Agriculture and Agri-Food Canada
Agriculture et Agroalimentaire Canada

Legacy soil survey data mining for digital soil mapping in Prince Edward Island, Canada

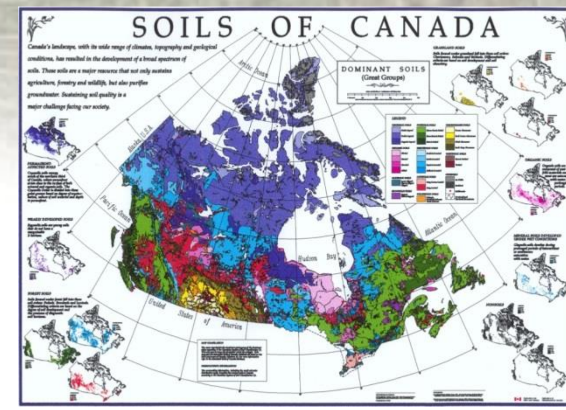
Xiaoyuan Geng, Juanxia He, Yefang Jiang, Bert VandenBygaart

Science and Technology Branch, Agriculture & Agri-Food Canada

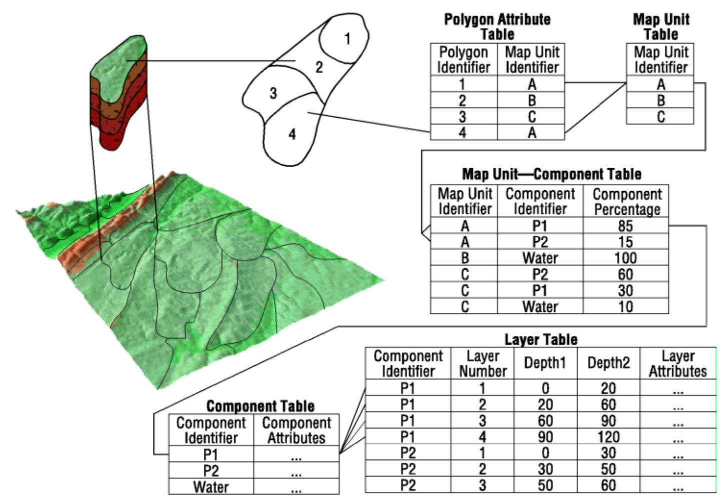
7th Global Digital Soil Mapping Workshop, 2016
June 27th to July 2nd , 2016, Aarhus, Denmark

Canada

State of the national soil and soil landscape data



- National Ecological Framework (ECO)
- Soil Landscapes of Canada (SLC)
- Canada Land Inventory (CLI)
- Detailed Soil Surveys (DSS)
- Site (pedon) data
- Soil Classification System for Canada
- National soil carbon database
- http://sis.agr.gc.ca/cansis



Methods for future national soil data provision

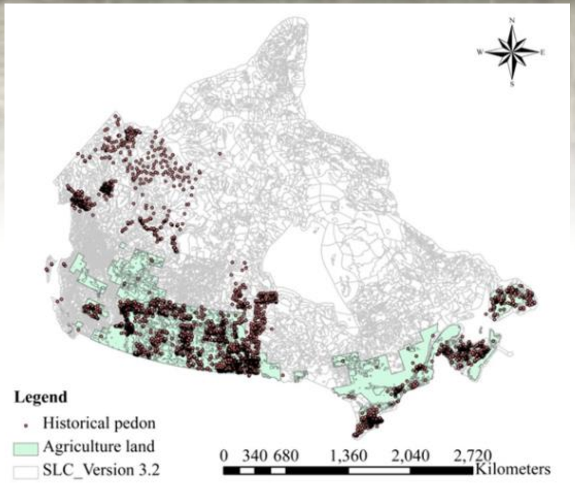
Geostatistic based approaches

Kriging and Co-Kriging
GLM etc.

Knowledge-based inference

Classification & Regression Tree
Random Forest
Fuzzy Set and Fuzzy Logic
Neural Networks
Bayesian Networks
Support Vector Machine (SVM)

Two approaches are not mutually Exclusive.



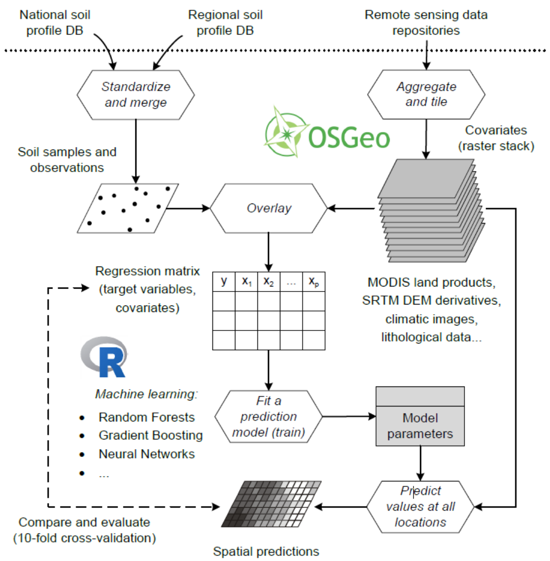
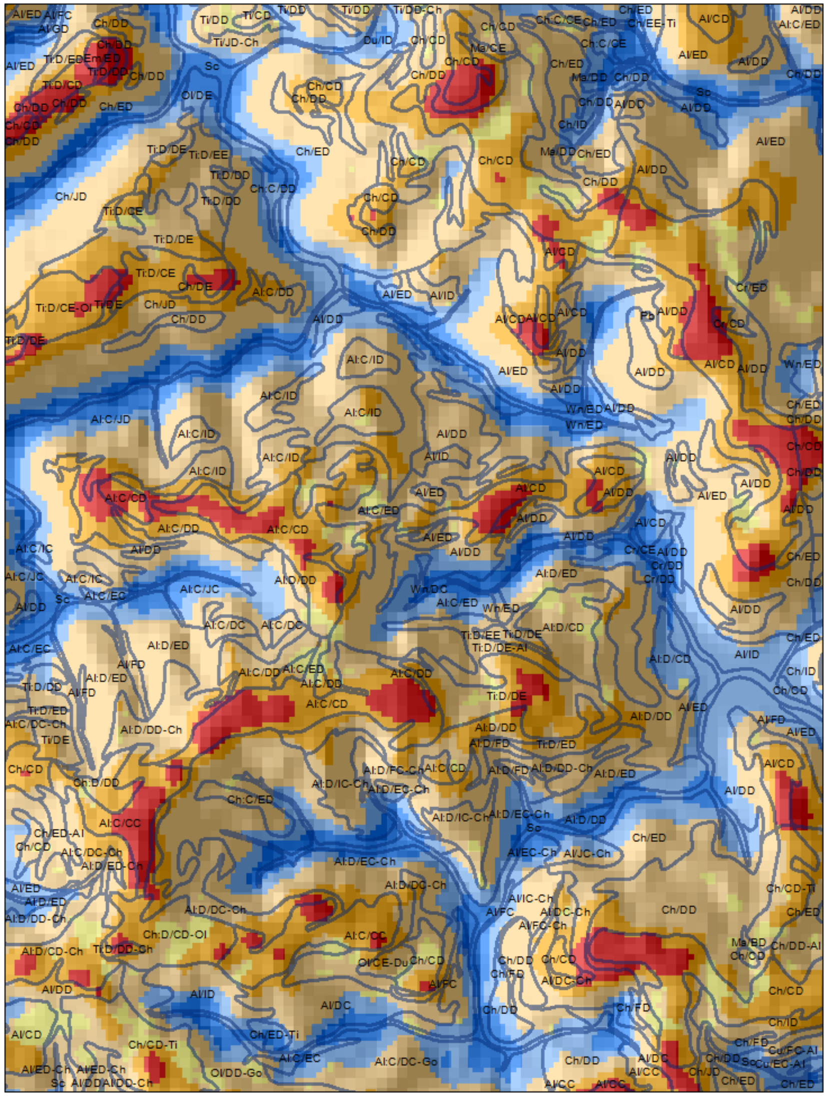


Diagram source: Hengl et al., 2016

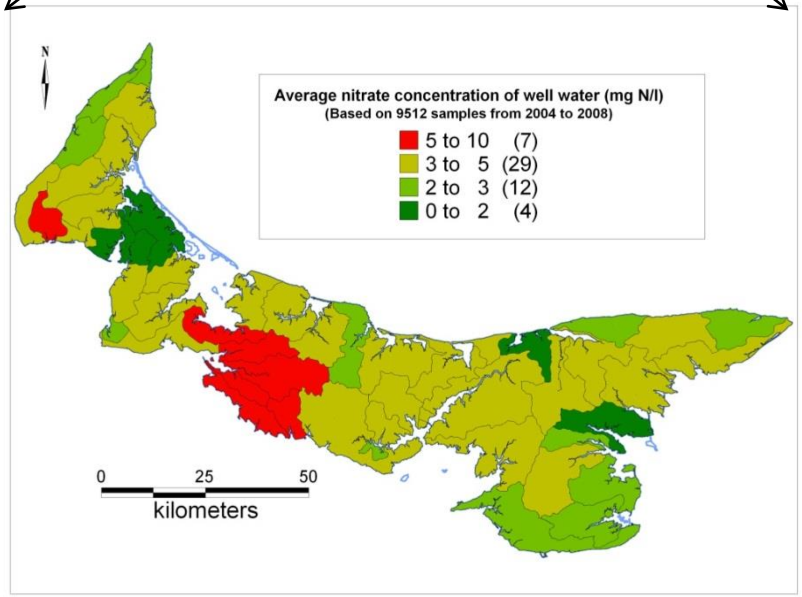
Hypothesis and legacy data mining

Any location within each of the single component polygons of the detailed soil survey can be used to represent a spatial location of the associated soil component or type for that polygon.

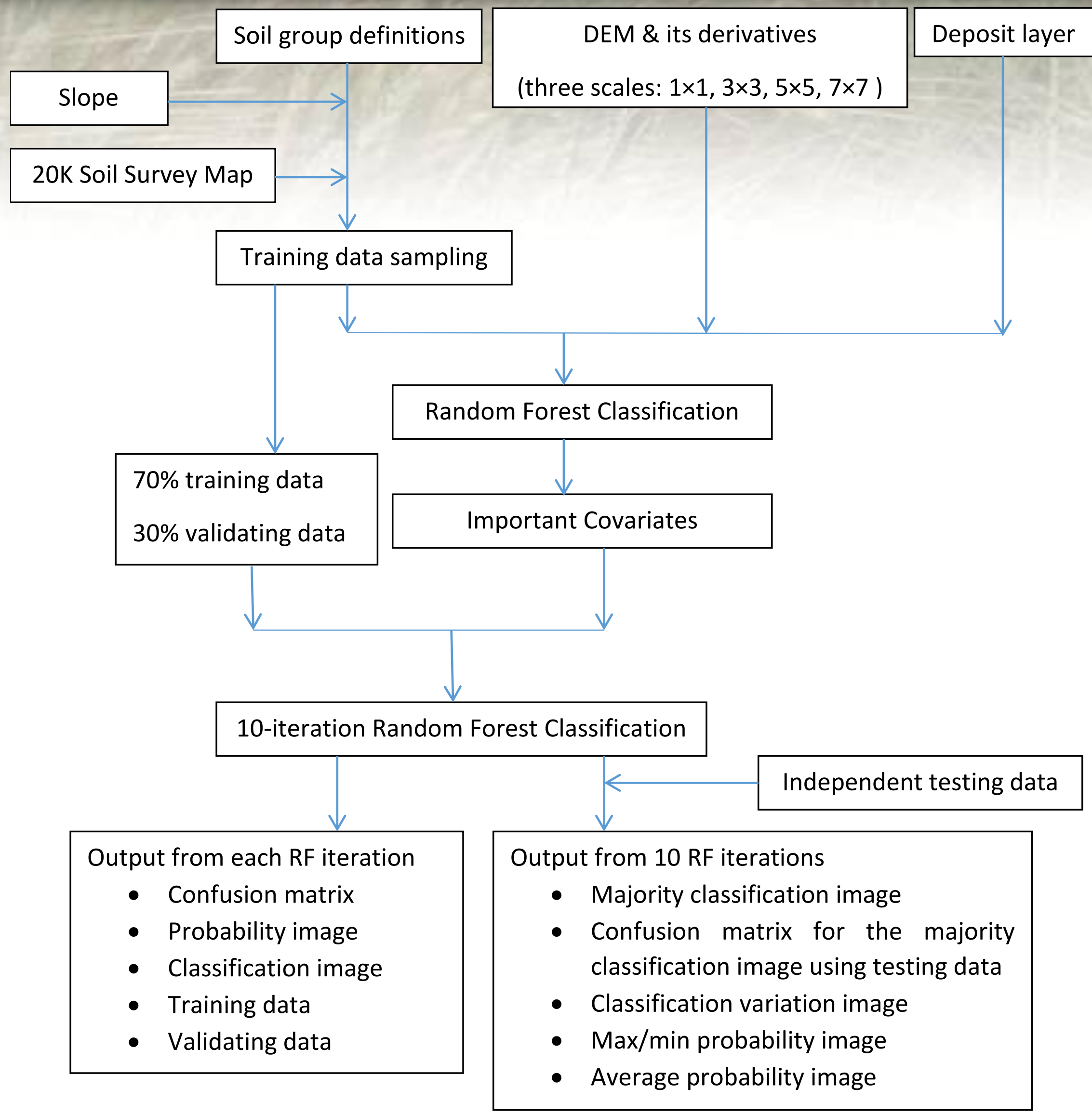


Location and requirements

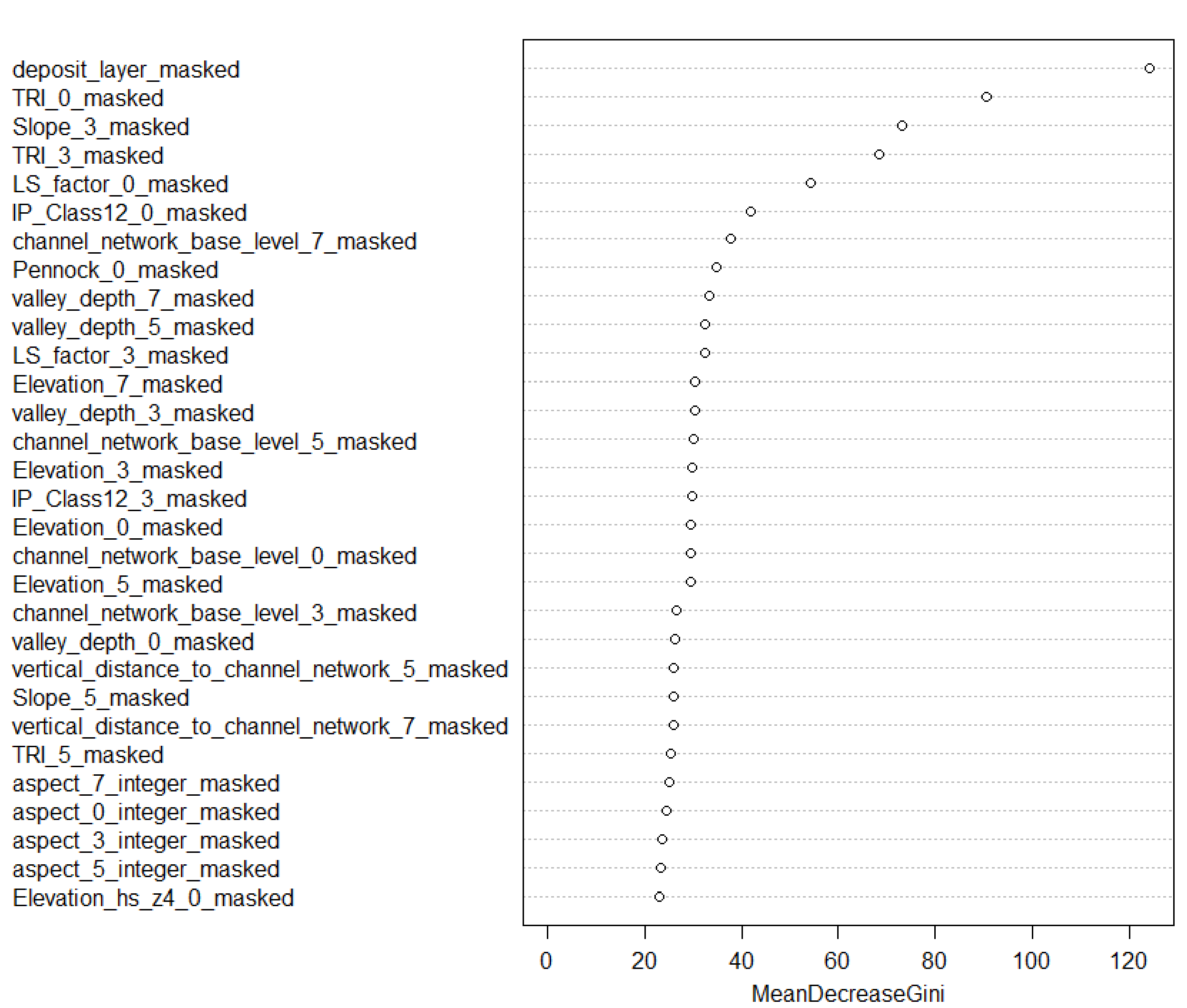
- Intensive crop production on permeable soils in sloping landscapes.
- High risk of groundwater contamination by nutrients and agri-chemicals.
- Loss of productivity and water course siltation due to soil erosion.
- Competition between irrigation and environmental water uses.



Data and methods



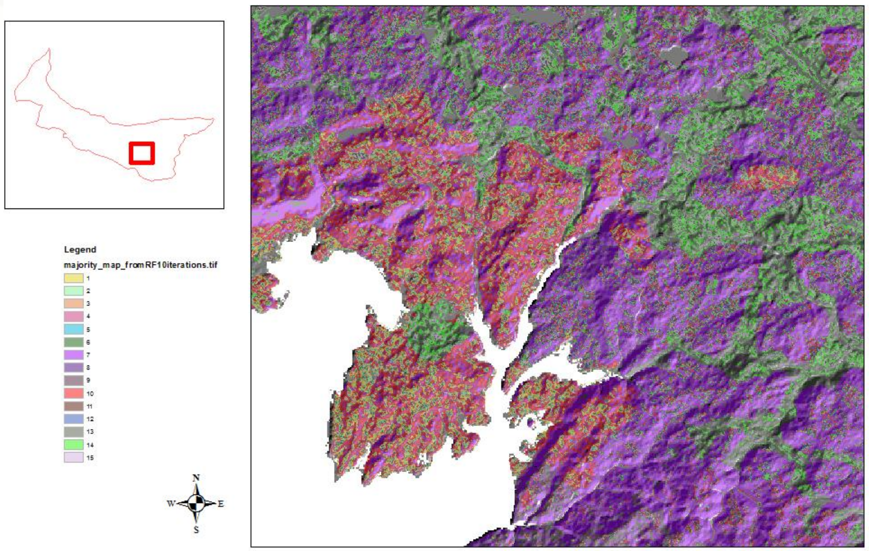
Data and methods: multi-scale feature reduction



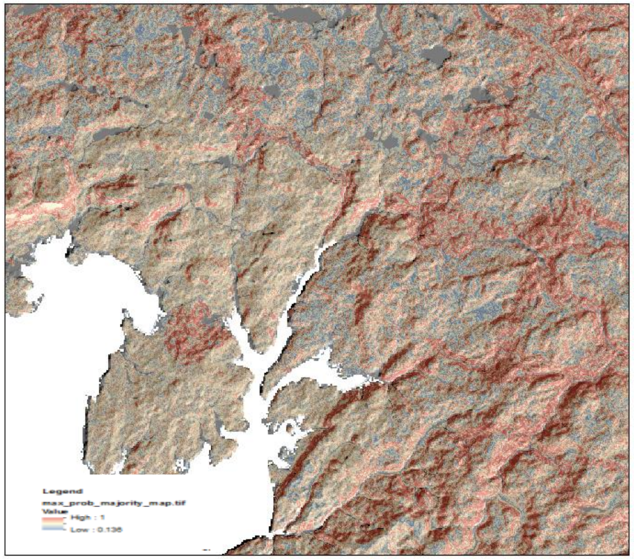
Covariates selected: Surficial geological material, Topographic Rugedness Index (TRI), Slope gradient and TRI at 90m resolution, and LS_factor.

Results and discussions

- Overall accuracy is less than 25% with fully random sampling
- Overall accuracy is increased to 40% with simple sampling constraints
- Both soil type and probability maps are available



Soil Class Map Based on 10 Iterations of RF



Maximum Probability Map Based on 10 Iterations of RF

	1	2	3	4	5	6	7	8	9	10	Average
Overall Accuracy	0.43	0.43	0.43	0.43	0.45	0.42	0.43	0.44	0.43	0.43	0.432
Kappa	0.39	0.39	0.39	0.38	0.41	0.37	0.4	0.4	0.39	0.38	0.39

Results and discussions continued

Surficial geology is most defining soil distribution across the landscapes in PEI

Soil types mapped and reported via legacy soil survey need to be examined and regrouped

Machine learning based approach is more feasible in Canada

Independent validation data set(s) are vital

Repeatable methods as new training points and co-variants becoming available

Sources of training information for machine learning are many, but needs expert analysis

What's next?

National vs. case specific (business driven) DSM

- Across various resolutions (250m to 10m)
- Training data and data mining
- Canadian peatland mapping and carbon stocks
- Changing environment and permafrost soils
- Ensemble and multi-fold machine learning

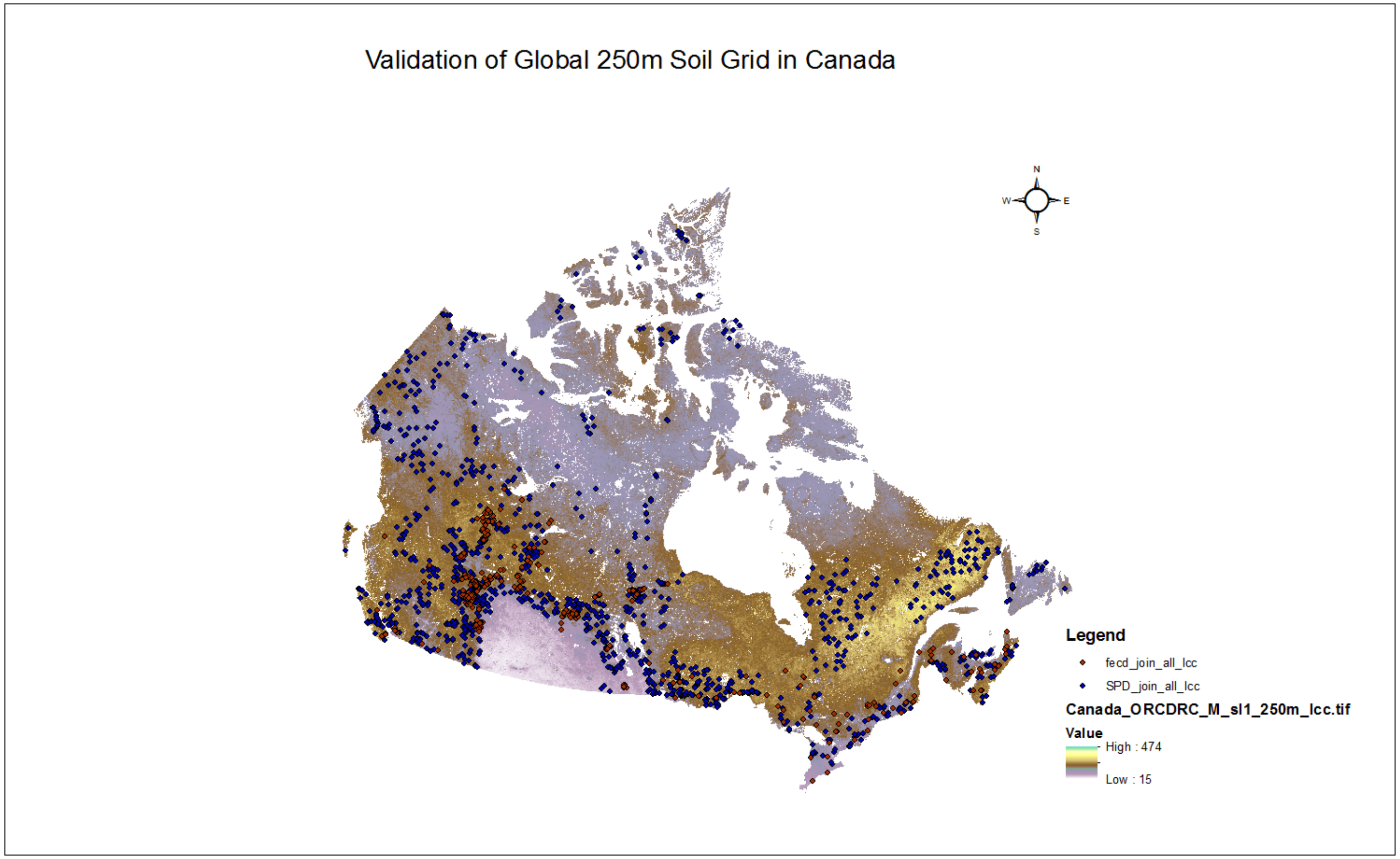
From soil type to soil properties

- Inference from soil properties vs via soil type
- Representative data with residual Kriging

Validation and integrated use

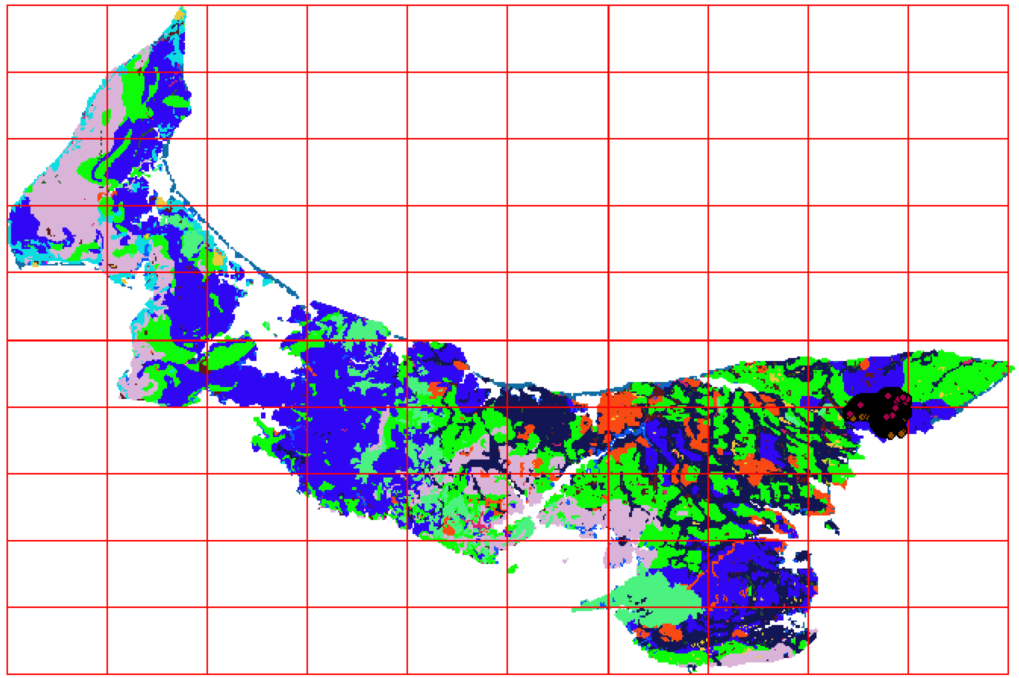
- Necessary field inspection and sampling
- Sediment loading and nutrients management
- BMPs research, design and evaluation

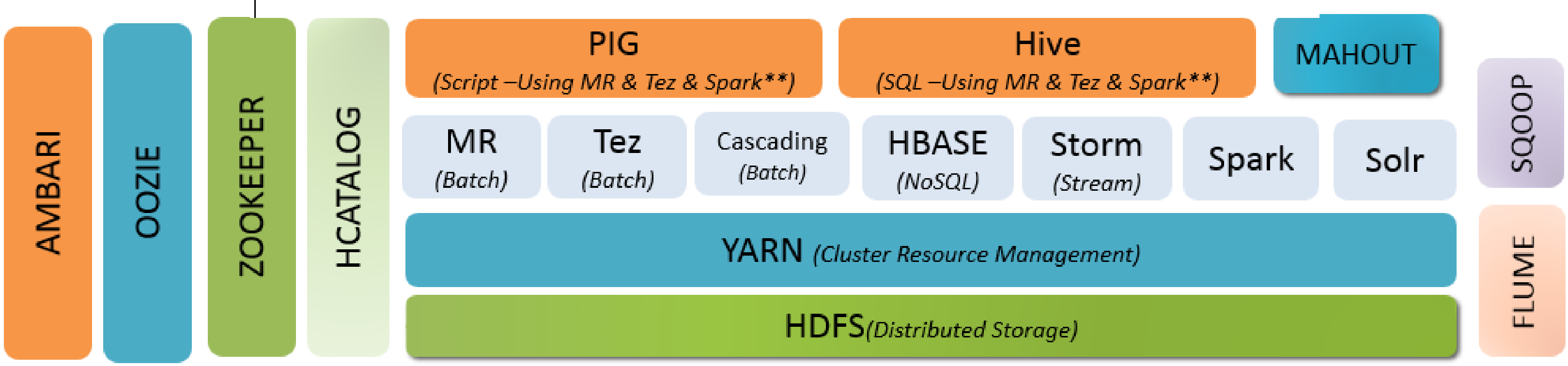
International collaboration and partnership



Hengl, T., J. M. Jesus, G. B . M. Heuvelink, M. R. Gonzalez, M. Kilibarda, A. Blagoti, W. Shangguan, M. N. Wright, X. Geng, B. Bauer-Marschallinger, M. A. Guevara, R. Vargas, R. A. MacMillan, N.H. Batjes, J.G.B. Leenaars, I. Wheeler, S. Mantel, B. Kempen, 2016. SoilGrids250m: global gridded soil information based on Machine Learning

Big data algorithms and advanced computing





Thank you!

Contact: xiaoyuan.geng@agr.gc.ca
Tel. 613-759-1895

R and RGDAL based open environment

```
# imput:
#   1: working directory where contains all the covariates
#   2: training data
#   3: covariate layers, tif format
# output:
#   1: classification result using RF models
#   2: confusion matrix for each iteration derived from testing points.
#   3: training and validation data (shapefile) for each iteration
#   4: variable importance
#   5: Confusion errors derived from RF
………
for (j in 1:10){
 #step 2.1: to randomly sample 70% training points per class to implement RF and the rest to compute confusion matrix
 i=1
 subset.0=subset(points,points$GroupID==levels(points$GroupID)[i])
 training=subset.0[sample(1:nrow(subset.0),ceiling(length(subset.0)*0.7),replace=FALSE),] #spatialPointsDataFrame
 validation=subset(subset.0,!subset.0$ID %in% training$ID)

 for (i in 2:length(levels(points$GroupID))) {
  subset.0=subset(points,points$GroupID==levels(points$GroupID)[i])
  #str(subset.0)
  training.sampled=subset.0[sample(1:nrow(subset.0),ceiling(length(subset.0)*0.7),replace=FALSE),] #spatialPointsDataFrame
  validation.sampled=subset(subset.0,!subset.0$ID %in% training.sampled$ID)
  training=spRbind(training.sampled,training)
  validation=spRbind(validation.sampled,validation)
 }
 …….
 writeOGR( training,dsn=wd,layer=paste("training_",toString(j),sep=""),driver="ESRI Shapefile",overwrite_layer =TRUE )
 writeOGR( validation,dsn=wd,layer=paste("testing_",toString(j),sep=""),driver="ESRI Shapefile",overwrite_layer =TRUE )
 ……
```