Accurate digital mapping of rare soils

Colby Brungard[1,], Budiman Minasny[2]

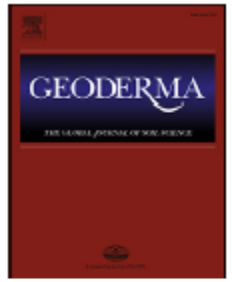
[1]*Department of Plants, Soils, and Climate, Utah State University, United States*

[2]*Faculty of Agriculture & Environment, The University of Sydney, NSW 2006, Australia*

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]

Background:
Soil class prediction from a DSM study, Wyoming, USA

U.S. Soil Taxonomy Subgroup Classes	Pedons[a]	% of total[b]	Producer's accuracy %
Ustic Haplargid	26	46	86
Ustic Torriorthent	21	37	83
Badland	6	10	50
Ustic Paleargid	2	4	0
Ustic Torrifluvent	2	4	0
Total	57	100	

Majority classes = good prediction

Minority/rare classes = poor prediction

[a] Total number of pedons per subgroup class.

[b] Percent of total observations represented by each subgroup class.

Table modified from: Brungard, C.W., Boettinger, J.L., Duniway, M.C., Wills, S.A., Edwards Jr., T.C. 2015. Machine learning for predicting soil classes in three semi-arid landscapes, Geoderma, Volumes 239–240

Problem:

- Rare soil: a soil taxonomic class with few observations
- Accurate modeling of rare soils is needed for
  - Rare plant distribution (e.g., *Schoenocrambe suffrutescens*) (Baker et al., 2016)
  - Vineyard suitability (e.g. Terra Rossa)
  - Targeted land management (e.g., wetland restoration)

- However; rare soils are difficult to accurately predict

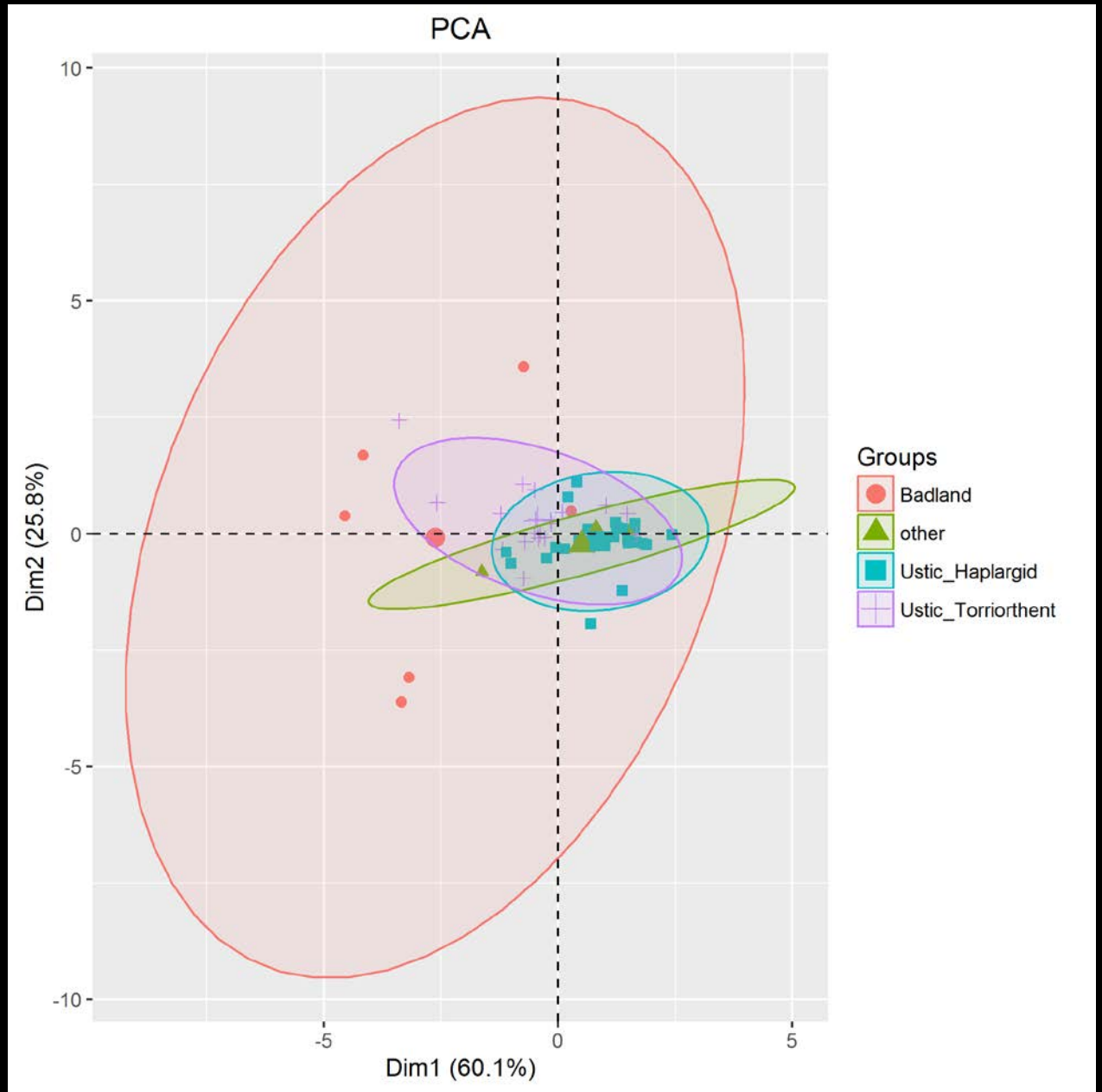
Why are some soils classes rare?

- Endemic soils
  - Soils restricted to a particular geographic area based on a unique combination of soil-forming factors (Bockheim, 2005)
  - Likely small areas, infrequently sampled
- Sampling issues
  - Biased sampling
  - Arbitrary study area boundaries

Why are rare soils difficult to predict?

1) Conceptual reasons
   - Different defining and predicting variables
     - we define soil classes with a set of soil profile variables, but try to model them with environmental variables
   - Soil classes difficult to separate in feature space

Soil Classes Difficult to Separate in Feature Space



Why are rare soils difficult to predict?

2) Practical reasons

- Accurate classification requires many observations
  - A few classes have the majority of the observations
    - Good prediction accuracy
  - Rare soil classes (minority classes) have few observations
    - Poor prediction accuracy

Possible solutions to practical problems

- Decrease the number of classes
  - Combine minor classes with similar classes or into ‘other’ class
    - Can exclude rare classes
    - Should be done for classes with fewer observations than covariates
- Use a weighting/cost scheme
  - Useful, but cannot produce class probability predictions (often most interesting)
- Increase number of observations in rare classes

How to increase observation numbers in rare classes?

- Expert knowledge
  - Case-based reasoning (Shi et al., 2004)
  - Expert knowledge used to generate synthetic observations
- Data-driven approach
  - Synthetic minority-class oversampling (e.g., SMOTE)
  - Synthetic observations generated by similarity to existing observations
  - Currently only for two-class problems (soil datasets are multi-class)

Synthetic minority-class oversampling for multi-class soil datasets - SIMONA

- Goal: Generate equal number of observations in each soil class
  - Synthetic samples similar to field samples in geographic and feature space
- An alternative algorithm from Abdi and Hashemi (2015) which only sampled the feature space

SIMONA - Synthetic minority class oversampling for multiclass soil datasets

1. Calculate weights for each minority class observation
   - Weights based on *n* surrounding points with matching class label. Weight higher for more surrounding observations with matching class labels
2. Randomly choose one minority class observation by weight
3. Extract all covariate values around observation in a *w*-meter buffer
4. Calculate similarity between observation and covariates with Gower's similarity index (including categorical variables)
5. Randomly choose one of the *p* most similar samples
6. Add to synthetic sample set
7. Repeat until number of original + synthetic samples = the number of observations in the majority class

Preliminary Case Study

- Area: 296 km^2 – North-central Wyoming, USA
- Elevation: ~ 1500 m
- Geology: mudstone, sandstone, conglomerate, limestone, shale and coal
- Climate: cool and dry
- Land use: oil, gas, coal production

Photo courtesy of elifino57: https://ssl.panoramio.com/user/6155411

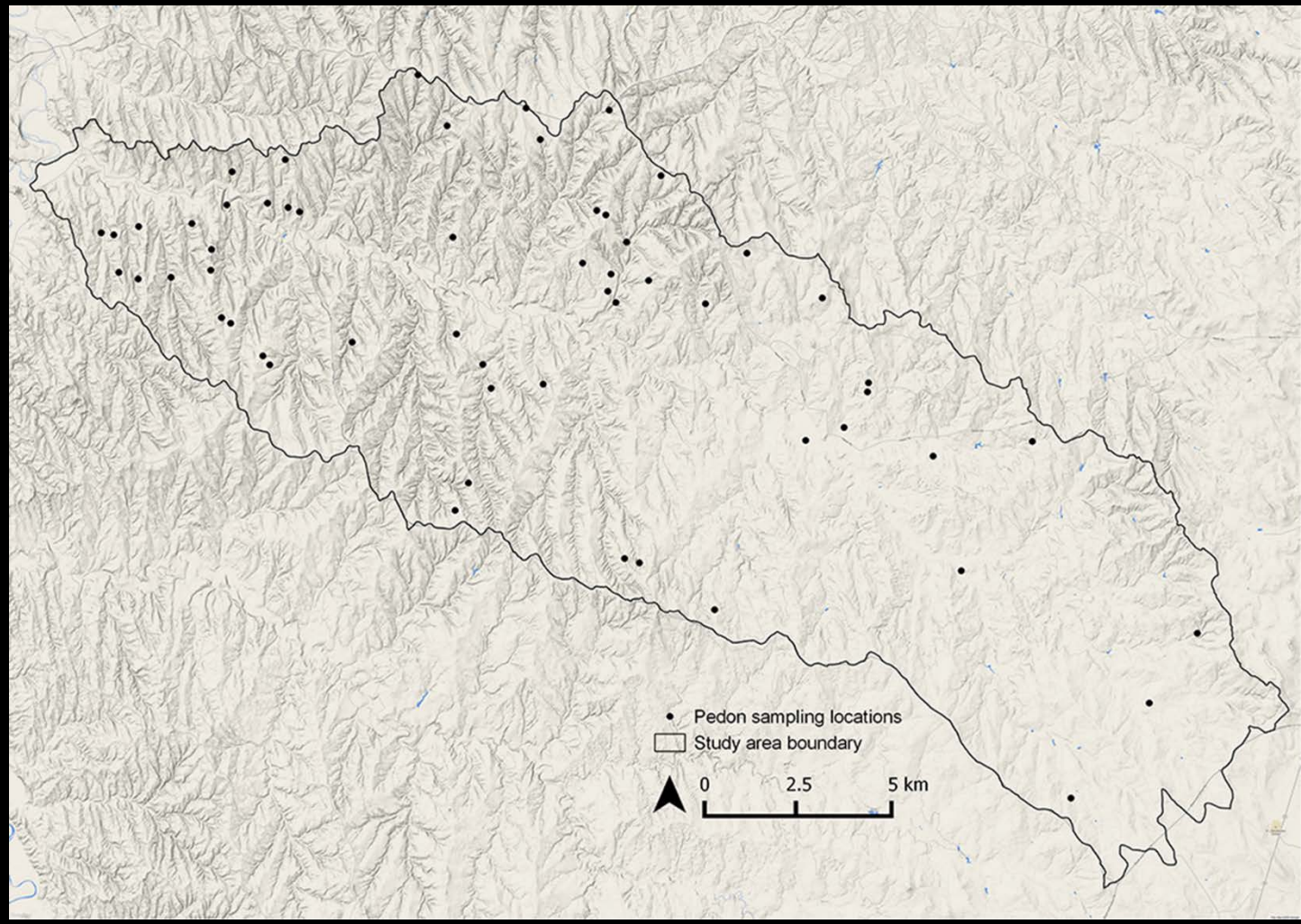
Sampling

- 57 pedons
- 5 subgroup* classes
  - Imbalanced - 83% of observations in two classes

U.S. Soil Taxonomy Subgroup Classes	Pedons[a]	% of total[b]
Ustic Haplargid	26	46
Ustic Torriorthent	21	37
Badland	6	10
Ustic Paleargid	2	4
Ustic Torrifluvent	2	4
Total	57	100

[a] Total number of pedons per subgroup class.

[b] Percent of total observations represented by each subgroup class.



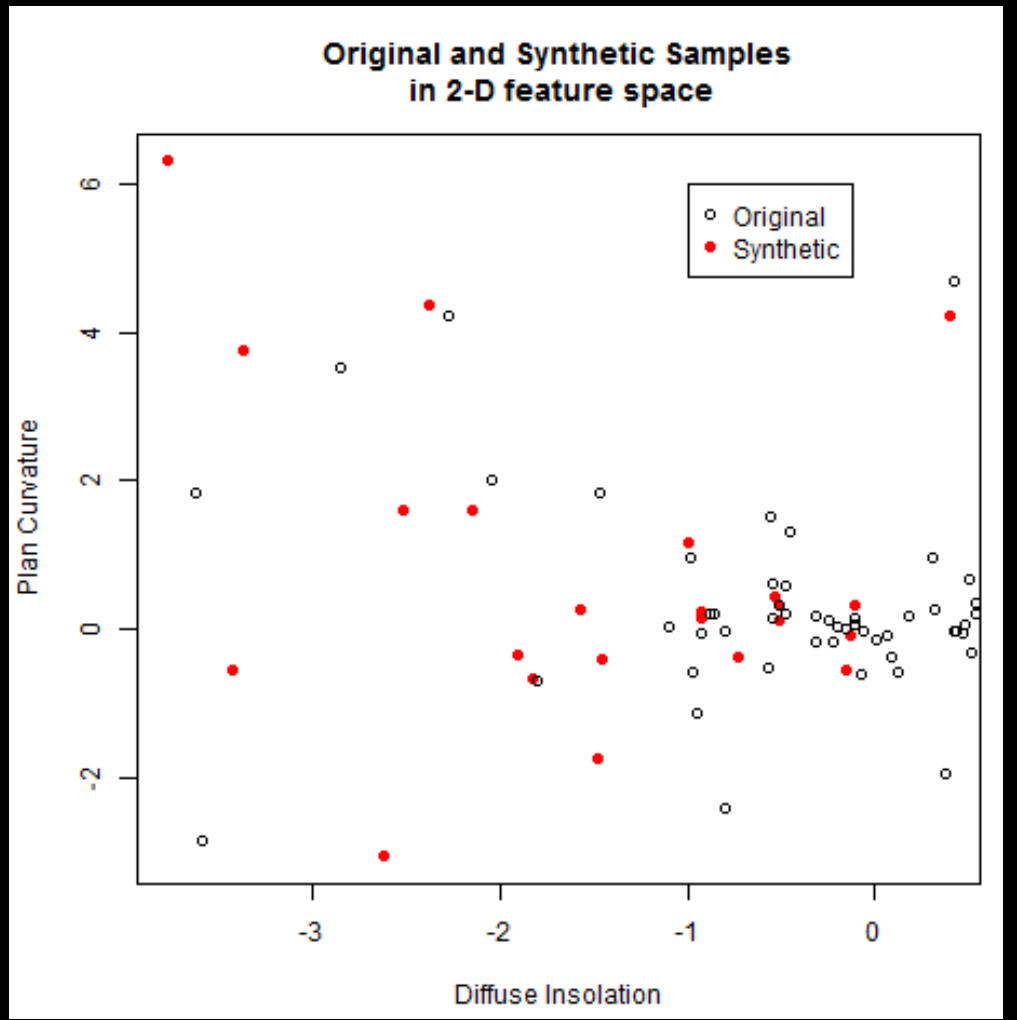
* Subgroups according to US Soil Taxonomy

SIMONA

- **Covariates:** Plan curvature, Diffuse Insolation, Landsat band ratio 5/2, Catchment Slope

U.S. Soil Taxonomy Subgroup Classes	Original	Synthetic	Original + Synthetic
Ustic Haplargid	26	0	26
Ustic Torriorthent	21	5	26
Badland	6	20	26
Minor[a]	4	22	26
Total	57		104

[a] Combined 3 minor classes with < = 2 observations



Modelling

- CART and Neural Network models
  - Original samples
  - Original + Synthetic samples
- Case weights
  - Original samples assigned high class weight (1)
  - Synthetic samples assigned low case weight (0.25)
- Model accuracy
  - Bootstrap sampling repeated 10 times
    - Overall accuracy
    - Kappa
    - Confusion matrices

Results: Model Accuracy

Overall Accuracy*

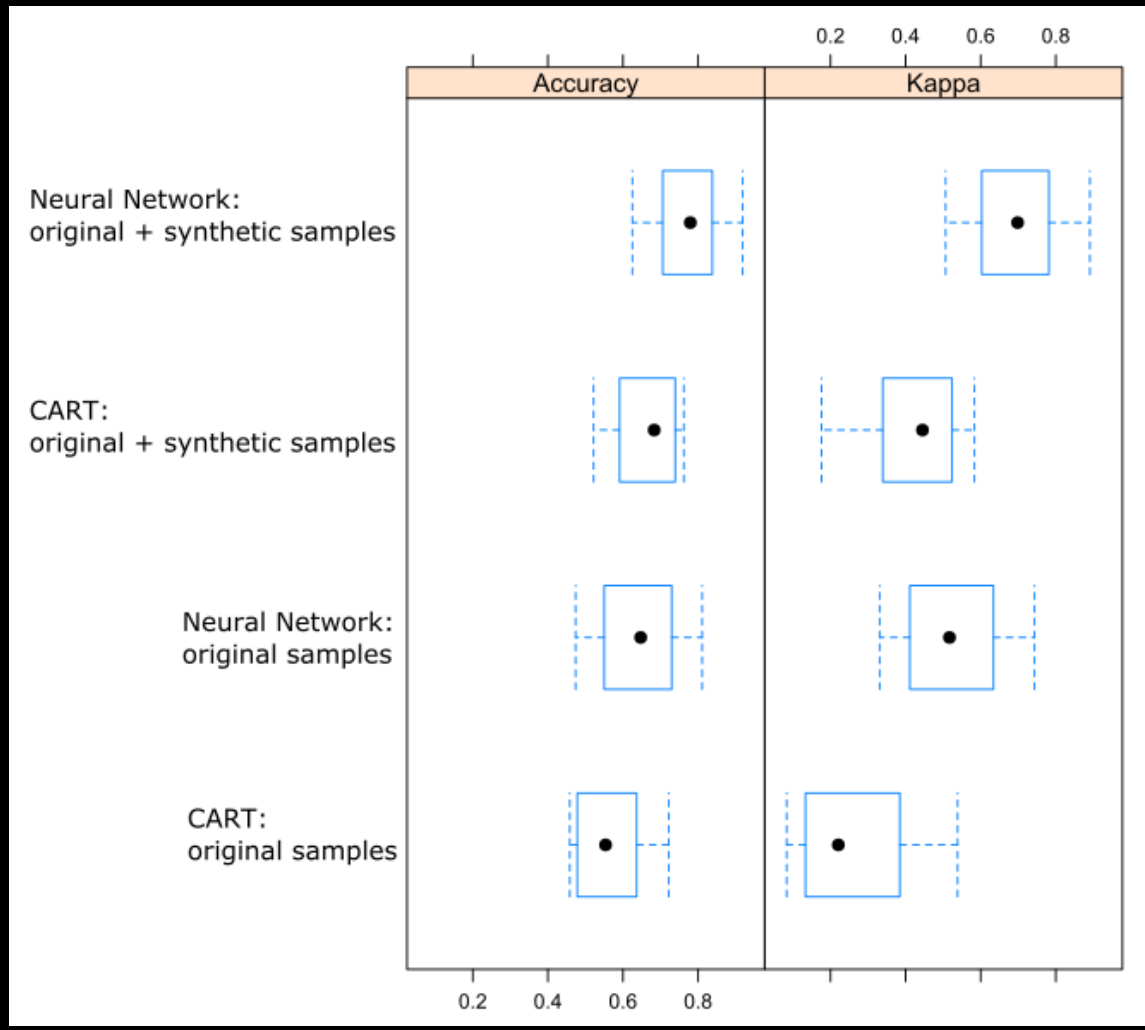
Model	Original Samples	Original + Synthetic
CART	0.57	0.64
Neural Net	0.67	0.77

* mean of bootstrap sampling repeated 10 times

Kappa*

Model	Original Samples	Original + Synthetic
CART	0.26	0.52
Neural Net	0.43	0.70

* mean of bootstrap sampling repeated 10 times



Results

- Confusion Matrices from bootstrap sampling (neural network model only)
  - Accurate models have large diagonal and small off-diagonal values

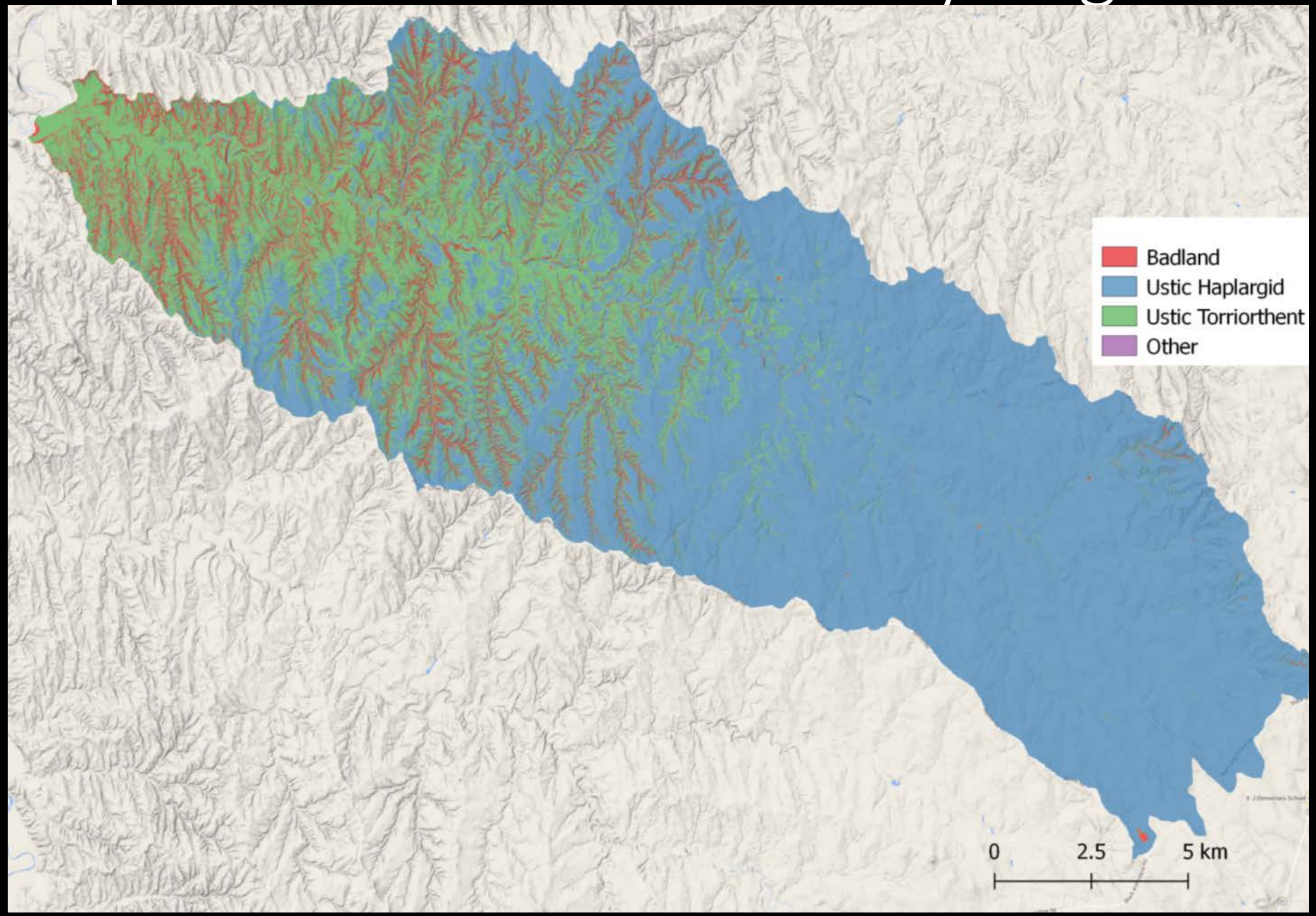
Original samples

		Actual			
		Badland	Ustic Haplargid	Ustic Torriorthent	Minor
Predicted	Badland	1	0	0	1
	Ustic Haplargid	3	40	7	4
	Ustic Torriorthent	5	11	26	3
	Minor	0	0	0	0

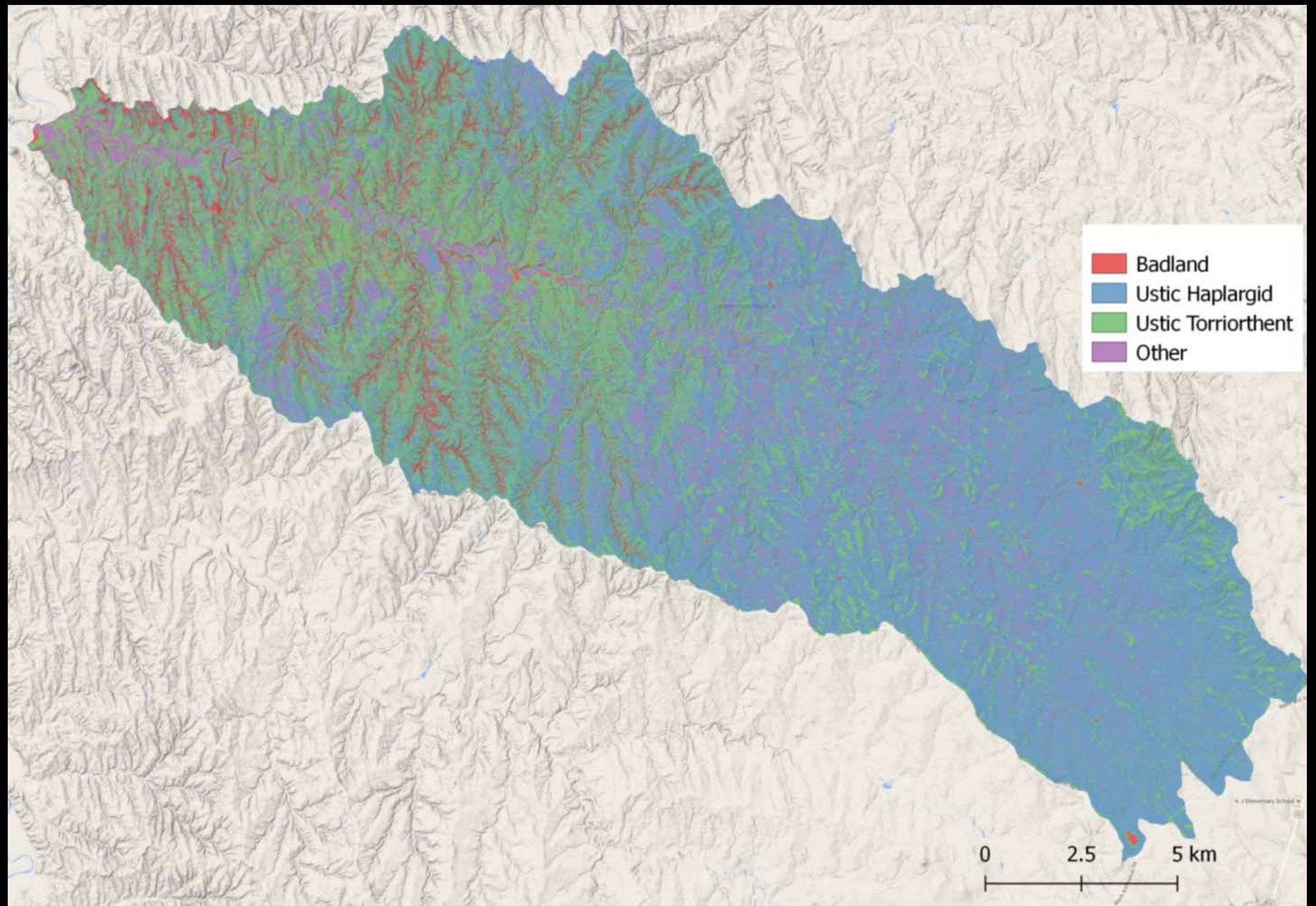
Original + synthetic

		Actual			
		Badland	Ustic Haplargid	Ustic Torriorthent	Minor
Predicted	Badland	21	0	3	0
	Ustic Haplargid	1	17	5	2
	Ustic Torriorthent	1	5	18	1
	Minor	0	3	2	21

Results: Spatial Prediction with only original samples



Results: Spatial Prediction with synthetic samples



Conclusions

- SIMONA (Synthetic minority class oversampling for multiclass soil datasets) is promising for increasing rare soil class prediction accuracy
  - Need to further test with independent validation data set
- May provide a means to increase prediction accuracy when additional field sampling is limited

References

- Abdi, L., and S. Hashemi. 2015. To combat multi-class imbalanced problems by means of over-sampling and boosting techniques. Soft Computing, 19:12. pp 3369-3385.
- Baker, J.B., Fonnesbeck, B.B., and J.L., Boettinger. 2016. Modeling Rare Endemic Shrub Habitat in the Uinta Basin Using Soil, Spectral, and Topographic Data. Soil Science Society of America Journal, 80. pp. 395–408
- Bockheim, J.G. 2005. Soil endemism and its relation to soil formation theory. Geoderma, 129. pp. 109–124
- Shi, X., Zhu, A-X., Burt, J.E., Qi, F., and D. Simonson. 2004. A Case-based Reasoning Approach to Fuzzy Soil Mapping. Soil Sci. Soc. Am. J. 68. pp. 885–894