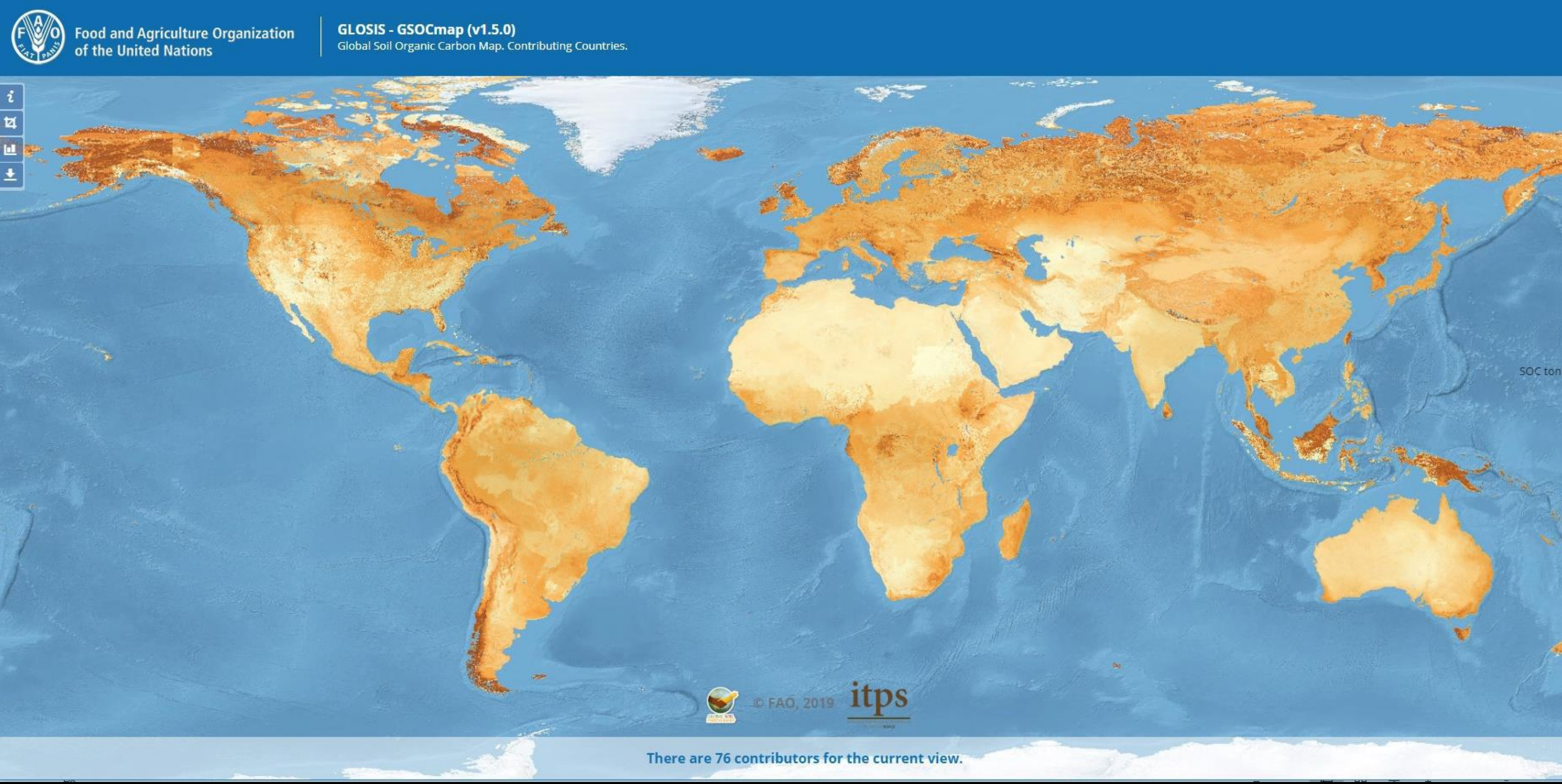
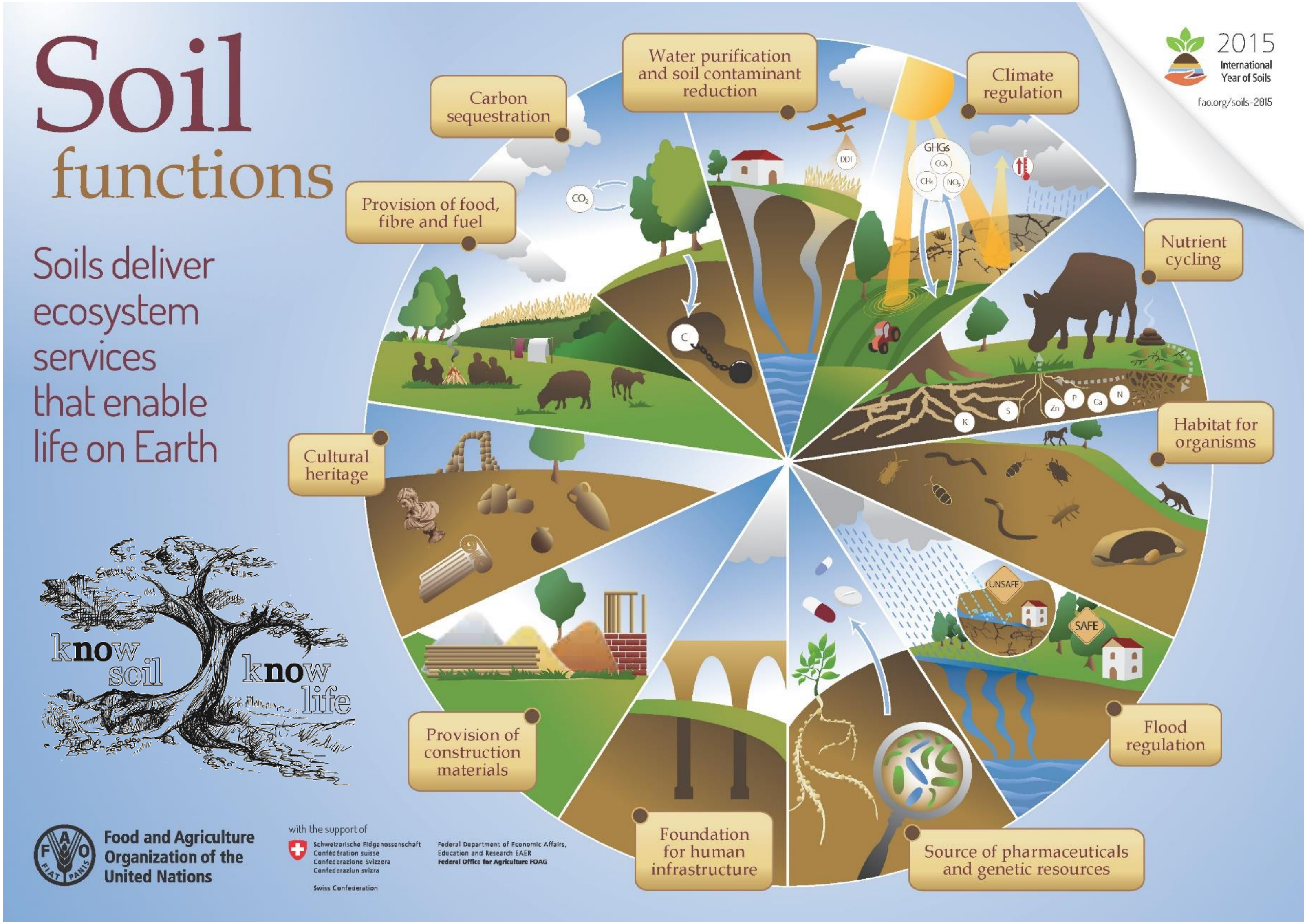


Digital Soil Mapping [From 0 TO 100]

Ruhollah Taghizadeh
Soil Science and Geomorphology

E-mail: *ruhollah.taghizadeh-mehrjardi@mnf.uni-tuebingen.de*

Why Soil Matters



Soil Degradation

Soil erosion

Land clearing

Soil salinization

Soil compaction

Soil maps

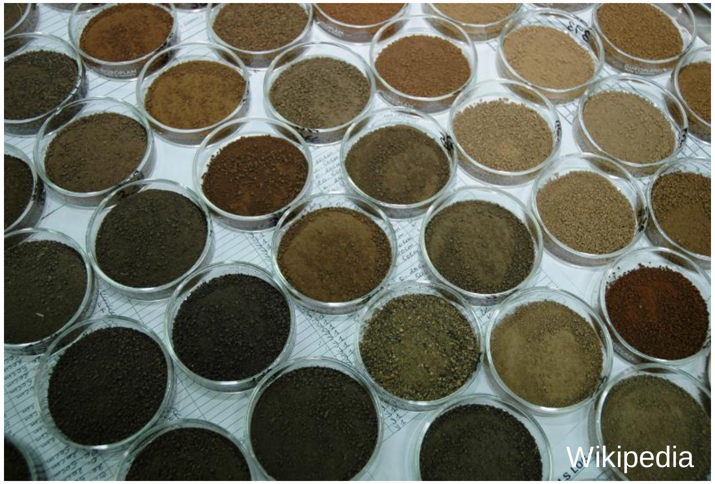
are an essential tool in achieving sustainable use of the land

Soil Mapping Steps

(1) study area

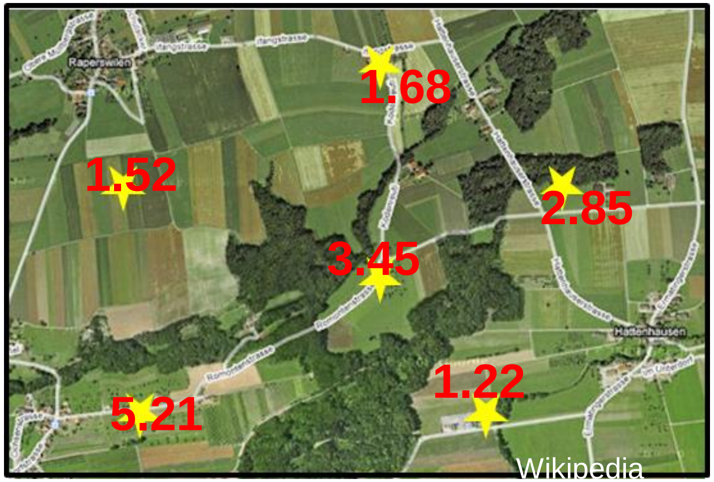
(2) sampling plan

(3) field data collection



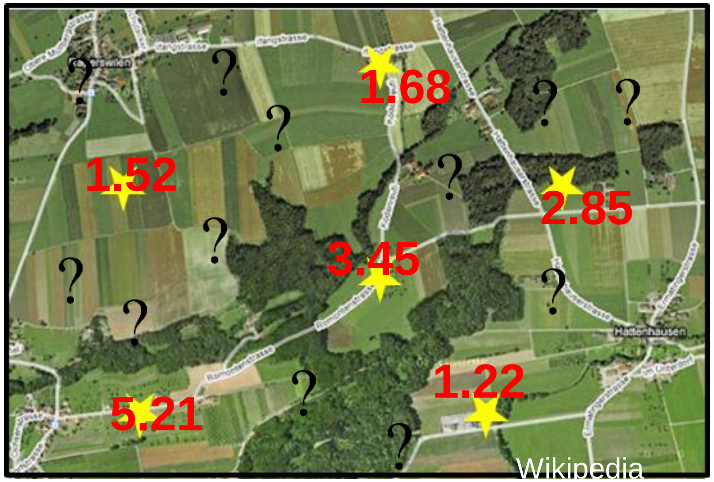
(4) soil analysis

Soil Mapping Steps



(5) soil point data

Soil Mapping Steps

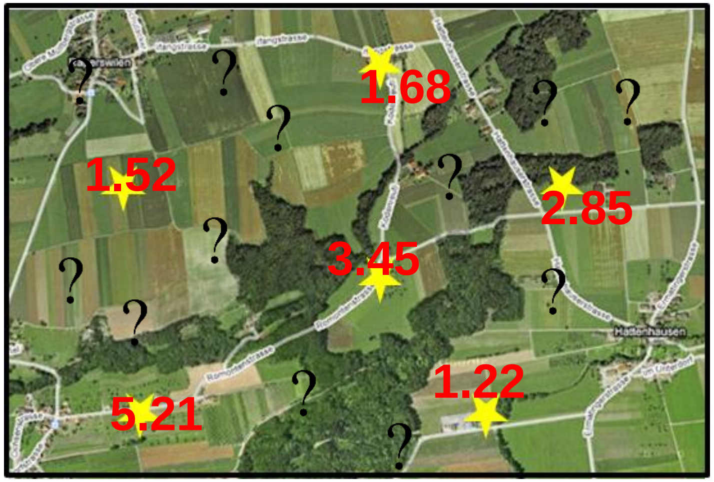


(5) soil point data

(6) create soil map

Dent & Young, 1981; Zhu, et al. 2001

Soil Mapping Steps

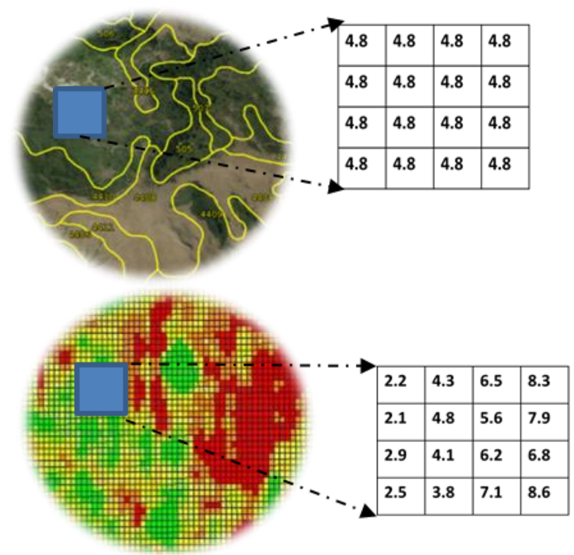


(5) soil point data

(6) create soil map

Soil mapping

- ➢ **C**onventional **S**oil **M**apping
- ➢ **D**igital **S**oil **M**apping



4.8	4.8	4.8	4.8
4.8	4.8	4.8	4.8
4.8	4.8	4.8	4.8
4.8	4.8	4.8	4.8

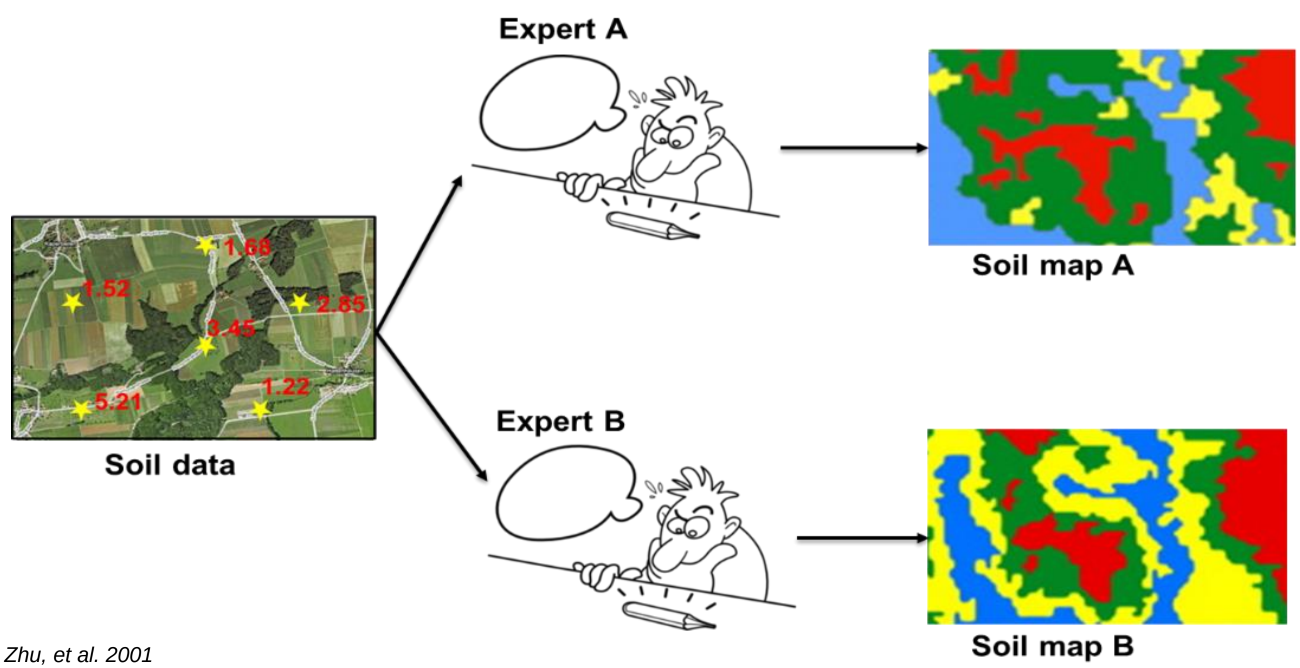
2.2	4.3	6.5	8.3
2.1	4.8	5.6	7.9
2.9	4.1	6.2	6.8
2.5	3.8	7.1	8.6

Dent & Young, 1981; Zhu, et al. 2001

Conventional Soil Maps

Limitations:

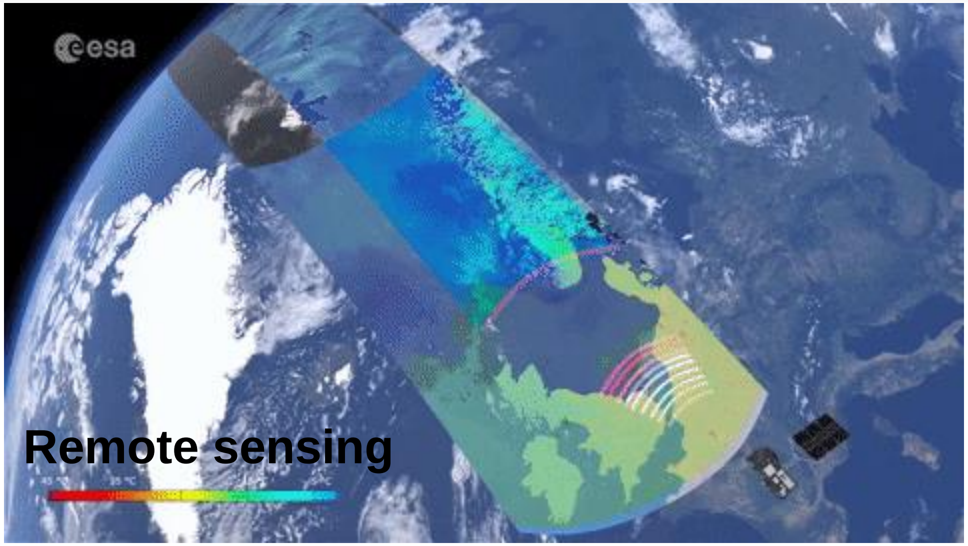
1. Don’t consider the continuous nature of soil variability.
2. Are expensive and time-consuming
3. Are not suitable for quantitative modelling purposes
4. Depend on the experience and skills of the soil surveyor.
5. Unknown accuracy



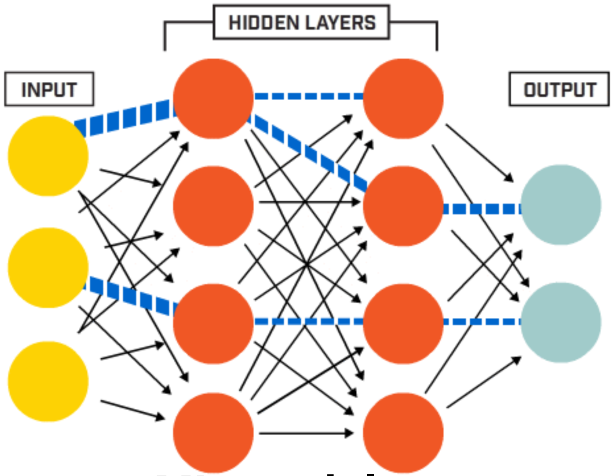
Dent & Young, 1981; Zhu, et al. 2001

From Paper to Pixels

Conventional to Digital

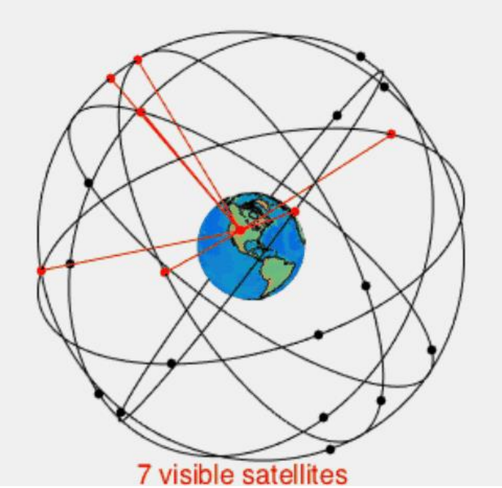


Remote sensing



ML models

Unmanned aerial vehicle



GPS and GIS

Introducing the SCORPAN Model

ELSEVIER

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

SCORPAN model

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

S : Soil, at a specific point in space and time: soil classes, Sc or soil attributes, Sa

From Jenny's Equation

c : climate, climate properties of the environment;
o : organisms, vegetation;
r : topography, landscape attributes;
p : parent material, lithology;
a : age or time factor;

Additions:

s : soil, prior knowledge of the soil at a point;
n : space, relative spatial position;
ε : auto-correlated random spatial variation.

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

McBratney et al., 2003

SCORPAN model

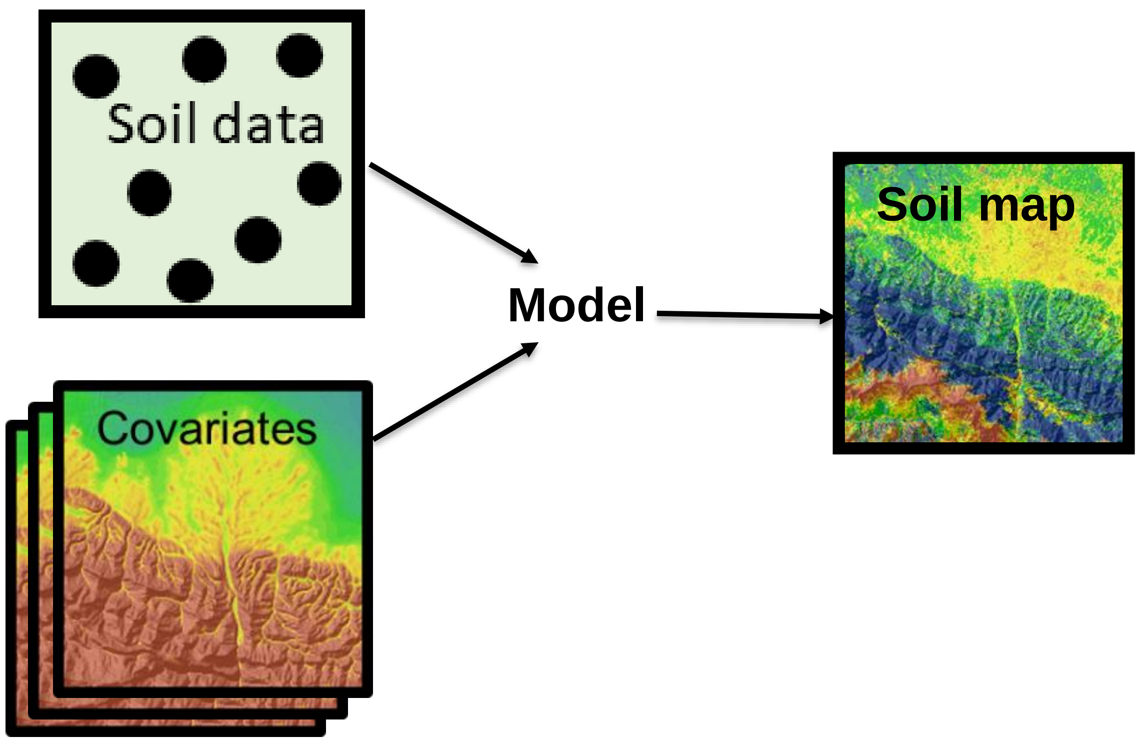
$$S = f\,(s,\ c,\ o,\ r,\ p,\ a,\ n) + \varepsilon$$

Soil data

Model

Covariates

spatially
dependent
residuals

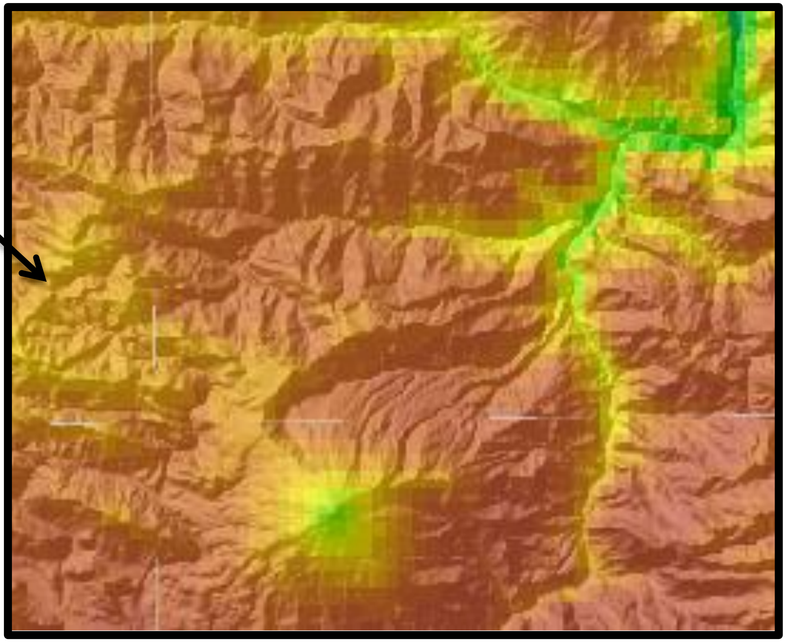


SCORPAN model

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

Soil data Model Covariates

elevation



Study area

SCORPAN model

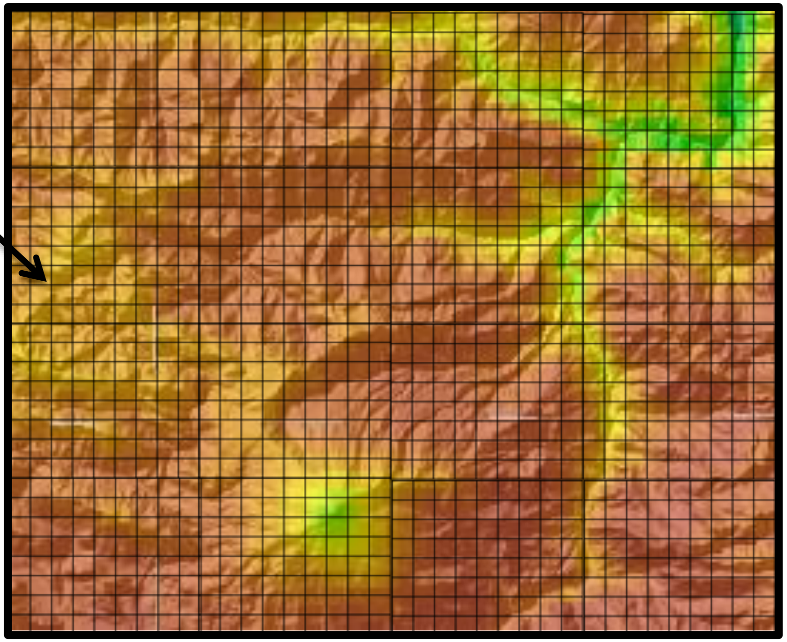
$$\boxed{S} = \boxed{f}\boxed{(s,\ c,\ o,\ r,\ p,\ a,\ n)} + \varepsilon$$

Soil data

Model

Covariates

elevation



Study area

SCORPAN model

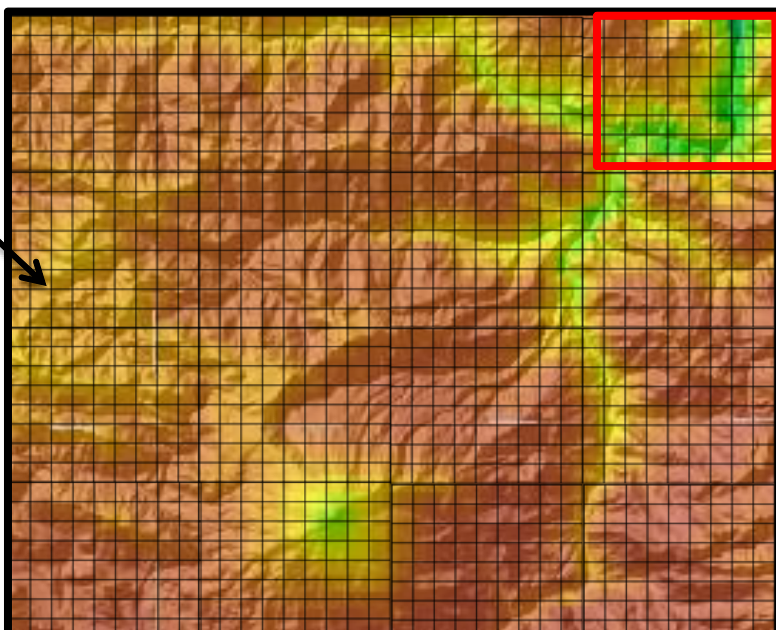
$$S = f(s, c, o, r, p, a, n) + \varepsilon$$

Soil data

Model

Covariates

elevation



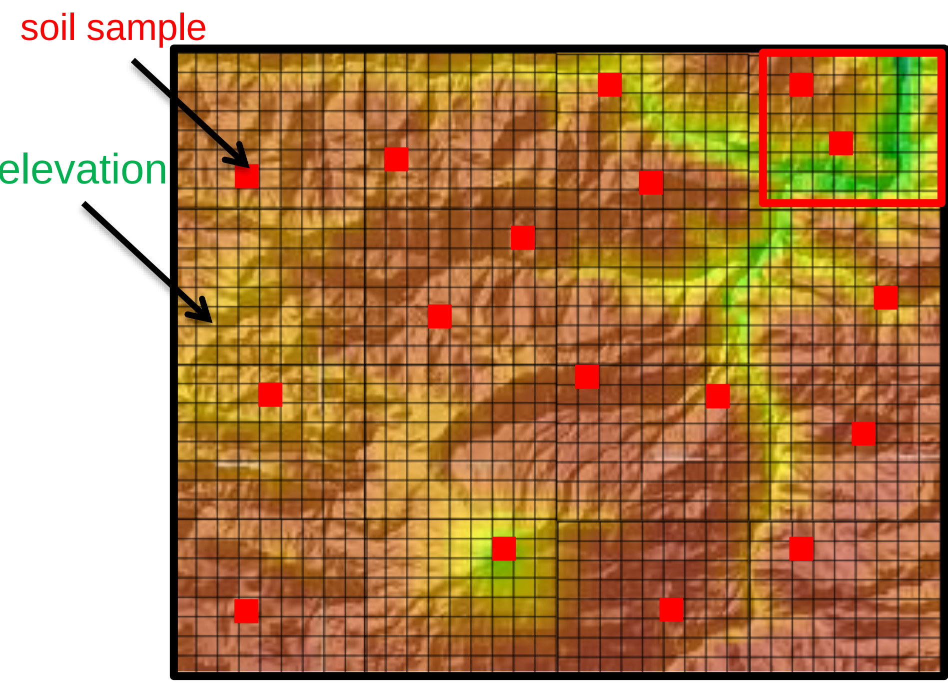
Study area

86	65	68	53	18	8	5	9	21
80	76	76	20	24	8	6	11	76
46	75	84	28	6	9	24	50	67
67	86	15	27	7	5	80	82	86
33	38	12	15	5	58	71	75	86
86	16	19	15	18	28	32	35	84
18	7	14	13	22	48	53	59	78
11	8	15	31	28	58	58	74	72

SCORPAN model

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

Soil data Model Covariates



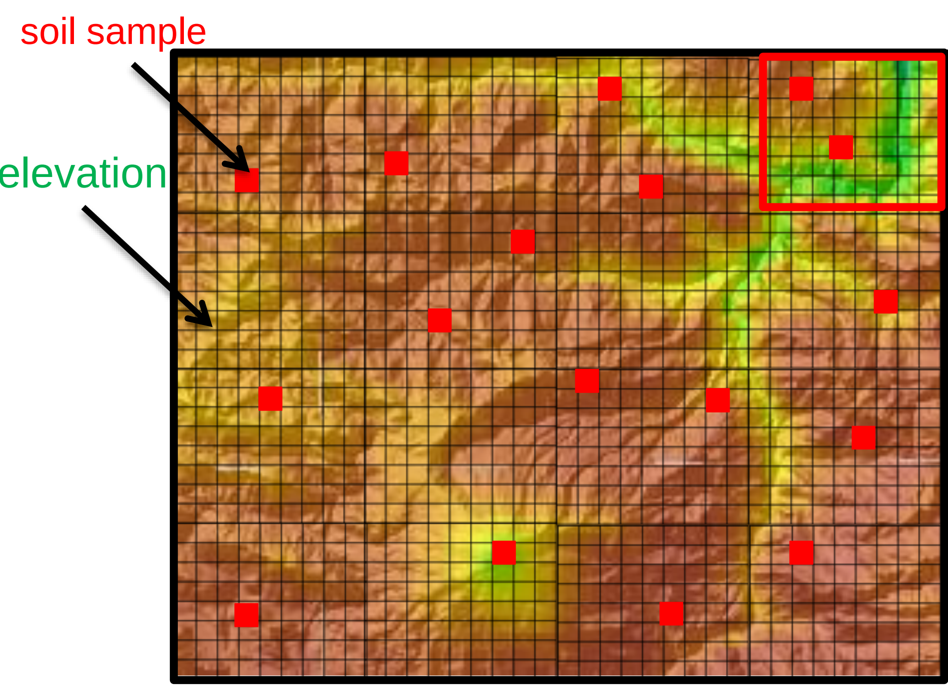
Study area

86	65	68	53	18	8	5	9	21
80	**76**	76	20	24	8	6	11	76
46	75	84	28	6	9	24	50	67
67	86	15	27	7	5	80	82	86
33	38	12	**15**	5	58	71	75	86
86	16	19	15	18	28	32	35	84
18	7	14	13	22	48	53	59	78
11	8	15	31	28	58	58	74	72

SCORPAN model

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

Soil data Model Covariates



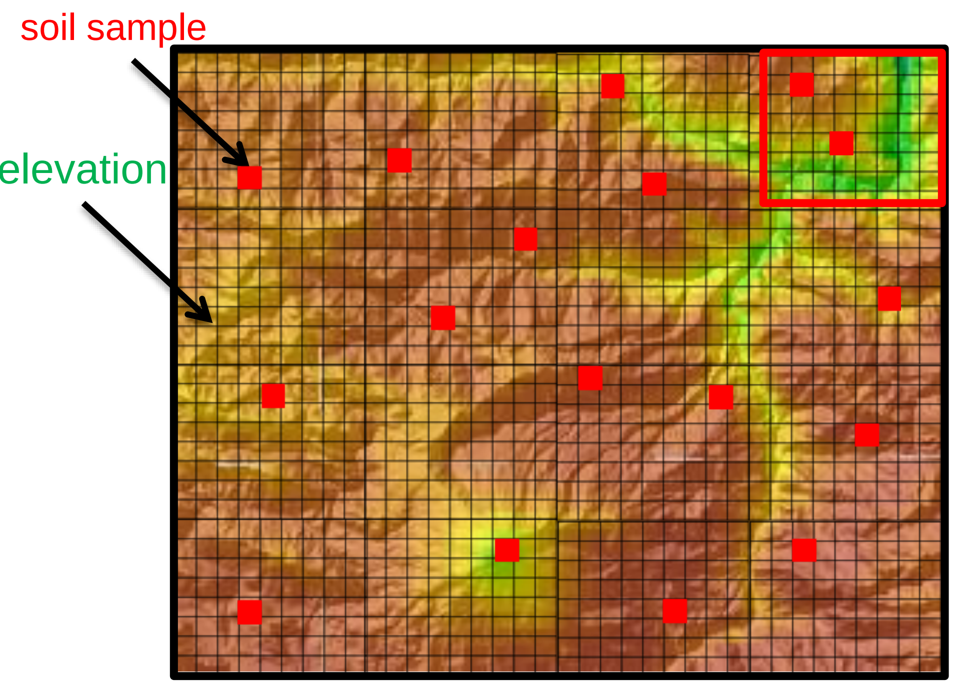
Study area

86	65	68	53	18	8	5	9	21
80	**76**	76	20	24	8	6	11	76
46	75	84	28	6	9	24	50	67
67	86	15	27	7	5	80	82	86
33	38	12	**15**	5	58	71	75	86
86	16	19	15	18	28	32	35	84
18	7	14	13	22	48	53	59	78
11	8	15	31	28	58	58	74	72

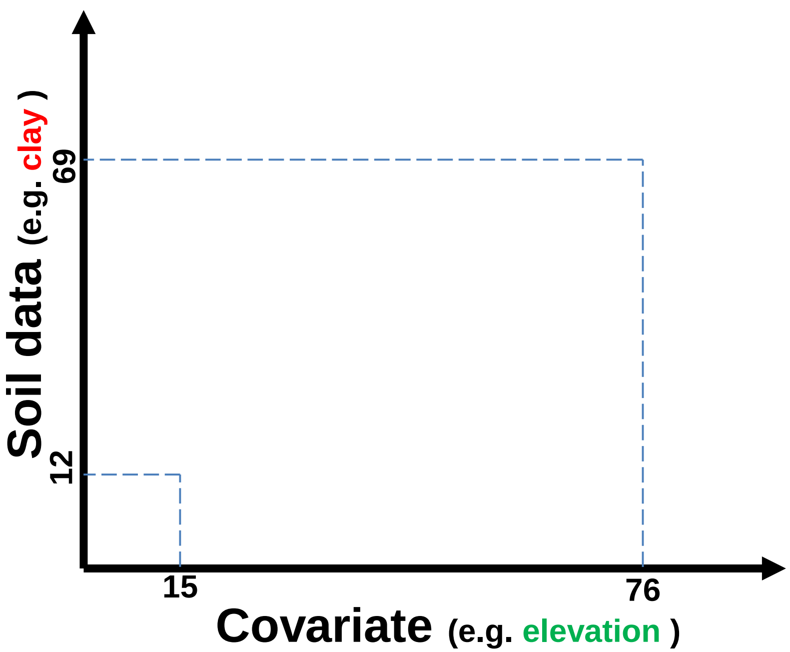
Samples	X	Y	Clay	Elevation
1	765342	2658345	**69**	**76**
2	765269	2695483	**12**	**15**

SCORPAN model

$$\underbrace{\boxed{S}}_{\text{Soil data}} = \underbrace{\boxed{f}}_{\text{Model}} \underbrace{\boxed{(s,\ c,\ o,\ r,\ p,\ a,\ n)}}_{\text{Covariates}} + \varepsilon$$



Study area



SCORPAN model

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

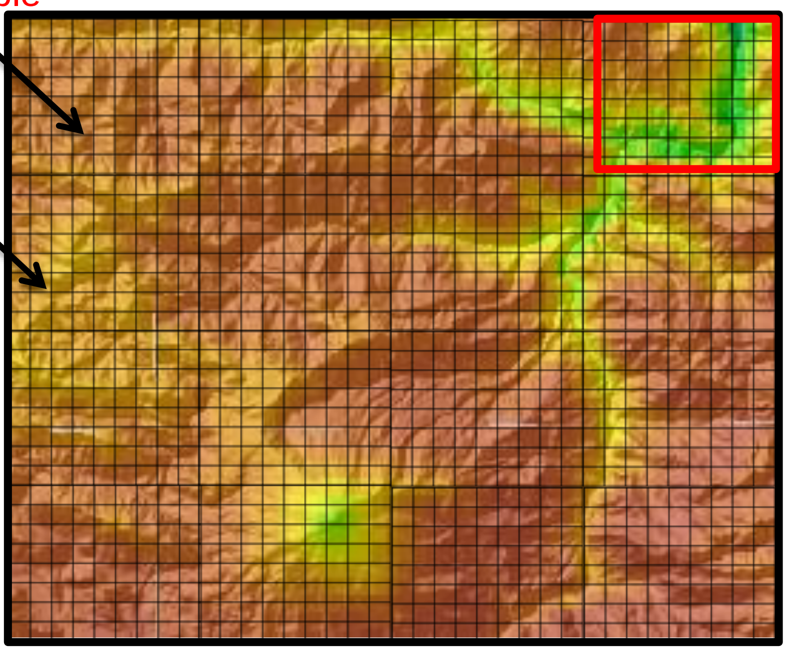
Soil data

Model

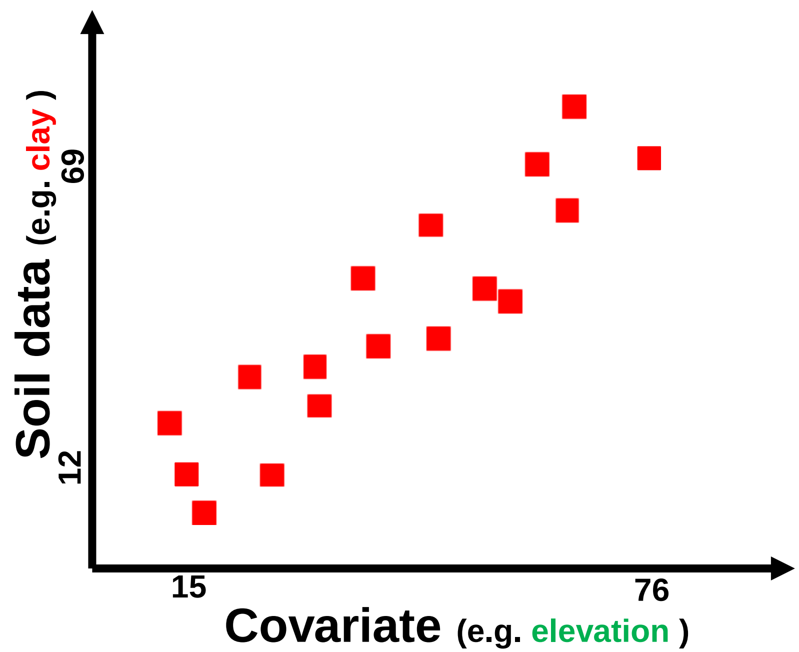
Covariates

soil sample

elevation



Study area



SCORPAN model

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

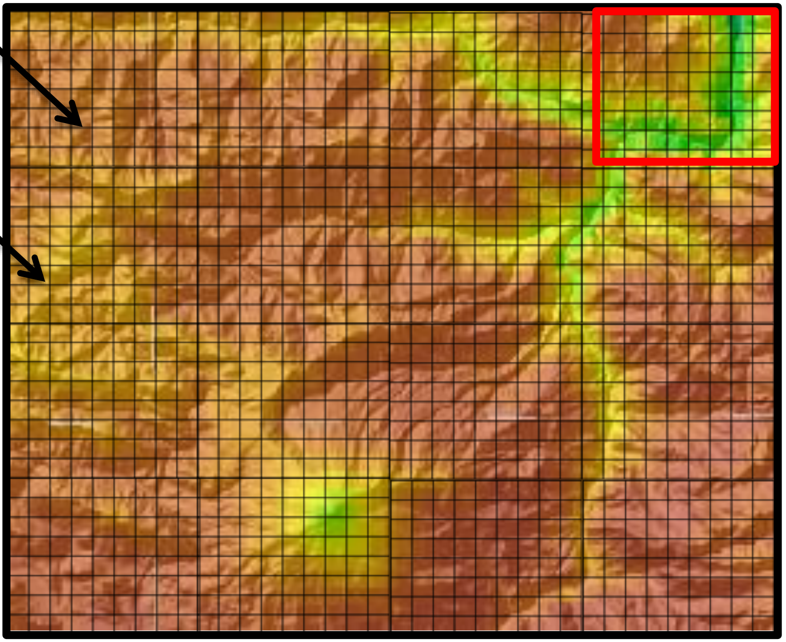
Soil data

Model

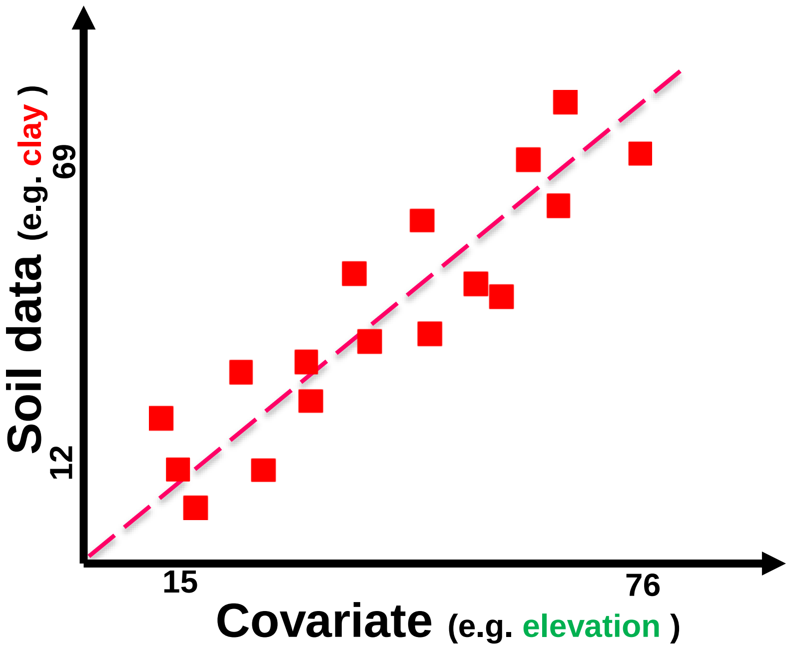
Covariates

soil sample

elevation



Study area



SCORPAN model

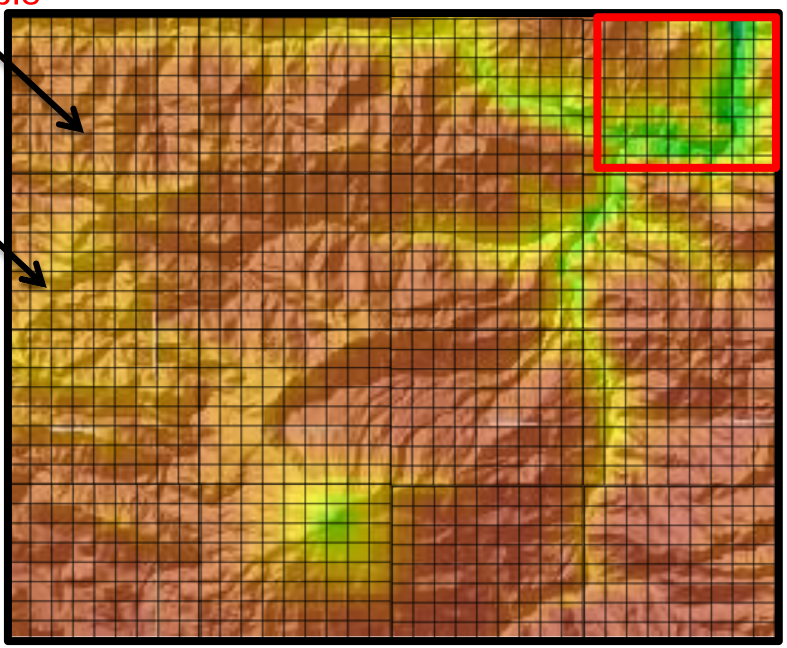
$$\boxed{S} = \boxed{f}\boxed{(s, \quad c, \quad o, \quad r, \quad p, \quad a, \quad n)} + \varepsilon$$

Soil data Model Covariates

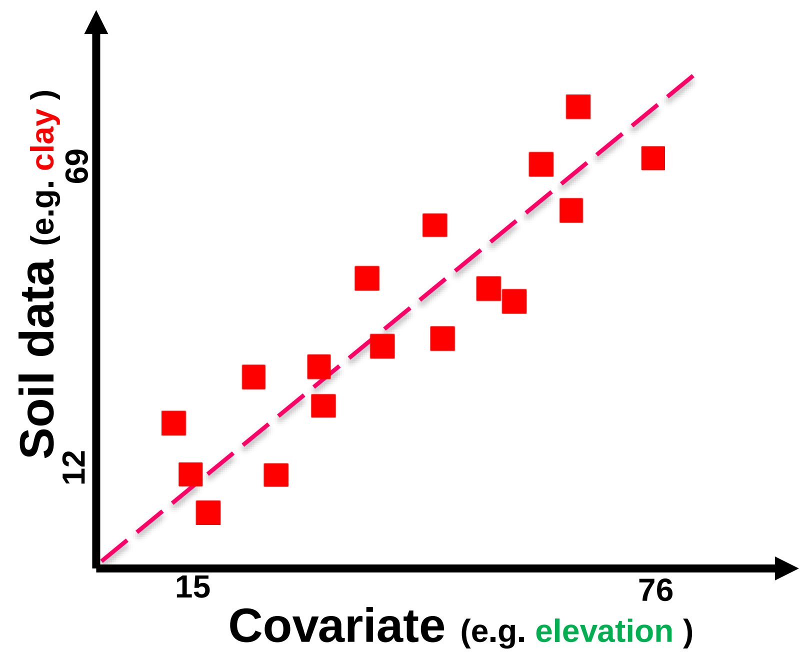
Soil clay = 0.024 + 0.83 × elevation

soil sample

elevation



Study area

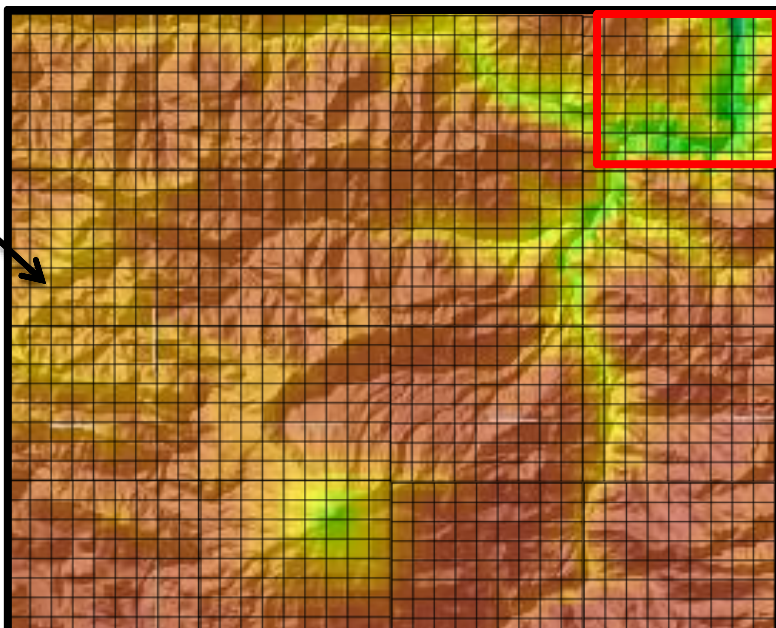


SCORPAN model

Soil clay = 0.024 + 0.83 × elevation

0.024 + 0.83 × 86 → **71**

elevation



Study area

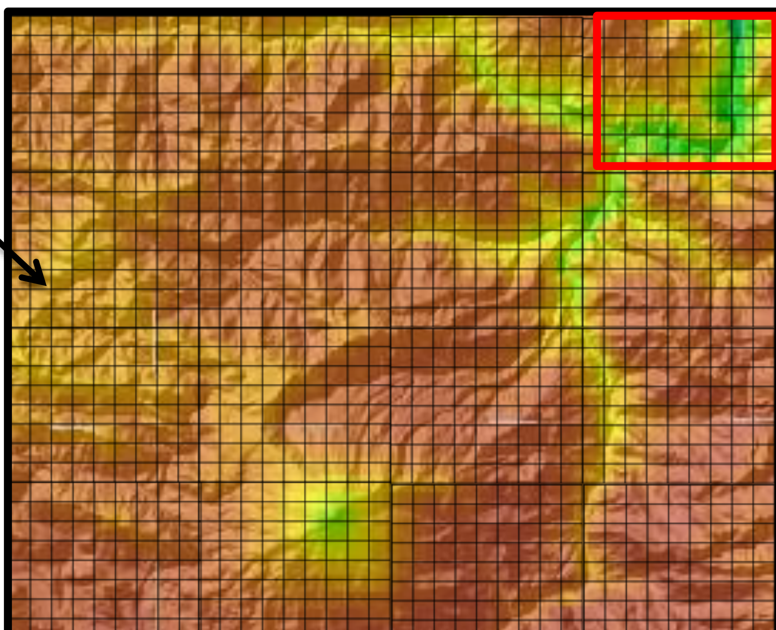
71	65	68	53	18	8	5	9	21
80	76	76	20	24	8	6	11	76
46	75	84	28	6	9	24	50	67
67	86	15	27	7	5	80	82	86
33	38	12	15	5	58	71	75	86
86	16	19	15	18	28	32	35	84
18	7	14	13	22	48	53	59	78
11	8	15	31	28	58	58	74	72

SCORPAN model

Soil clay = 0.024 + 0.83 × elevation

0.024 + 0.83 × 65 → 54

elevation



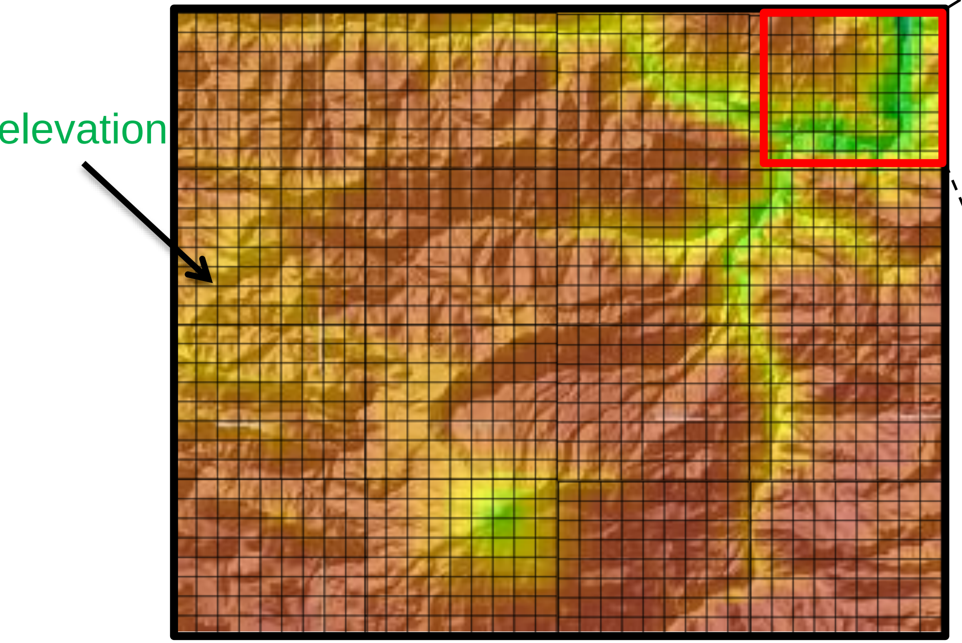
Study area

71	54	68	53	18	8	5	9	21
80	76	76	20	24	8	6	11	76
46	75	84	28	6	9	24	50	67
67	86	15	27	7	5	80	82	86
33	38	12	15	5	58	71	75	86
86	16	19	15	18	28	32	35	84
18	7	14	13	22	48	53	59	78
11	8	15	31	28	58	58	74	72

SCORPAN model

Soil clay = 0.024 + 0.83 × elevation

0.024 + 0.83 × 68 → 56



Study area

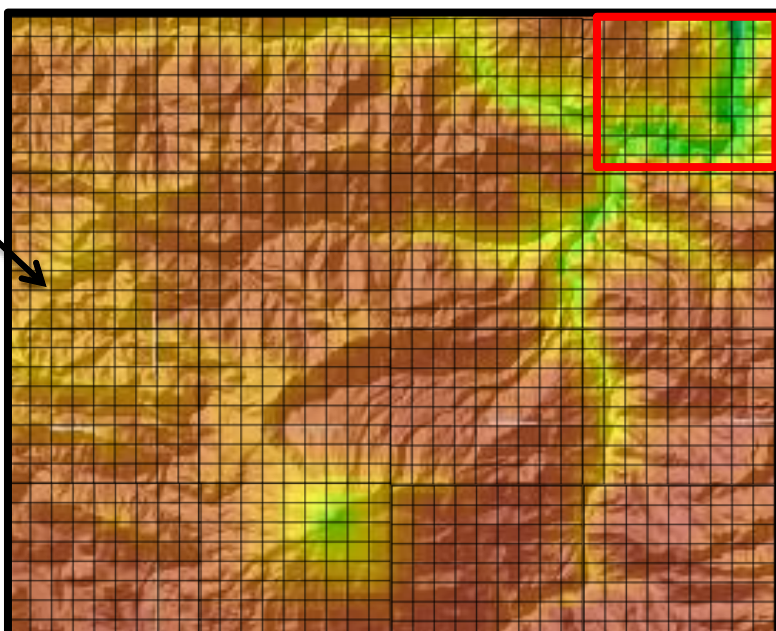
71	54	56	53	18	8	5	9	21
80	76	76	20	24	8	6	11	76
46	75	84	28	6	9	24	50	67
67	86	15	27	7	5	80	82	86
33	38	12	15	5	58	71	75	86
86	16	19	15	18	28	32	35	84
18	7	14	13	22	48	53	59	78
11	8	15	31	28	58	58	74	72

SCORPAN model

Soil clay = 0.024 + 0.83 × elevation

0.024 + 0.83 × 53 → 44

elevation



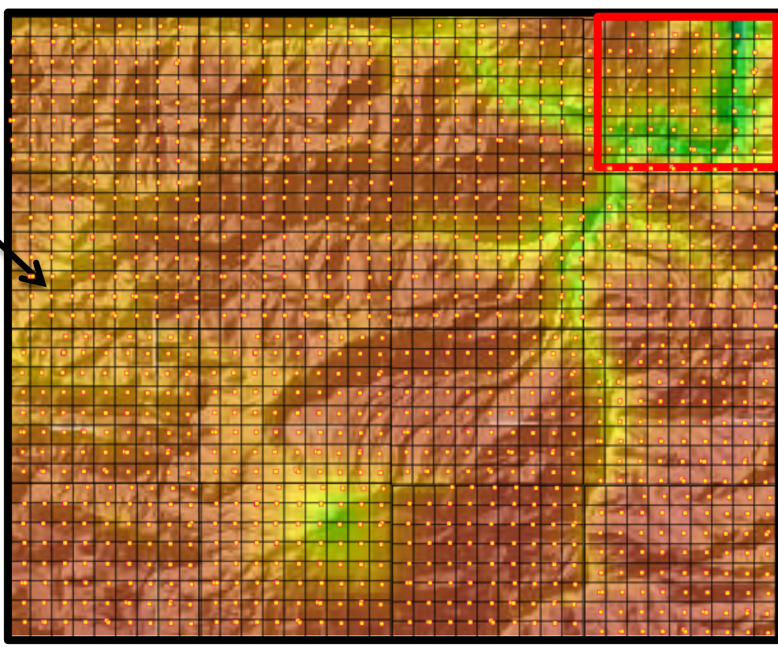
Study area

71	54	56	44	18	8	5	9	21
80	76	76	20	24	8	6	11	76
46	75	84	28	6	9	24	50	67
67	86	15	27	7	5	80	82	86
33	38	12	15	5	58	71	75	86
86	16	19	15	18	28	32	35	84
18	7	14	13	22	48	53	59	78
11	8	15	31	28	58	58	74	72

SCORPAN model

Soil clay = 0.024 + 0.83 × elevation

elevation



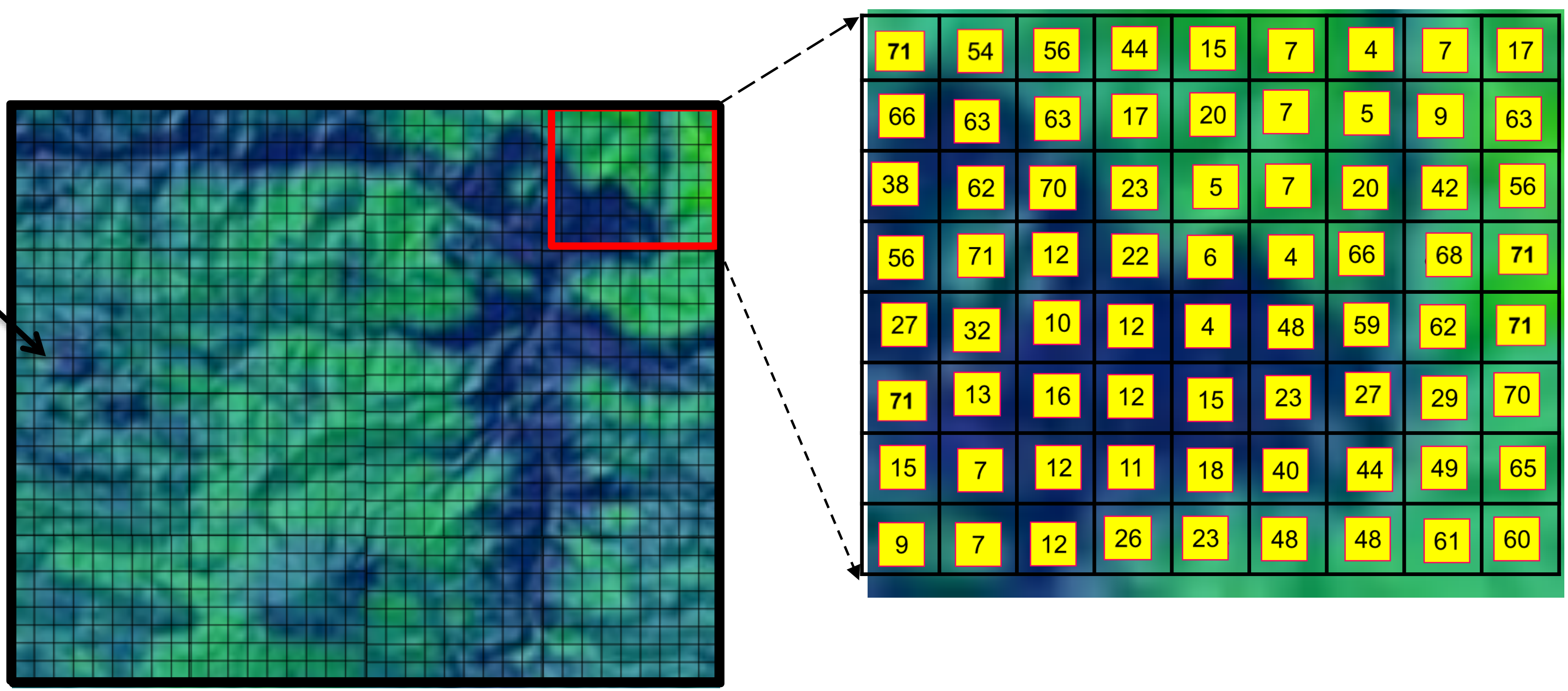
Study area

71	54	56	44	15	7	4	7	17
66	63	63	17	20	7	5	9	63
38	62	70	23	5	7	20	42	56
56	71	12	22	6	4	66	68	71
27	32	10	12	4	48	59	62	71
71	13	16	12	15	23	27	29	70
15	7	12	11	18	40	44	49	65
9	7	12	26	23	48	48	61	60

SCORPAN model

Soil clay = 0.024 + 0.83 × elevation

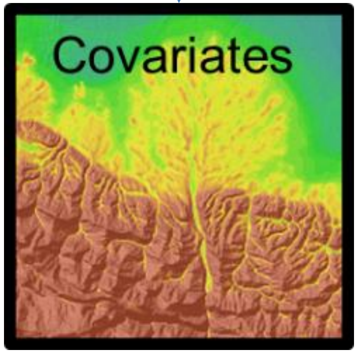
Clay map



Study area

SCORPAN Model

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



Environmental Covariates

s
c
o
r
p
a
n

Possible sources of information to represent seven scorpan factors

scorpan factor	Possible representatives
s	Legacy soil maps, point observations, expert knowledge
c	Temperature and precipitation records
o	Vegetation maps, species abundance maps, yield maps, land use maps
r	Digital elevation model, terrain attributes
p	Legacy geology maps, gamma radiometric information
a	Weathering indices, geology maps
n	Latitude and longitude or easting and northing, distance from landscape features, distance from roads, distance from point sources of pollution

- Proximal sensing, Remote sensing, and digital elevation models are three prominent examples of high-resolution environmental covariates can be used to represent various scorpan factors

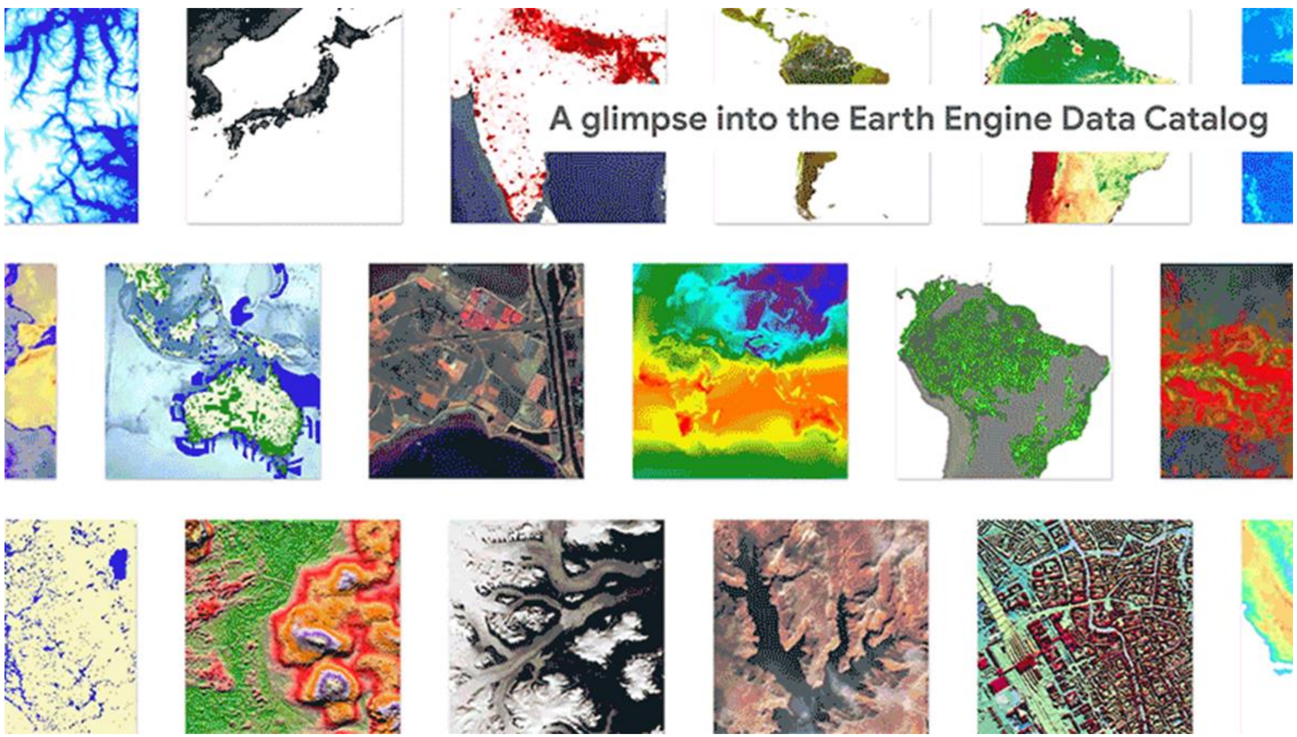
McBratney et al., 2003

Where Can I Obtain the Covariates?

s
c
o
r
p
a
n

Google Earth Engine

rgee

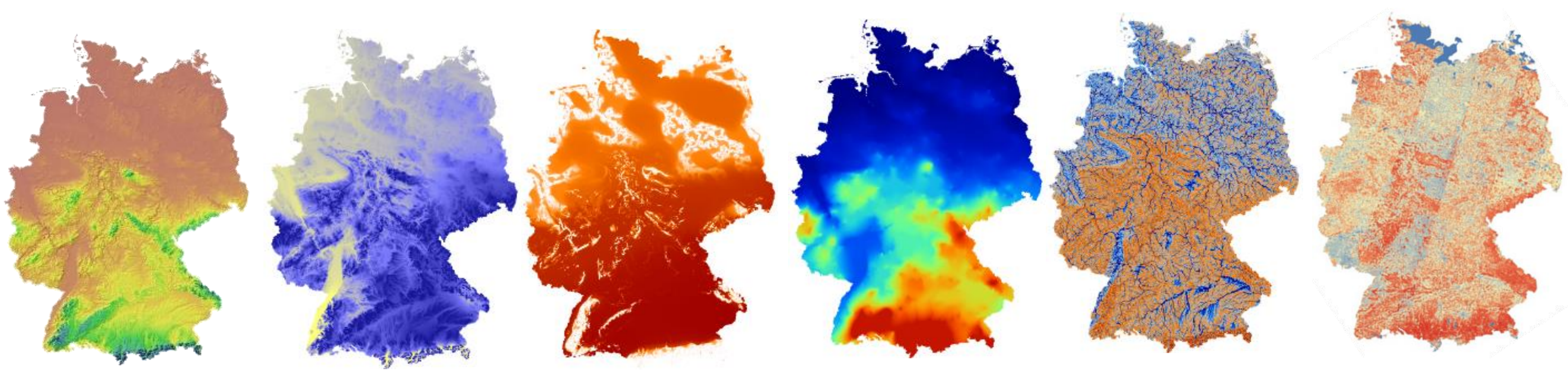


Technical and Practical Notes

s
c
o
r
p
a
n

Covariate data sources

- SRTM digital elevation model (30 m resolution)
- WorldDEM digital elevation model (12 m resolution)
- Landsat-8 satellite images (30 m resolution)
- Sentinel-2 satellite images (10, 20, and 60 m resolution)
- MODIS satellite images (250, 500, and 1000 m resolution)
- Global Land Cover maps (30 m resolution)
- JAXA’s ALOS radar images (20 m resolution)
- Monthly precipitation images (1000 m resolution)
- Geology maps (polygon)



Technical and Practical Notes

s
c
o
r
p
a
n

Preparing covariate layers

1. Converting polygon maps to rasters

2. Downscaling or upscaling rasters to a common resolution,

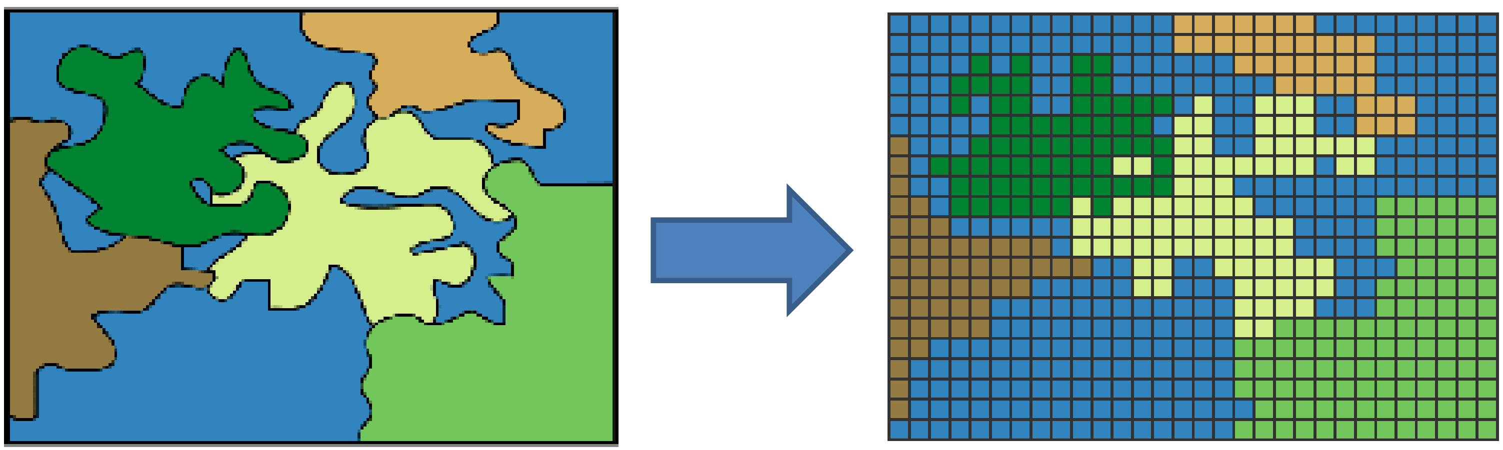
3. Filtering out missing pixels/reducing noise and multicolinearity problems,

4. Overlaying raster stacks and points

Technical and Practical Notes

s
c
o
r
p
a
n

1. Converting polygon maps to rasters



Polygon features

Raster polygon features

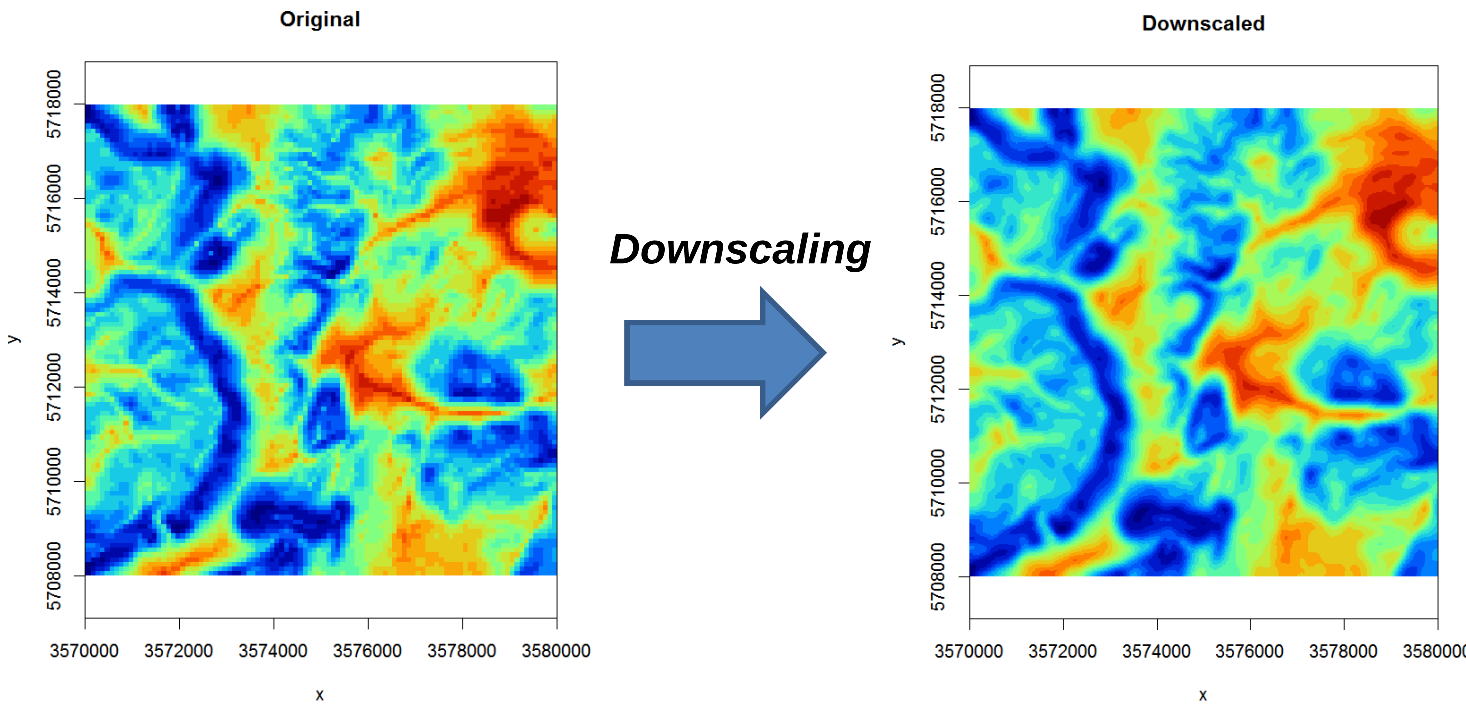
soilmapper.org
desktop.arcgis.com

Technical and Practical Notes

s c o r p a n

2. Downscaling or upscaling rasters

- To adjust the resolution of some covariates that have either too coarse or too fine a resolution compared to the target resolution
- The process of bringing raster layers to a common grid resolution is also known as resampling

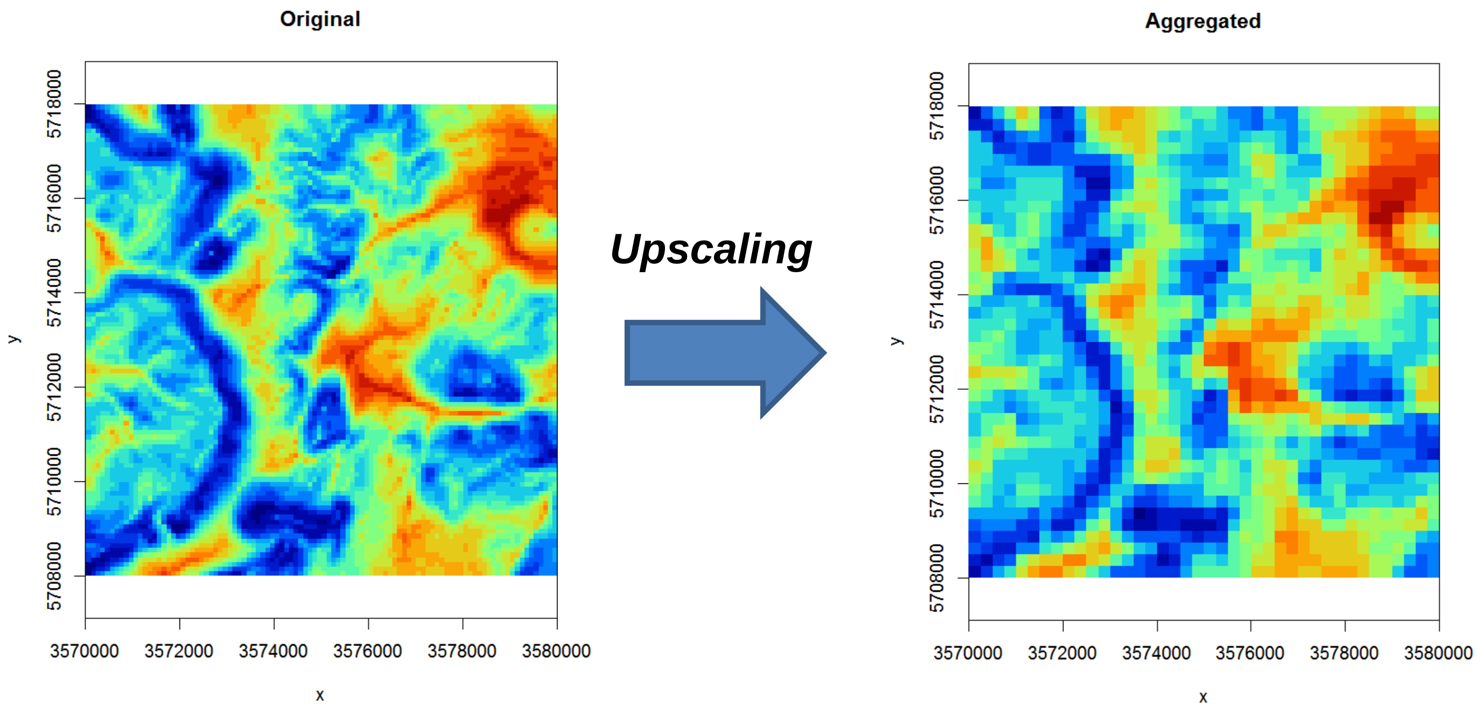


Technical and Practical Notes

s
c
o
r
p
a
n

2. Downscaling or upscaling rasters

- To adjust the resolution of some covariates that have either too coarse or too fine a resolution compared to the target resolution
- The process of bringing raster layers to a common grid resolution is also known as resampling



Technical and Practical Notes

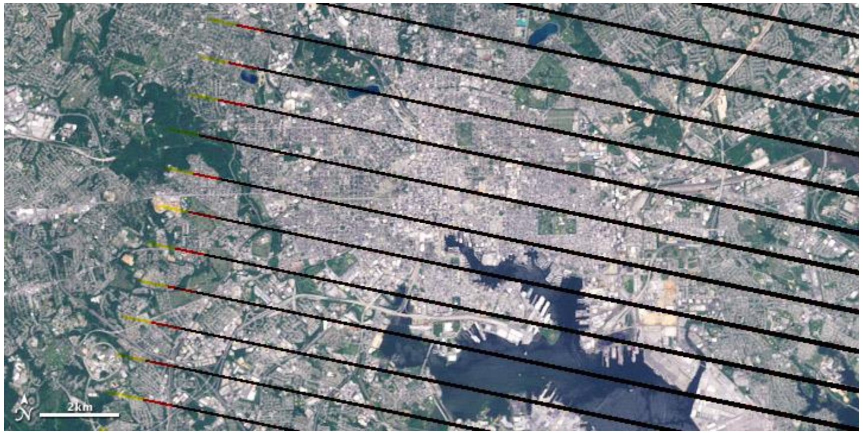
s
c
o
r
p
a
n

3. Filtering out missing pixels and artifacts

- could cause serious problems for producing soil maps as the missing pixels and artifacts would propagate to predictions: if only one layer in the raster stack misses values then predictive models might drop whole rows in the predictions even though data is available for 95% of rows.

- Missing pixels can be efficiently filtered by using for example the gap filling functionality available in the SAGA GIS

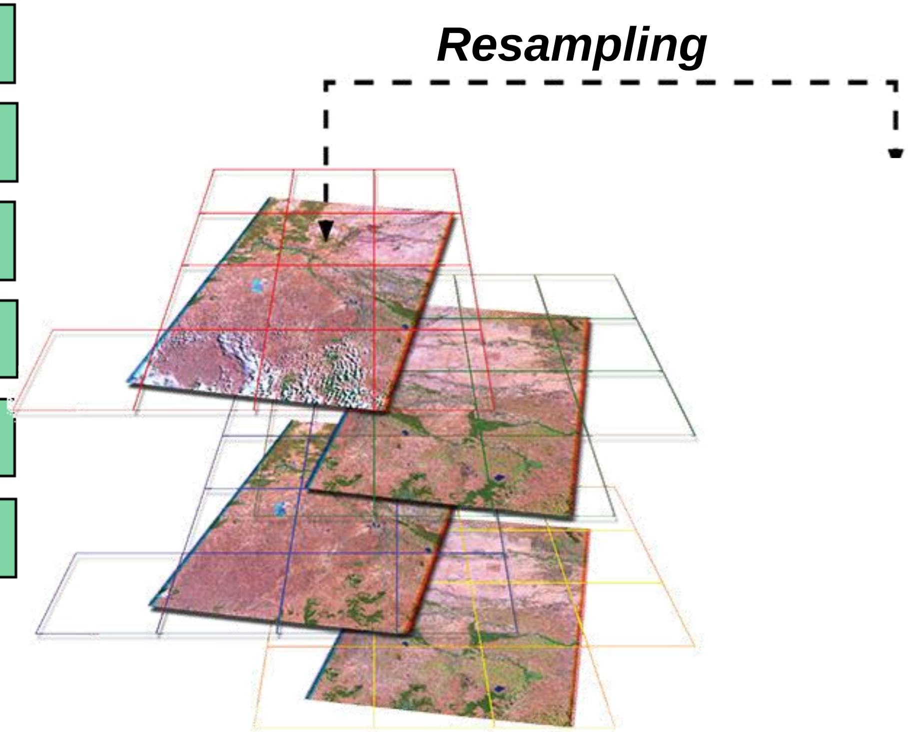
- Another way to filter the missing pixels, to reduce noise and to reduce data overlap is to use Principal Components transformation of original data



Technical and Practical Notes

s
c
o
r
p
a
n

Raster stacks

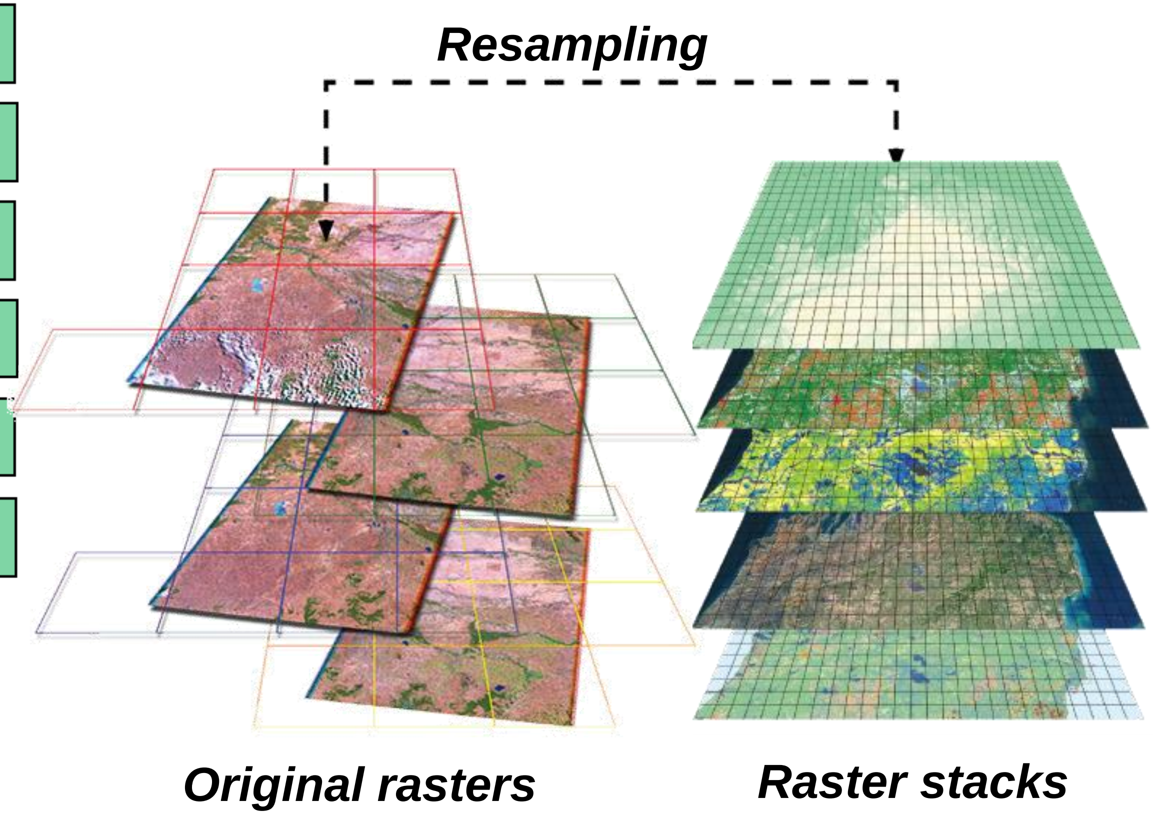


Original rasters

Technical and Practical Notes

s
c
o
r
p
a
n

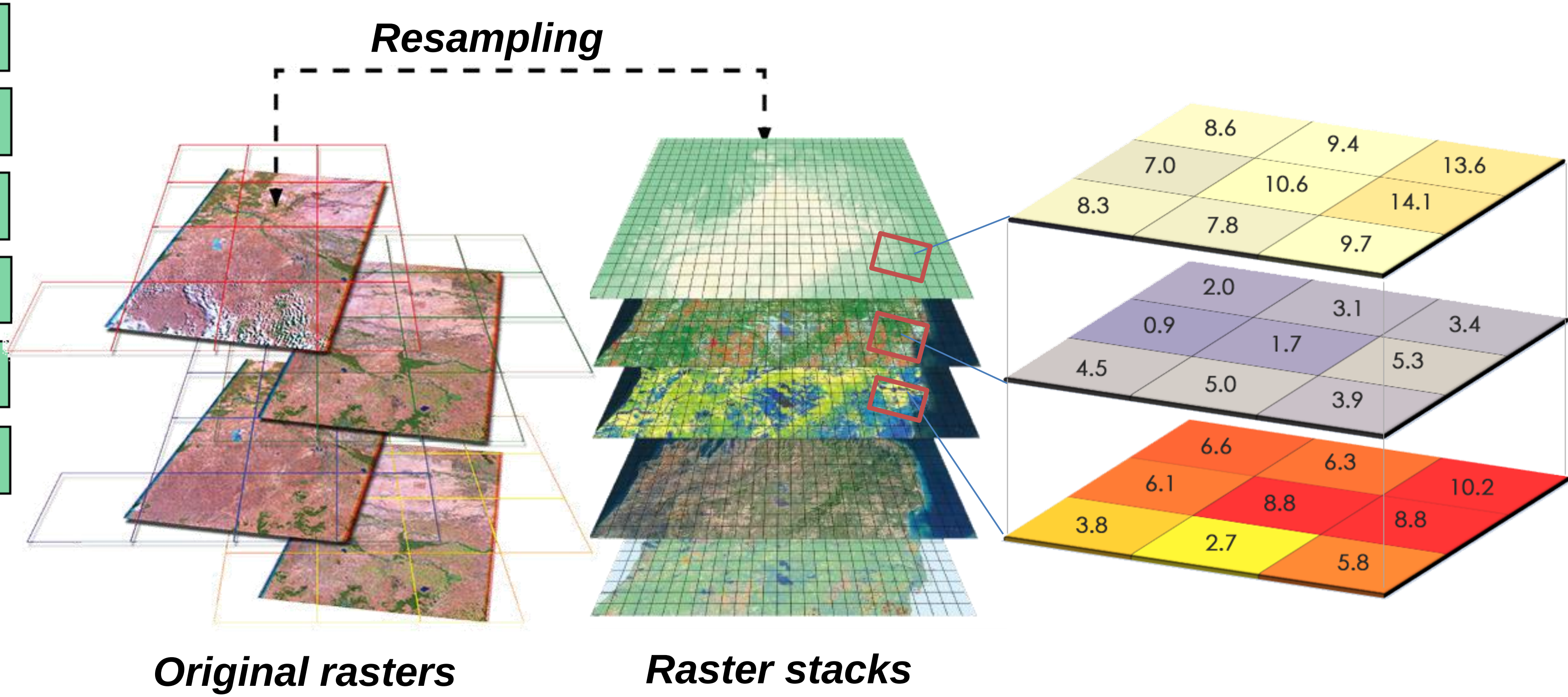
Raster stacks



Technical and Practical Notes

s
c
o
r
p
a
n

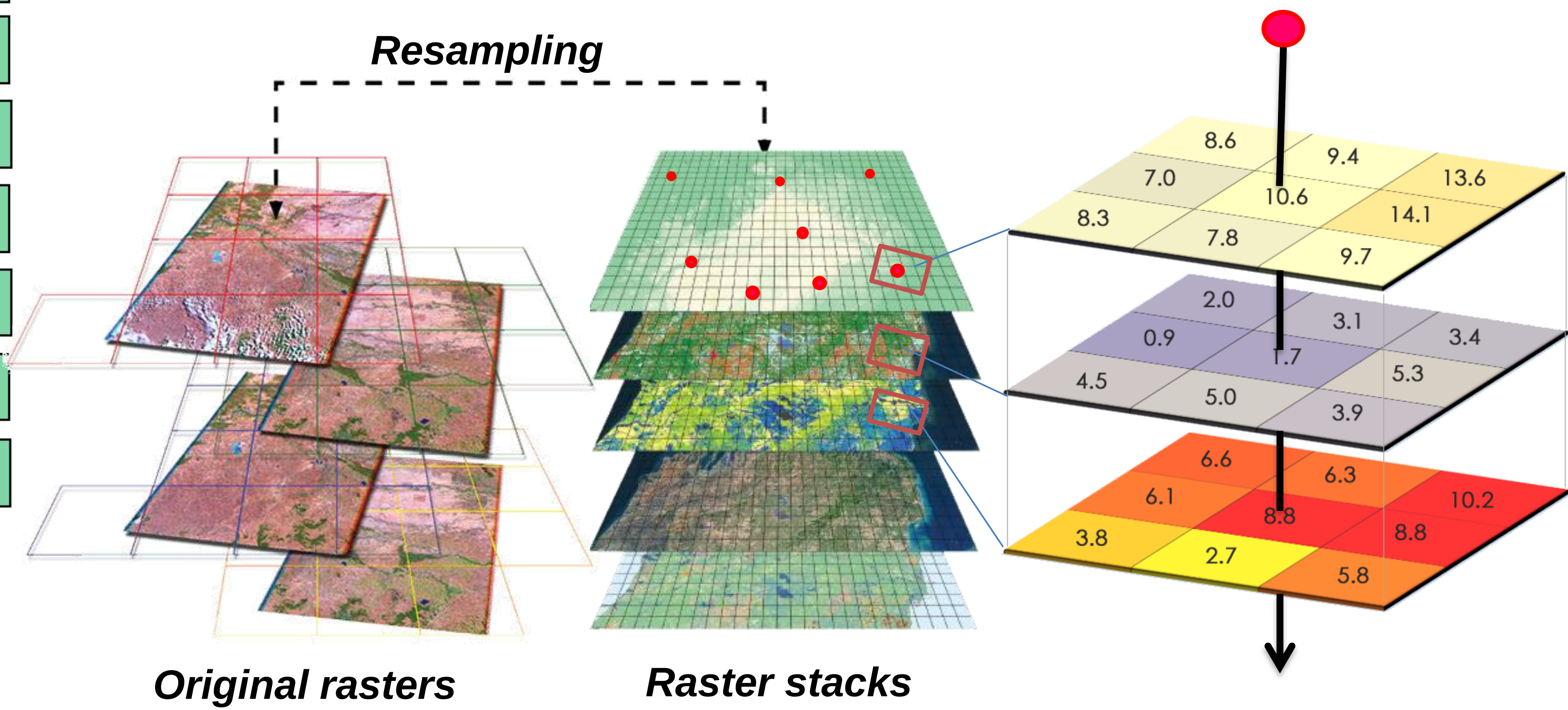
Raster stacks



Technical and Practical Notes

s
c
o
r
p
a
n

4. Overlaying raster stacks and points

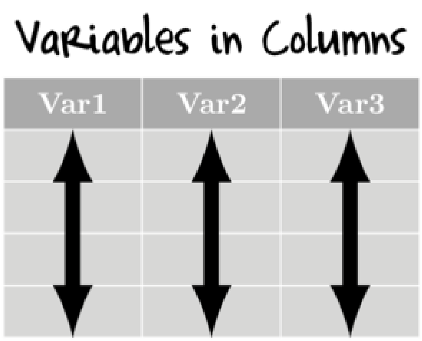


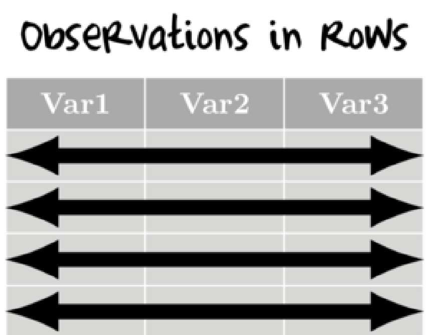
Technical and Practical Notes

Geo-database
Tidy data
(data.frame)

Predictors
Covariates
Independent variables
Features

X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15	Class
-0.60	3.06	-5.77	0.69	1.46	1.22	1.94	0.61	-0.07	-2.02	-0.87	0.71	0.89	-0.91	0.21	a
0.61	1.61	0.21	2.74	0.46	1.29	0.13	0.55	-0.60	-0.53	0.70	-0.41	-0.62	1.17	-0.32	b
3.81	0.99	-1.21	0.30	0.11	-0.34	-0.29	0.15	-0.09	1.30	0.08	0.66	-0.44	0.45	-0.13	a
0.50	0.14	-0.30	0.64	-1.28	0.28	0.76	-0.61	0.10	-0.07	0.41	-0.53	-0.62	0.06	-0.52	c
1.46	0.23	0.17	0.53	-0.94	0.37	0.64	-0.83	0.20	-0.06	0.18	-0.35	-0.64	0.39	-0.22	a
3.07	-0.19	1.31	0.84	-0.66	1.05	0.94	-1.12	-0.23	-0.31	0.45	-0.54	-0.16	0.38	0.53	a





Response variable
Target variable
Dependent variable
(Soil clay or soil types)

Supervised Learning

Supervised learning

- For each covariate value $\mathbf{x_i}$ there is also a response value $\mathbf{y_i}$
- → e.g. random forest
- → what we usually do for soil mapping

X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15	Class
-0.60	3.06	-5.77	0.69	1.46	1.22	1.94	0.61	-0.07	-2.02	-0.87	0.71	0.89	-0.91	0.21	a
0.61	1.61	0.21	2.74	0.46	1.29	0.13	0.55	-0.60	-0.53	0.70	-0.41	-0.62	1.17	-0.32	b
3.81	0.99	-1.21	0.30	0.11	-0.34	-0.29	0.15	-0.09	1.30	0.08	0.66	-0.44	0.45	-0.13	a
0.50	0.14	-0.30	0.64	-1.28	0.28	0.76	-0.61	0.10	-0.07	0.41	-0.53	-0.62	0.06	-0.52	c
1.46	0.23	0.17	0.53	-0.94	0.37	0.64	-0.83	0.20	-0.06	0.18	-0.35	-0.64	0.39	-0.22	a
3.07	-0.19	1.31	0.84	-0.66	1.05	0.94	-1.12	-0.23	-0.31	0.45	-0.54	-0.16	0.38	0.53	a

Supervised learning

- – Regression: continuous responses, e.g. soil clay content, rainfall
- – Classification: categorical responses (binary or multinomial), e.g. soil type

Available Models

CART (Boosting, Bagging), BART, MART, Random Forest, C5.0

Tree-Based

Linear-Models

Multiple linear regression, Generalized linear model, Generalized additive model, Logistic Regression

Cubist, Logistic Model Tree

Model-Trees

Expert-Knowledge

Fuzzy Inference Systems, Rule Induction Algorithms

Artificial Neural Networks, Convolutional Neural Networks

Neural Networks

Support Vector Machines

Linear SVM, Polynomial SVM, Radial Basis Function SVM

k Nearest Neighbors, Nearest Shrunken Centroid

Distance-Based Learners

There are so many!! 

Available Models in R

- The caret package is a set of functions that attempt to streamline the process for creating predictive models. The package contains tools for:

1. data splitting
2. pre-processing
3. feature selection
4. model tuning using resampling
5. variable importance estimation

6 Available Models

The models below are available in `train`. The code behind these protocols can be obtained using the function `getModelInfo` or by going to the github repository.

Show 238 entries

Search:

Model	*method* Value	Type	Libraries	Tuning Parameters
AdaBoost Classification Trees	adaboost	Classification	fastAdaboost	nIter, method
AdaBoost.M1	AdaBoost.M1	Classification	adabag, plyr	mfinal, maxdepth, coeflearn
Adaptive Mixture Discriminant Analysis	amdai	Classification	adaptDA	model
Adaptive-Network-Based Fuzzy Inference System	ANFIS	Regression	frbs	num.labels, max.iter
Adjacent Categories Probability Model for Ordinal Data	vglmAdjCat	Classification	VGAM	parallel, link

http://topepo.github.io/caret/index.html

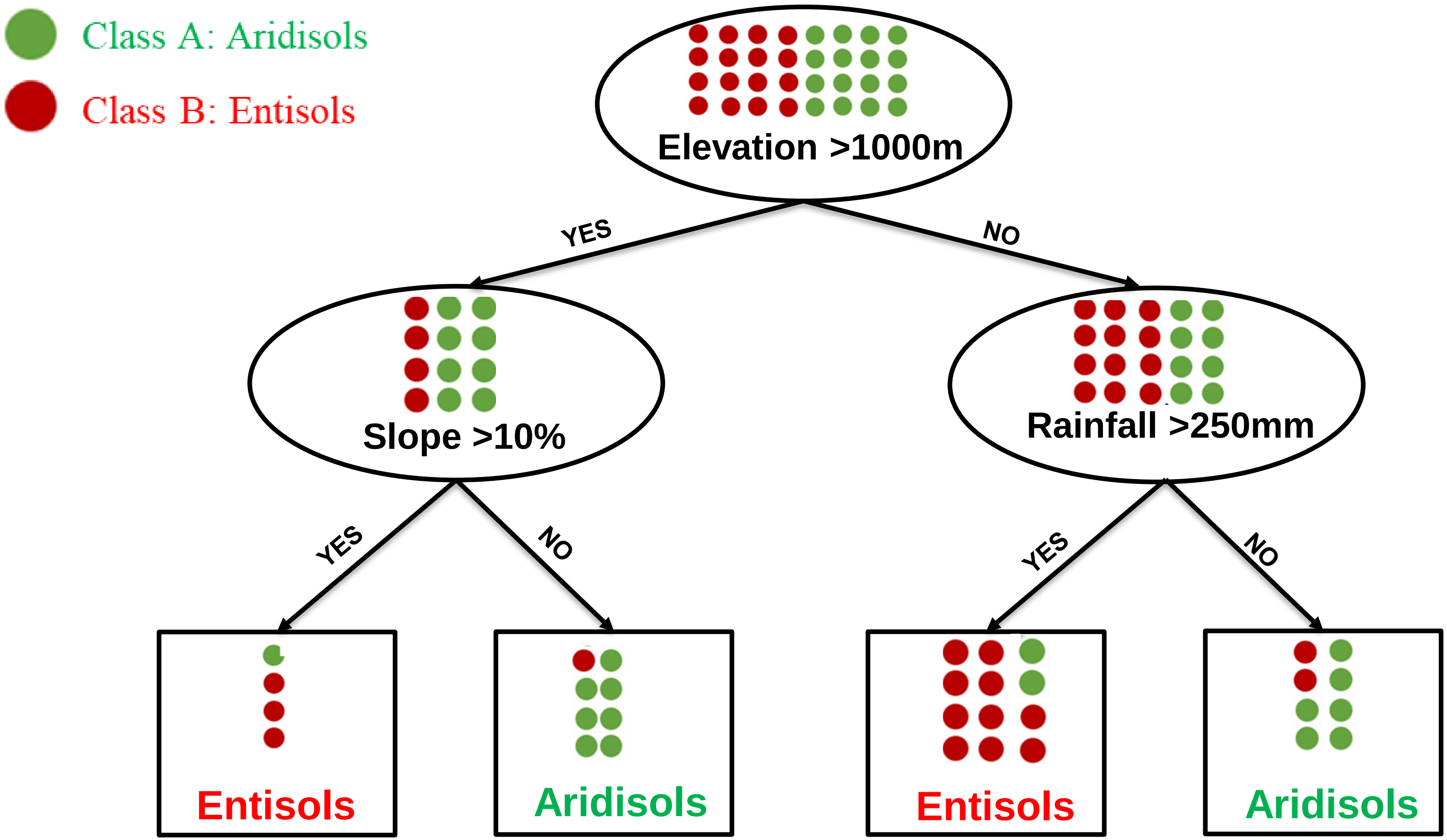
From Available Models

Decision Tree

Random Forest

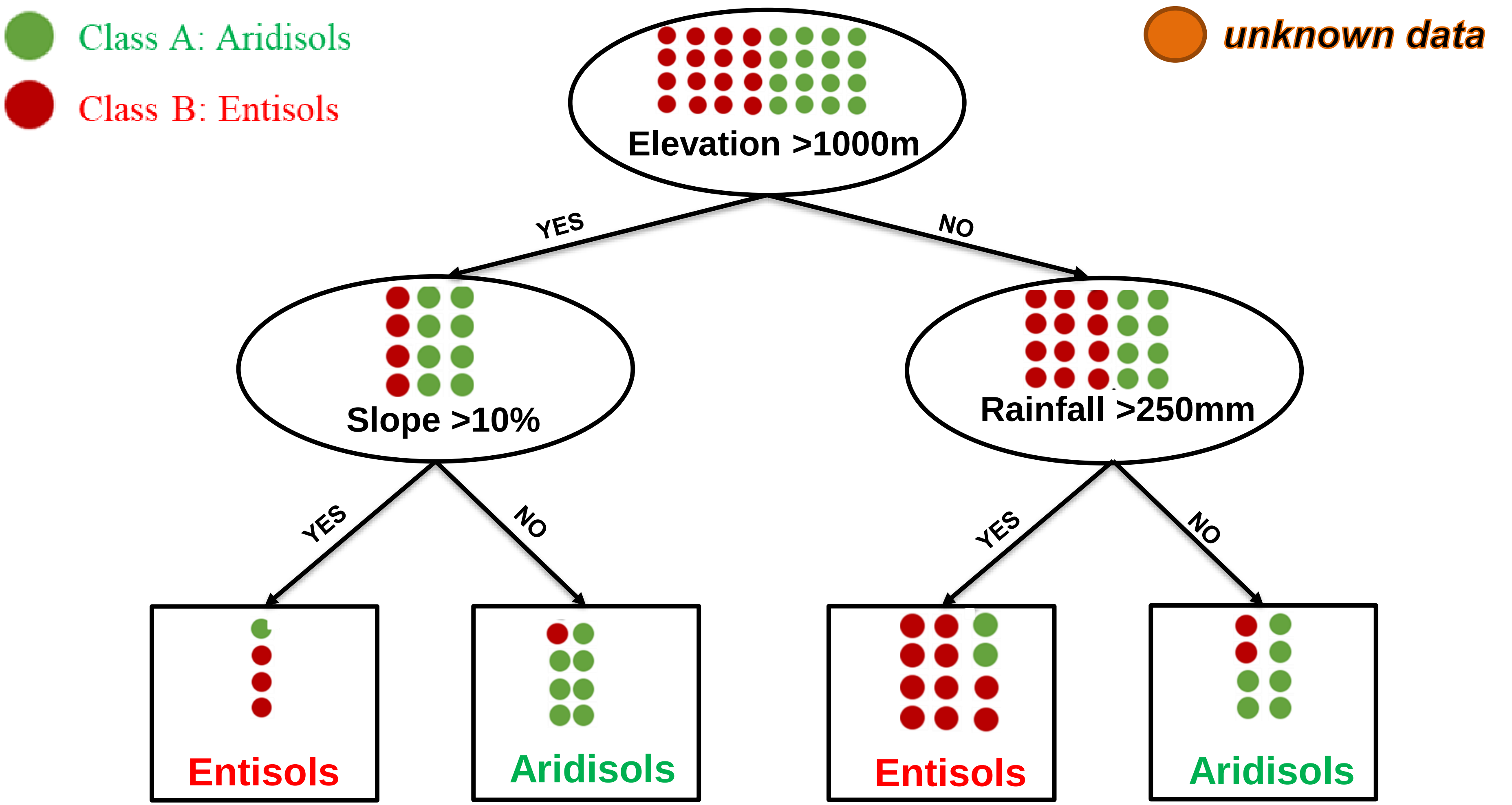
Decision Tree

Decision Tree: example II



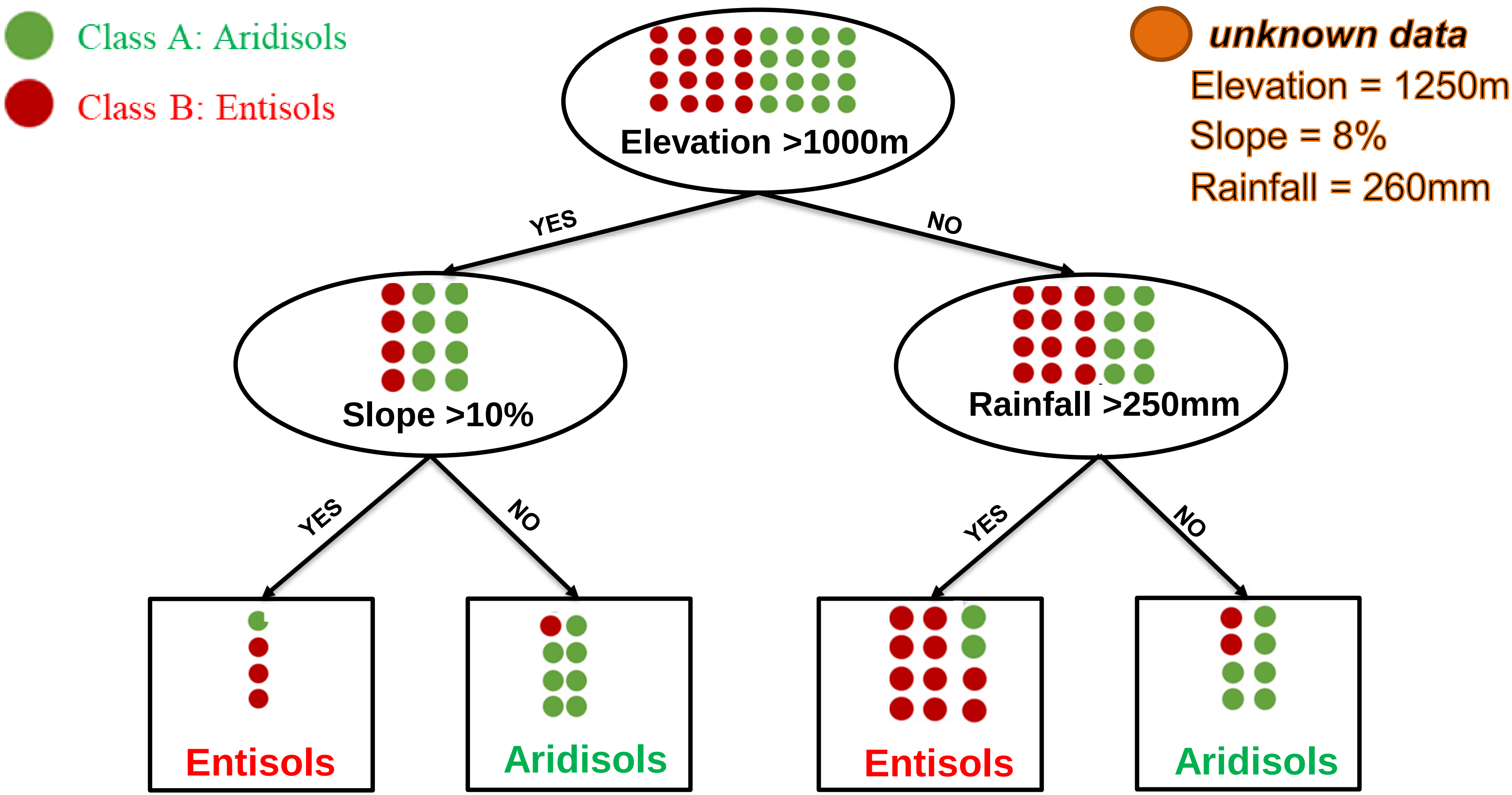
A decision tree model to predict Entisols and Aridisols

Decision Tree: example II



A decision tree model to predict Entisols and Aridisols

Decision Tree: example II



A decision tree model to predict Entisols and Aridisols

Decision Tree: example II

Class A: Aridisols

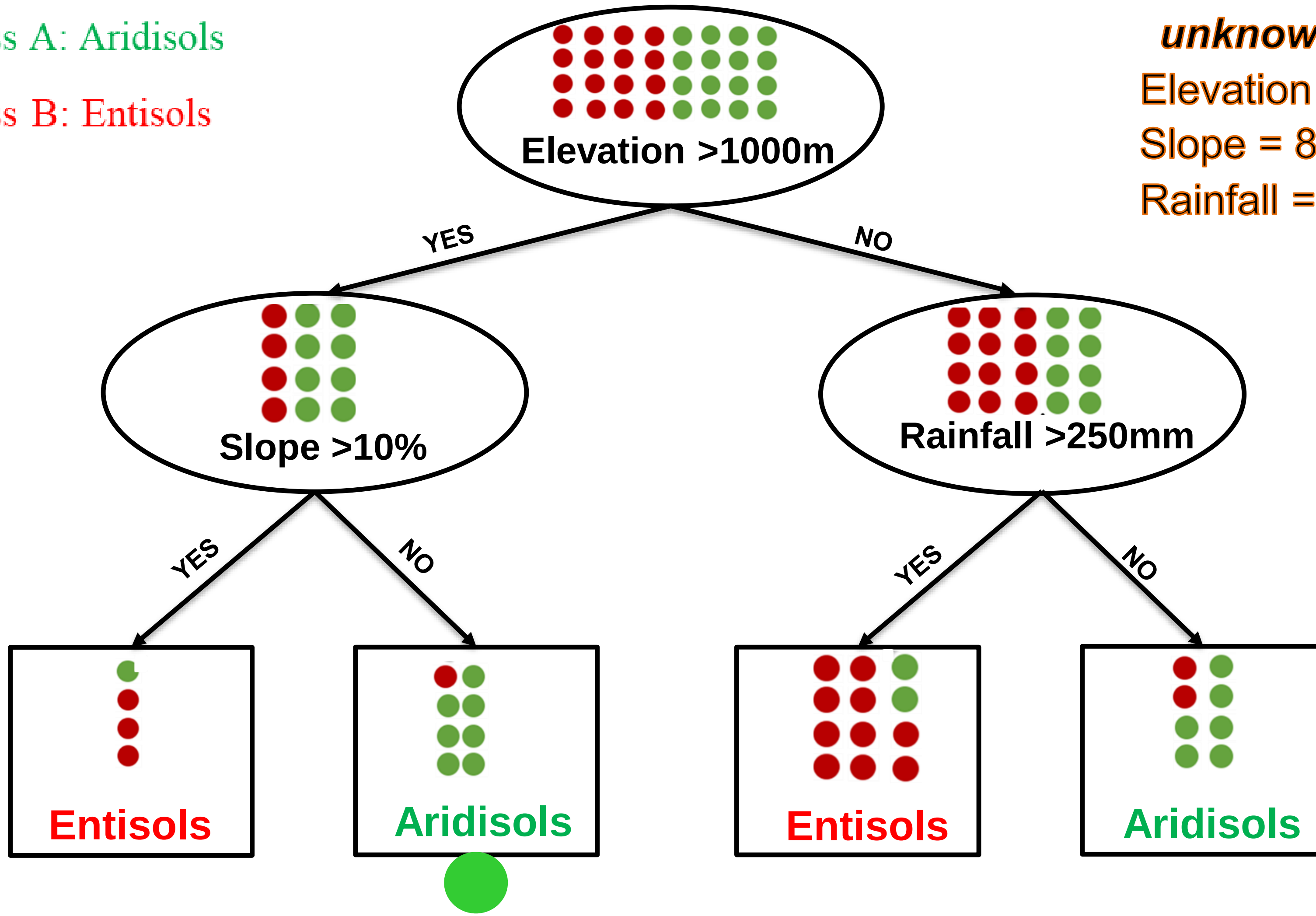
Class B: Entisols

unknown data

Elevation = 1250m

Slope = 8%

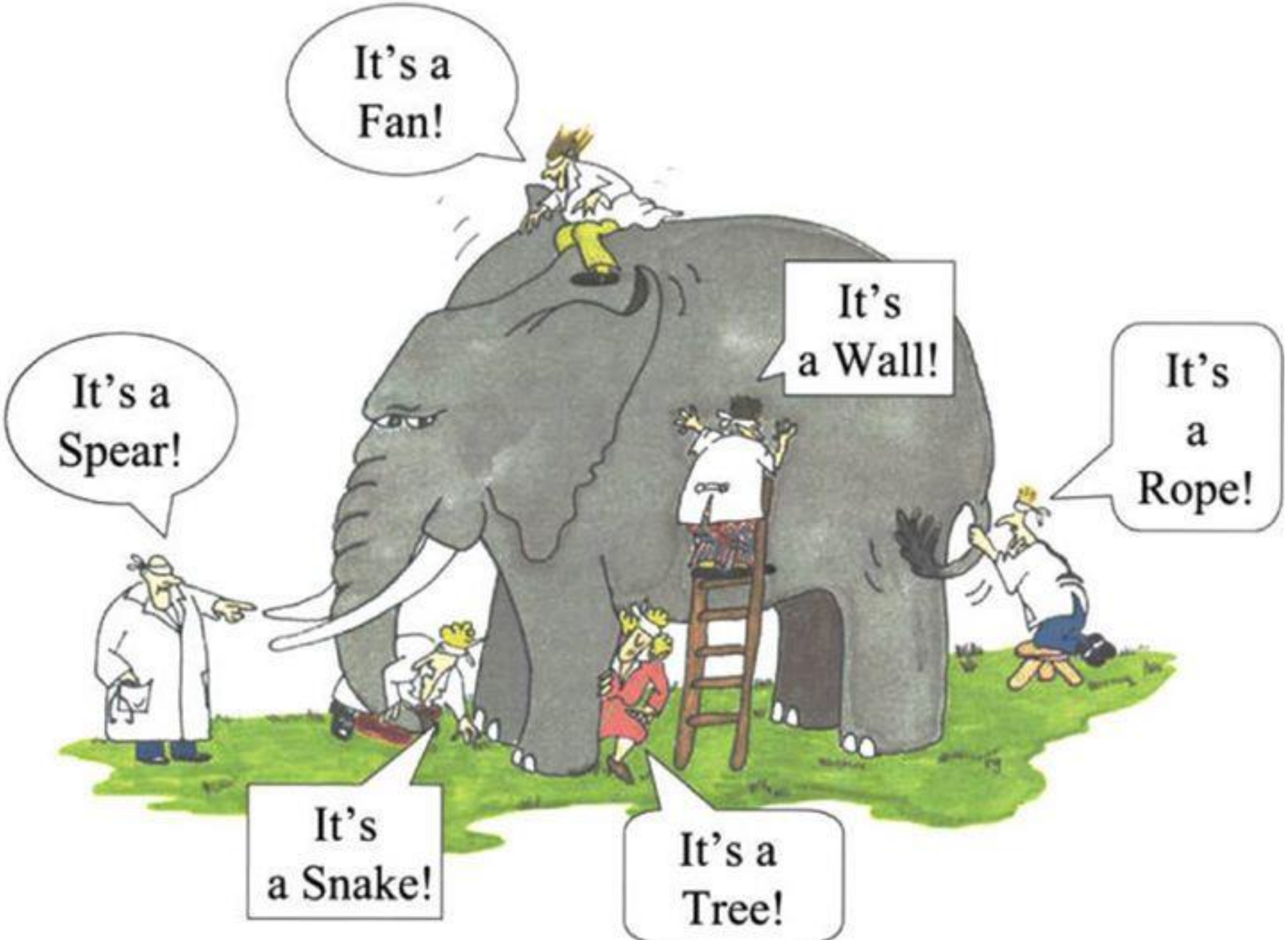
Rainfall = 260mm



A decision tree model to predict Entisols and Aridisols

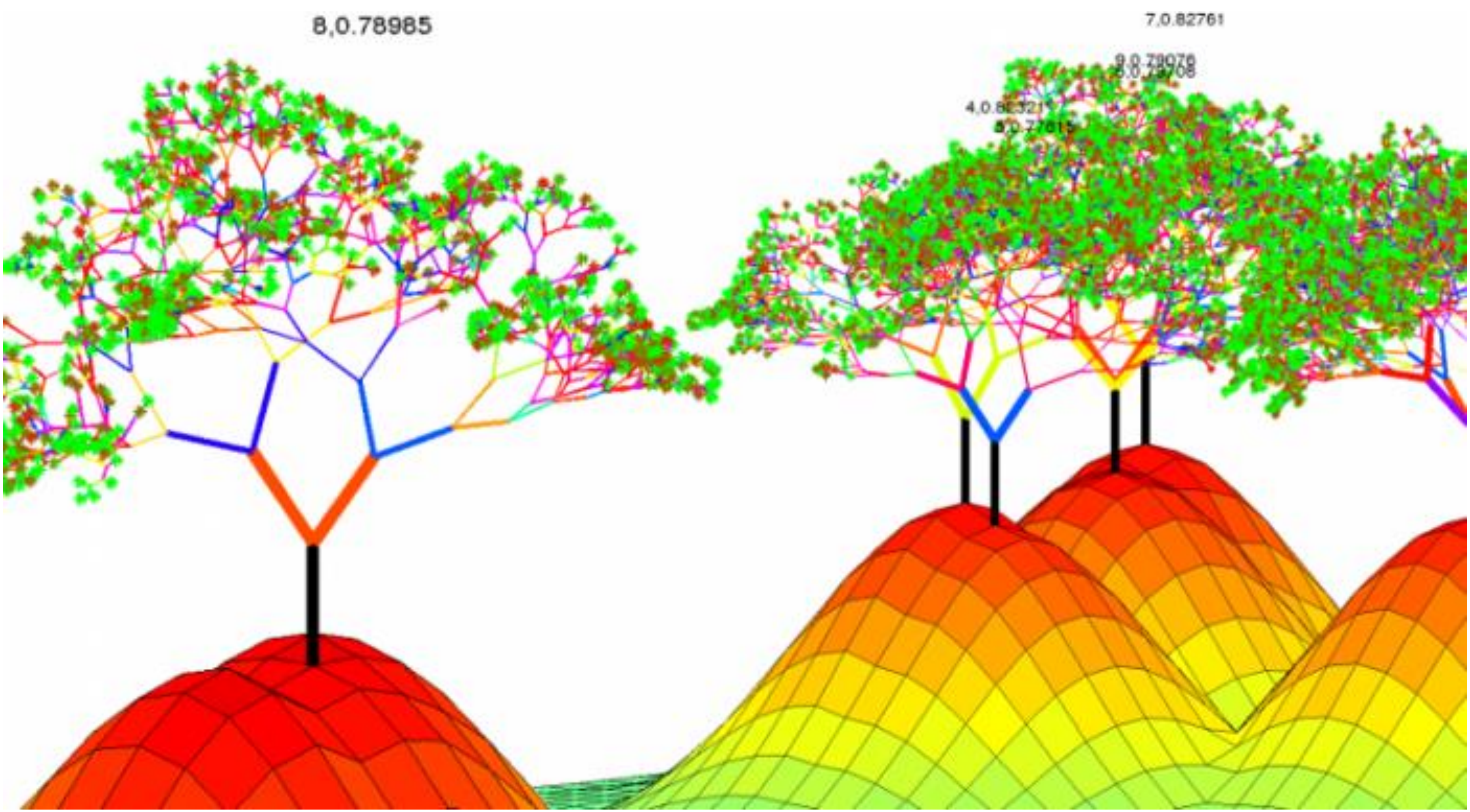
https://www.kdnuggets.com/

Ensemble Methods



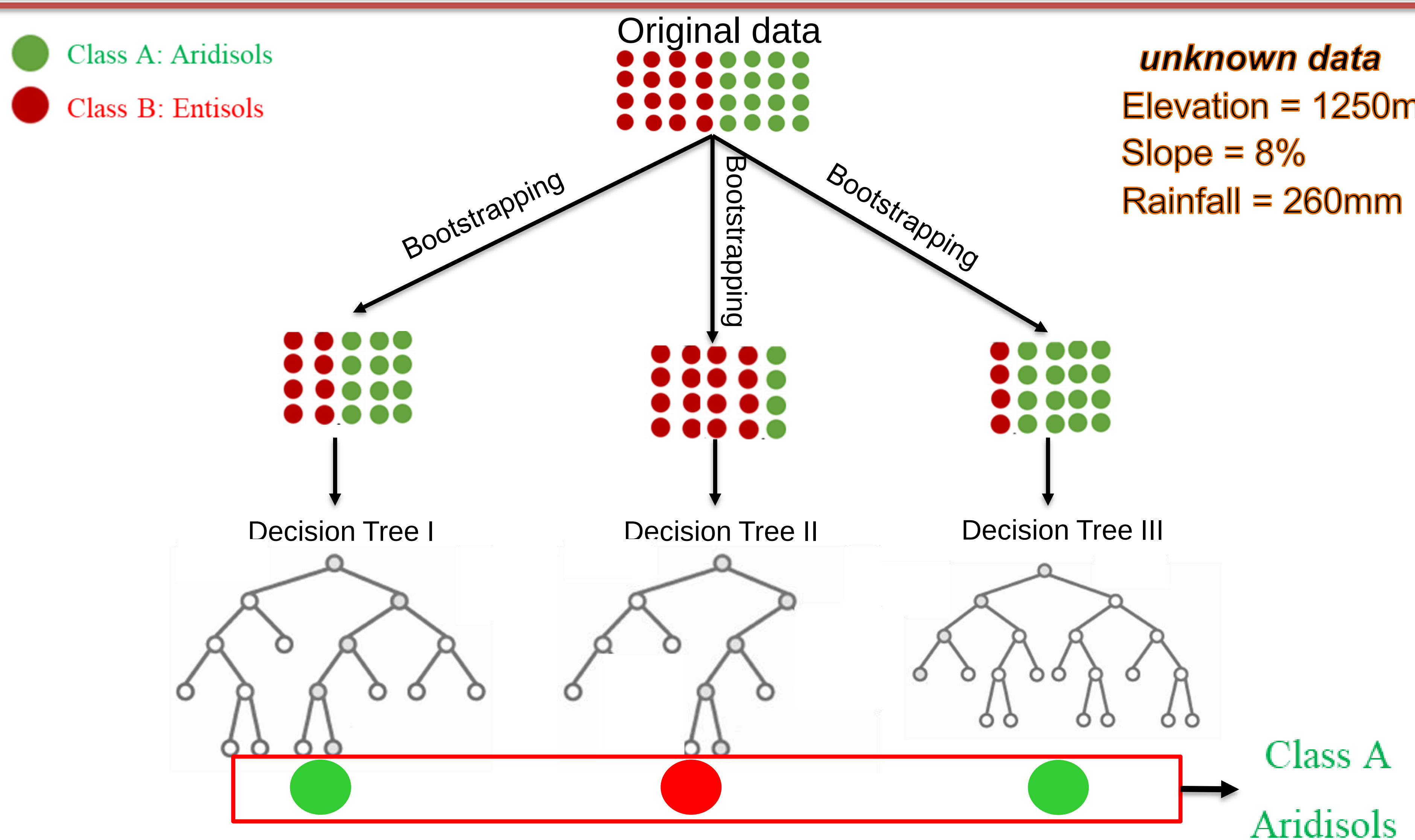
Source: Patheos.com

Random Forest



https://towardsdatascience.com/random-forest-learning-essential-understanding-1ca856a963cb

Bagging: example I



How Does It Work?

- Random Forest:

  Bagging

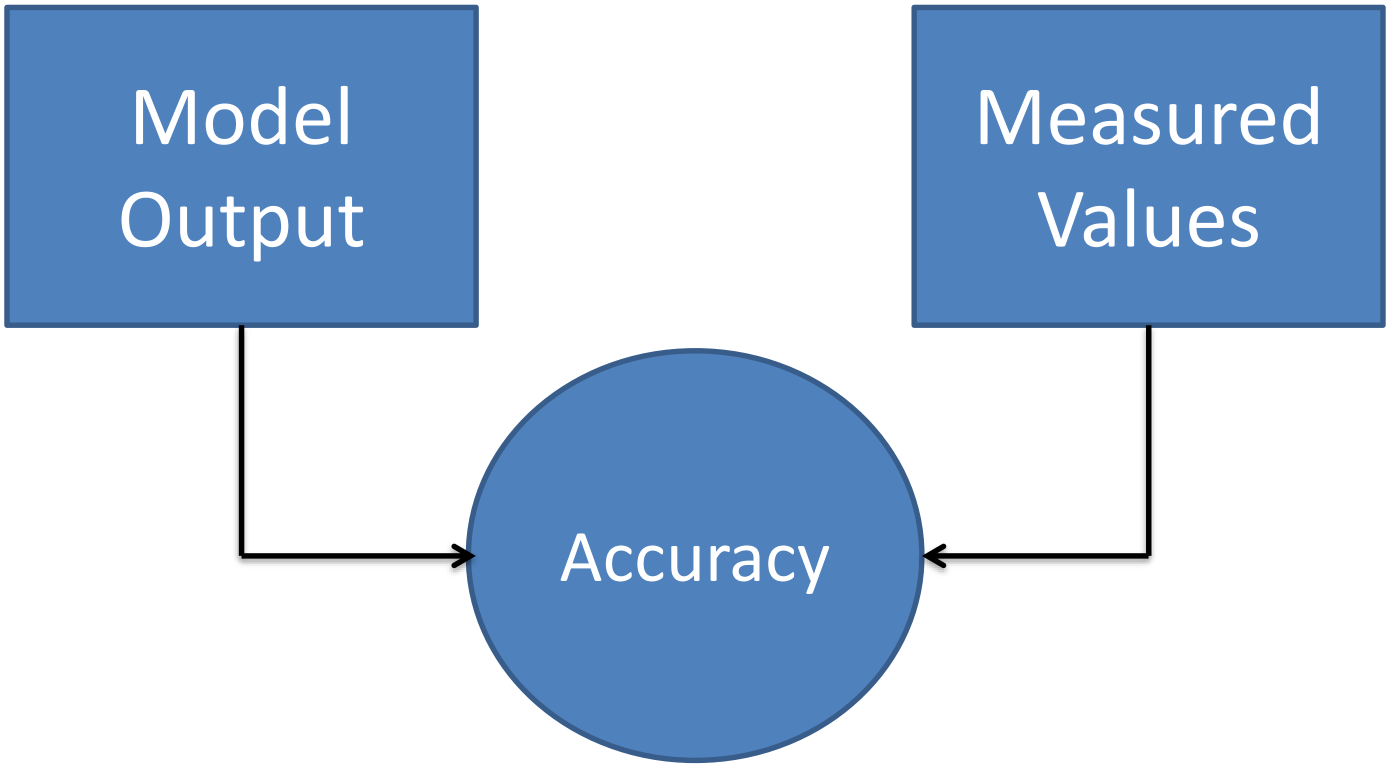
  - Draw a bootstrap sample:
    - Random selection of 2/3 of the training data; repeat *n* times.
    - Grow an unpruned tree to each bootstrap sample

  - Random predictor selection: for each split in each tree a random subset is selected from the predictor variables. The best split is chosen from among the selected predictors.

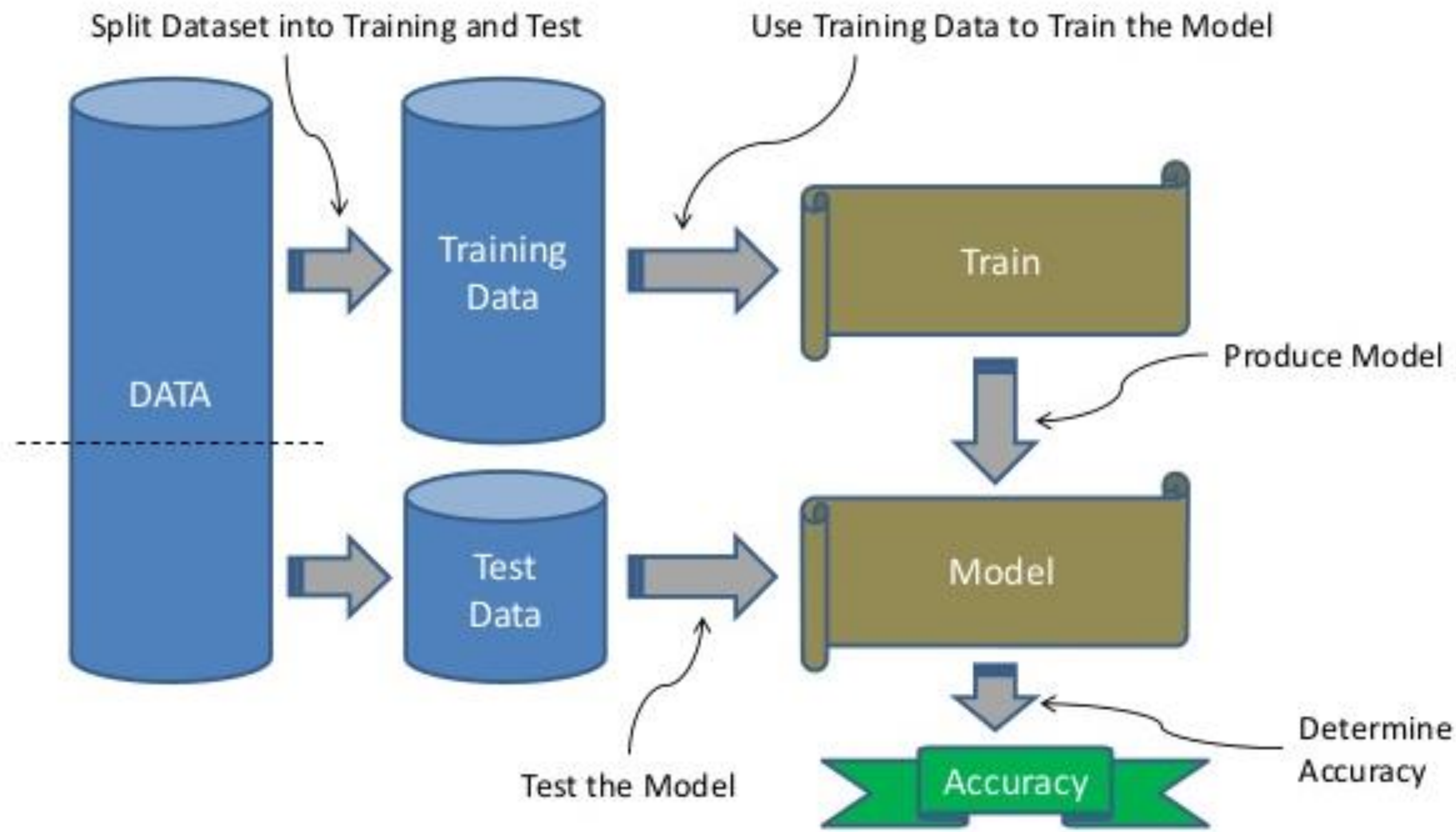
  - Predict new data by aggregating the predictions of the n trees.
    - Average for continuous variables
    - Majority vote for categorical variables

Accuracy Assessment of Models

- Assess accuracy of a model output is one of the most important steps in digital soil mapping

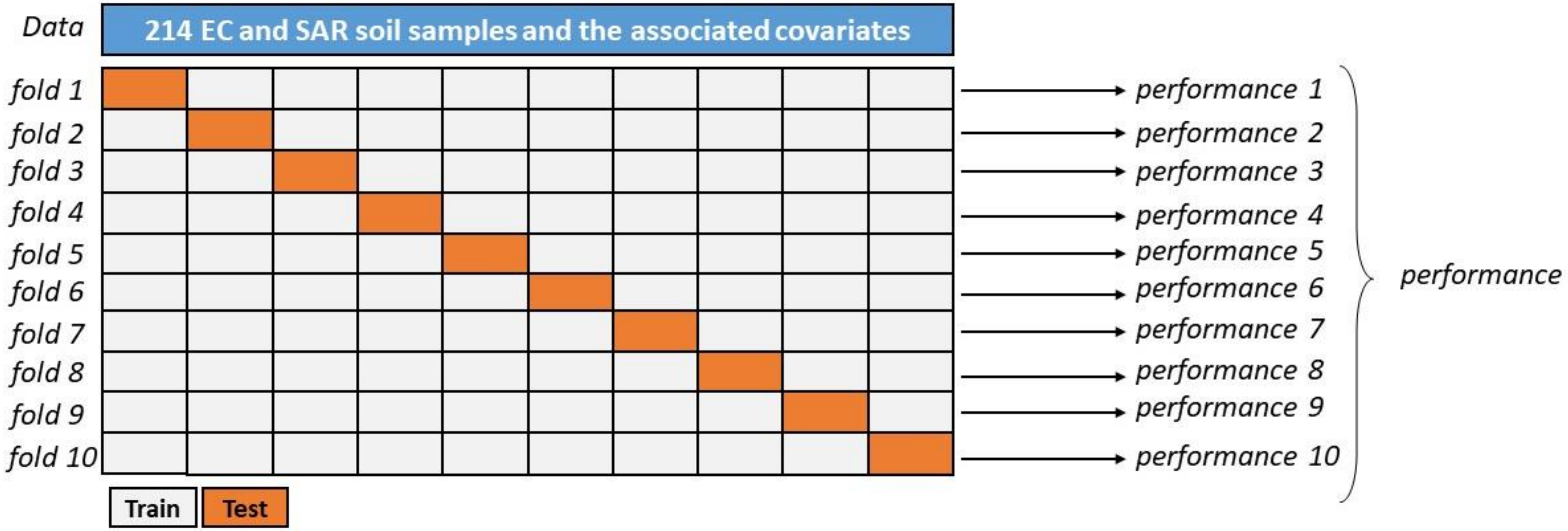


Training and Testing the Models

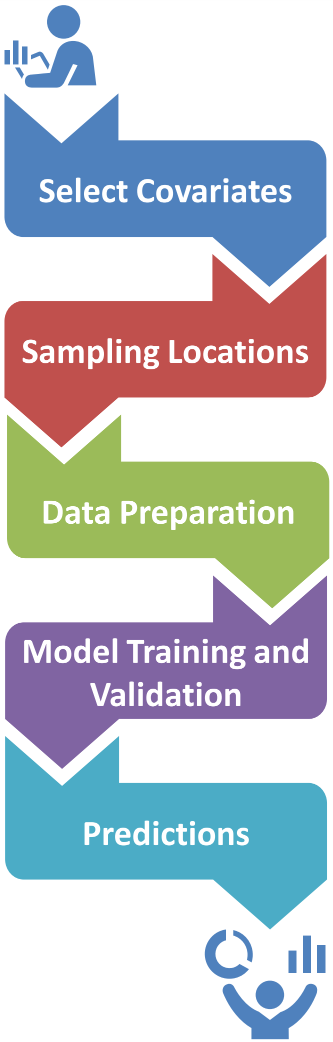


K-Folds Cross Validation

- K-Folds Cross Validation: k = 10



Take-Home Message



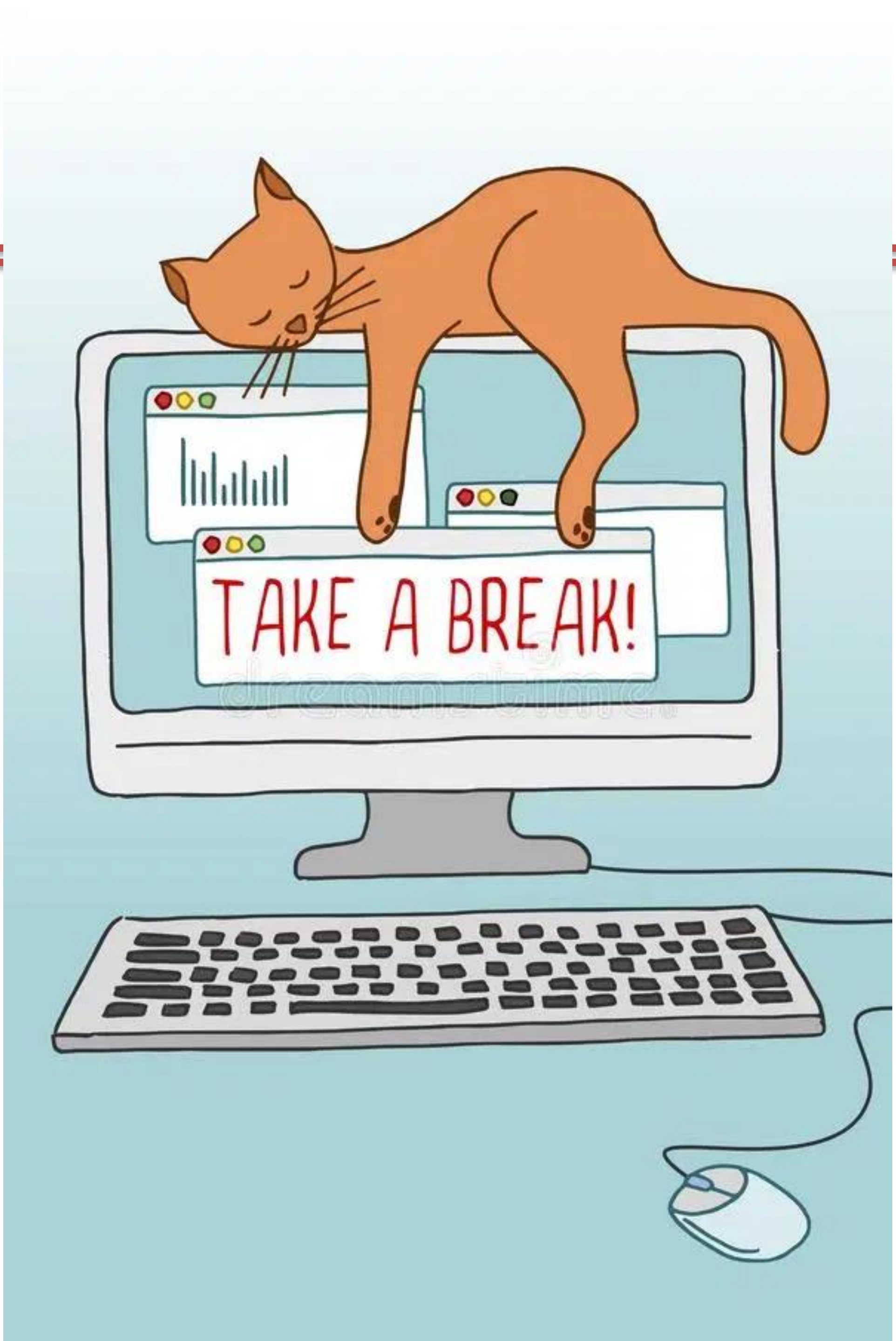
1. Environmental covariates, relevant as predictors of soil property/class, are derived from remote sensing, digital elevation, climatic datasets, …

2. Soil samples are collected at the specified locations (e.g., Latin hypercube sampling) and soil property is measured in the laboratory.

3. Intersecting the covariates with the soil point observations.

4. Machine learning models (e.g., random forest) are trained using training data, and accuracy assessment is carried out using the test data set.

5. The ML models are applied to the entire study area in order to produce a soil property/class map.


TAKE A BREAK!

Practical part

- ***1. Import and Inspect Soil Data***
- ***2. Prepare Covariates***
- ***3. Intersecting the Covariates with the Soil Data***
- ***4. Split Data: Train and Test***
- ***5. Train a Random Forest***
- ***6. Test a Random Forest***
- ***7. Covariate Importance***
- ***8. Predict Soil Map***
- ***9. Uncertainty Analysis***

Thank You For Your Attention

Ruhollah Taghizadeh
Soil Science and Geomorphology

E-mail: *ruhollah.taghizadeh-mehrjardi@mnf.uni-tuebingen.de*