



Agriculture and
Agri-Food Canada

Agriculture et
Agroalimentaire Canada



Use of Data-Driven and Knowledge-Driven Methods In Disaggregating Soil Survey Maps, Okanagan Basin, Southern British Columbia.

Scott Smith, Eve Flager, Bahram Daneshfar, Grace Frank
and Chuck Bulmer

Spatial Predictions Symposium, CSSS-ASSS annual meeting, San Antonio, TX, Oct 18, 2011

Canada

Presentation Outline

- Background – objectives, approach, team
- Method- techniques, predictors, knowledge base
- Results – hardened map, difference from legacy map
- Accuracy assessment
- Discussion –scaling-up issues, use of legacy maps, lessons learned.

Background - introduction

Context

- Okanagan Basin is a semi-arid region with growing demand for a limited water supply

Objective

- Disaggregate existing legacy soil maps to provide raster-based soil attributes to modeling effort

Approach

Select a sub-watershed to test several methods of disaggregation

Assemble multidisciplinary team to undertake the work



Background – study area

- 75,000 ha sub-watershed of the Okanagan River, a tributary to the Columbia River
- Elevation range from 350m asl at Okanagan Lake to 2000 m on highest ridges
- Uplands are commercial forest land, valley floor is irrigated horticulture
- Annual precipitation ranges from 300 mm to 900mm
- Mean annual temperature varies from 11°C to 3°C
- Xerolls to Cryods

Methods: Two main groups of methods for predictive mapping

- **Knowledge-Driven methods**

Relationships between the target variable and predictor variables are defined based on expert opinion.

- **Data-Driven methods (quantitative empirical modeling)**

Relationships between the target variable and predictor variables are quantified by the method, based on the data and then used for prediction.

Methods

- Fuzzy logic inference engine - ARC SIE (modified)
- Expert knowledge rule set (Scott's folly)
- Logistic Regression
- Weights of Evidence (WOFE)
- Hybrid method using contrast values from WOFE to define rule curves in ARC SIE

Methods - predictors

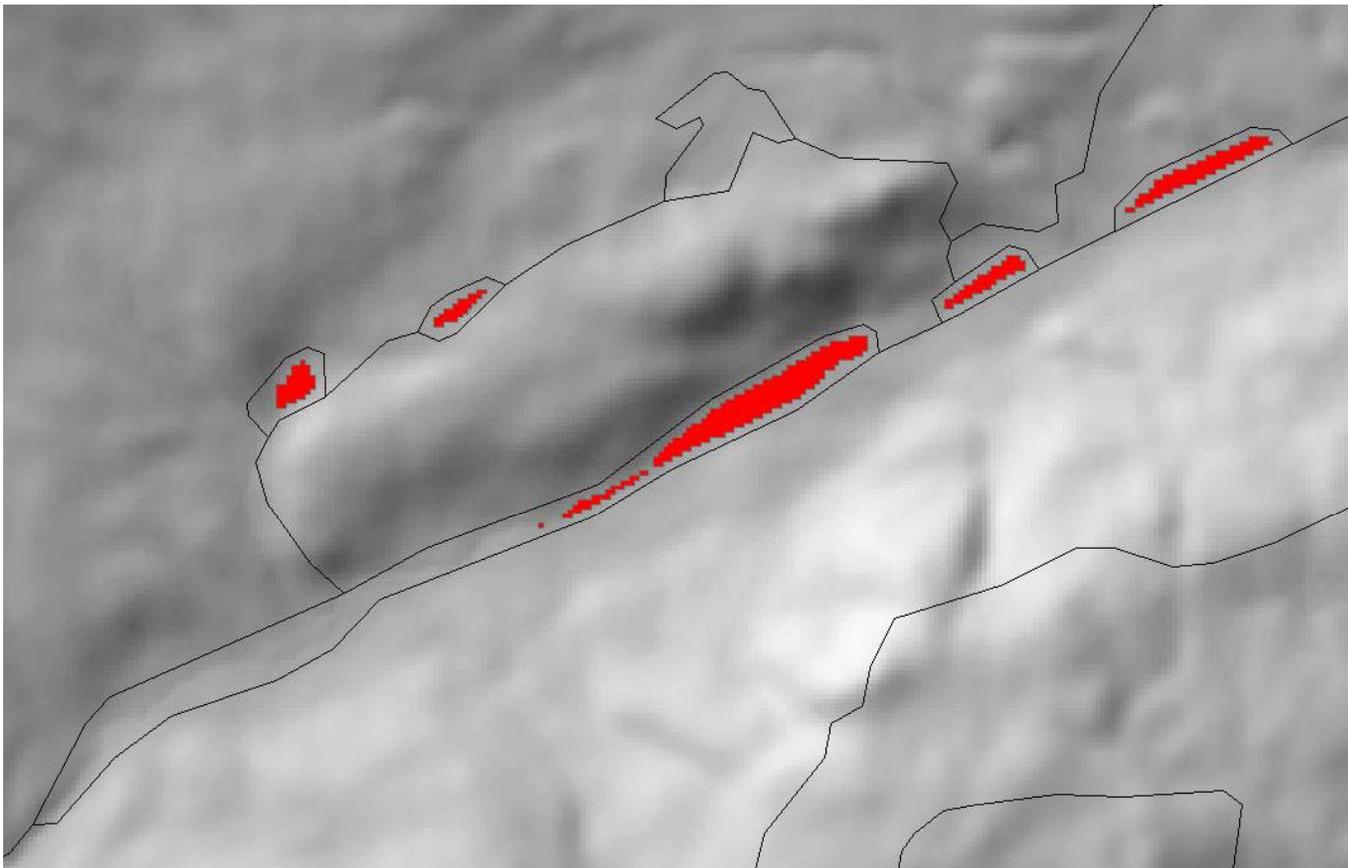
- Target variable – 23 soil series mapped in study area

Feature Type	Dataset	Data Source
Topographic	<ul style="list-style-type: none"> • Elevation • Slope Percent • Aspect • Relative Heights Slope Position (RHSP) • LandMapR Landform Classes • Topographic Position Index (TPI) • Topographic Position Index (TPI) Landform Classification • Stream Network 	<ul style="list-style-type: none"> • CDED DEM • Terrain Resource Inventory Mapping (TRIM) Stream Network
Vegetation/Climate	<ul style="list-style-type: none"> • BEC Zones and Sub-Zones • CIRCA 2000 Land Cover 	<ul style="list-style-type: none"> • 1:50,000 polygons downloaded from Ministry of Environment • 30m CIRCA 2000 raster downloaded from Geogratis
Parent Material	<ul style="list-style-type: none"> • Surficial Material 	<ul style="list-style-type: none"> • 1:20,000 polygons provided by Ministry of Environment

Methods - Soil Polygon Refinement for Training Data Sampling

- Used soil polygons from the soil polygon coverage
- Used a polygon for a particular soil series if it consisted of at least 70% of that soil series
- Inverse buffered the polygons by 50m
- Refined the polygon using BEC zone, parent material, slope, and elevation
- For each soil series, 200 random training points were generated from the refined polygons for WOFE calculations

CXZ



Methods

Weights of Evidence Terms

- **Weights for patterns**
 - **W+** weight for inside the pattern
 - **W-** Weight for outside the pattern
 - **0** Weights for areas of no data
- **Contrast** : a measure of the spatial association of pattern with sites
- **Studentized Contrast**: a measure of the significance of the contrast

$$\text{Contrast: } C = W^+ - W^-$$

Extracting C values for TPI

Table

Bov_TPI200

OID	CLASS	AREA_SQ_KM	AREA_UNITS	NO_POINTS	WPLUS	S_WPLUS	WMINUS	S_WMINUS	CONTRAST	S_CONTRAST	STUD_CNT	GEN_CLASS	WEIGHT	W_STD
0	1	18.1225	1812.25	0	0	0	0	0	0	0	0	1	0	0
1	2	131.8475	13184.75	0	0	0	0	0	0	0	0	2	0	0
2	3	335.26625	33526.625	3	-3.3855	0.5774	0.57	0.0714	-3.9555	0.5818	-6.799	3	0	0
3	4	217.50375	21750.375	71	0.2145	0.1189	-0.101	0.0882	0.3155	0.148	2.1316	4	0	0
4	5	56.036875	5603.6875	126	2.1638	0.0901	-0.9191	0.1163	3.0829	0.1471	20.9537	5	0	0

(0 out of 5 Selected)

Table

tpi200_zonal_stats

Rowid	VALUE	COUNT	AREA	MIN	MAX	RANGE	MEAN	STD	SUM
1	1	28996	18122500	-44.22522	-12.085083	32.140137	-18.083647	5.595839	-524353.38
2	2	210956	131847500	-12.084595	-3.406738	8.677856	-5.955449	2.129738	-1256337.6
3	3	536426	335266240	-3.406677	1.414307	4.820984	-0.826904	1.327076	-443572.63
4	4	348006	217503740	1.414368	7.199463	5.785095	3.623603	1.575064	1261035.6
5	5	89659	56036876	7.199707	37.733154	30.533447	10.737958	3.56502	962754.63

(1 out of 5 Selected)

Inference

Attribute Rule

Limiting Feature

Feature	Limiting
rhsp	<input checked="" type="checkbox"/>
slp_unf	<input type="checkbox"/>
tpi300200	<input type="checkbox"/>
tpi200	<input checked="" type="checkbox"/>
wetarea_hires	<input type="checkbox"/>
landcovertr	<input type="checkbox"/>
essforxv	<input checked="" type="checkbox"/>

Membership Value: 1

Weight Features

Spatial Setting >>

More Options

Do This

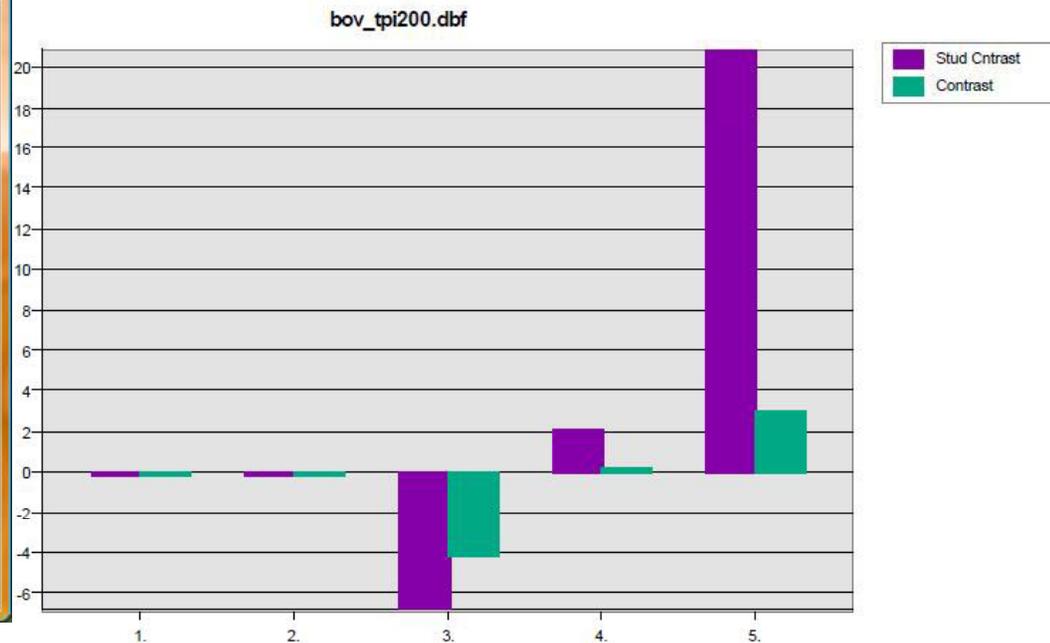
Do Batch

v1 7.2 v2 37.7

w1 5.8 w2 0

r1 2 r2 2

View Type: Data Range



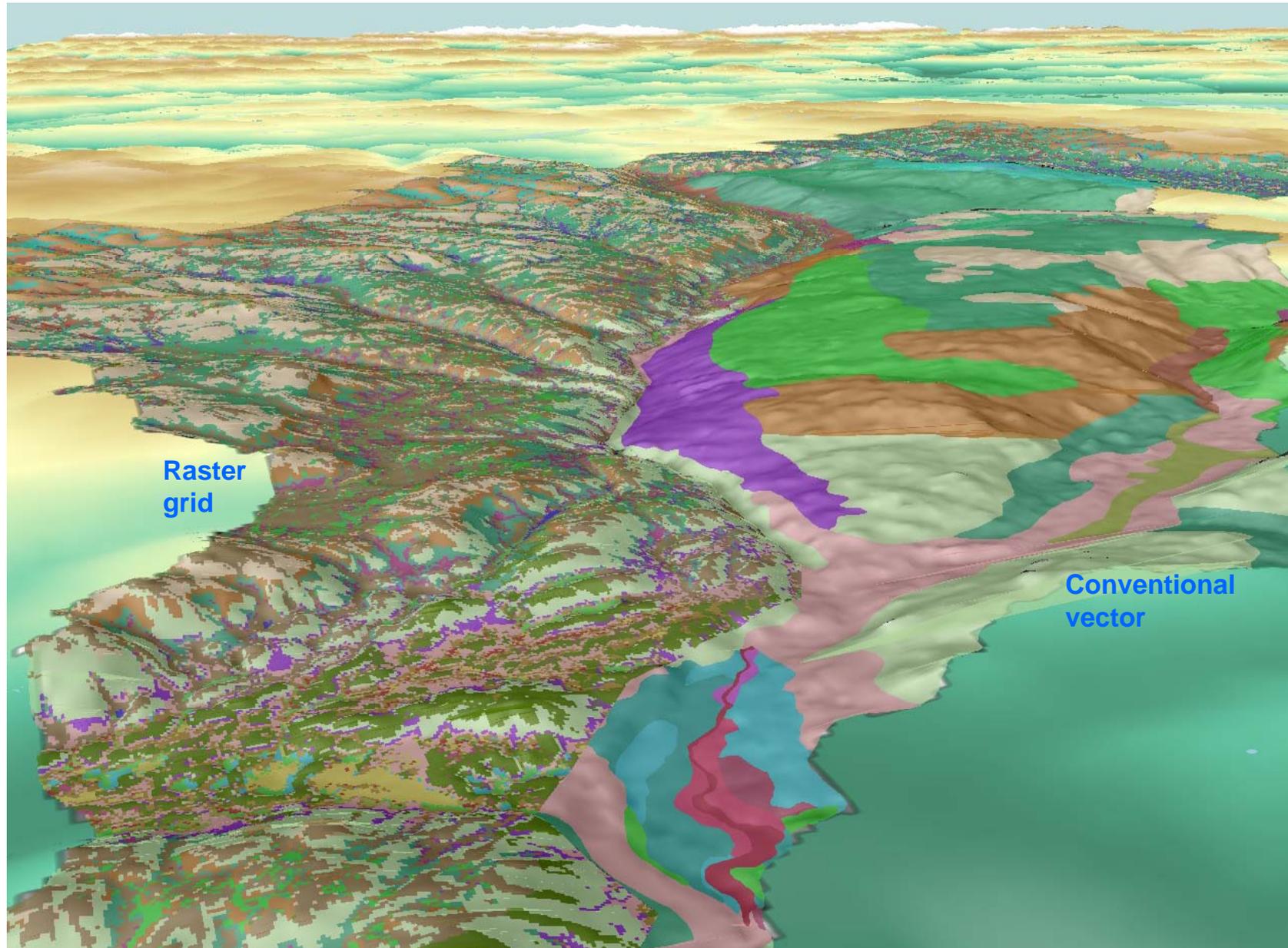
Methods – Validation

- Stratified study area by ecological zone then selected forestry roads within each zone, field sampled (identified soil series) at landscape positions along the right of way whenever change occurred.
- Two years of field work, total of 300 field checks, half to build the expert knowledge and rule set, half to validate the predictions
- Represents more field work than was used to create the legacy map

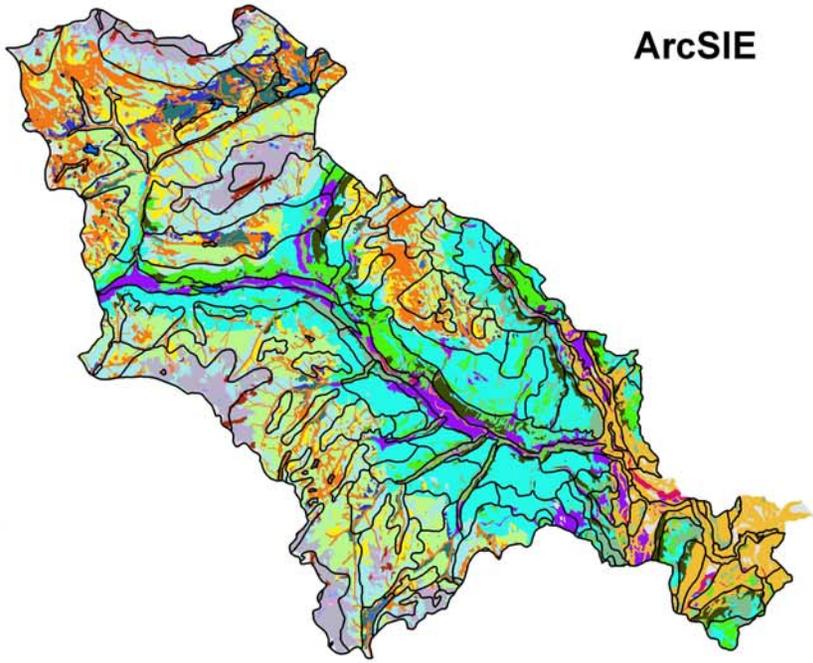
Methods – Run analyses

- Ran 5 methods with various interventions and modifications related to:
 - input data to ARC-SIE run,
 - use of selected predictors used in the Hybrid method,
 - setting probability and certainty limits in the WOFE and LG

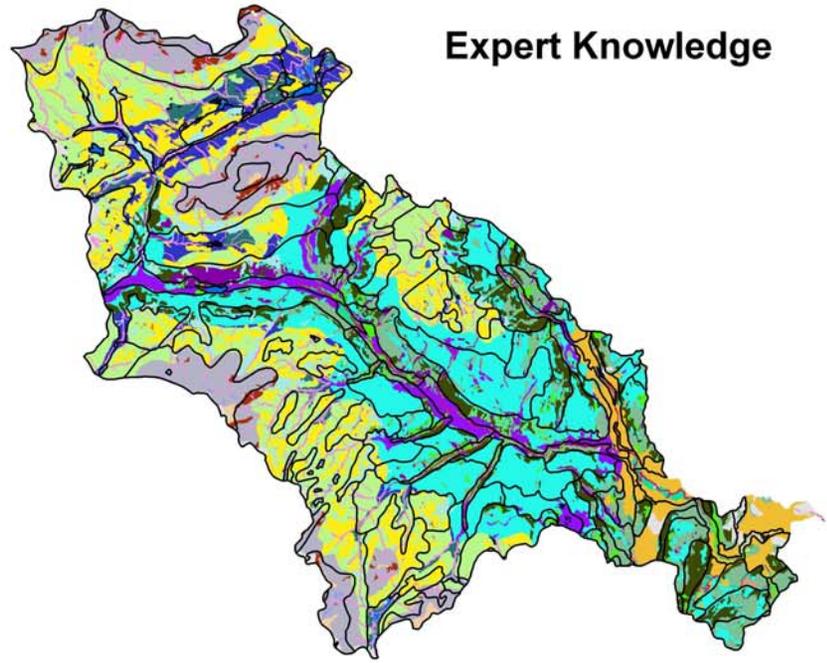
Results - Digital Soil Map vs Polygon Soil Map



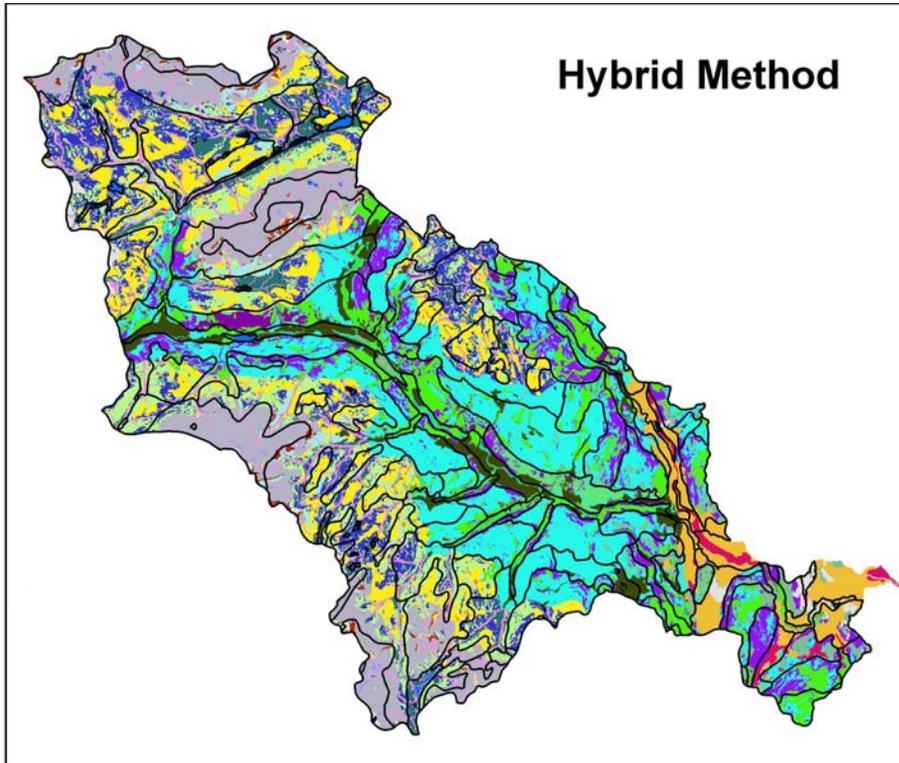
ArcSIE



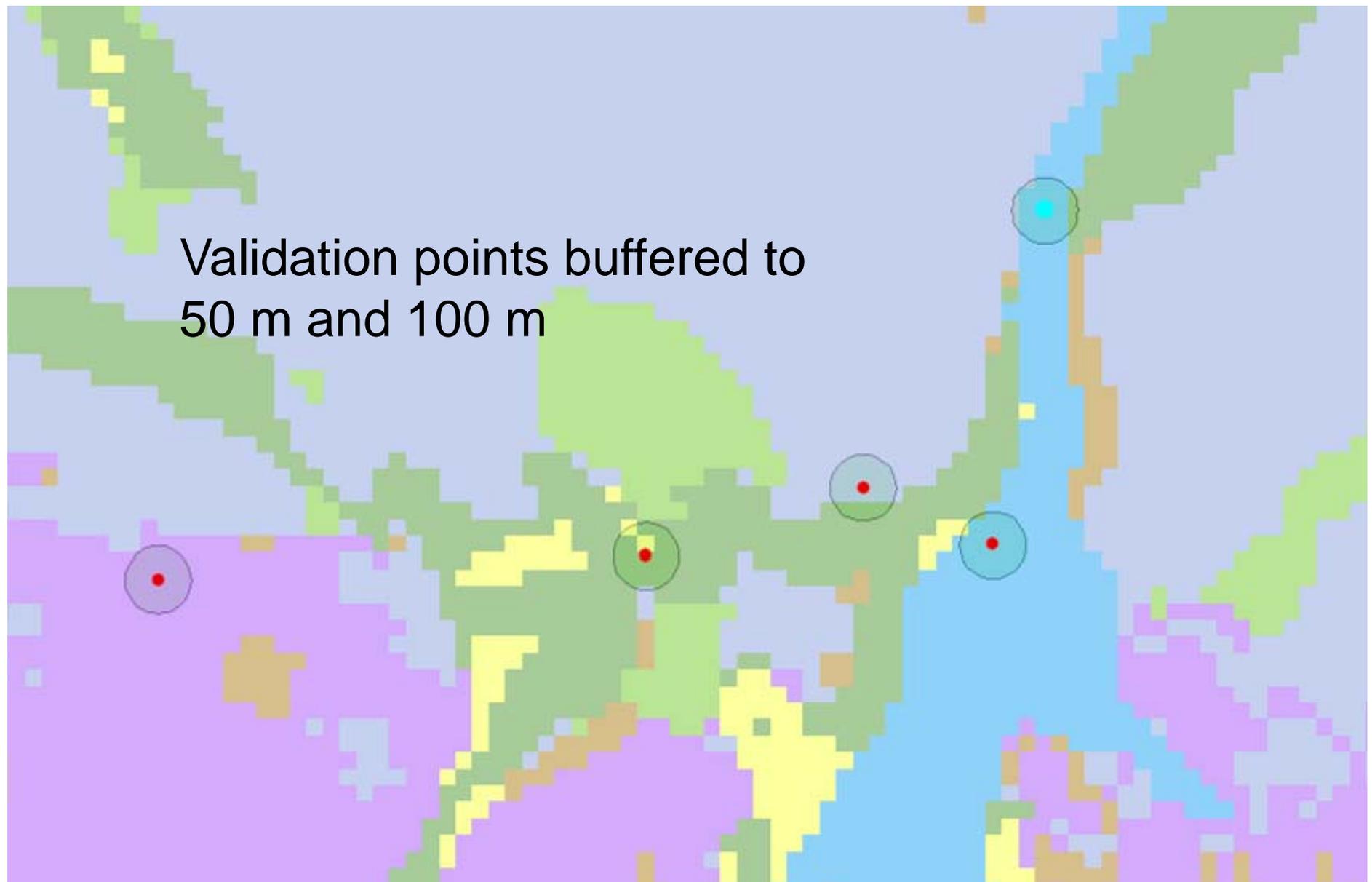
Expert Knowledge



Hybrid Method



Results - validation

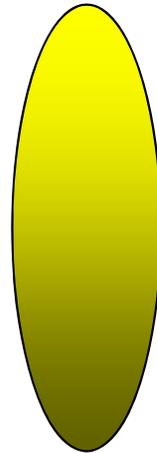


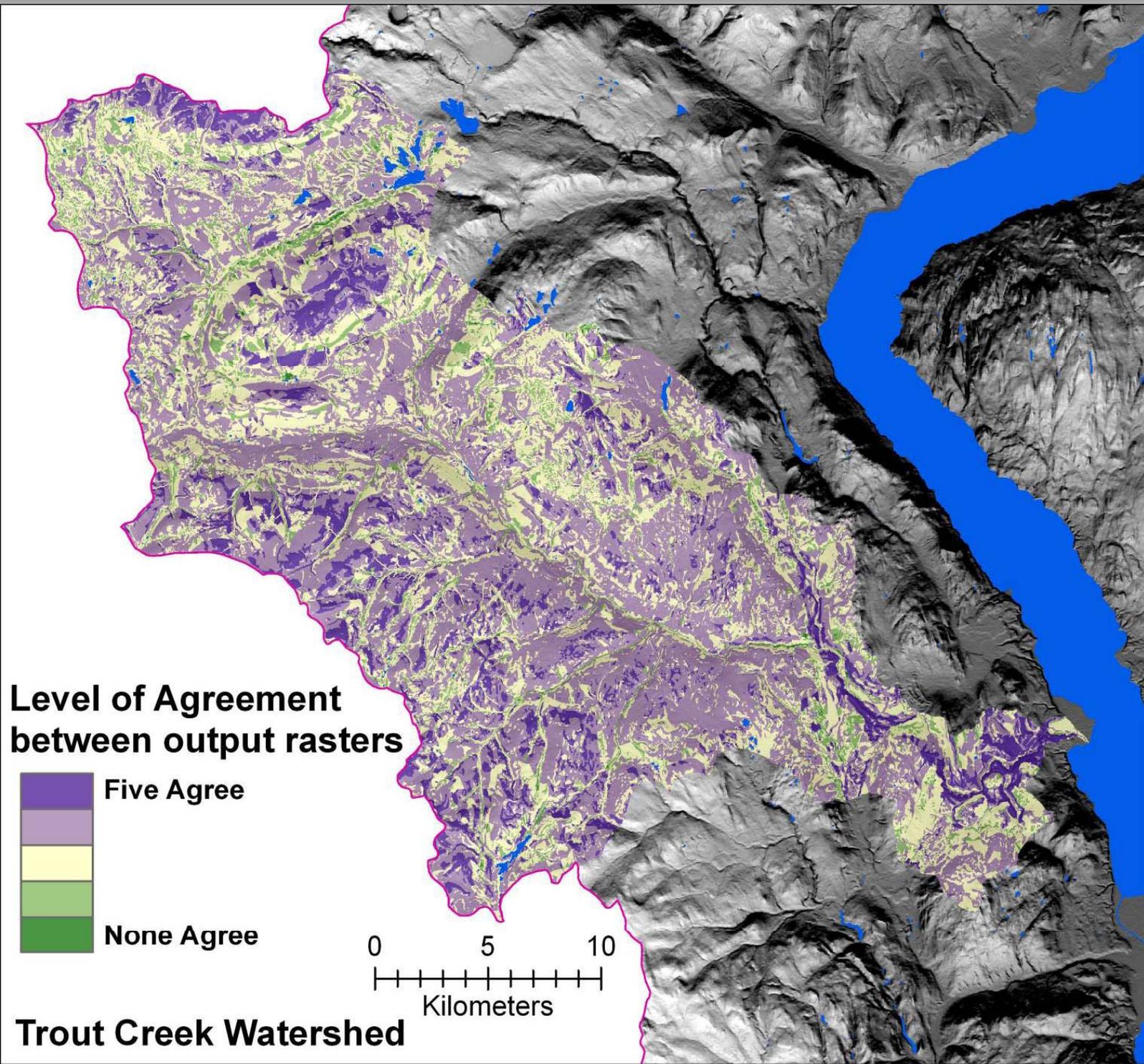
Results – % agreement observed vs predicted

Buffer	ARC SIE	Ex Rules	Hybrid	WOFE	LR
--------	---------	----------	--------	------	----

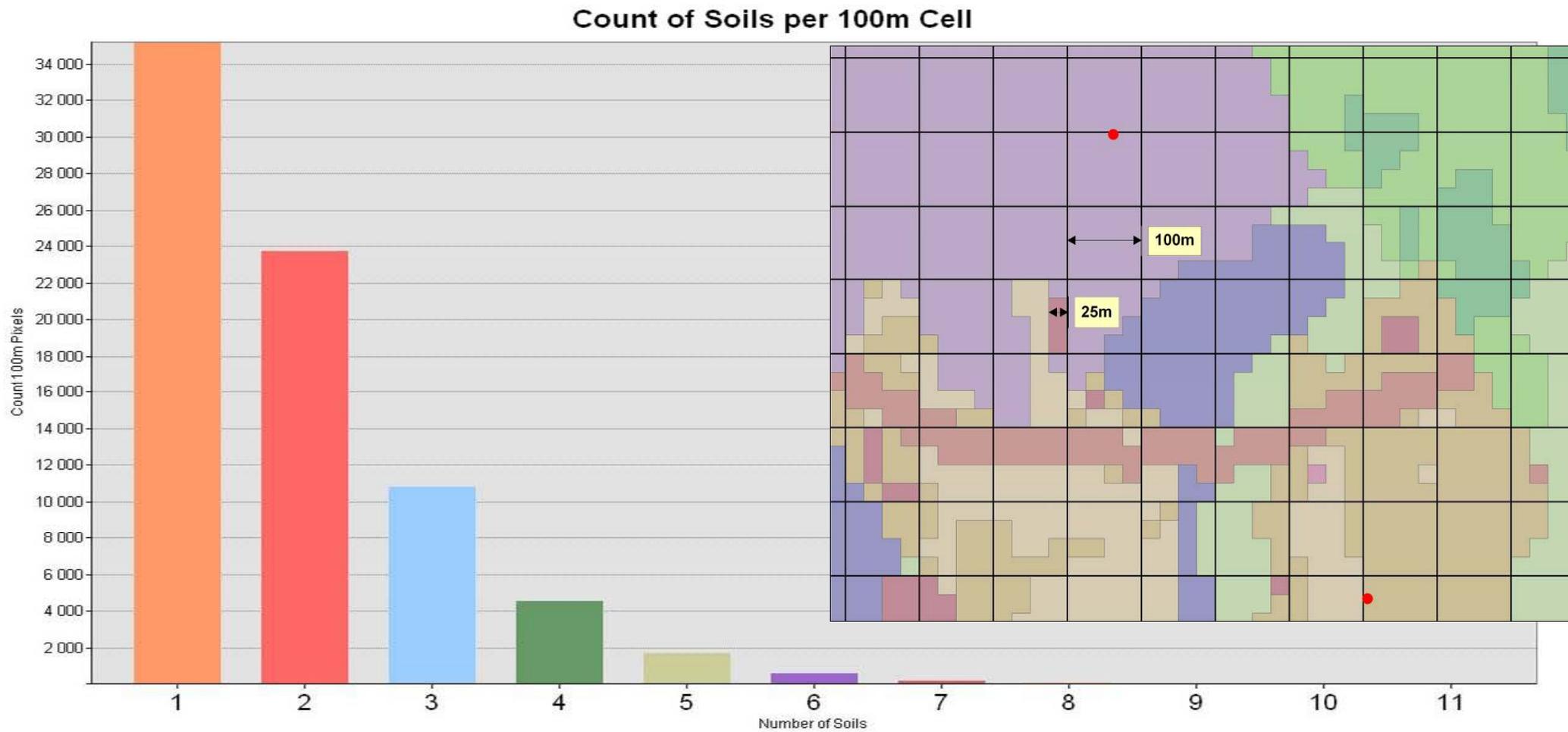
50 m	46	50		37	38
------	----	----	--	----	----

100 m	60	56		48	44
-------	----	----	--	----	----





Discussion – scaling up to 100 m cell



Discussion and Conclusions

- Use of legacy maps –problematic (correlation issues, difficult to identify taxonomic units in field)
- Keep it simple rule: in this case our simplest approach was not our best method, now where do we go?
- Lessons learned:
 - Predictors – a few good predictors better than many weak predictors
 - Expert knowledge – if exists for an area, then cost effective to capture, if not, very time-consuming to develop
 - No one method that satisfied all needs, but some cross over of data driven and knowledge driven systems produced our best results.



Thank you for your attention

Canada 