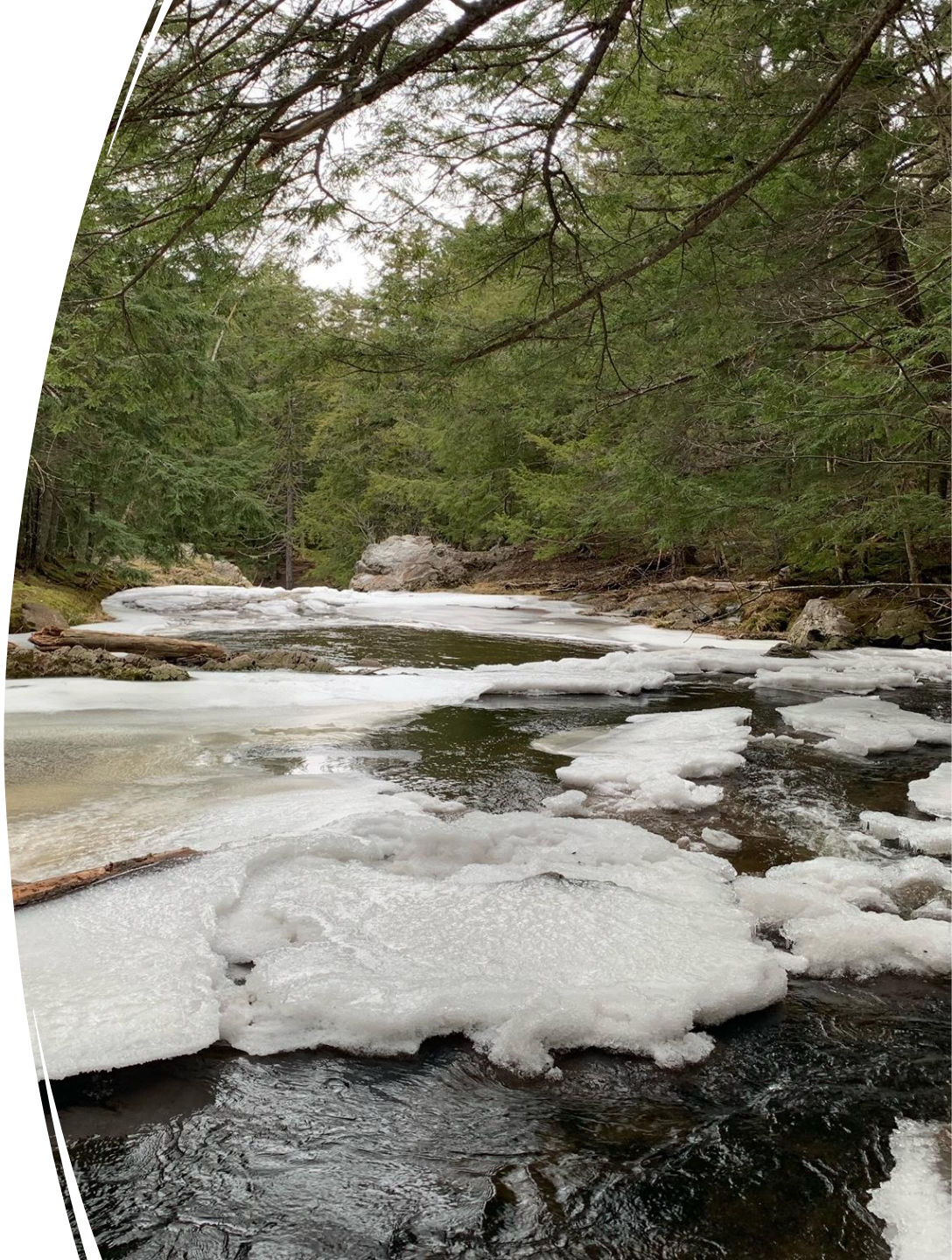


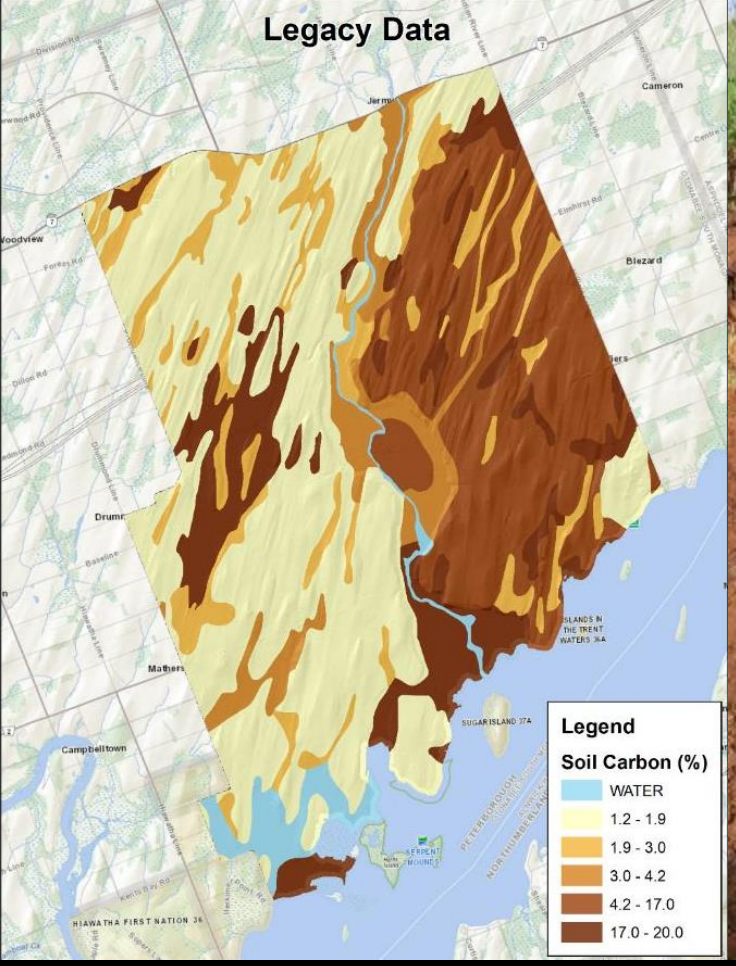
Case Study II: Summary

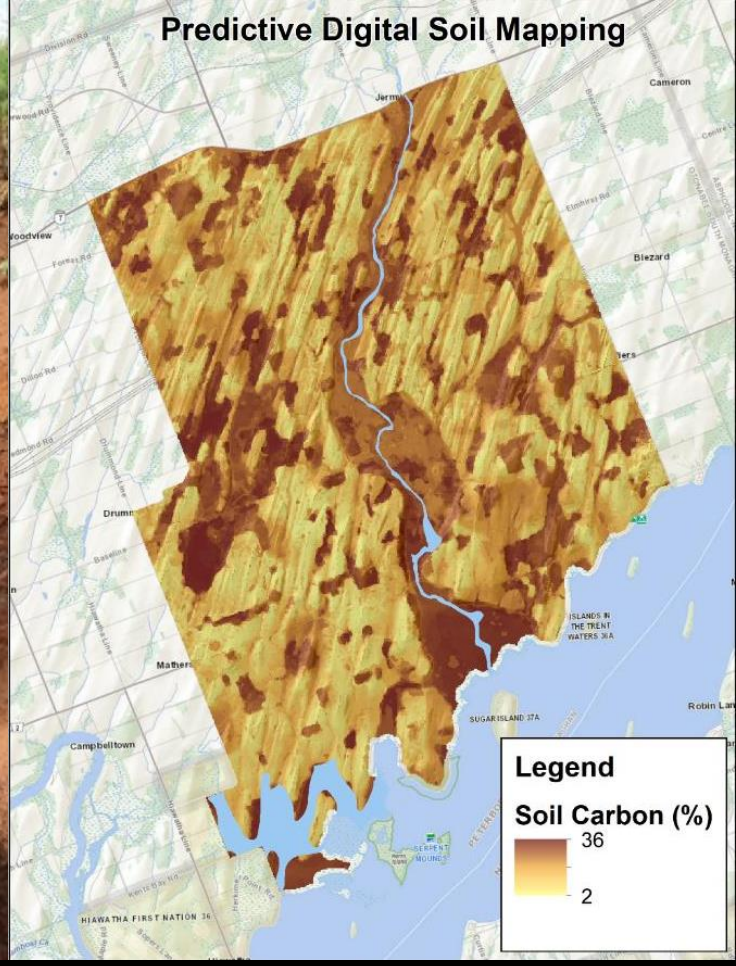
3. **Generate Probability Rasters:** Using the RF ensemble modeling approach, we are able to generate maps that show the probability of occurrence for each class.

4. **Generate Uncertainty Mapping:** We can calculate the entropy uncertainty using the probability rasters to see where prediction uncertainty is highest

4. **Variable Importance Analysis:** Using caret package, we will identify which variables contribute the most to the prediction of the target variable







5. Basics of Machine-Learning with Applications for Digital Soil Mapping

Canadian Digital Soil Mapping Workshop, 2020

CDSM WG
CANADIAN SOCIETY OF SOIL SCIENCE
SOCIÉTÉ CANADIENNE DE LA SCIENCE DU SOL
Ontario
UNIVERSITY of GUELPH
DALHOUSIE UNIVERSITY