

There are 76 contributors for the current view.

Digital Soil Mapping

Ruhollah Taghizadeh
Soil Science and Geomorphology

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

Content

❑ Part 1

- Conventional Soil Mapping
- Digital Soil Mapping
- Environmental Covariates

❑ Part 2

- Machine Learning Methods
- Decision Tree and Random Forest
- Model Validation and Uncertainty Analysis

Why Soil Matters

Soil functions

Soils deliver ecosystem services that enable life on Earth



Food and Agriculture Organization of the United Nations

with the support of
Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

Federal Department of Economic Affairs,
Education and Research EAER
Federal Office for Agriculture FOAG



2015
International Year of Soils
fao.org/soils-2015

Soil Degradation



Soil erosion



Land clearing



Soil salinization



Soil compaction

Soil Degradation



Dust storms form mainly as a consequence of **soil erosion** by windstorms

Increasing Population

THE GLOBAL DEMAND
FOR FOOD IS EXPECTED
TO INCREASE BY

60%
BY 2050



1 NO
POVERTY



2 ZERO
HUNGER



3 GOOD HEALTH
AND WELL-BEING



4 QUALITY
EDUCATION



5 GENDER
EQUALITY



6 CLEAN WATER,
AND SANITATION



7 AFFORDABLE AND
CLEAN ENERGY



8 DECENT WORK AND
ECONOMIC GROWTH



9 INDUSTRY, INNOVATION
AND INFRASTRUCTURE



10 REDUCED
INEQUALITIES



11 SUSTAINABLE
CITIES AND COMMUNITIES



12 RESPONSIBLE
CONSUMPTION
AND PRODUCTION



13 CLIMATE ACTION



17 PARTNERSHIPS
FOR THE GOALS



Soil maps

are an essential tool
in achieving
sustainable use of the
land

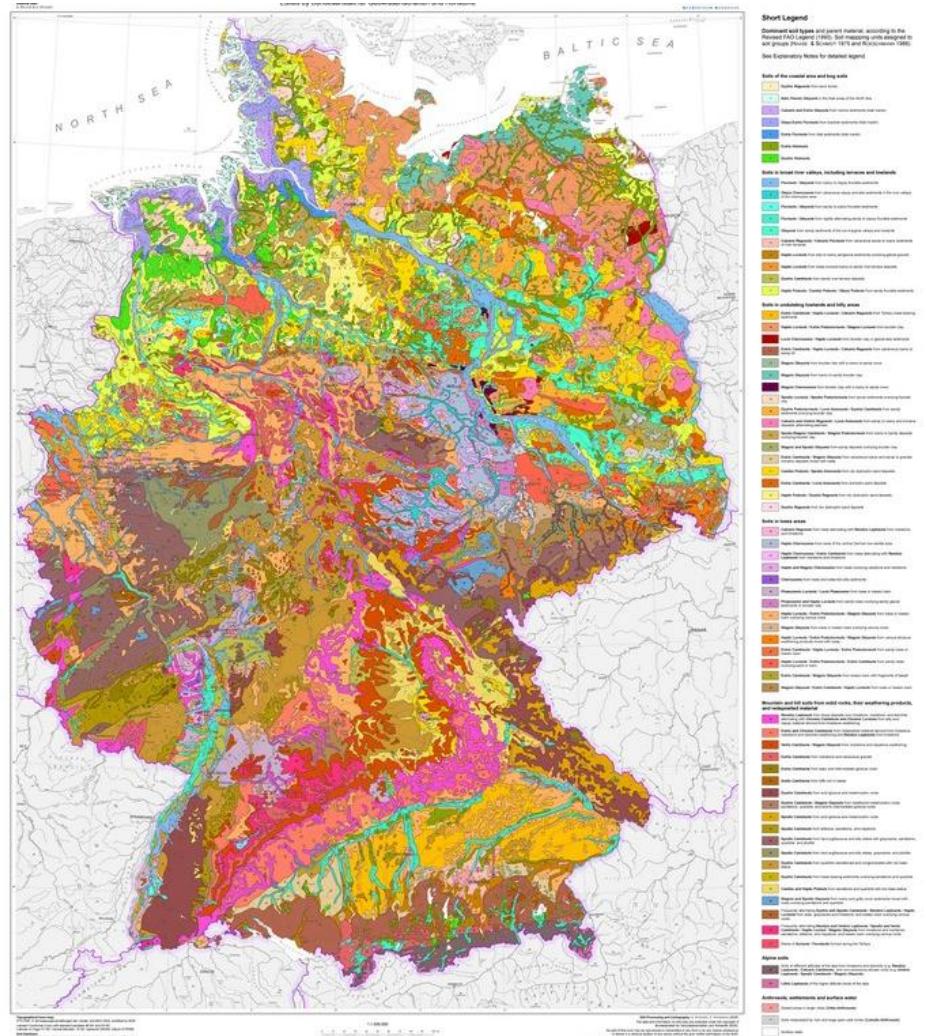
Soil Map

Definition:

Soil map is a **geographical representation** showing diversity of soil types and/or soil properties (soil pH, textures, organic matter, depths of horizons etc.) in the area of interest.

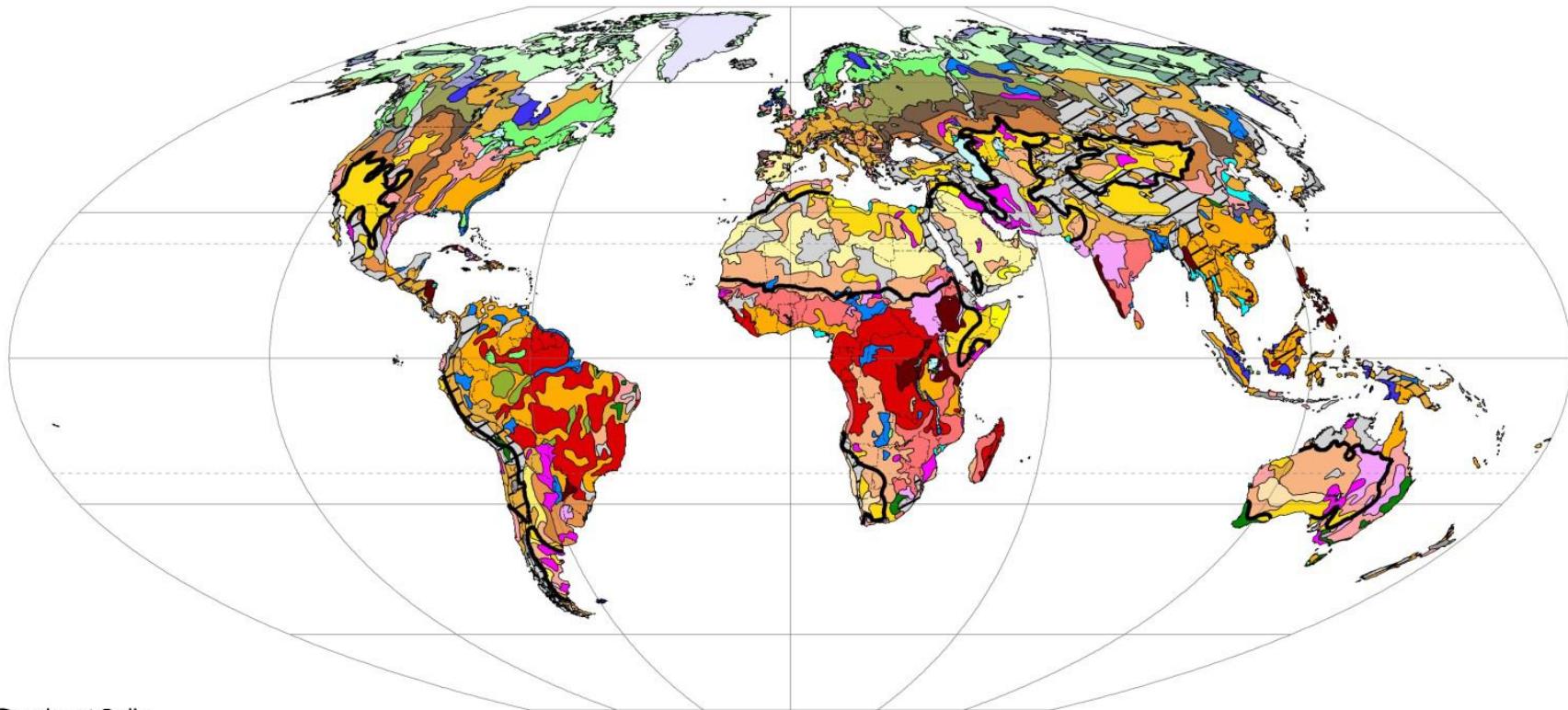
Users:

Decision makers, researchers/scientists, agronomists, farmers, etc.



Soil Map of the Federal Republic of Germany 1 : 1,000,000
(<https://www.bgr.bund.de>)

Global Soil Regions



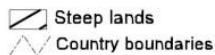
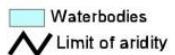
Dominant Soils

AC Acrisols, Alisols, Plinthosols
AB Albeluvicols, Luvisols
AN Andosols
AT Anthrosols
AR Arenosols
CL Calcisols, Cambisols, Luvisols
CA Calcisols, Regosols, Arenosols
CM Cambisols

CH Chernozems, Phaeozems
CR Cryosols
DU Durisols
FR Ferralsols, Acrisols, Nitisols
FL Fluvisols, Gleysols, Cambisols
GL Gleysols, Histosols, Fluvisols
GY Gypsisols, Calcisols
HR Histosols, Cryosols

HS Histosols, Gleysols
KS Kastanozems, Solonetz
LP Leptosols, Regosols
LR Leptosols, Cryosols
LX Lixisols
LV Luvisols, Cambisols
NT Nitisols
PH Phaeozems

PL Planosols
PT Plinthosols
PZ Podzols, Histosols
RG Regosols
SC Solonchaks, Solonetz
UM Umbrisols
VR Vertisols
gl Glaciers



Soil Mapping Steps



(1) study area



(2) sampling plan

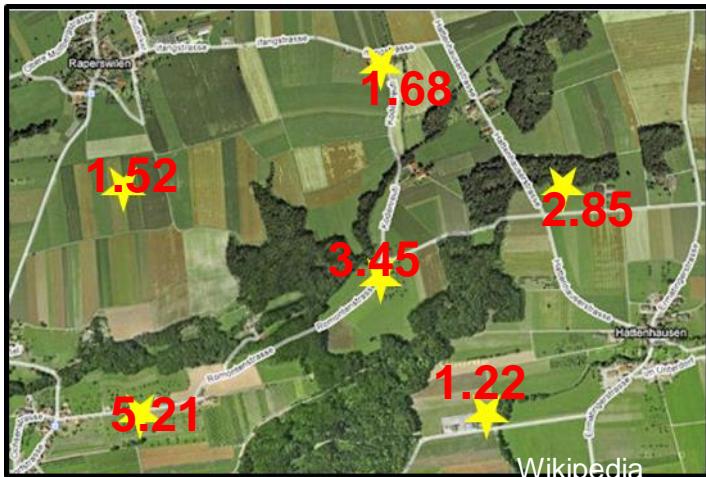


(3) field data collection



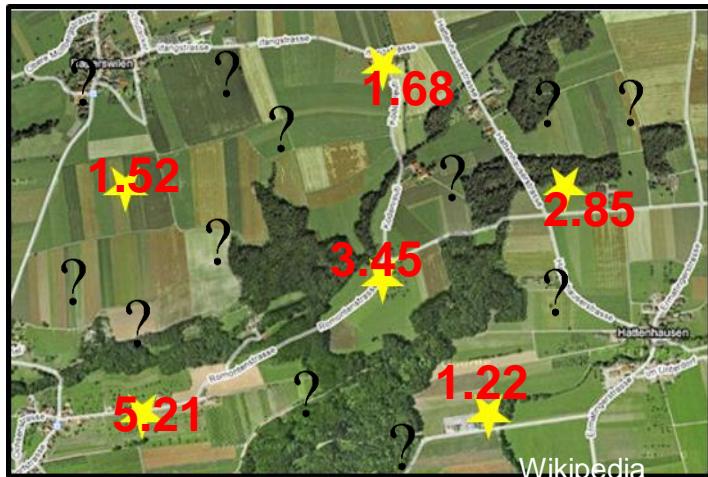
(4) soil analysis

Soil Mapping Steps



(5) soil point data

Soil Mapping Steps

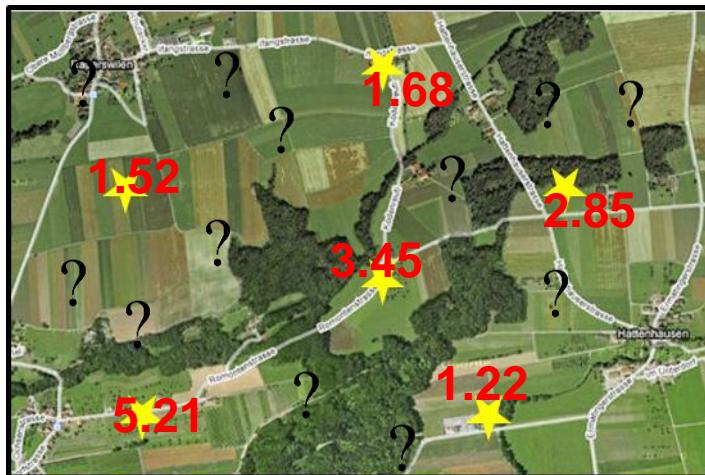


(5) soil point data



(6) create soil map

Soil Mapping Steps

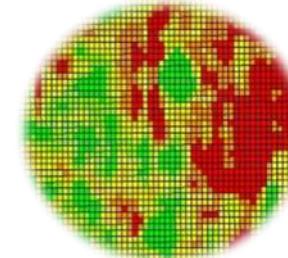


(5) soil point data

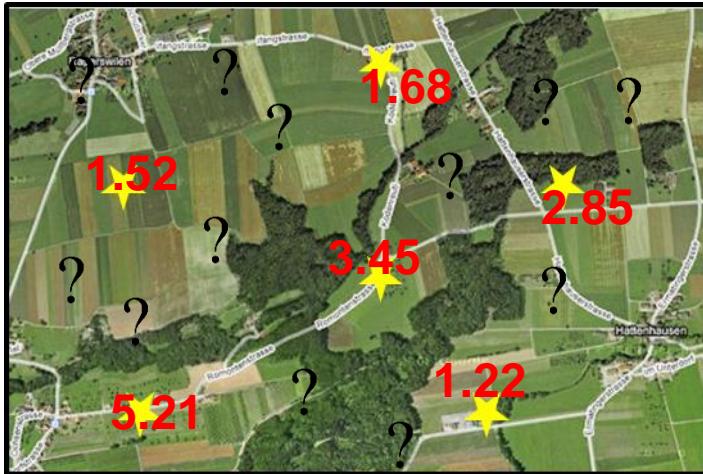


(6) create soil map

- Soil mapping
- Conventional Soil Mapping
 - Digital Soil Mapping



Soil Mapping Steps

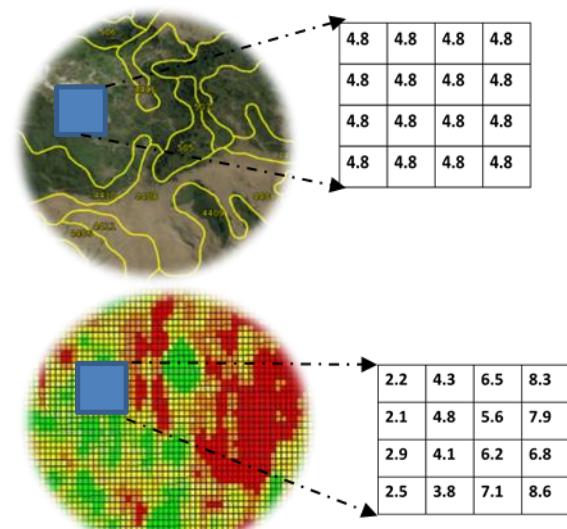


(5) soil point data

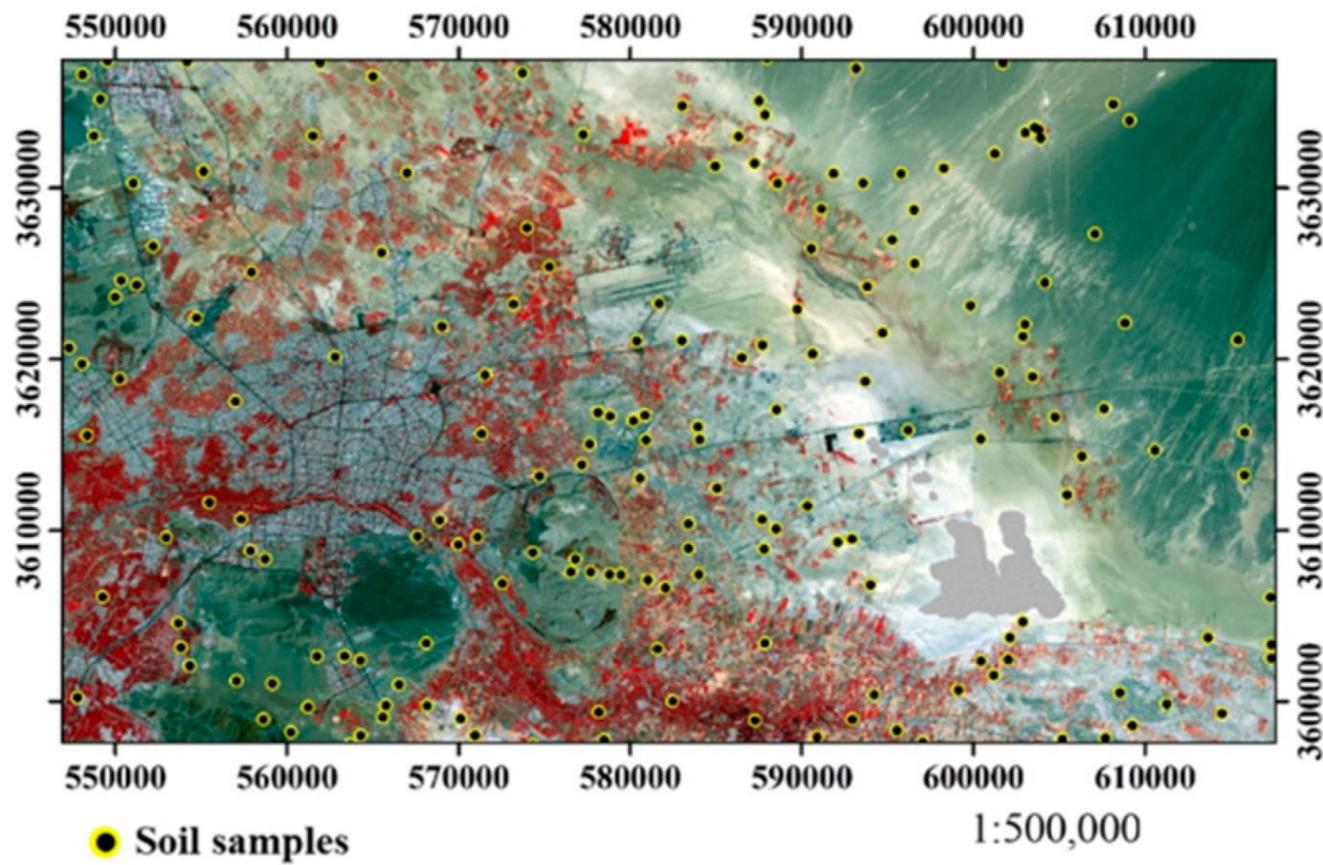


(6) create soil map

- Soil mapping
- Conventional Soil Mapping
 - Digital Soil Mapping

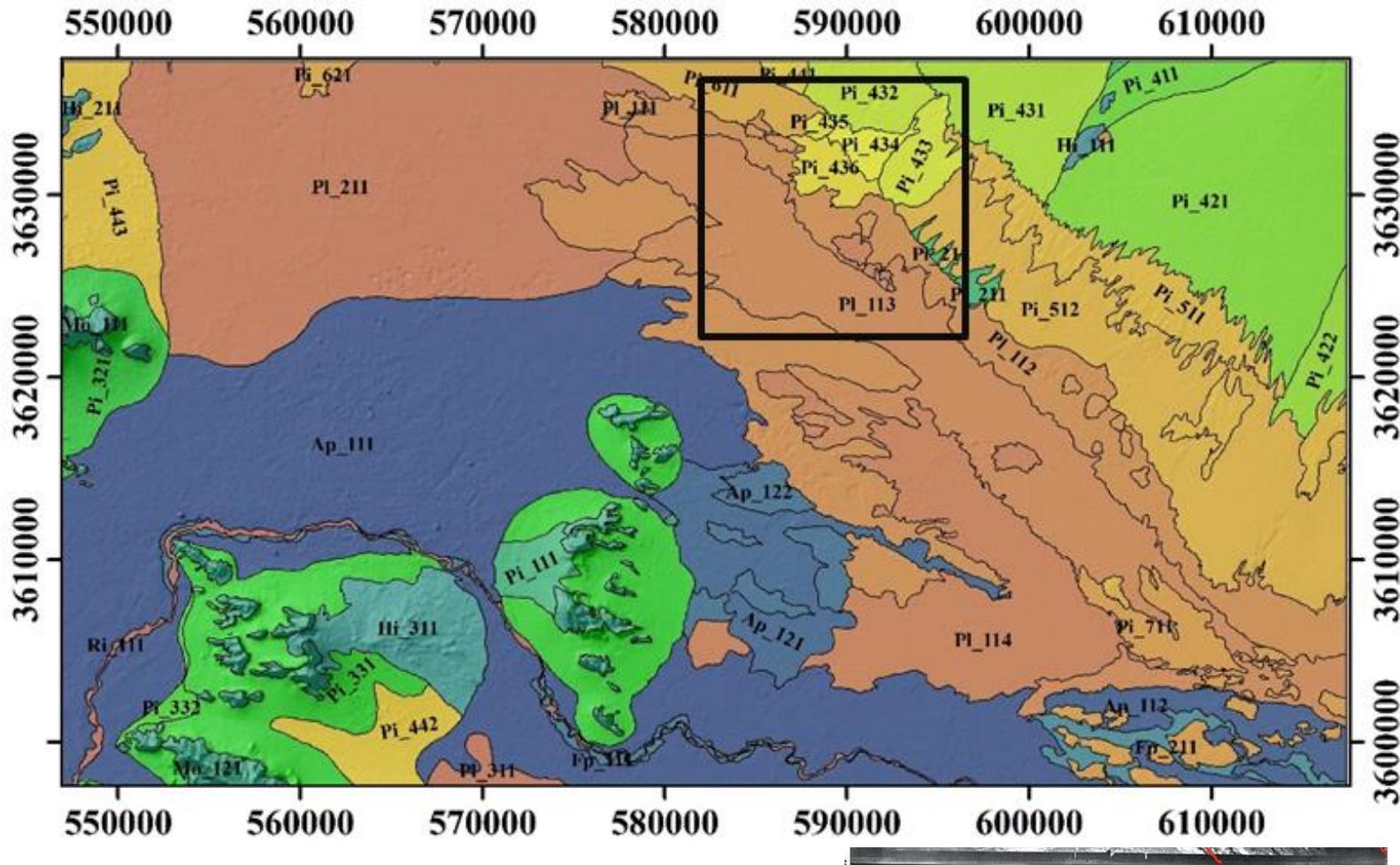


Conventional Soil Maps



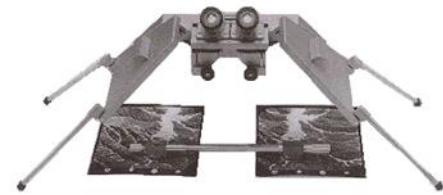
Location of study area in Iran and soil samples distribution

Conventional Soil Maps

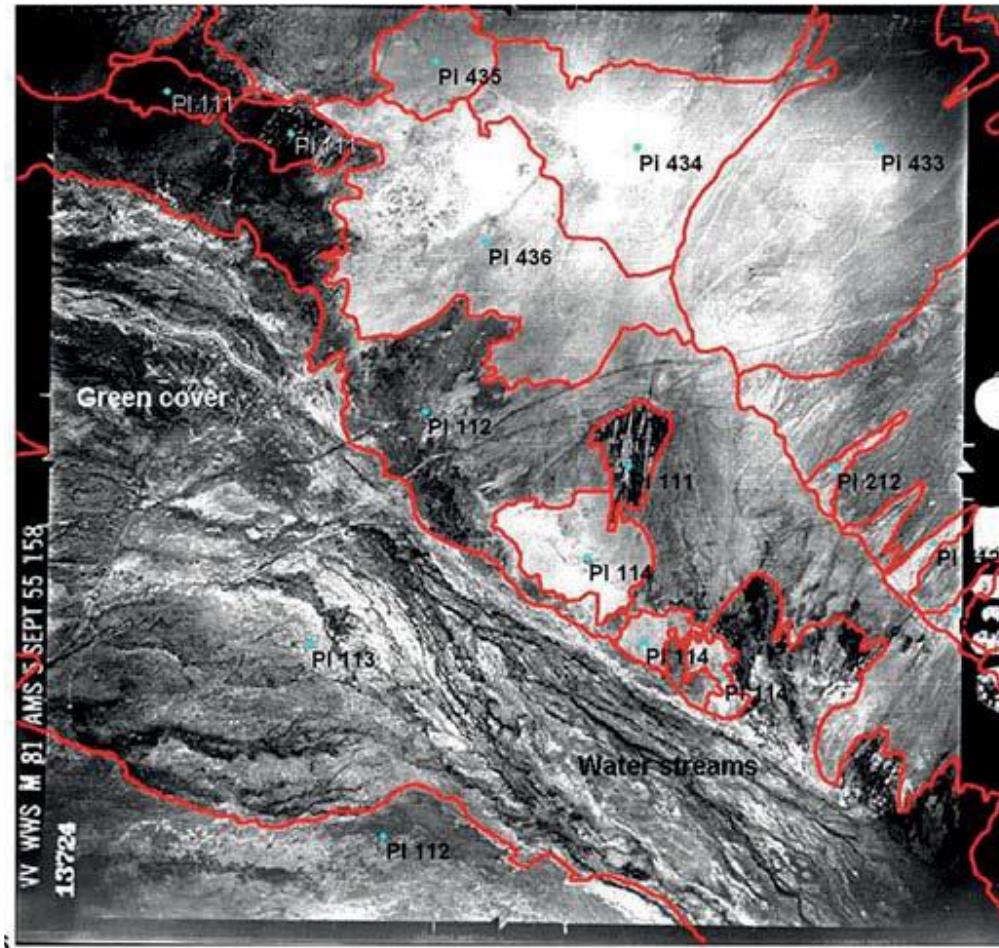


Soil map of Isfahan (Toomanian et al. 2006, Isfahan, Iran)

Conventional Soil Maps

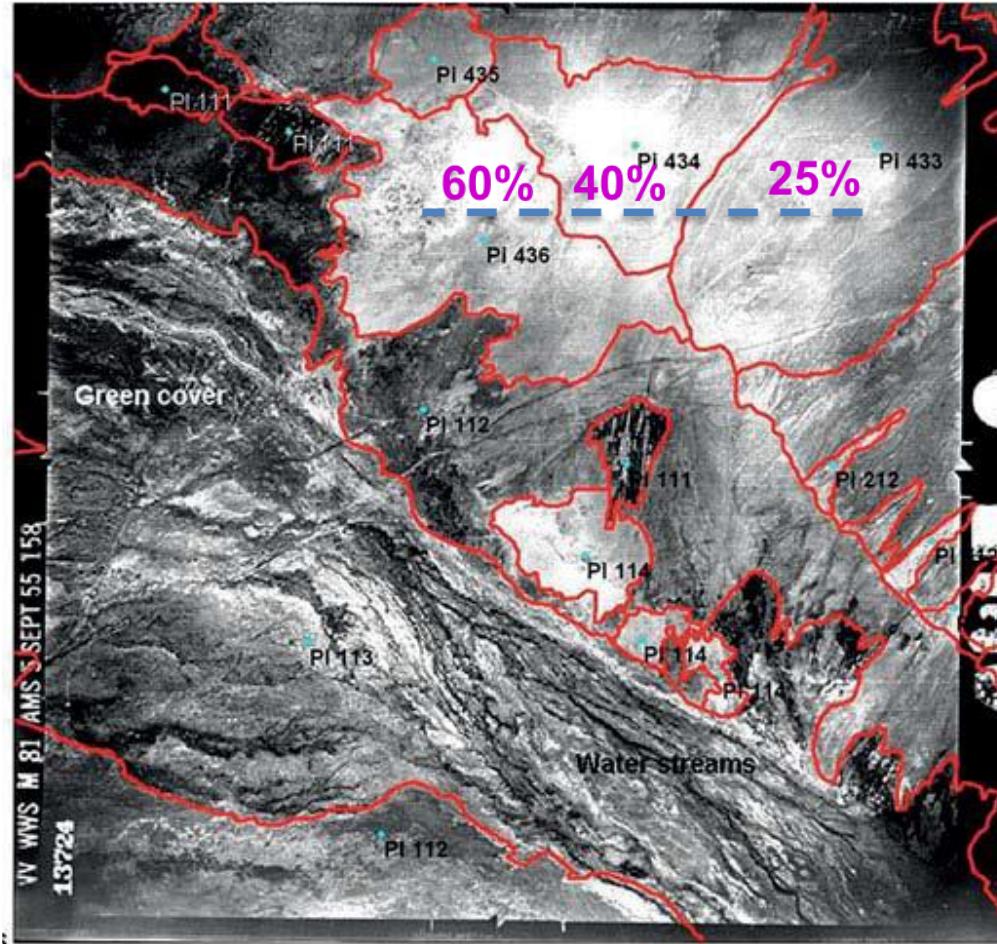


Stereoscopic viewer



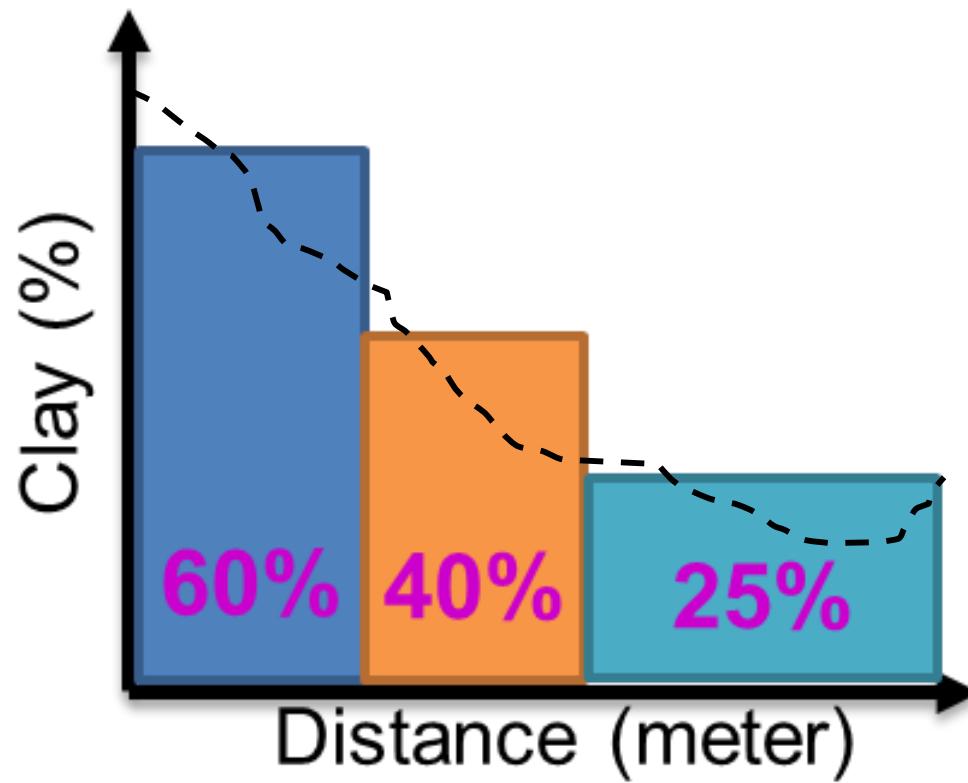
Soil units on the air photo (Toomanian et al. 2006, Isfahan, Iran)

Conventional Soil Maps



Soil units on the air photo (Toomanian et al. 2006, Isfahan, Iran)

Conventional Soil Maps



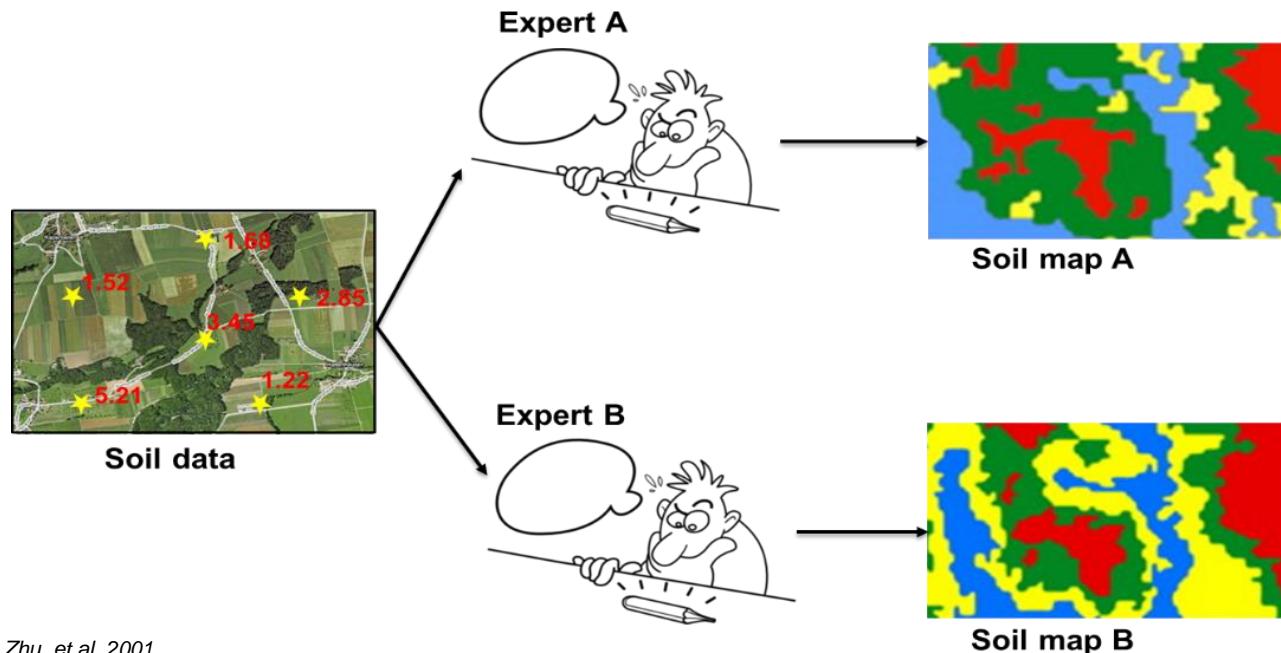
Soil spatial variability of soil clay content along the transect

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**

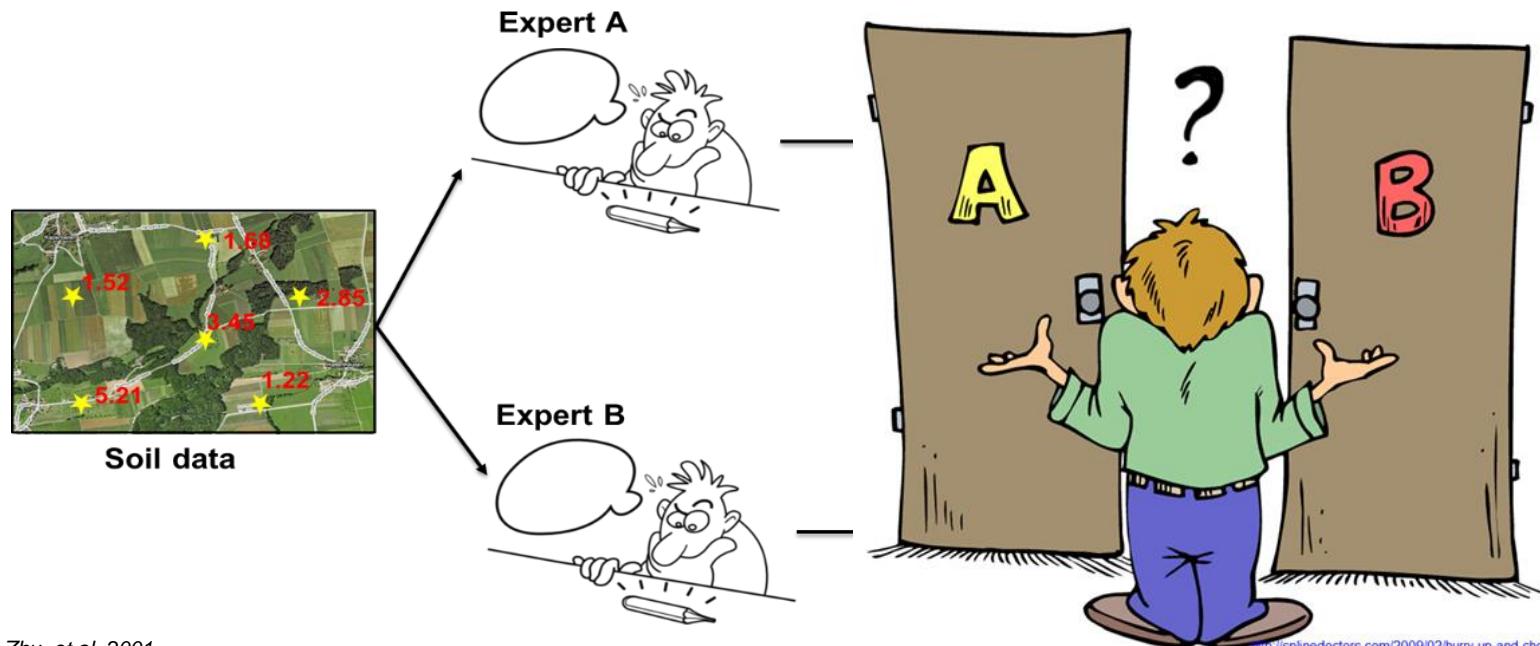


Conventional Soil Maps

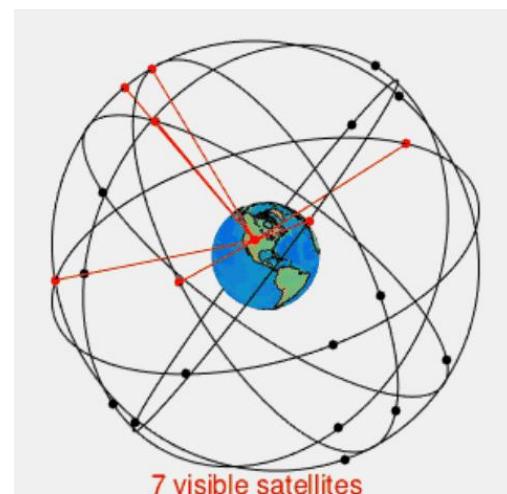
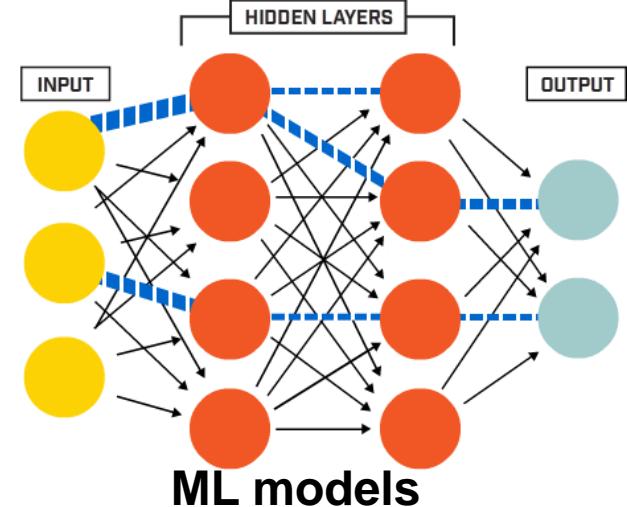
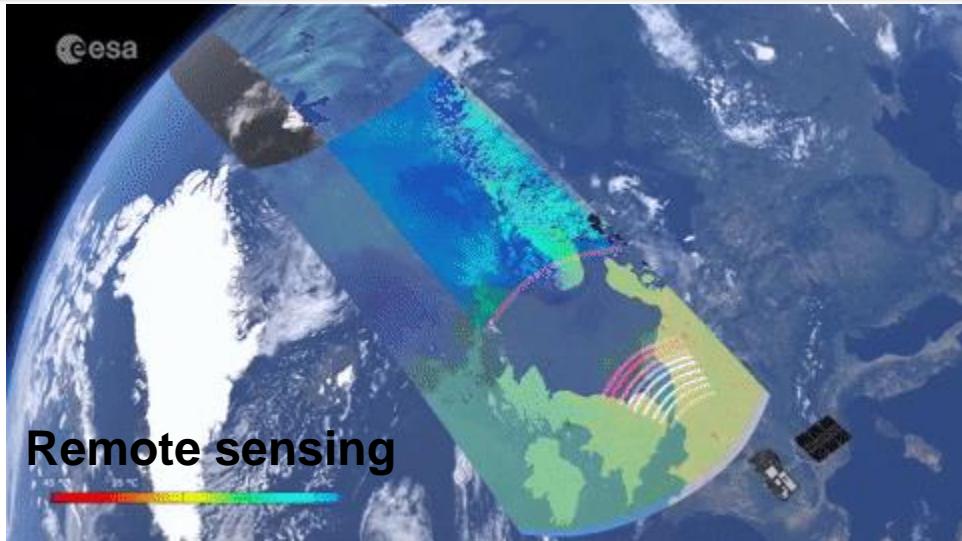


Limitations:

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**



Conventional to Digital



Digital Soil Mapping (DSM)

What is digital soil mapping?

- Digital Soil Mapping (DSM), predictive soil mapping, or pedometric mapping = computer assisted production of digital maps of soil type and soil properties, by use of mathematical and statistical **models** that combine information from **soil observations** with information contained in correlated environmental variables (**covariates**).
- DSM = the application of pedometric methods to **automate** the production of soil maps
- DSM is **NOT** digitising soil maps

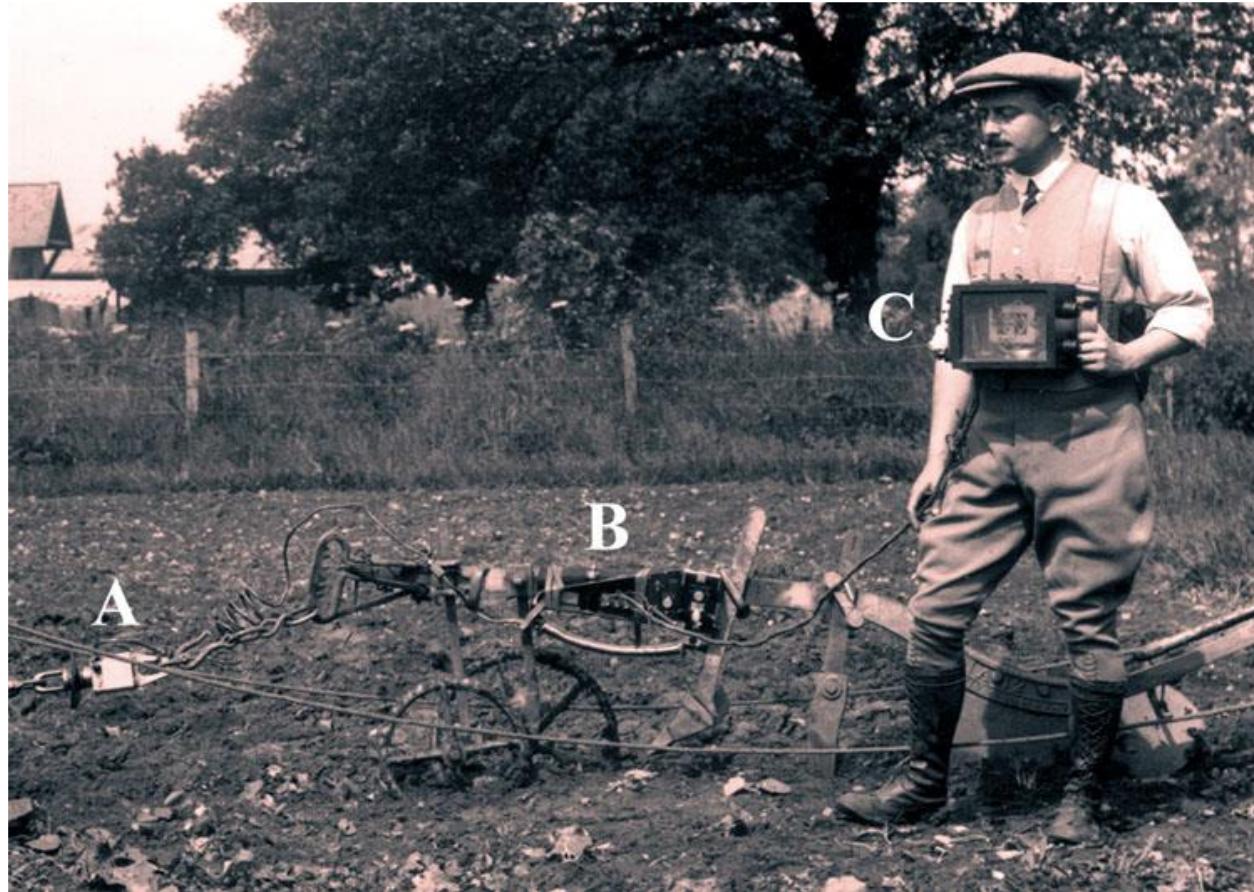
DSM: Some Historical Perspectives

In a broader sense, **automatic mapping** of soil information.

Achieved in many ways:

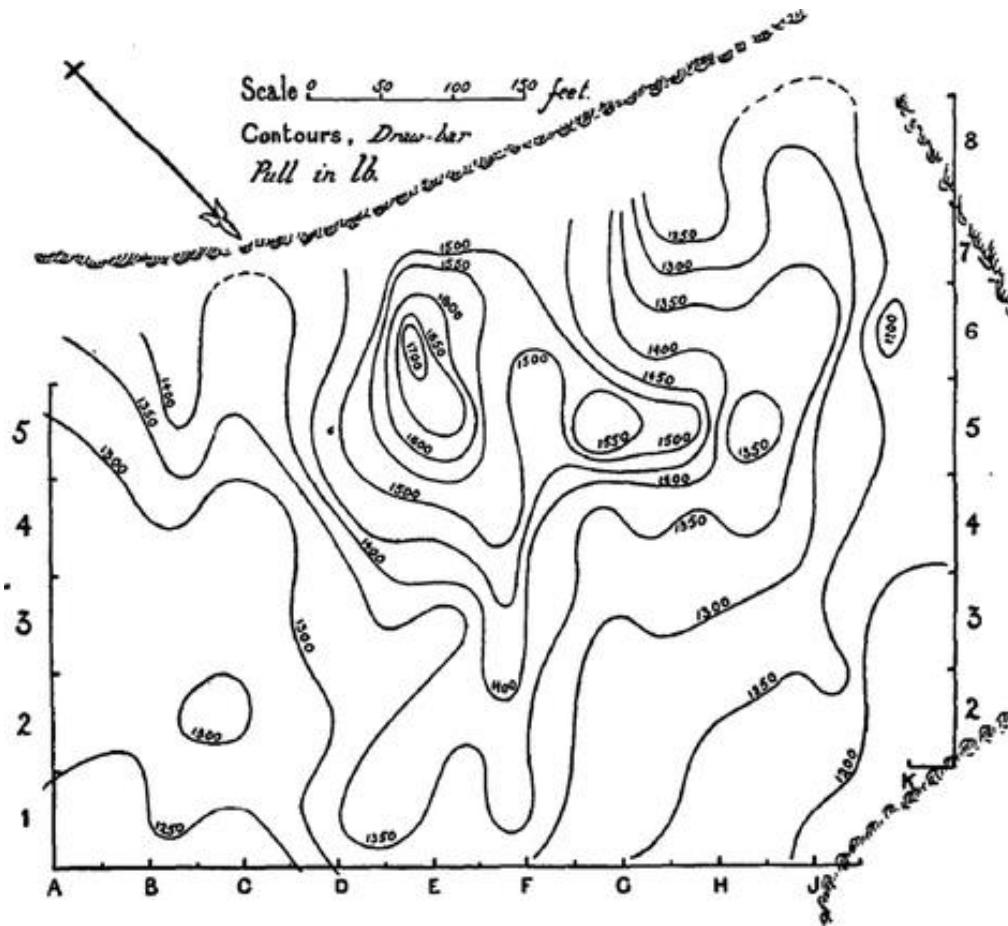
1. Collect dense observation (digital soil cartography)
2. Jenny clorpt functions
3. Interpolation (geostatistics)
4. Combination of clorpt & geostatistics

Collect Dense Soil Data ‘Automatically’



Studies in soil cultivation (Haines & Keen, 1925)

Collect Dense Soil Data ‘Automatically’



Studies in soil cultivation (Haines & Keen, 1925)

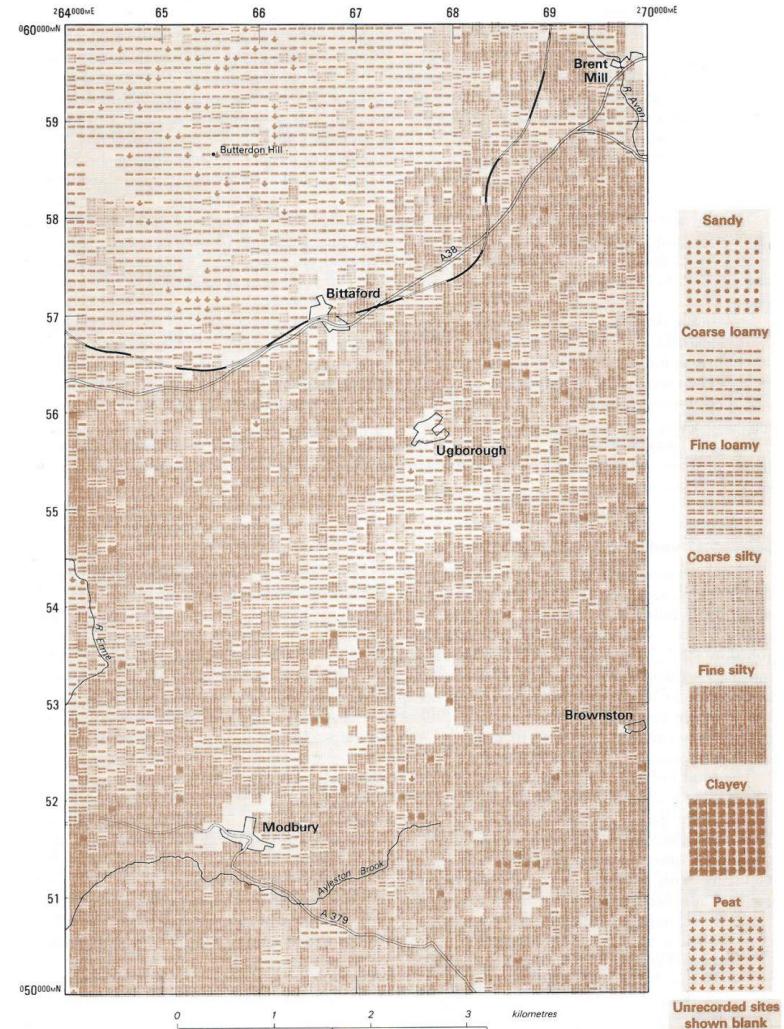
Collect Dense Soil Data ‘Automatically’



Physical 3D model of variation of soil mechanical resistance across the Broadbalk continuous wheat experiment (Haines & Keen, 1925)

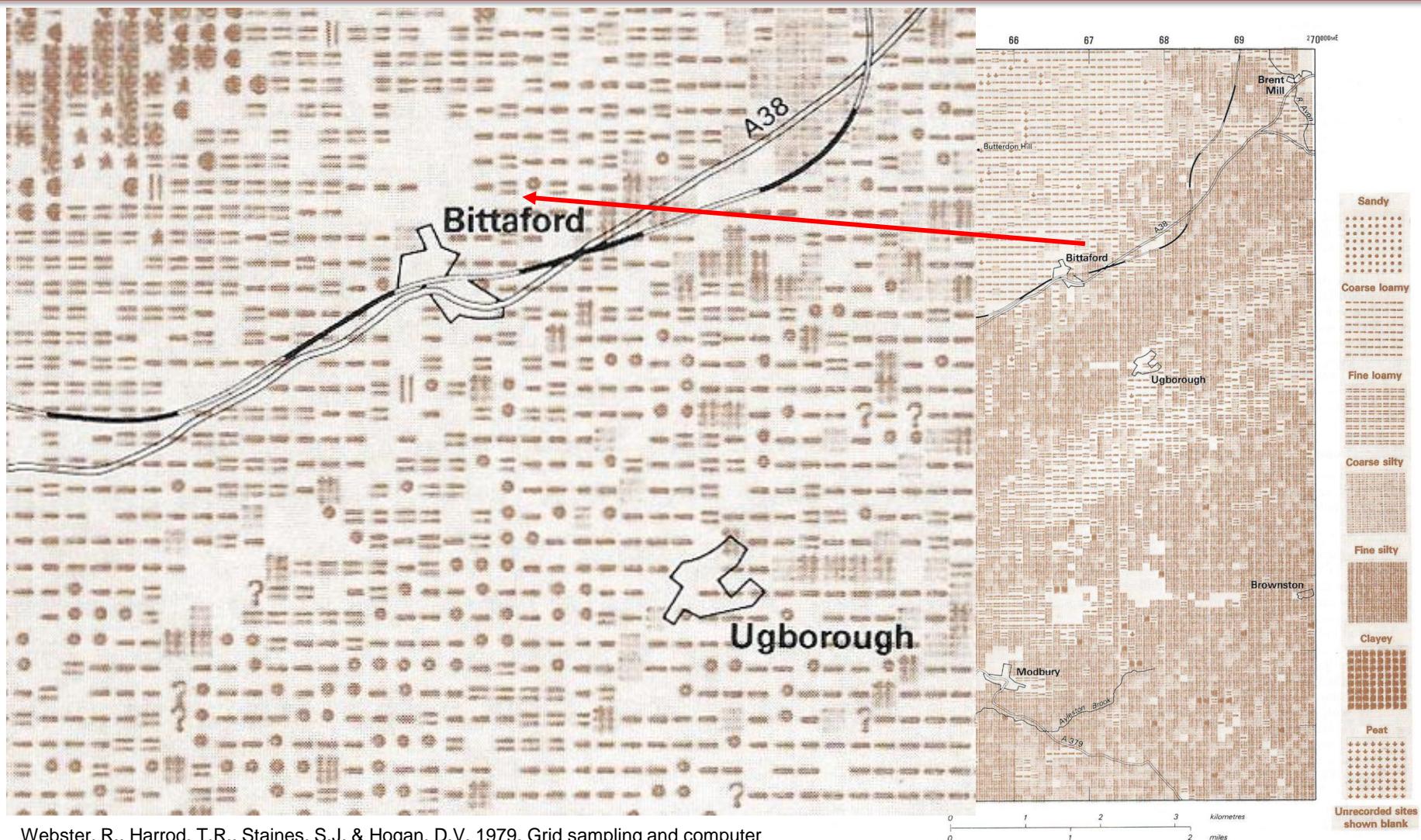
Collect Dense Soil Data ‘Manually’

- Soils of the Ivybridge area in the West of England.
- 60 km² inspected from auger borings on a 100-m square grid.
- 60 columns x 101 rows
- At each ‘pixel’ characteristics of the soil profile and site were recorded.



Webster, R., Harrod, T.R., Staines, S.J. & Hogan, D.V. 1979. Grid sampling and computer mapping of the Ivybridge area, Devon. Soil Survey Technical Monograph No 12, Soil Survey of England and Wales, Rothamsted Experimental Station, Harpenden.

Collect Dense Soil Data ‘Manually’



Webster, R., Harrod, T.R., Staines, S.J. & Hogan, D.V. 1979. Grid sampling and computer mapping of the Ivybridge area, Devon. Soil Survey Technical Monograph No 12, Soil Survey of England and Wales, Rothamsted Experimental Station, Harpenden.

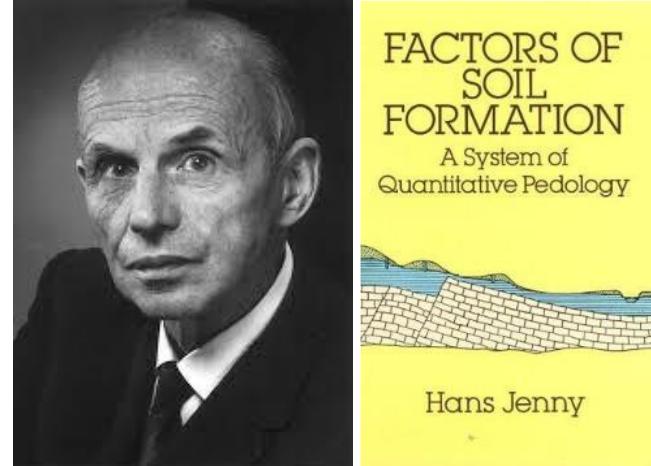
Jenny's (clorpt) Function

Soil forms under influence of **FIVE** soil forming factors:

1. climate
2. organisms
3. relief
4. parent material
5. Time

Each factor is an independent variable.

Jenny's (clorpt) equation (1941), soil forming factors:



$$S = f(c, l, o, r, p, t, \dots)$$

- Since Hans Jenny published his formulation in 1941, it has been used by researchers all over the world as a qualitative list for understanding the factors that may be important for producing the soil pattern within a region (**Conventional Soil Mapping**).

Jenny's (clorpt) Function

Soil Science Society of America Journal 

Division S-5—Soil Genesis, Morphology & Classification

Soil Attribute Prediction Using Terrain Analysis

1 Correction(s) for this article ▾

I. D. Moore  P. E. Gessler, G. A. Nielsen, G. A. Peterson

First published: 01 March 1993 | <https://doi.org/10.2136/sssaj1993.03615995005700020026x> | Citations: 86

 PDF  TOOLS  SHARE

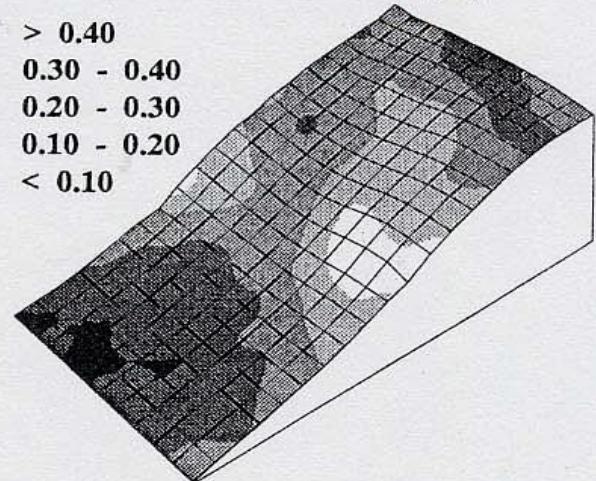
Abstract

This study is based on the hypothesis that catenary soil development occurs in many landscapes in response to the way water moves through and over the landscape. Furthermore, terrain attributes can characterize these flow paths and, therefore, soil attributes. Significant correlations between quantified terrain attributes and measured soil attributes were found on a 5.4-ha toposequence in Colorado. Slope and wetness index were the terrain attributes most highly correlated with surface soil attributes measured at 231 locations on a 15.24-m grid. Individually, they accounted for about one-

Moore et al. (1993) gave the first two-dimensional example using a set of terrain attributes derived from a digital elevation model on a 15-m grid to predict continuous soil properties such as A horizon thickness and p for a small catchment in Colorado.

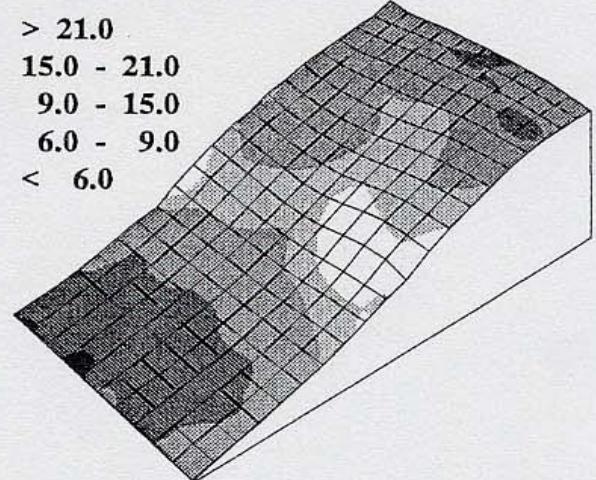
(b) Predicted A Horizon (m)

-  > 0.40
-  0.30 - 0.40
-  0.20 - 0.30
-  0.10 - 0.20
-  < 0.10



(d) Predicted Phosphorus (mg kg^{-1})

-  > 21.0
-  15.0 - 21.0
-  9.0 - 15.0
-  6.0 - 9.0
-  < 6.0



Interpolation (geostatistics)

- Neighbourhood law
- Predict values at unknown locations using values at measured locations
- Many interpolation methods: **kriging**, IDW, etc

Irrigation Science 1, 197 – 208 (1980)

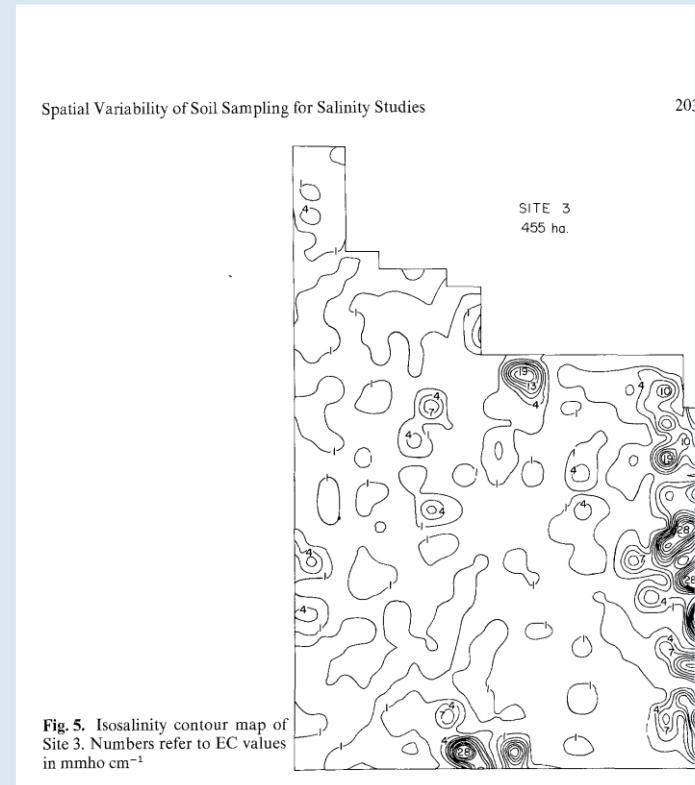
**Irrigation
Science**
© by Springer-Verlag 1980

**Spatial Variability of Soil Sampling
for Salinity Studies in Southwest Iran**

S. Hajrasuliha, N. Baniabbassi, J. Metthey, and D. R. Nielsen¹

Received January 21, 1980

Summary. This study addresses the technology for soil sampling of large agricultural fields which are inherently variable in both space and time. Three several hundred ha fields in southwest Iran initially sampled on an arbitrarily selected grid of 80 m to ascertain soil salinity levels were analyzed using both geostatistical and classical statistical methods. The results from two fields showed that the variance structure of the salinity observations were spatially dependent, and hence, geostatistical techniques allowed best linear unbiased estimates of salinity values interpolated between spatially observed sampling locations to yield contour lines of isosalinity. In the third field, salinity observations were found to be spatially independent, and hence, were analyzed by classical methods to yield the number of soil samples necessary to observe the fiducial limits within which it is expected that the true mean salinity exists for given levels of probability.





I will explain later

Geostatistics + clorpt

- In this approach 'clorpt' is used to predict the soil property of interest from environmental variables and kriging is used on the residuals (Regression Kriging).



Geoderma 67 (1995) 215–226

GEODERMA

Further results on prediction of soil properties from
terrain attributes: heterotopic cokriging and
regression-kriging

I.O.A. Odeh ^{a,b,*}, A.B. McBratney ^a, D.J. Chittleborough ^b

^a Co-operative Research Centre for Sustainable Cotton Production, Department of Agricultural Chemistry
and Soil Science, The University of Sydney, Sydney, N.S.W. 2006, Australia

^b The University of Adelaide, Department of Soil Science, Waite Campus, Glen Osmond, S.A. 5064, Australia

Received 1 February 1994; accepted 18 January 1995

Introducing the SCORPAN Model



Available online at www.sciencedirect.com



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;

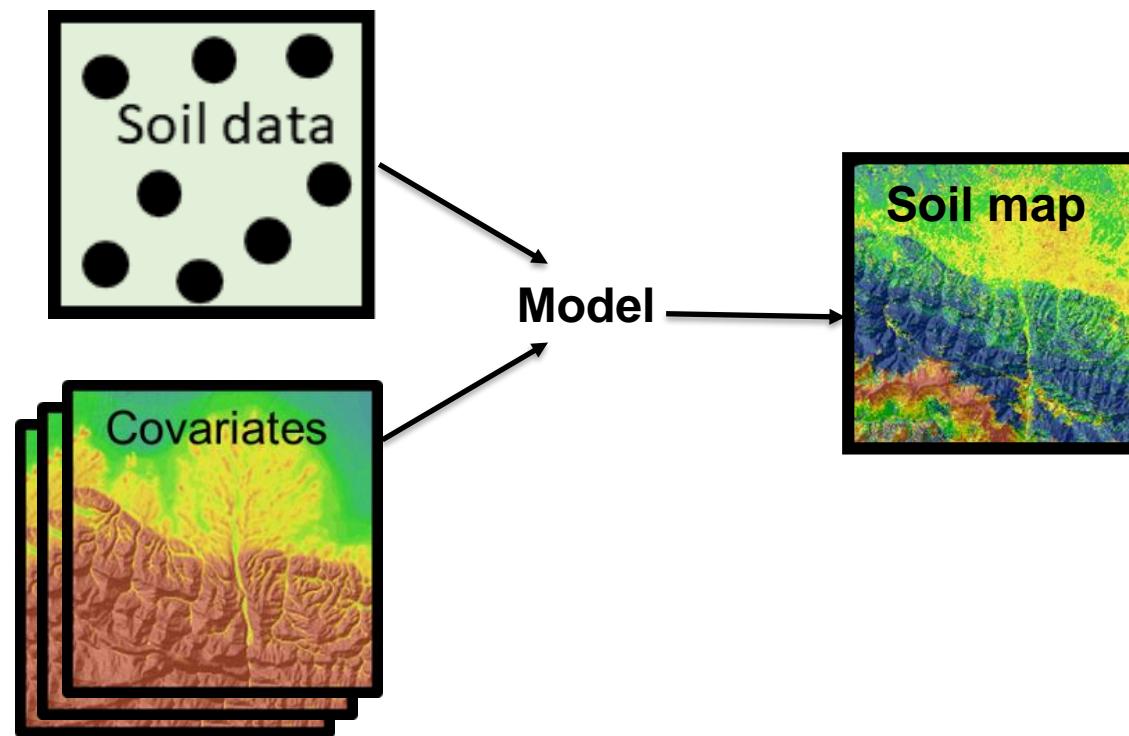
ε : auto-correlated random spatial variation.

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

SCORPAN model

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

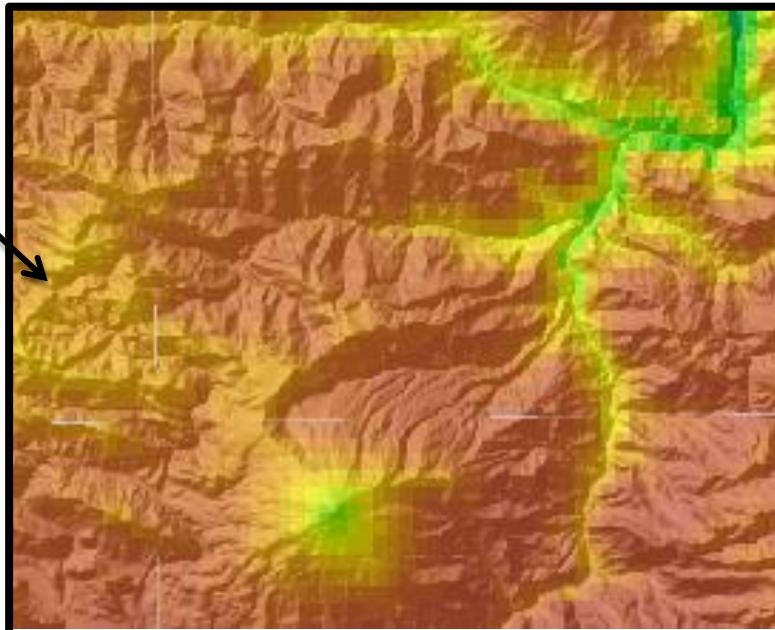
Soil data Model Covariates spatially dependent residuals



SCORPAN model

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

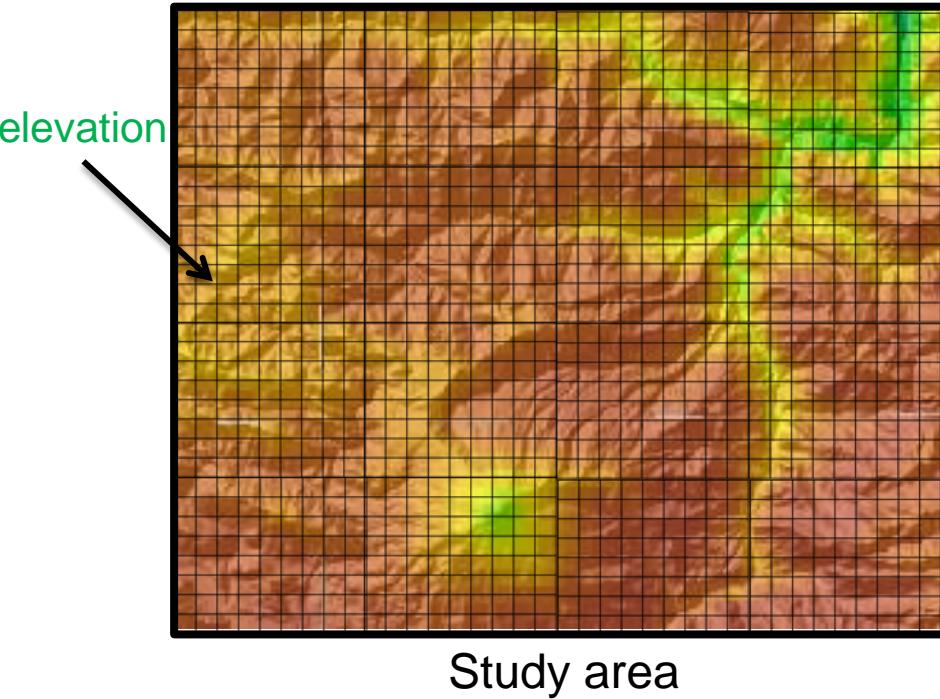
Soil data Model Covariates



SCORPAN model

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

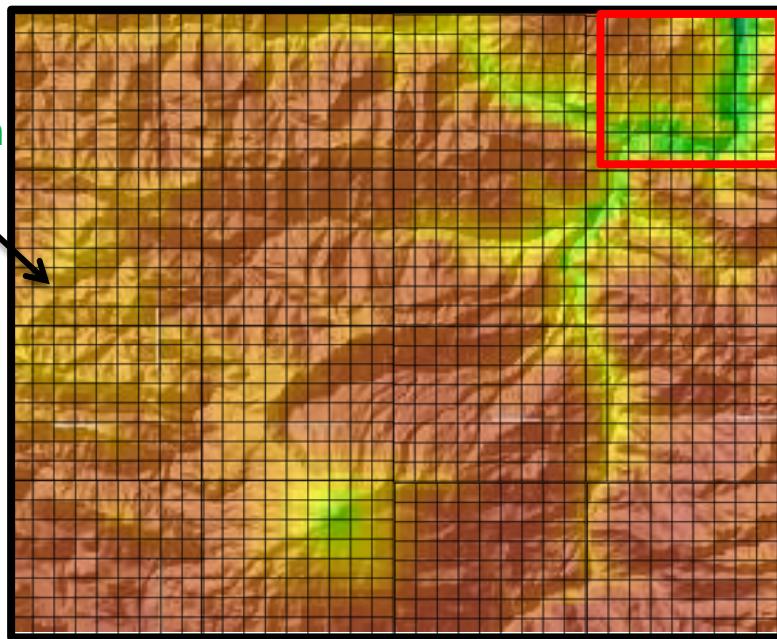
Soil data Model Covariates



SCORPAN model

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

Soil data Model Covariates

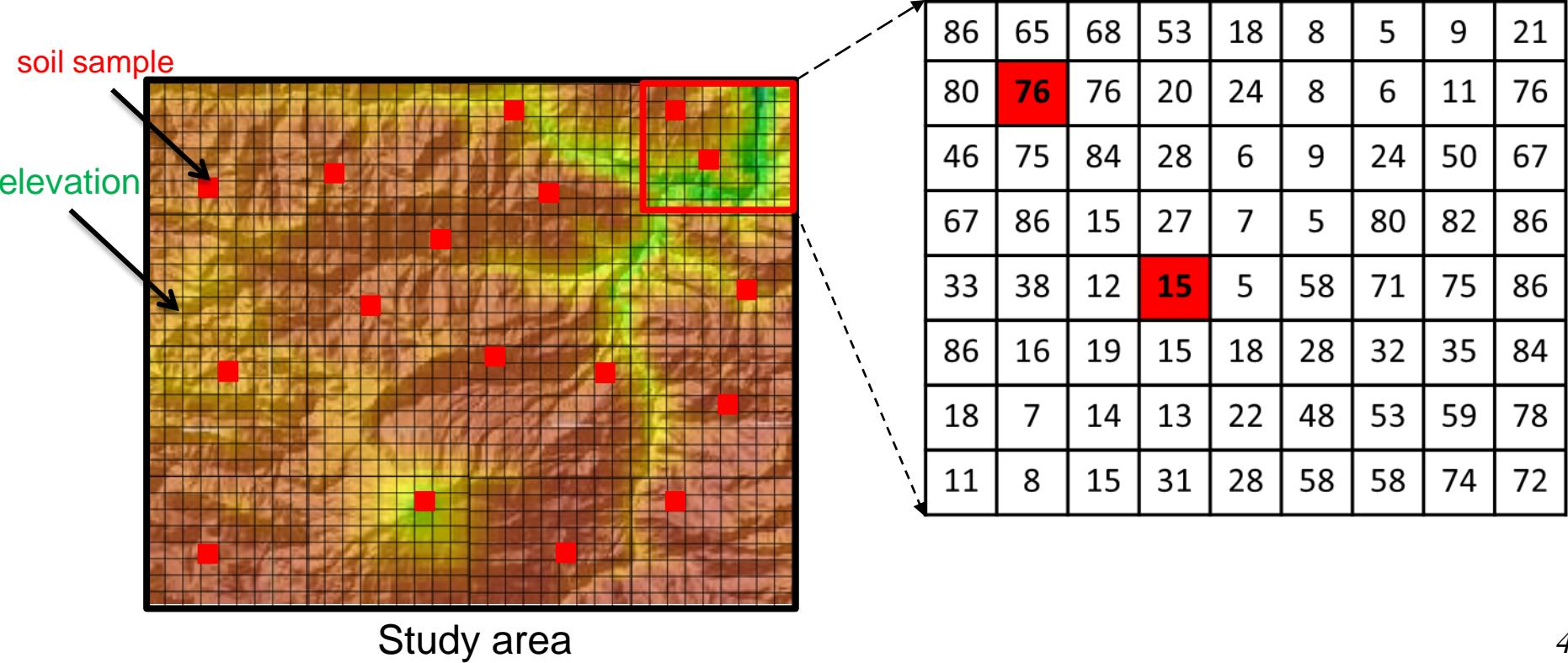


86	65	68	53	18	8	5	9	21
80	76	76	20	24	8	6	11	76
46	75	84	28	6	9	24	50	67
67	86	15	27	7	5	80	82	86
33	38	12	15	5	58	71	75	86
86	16	19	15	18	28	32	35	84
18	7	14	13	22	48	53	59	78
11	8	15	31	28	58	58	74	72

SCORPAN model

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

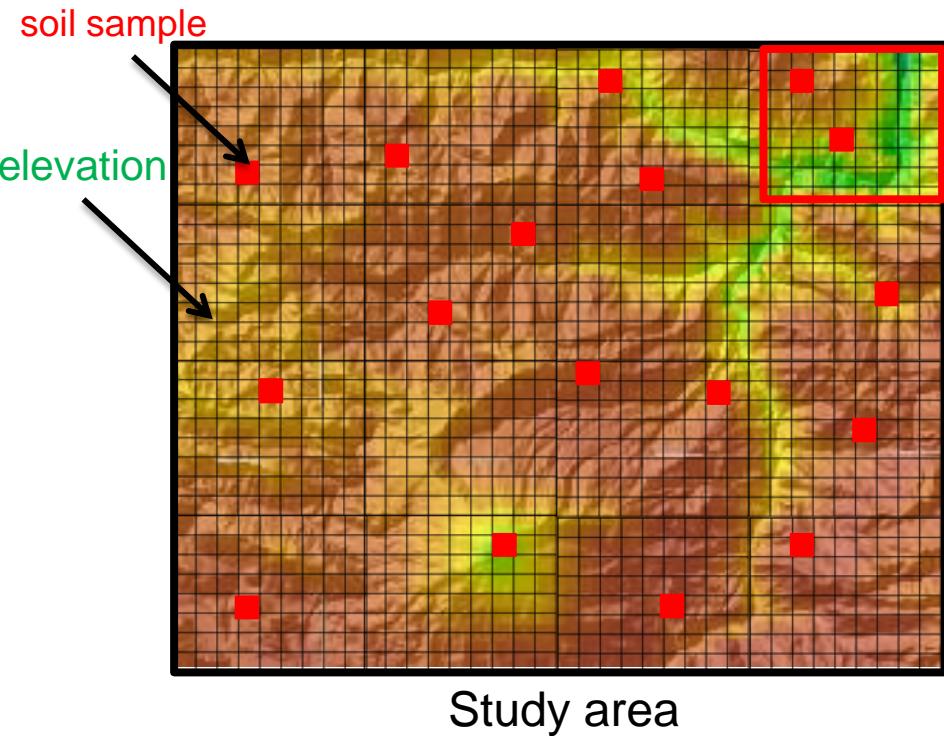
Soil data Model Covariates



SCORPAN model

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

Soil data Model Covariates



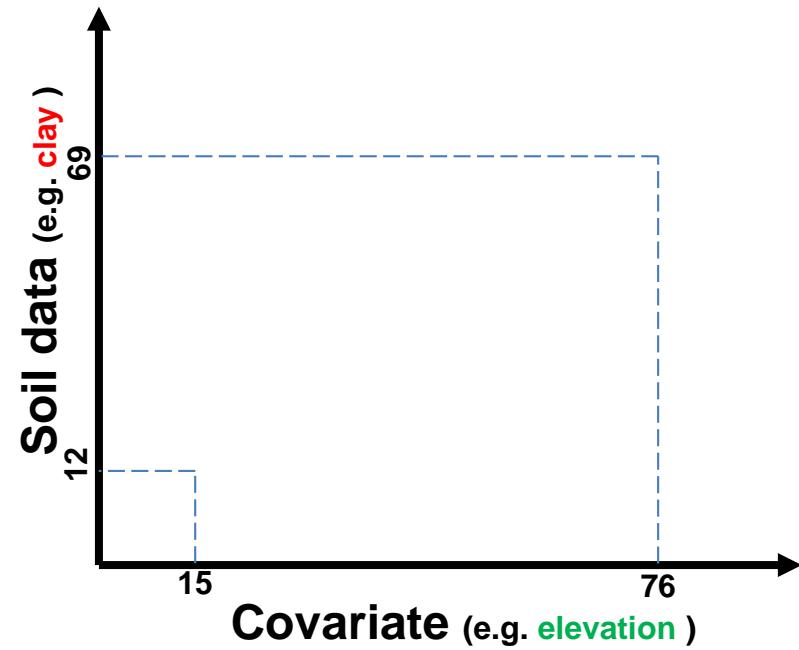
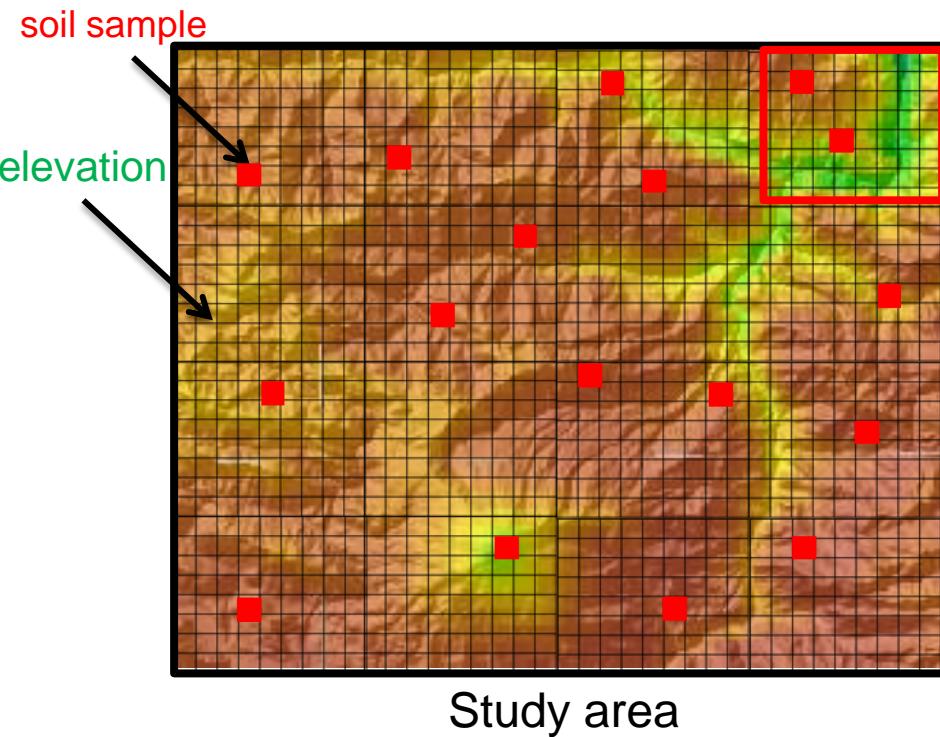
86	65	68	53	18	8	5	9	21
80	76	76	20	24	8	6	11	76
46	75	84	28	6	9	24	50	67
67	86	15	27	7	5	80	82	86
33	38	12	15	5	58	71	75	86
86	16	19	15	18	28	32	35	84
18	7	14	13	22	48	53	59	78
11	8	15	31	28	58	58	74	72

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

SCORPAN model

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

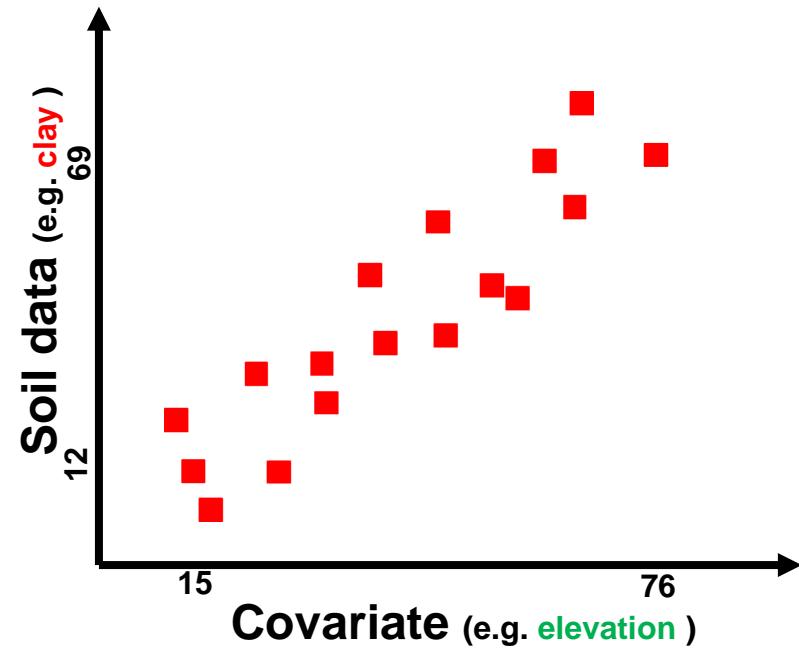
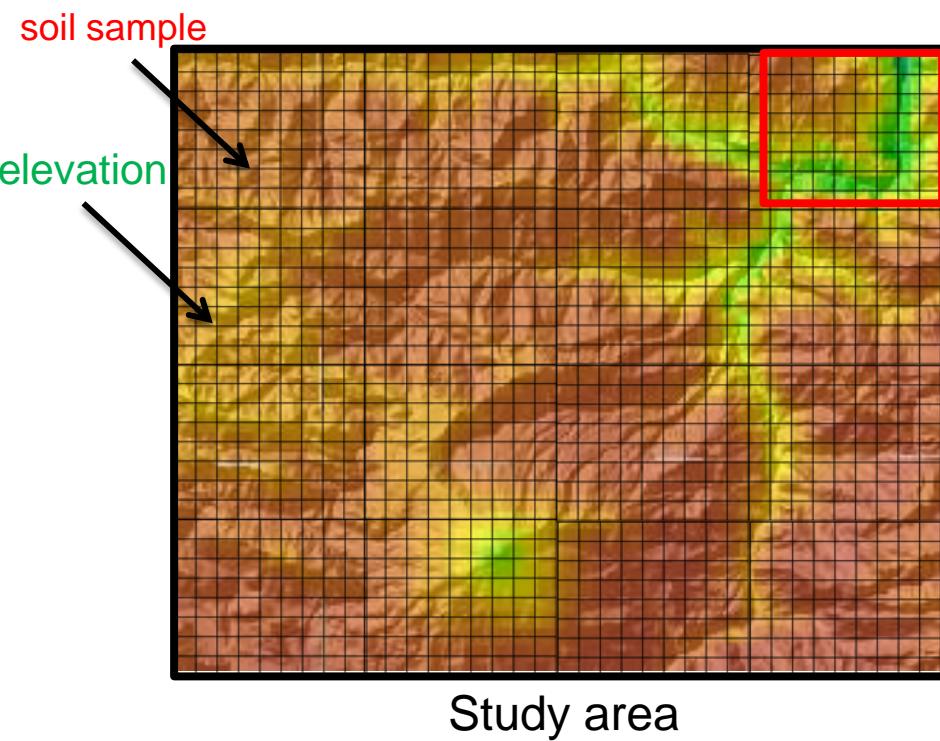
Soil data Model Covariates



SCORPAN model

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

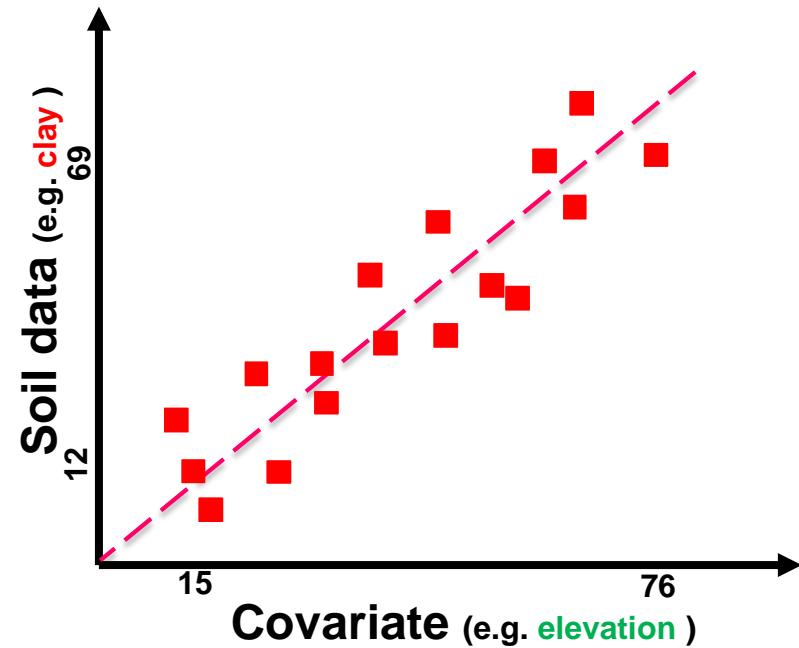
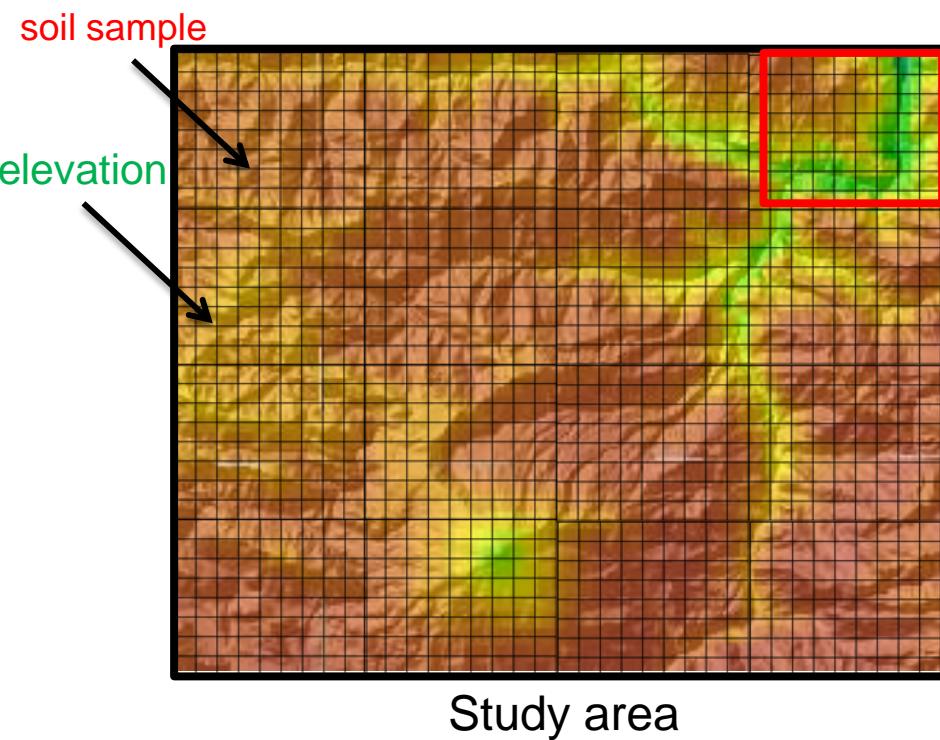
Soil data Model Covariates



SCORPAN model

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

Soil data Model Covariates

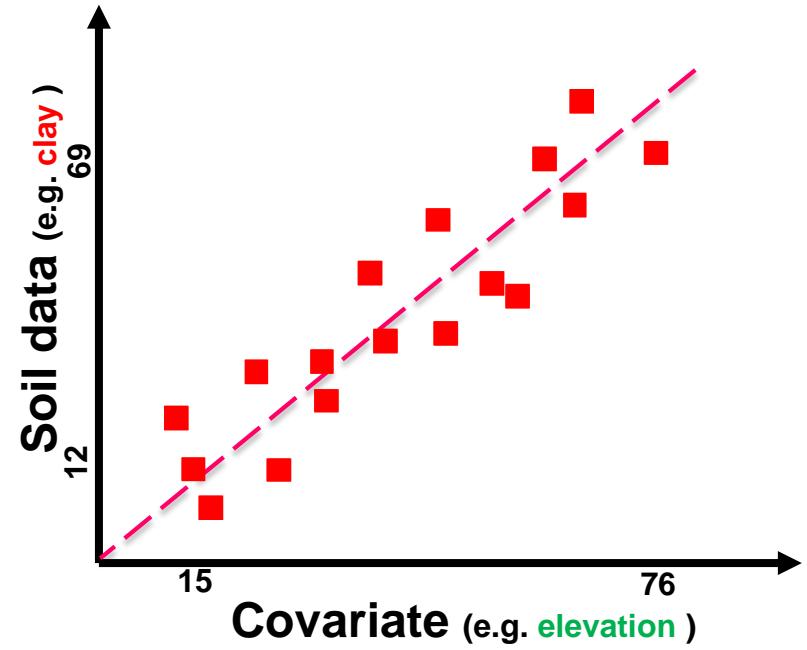
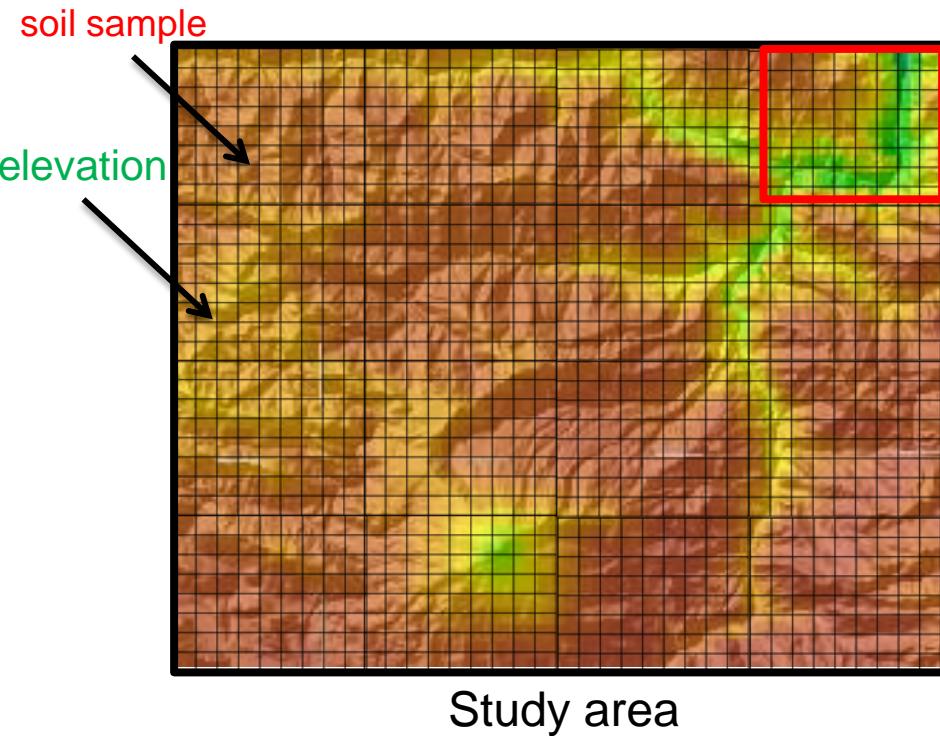


SCORPAN model

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

Soil data Model Covariates

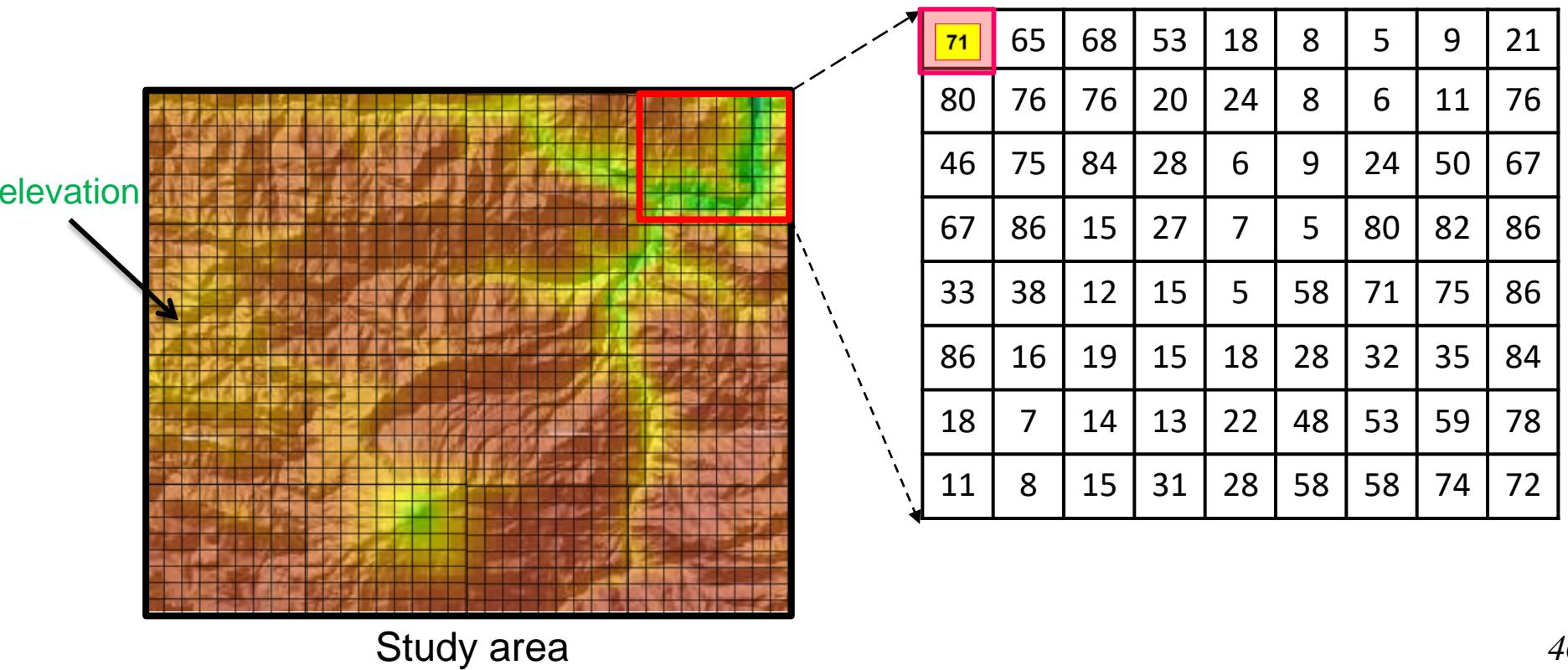
$$\text{Soil clay} = 0.024 + 0.83 \times \text{elevation}$$



SCORPAN model

$$\text{Soil clay} = \underline{0.024} + 0.83 \times \underline{\text{elevation}}$$

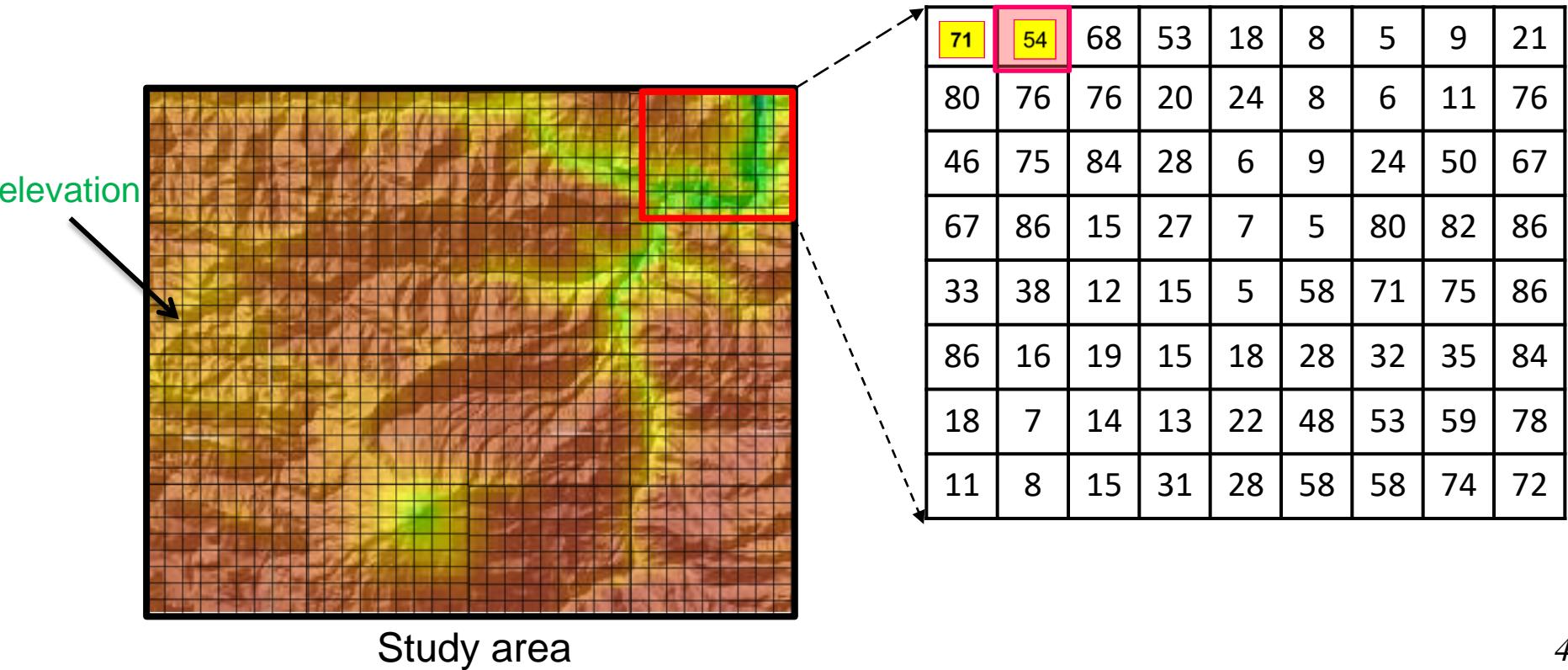
$$\underline{0.024 + 0.83 \times 86} \rightarrow \underline{71}$$



SCORPAN model

$$\text{Soil clay} = \underline{0.024 + 0.83 \times \text{elevation}}$$

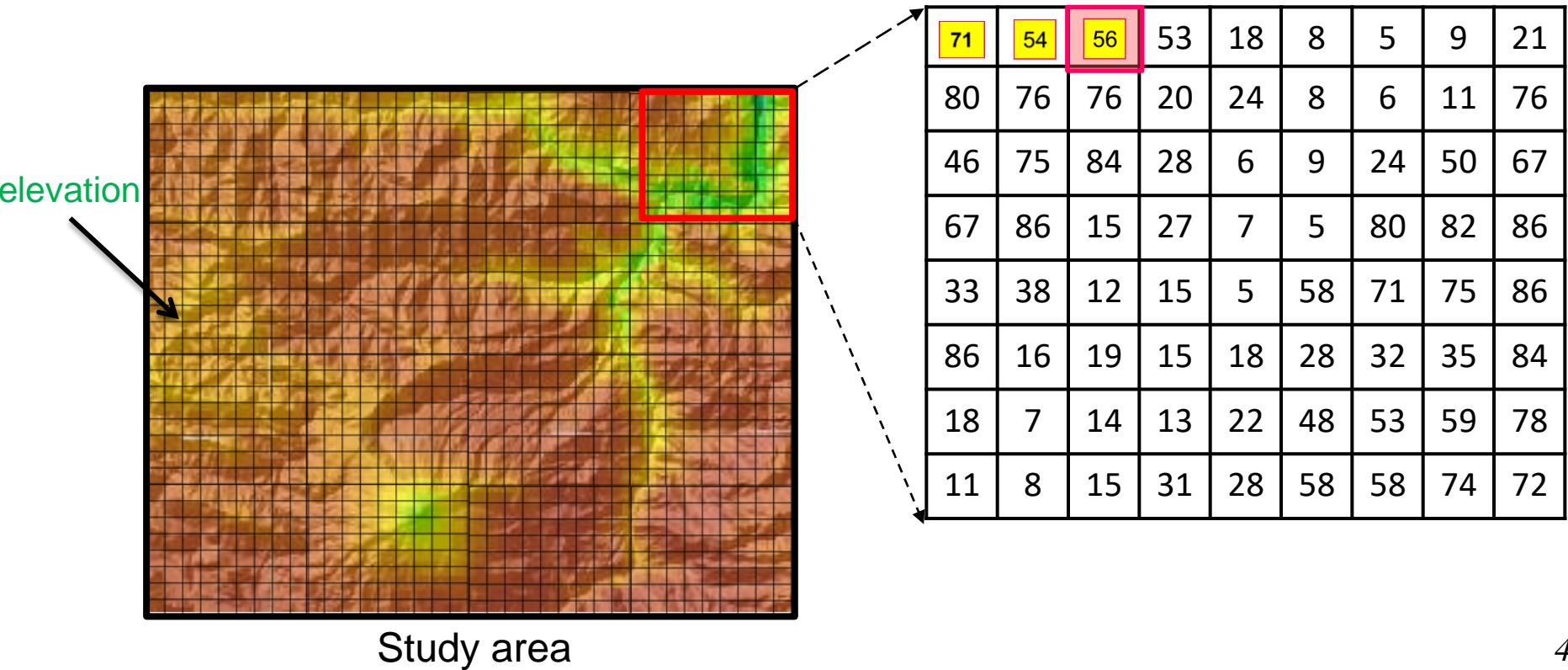
$$\underline{0.024 + 0.83 \times 65} \rightarrow \underline{54}$$



SCORPAN model

$$\text{Soil clay} = 0.024 + 0.83 \times \text{elevation}$$

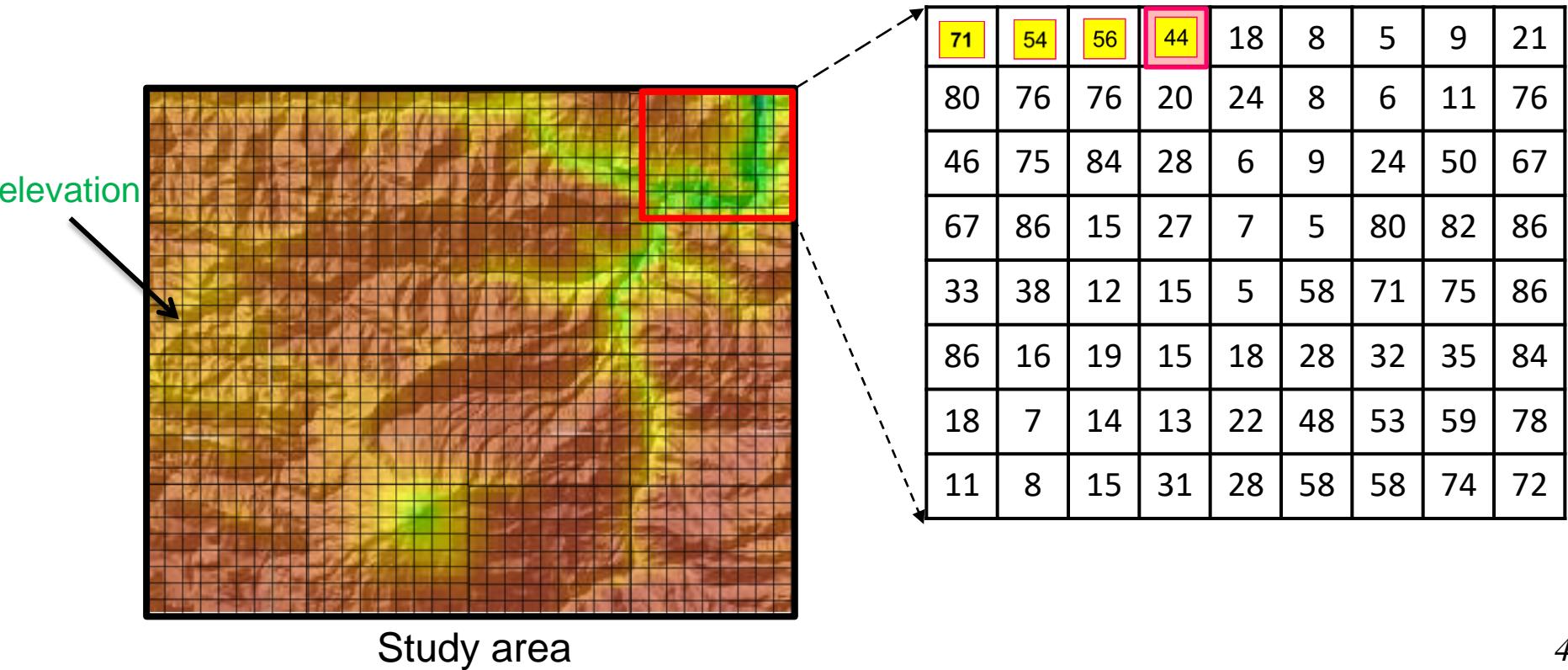
$$0.024 + 0.83 \times 68 \rightarrow 56$$



SCORPAN model

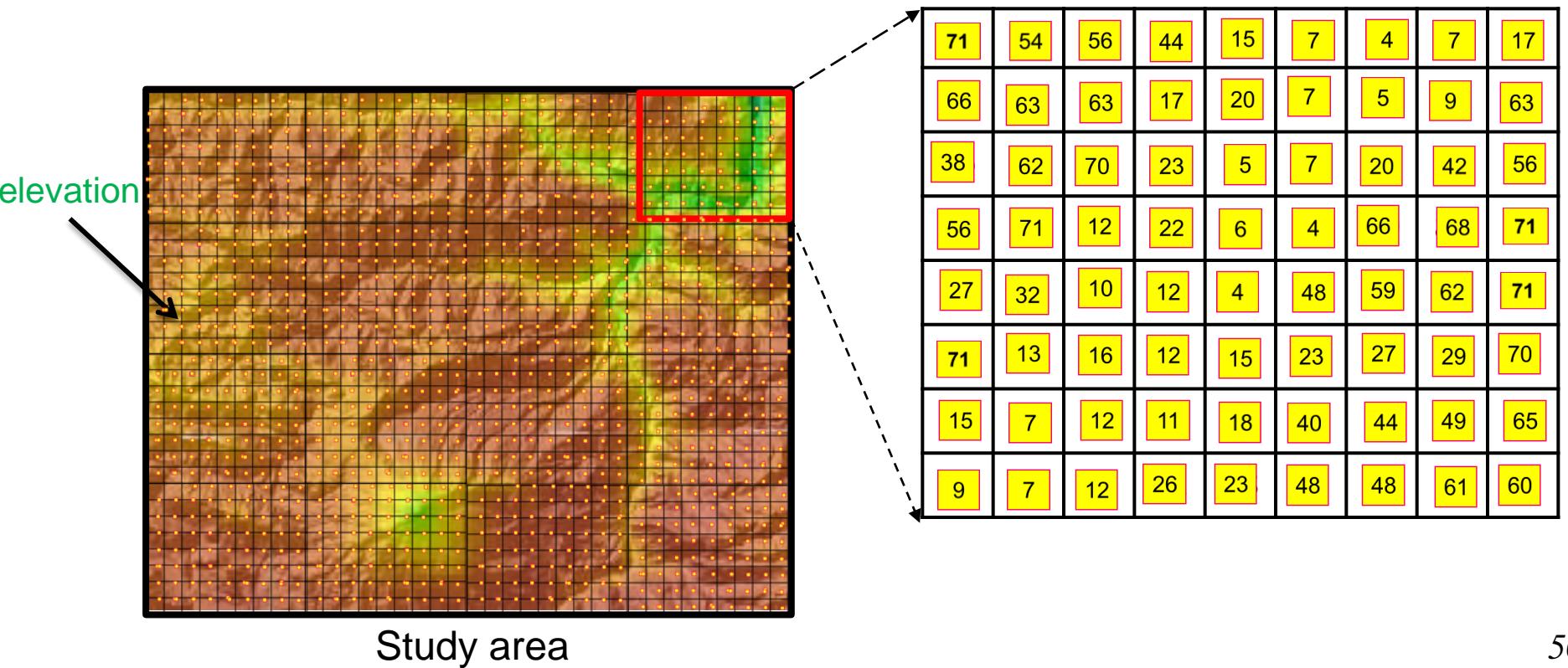
$$\text{Soil clay} = \underline{0.024 + 0.83 \times \text{elevation}}$$

$$\underline{0.024 + 0.83 \times 53} \rightarrow \underline{44}$$



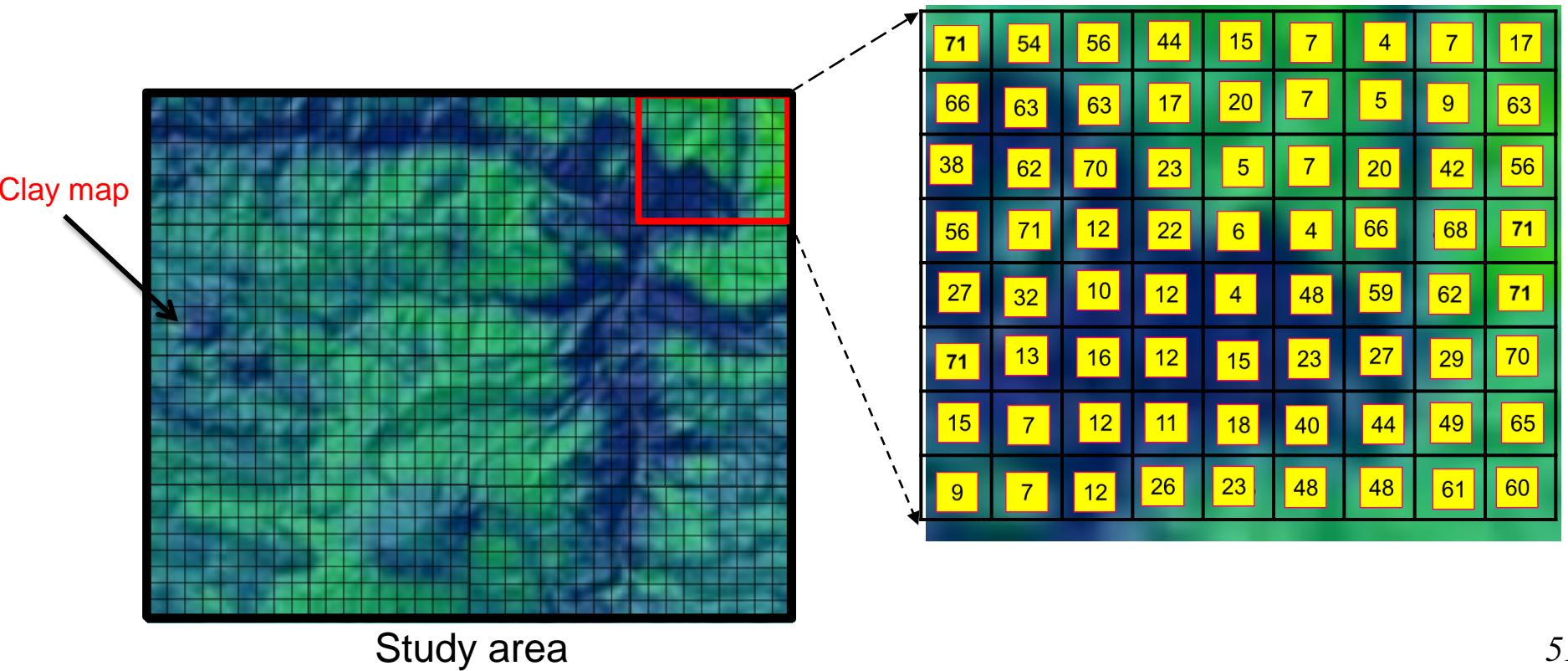
SCORPAN model

$$\text{Soil clay} = \underline{0.024 + 0.83 \times \text{elevation}}$$

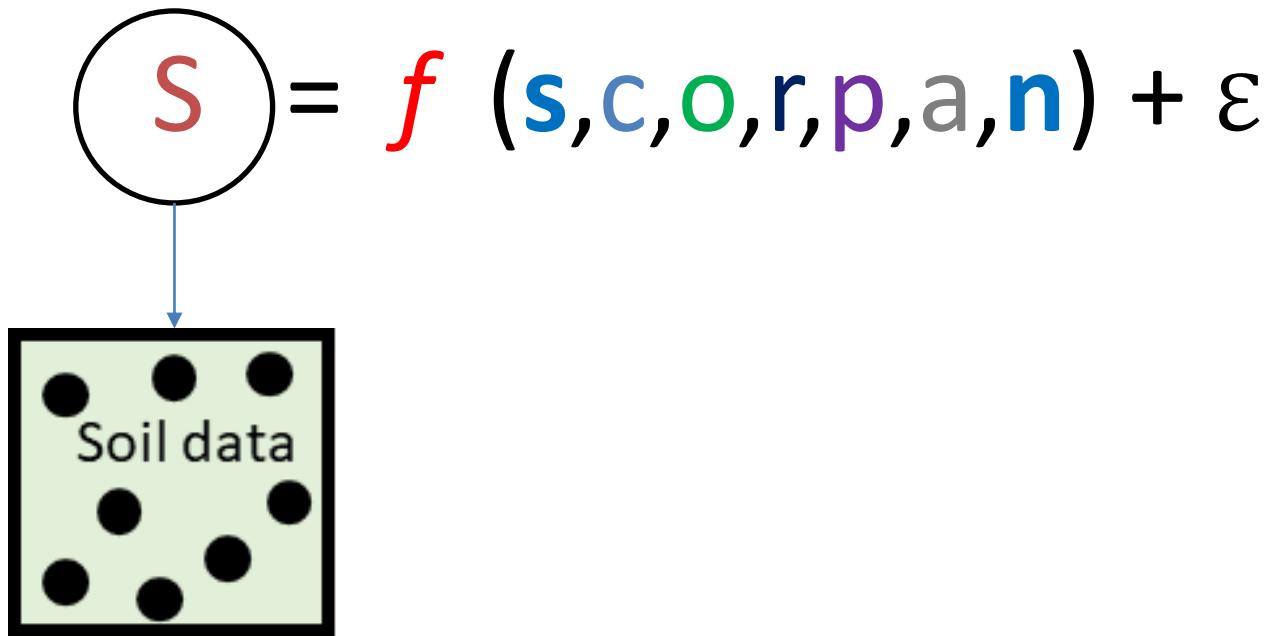


SCORPAN model

$$\text{Soil clay} = \underline{0.024 + 0.83 \times \text{elevation}}$$



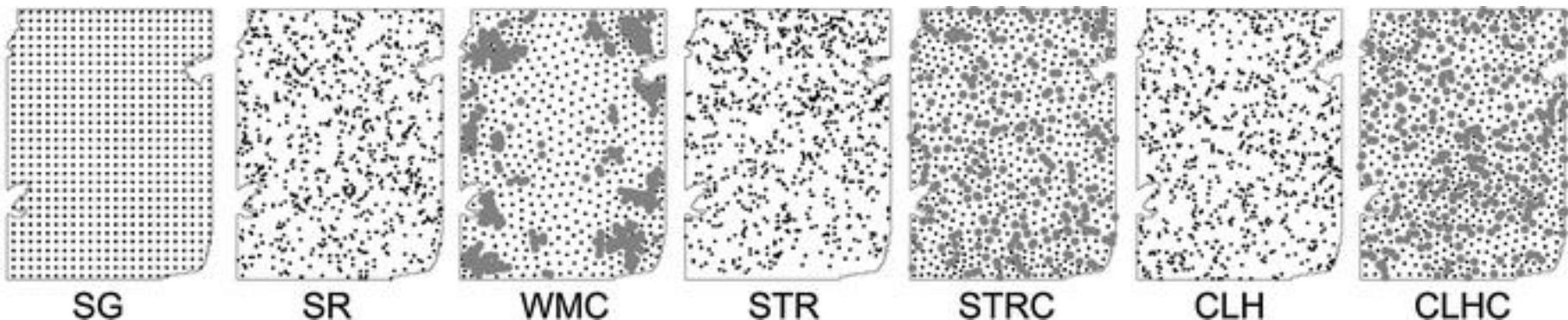
SCORPAN model



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

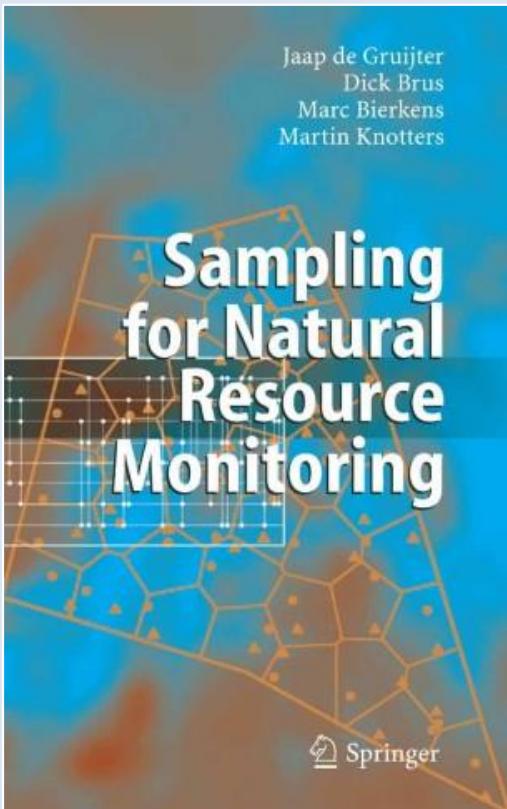
Soil Data (Sampling)

- Sampling is the process of selecting a sample from a population.
- Lack of funds for sampling and analysis is a major problem for any soil mapping project
- Soil sampling plan? Optimal?



SG square grid, SR simple random, WMC uniform distribution of point pairs for variogram estimation combined with spatial coverage, STR stratified random, STRC STR combined with spatial coverage, CLH conditional Latin hypercube, and CLHC CLH combined with spatial coverage (Zhao et al., 2015)

Sampling: Useful Resources



 ELSEVIER

Geoderma
Volume 338, 15 March 2019, Pages 464-480



Sampling for digital soil mapping: A tutorial supported by R scripts

D.J. Brus ^{a, b}✉
[Show more ▾](#)

<https://doi.org/10.1016/j.geoderma.2018.07.036>

[Get rights and content](#)

 ELSEVIER

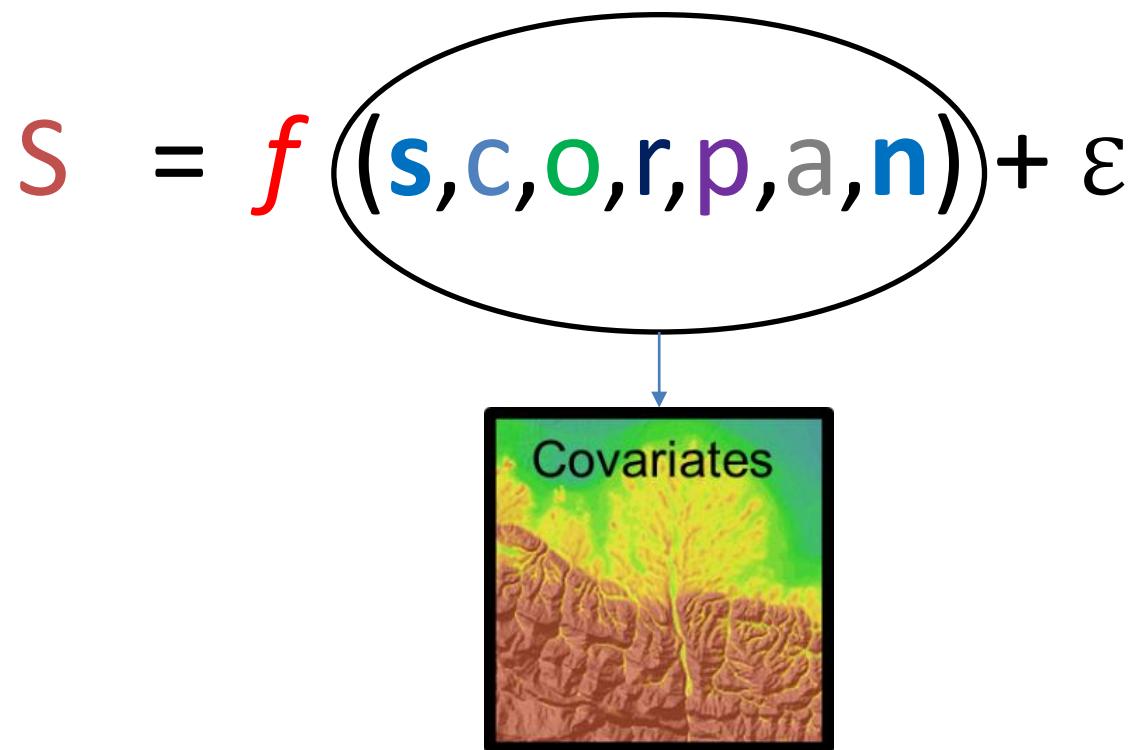
Computers & Geosciences
Volume 32, Issue 9, November 2006, Pages 1378-1388



A conditioned Latin hypercube method for sampling in the presence of ancillary information ★

Budiman Minasny ^a✉, Alex B. McBratney

SCORPAN Model



Environmental Covariates

s
c
o
r
p
a
n

Possible sources of information to represent seven scorpan factors

	<i>scorpan</i> factor	Possible representatives
	<i>s</i>	Legacy soil maps, point observations, expert knowledge
	<i>c</i>	Temperature and precipitation records
	<i>r</i>	Vegetation maps, species abundance maps, yield maps, land use maps
	<i>p</i>	Digital elevation model, terrain attributes
	<i>a</i>	Legacy geology maps, gamma radiometric information
	<i>n</i>	Weathering indices, geology maps
		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

Proximal Soil Sensing (PSS)

s
c
o
r
p
a
n

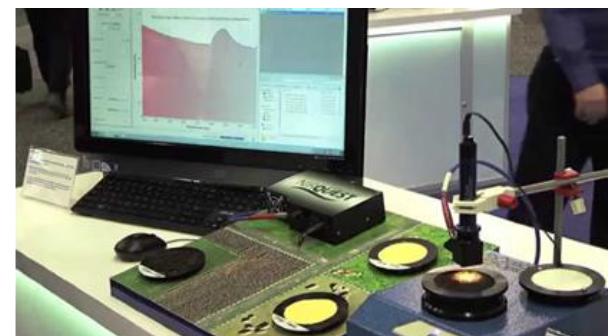
- Proximal soil sensing is a set of techniques developed to measure the physical, chemical and biological properties of soil when placing the sensor in contact, or at a proximal distance (within 2 m) to, the soil being characterized (Viscarra Rossel et al., 2011)
- The proximal soil sensors are more time- and cost effective than conventional laboratory analyses
- The proximal soil sensors facilitate the collection of larger amounts of (spatial) data



EM38



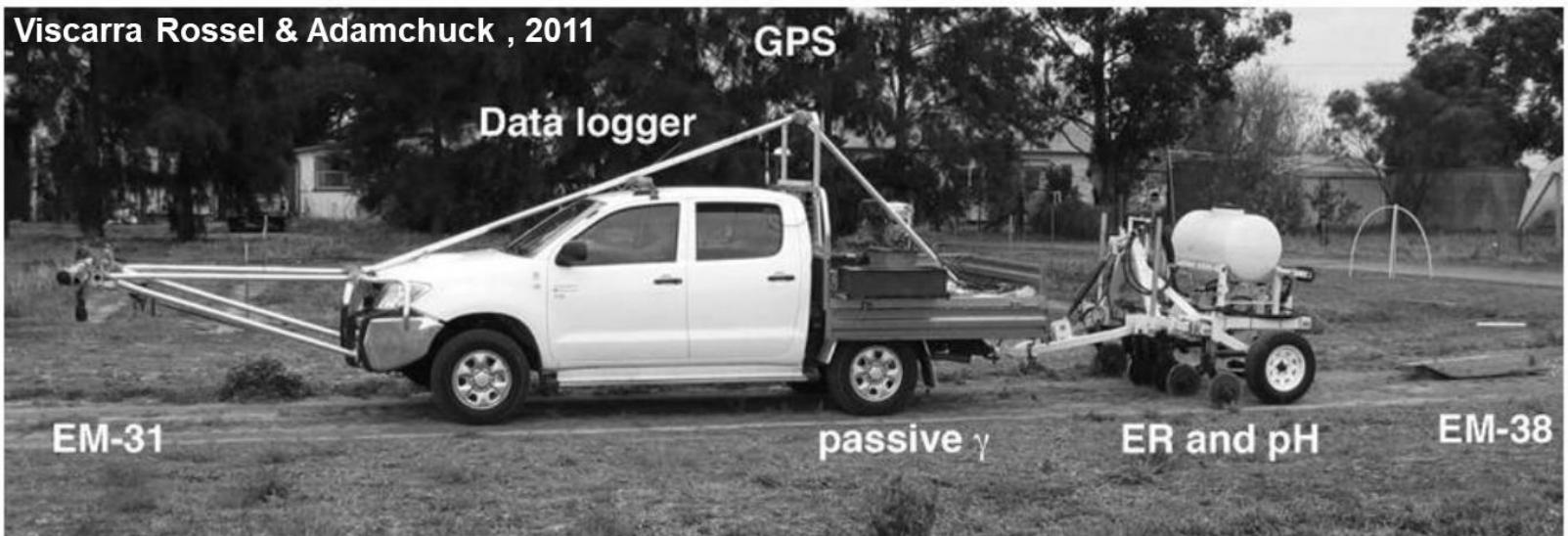
PXRF



NIR

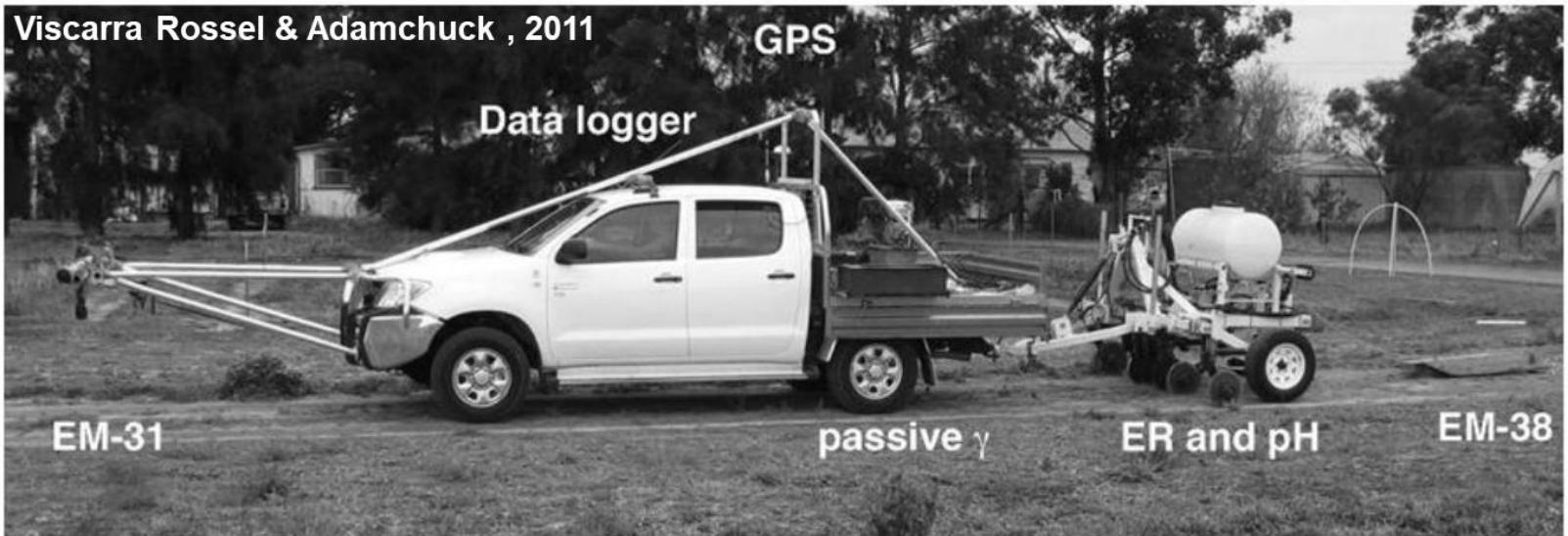
Proximal Soil Sensing: example

s
c
o
r
p
a
n



Proximal Soil Sensing: example

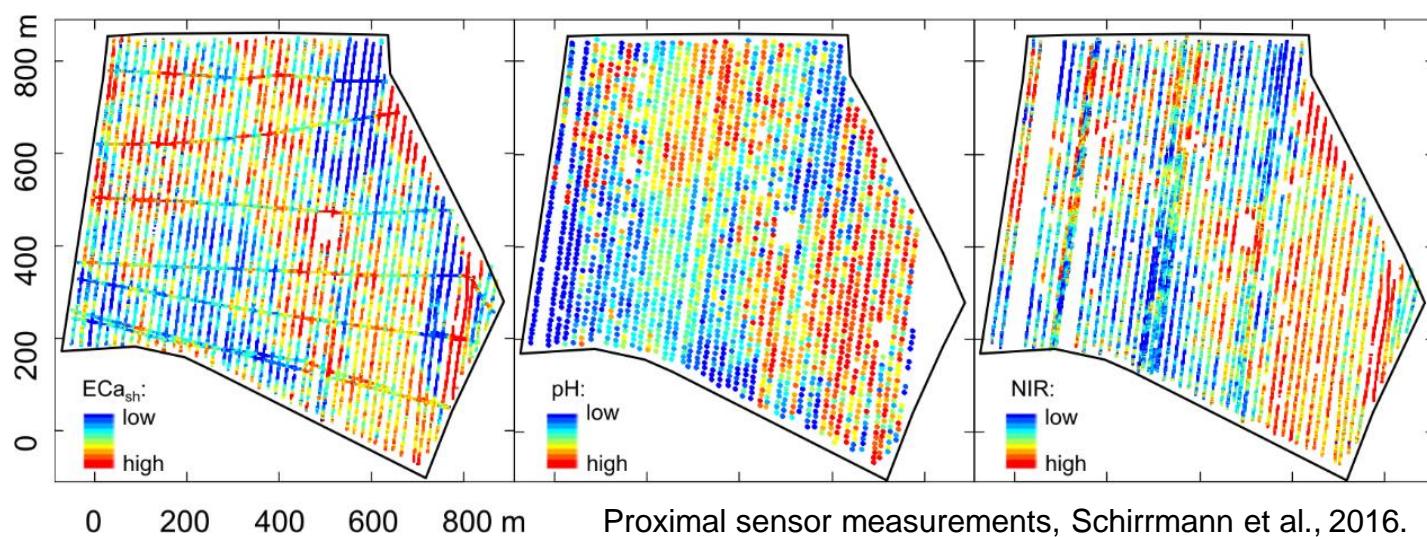
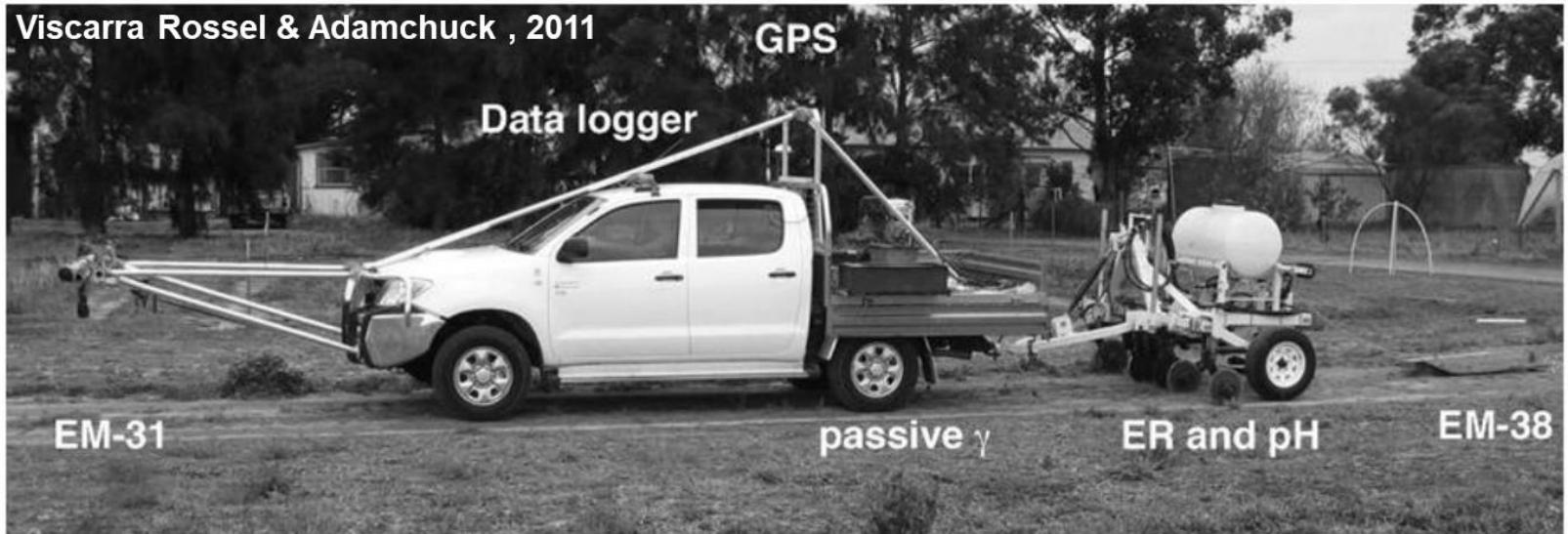
s
c
o
r
p
a
n



gifs.com

Proximal Soil Sensing: example

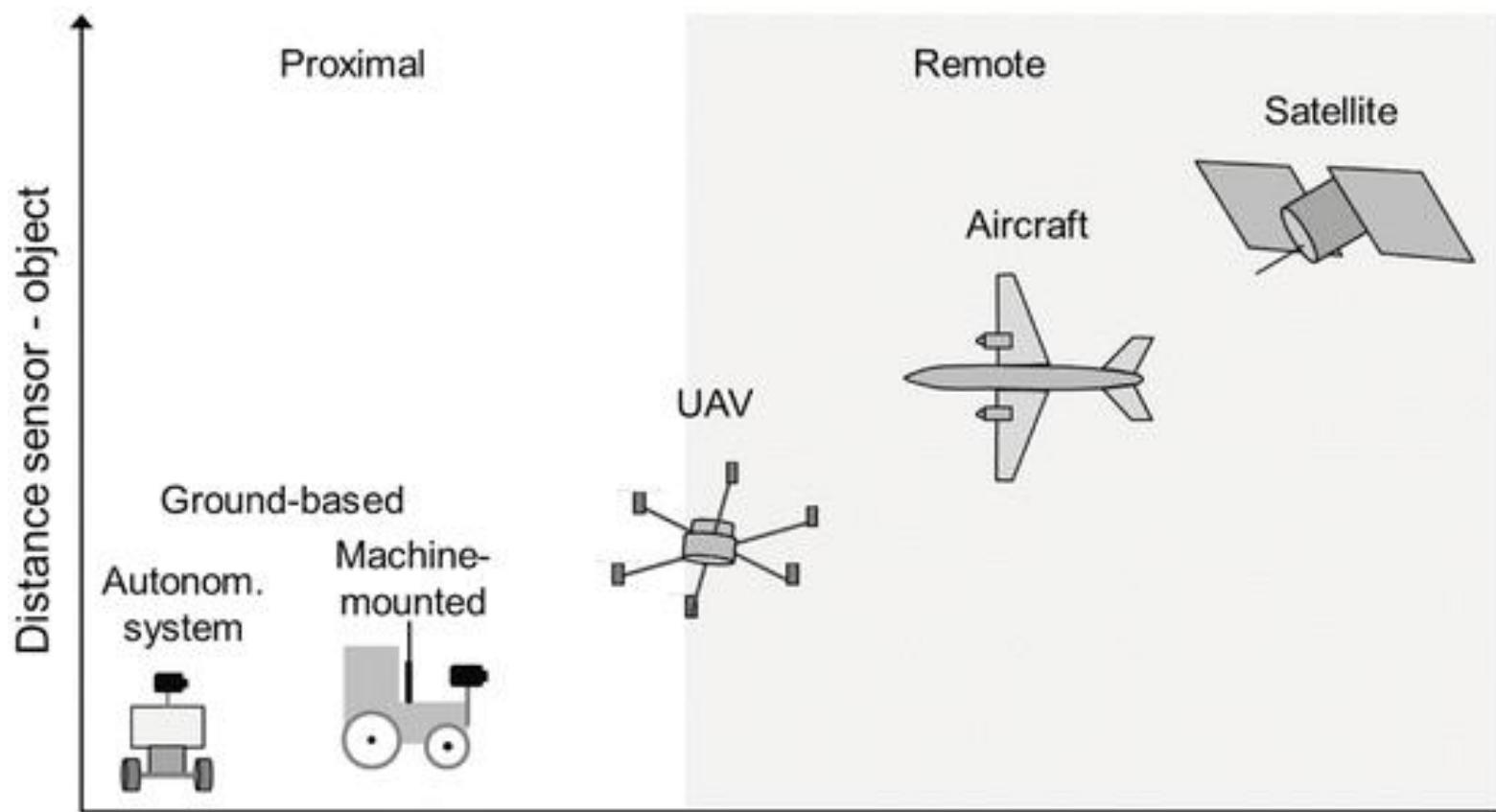
s
c
o
r
p
a
n



Remote Sensing (RS)

- The value of remote sensing over proximal sensing is that large spatial extents can be covered quickly with many estimates.

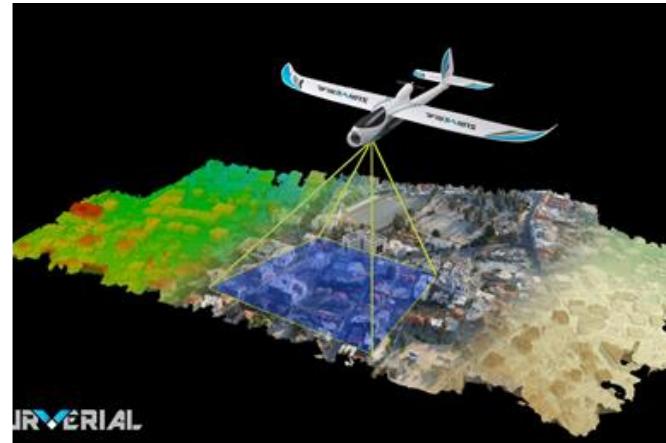
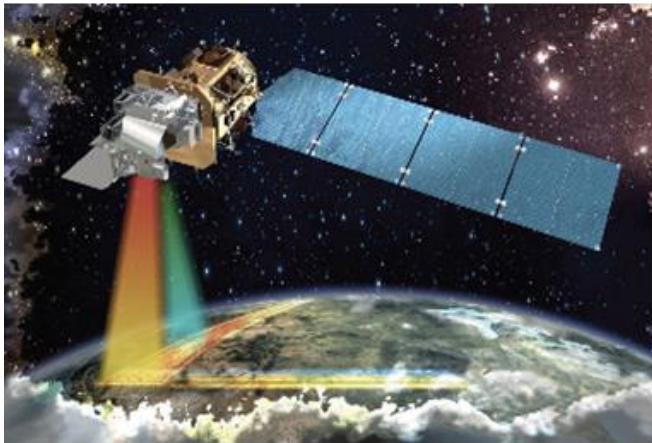
s
c
o
r
p
a
n



Remote Sensing

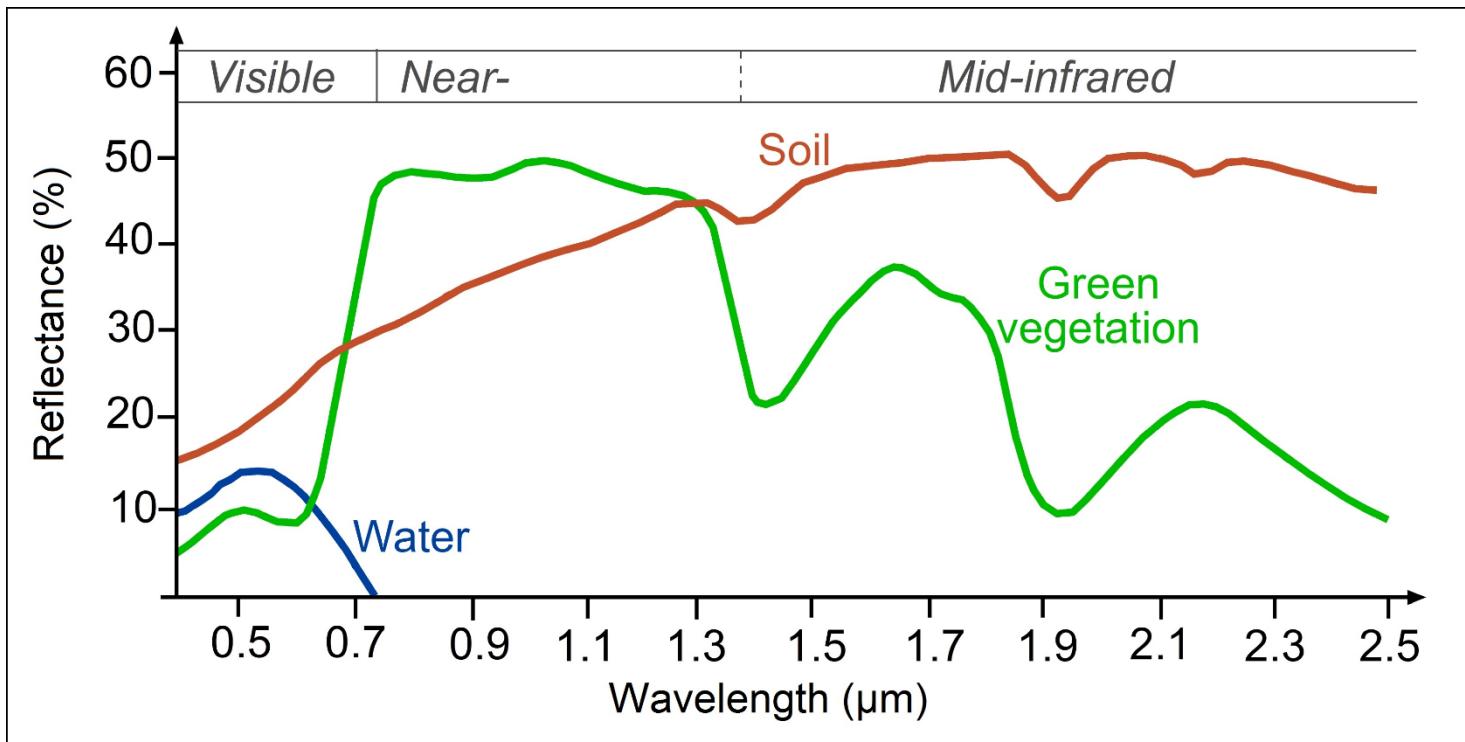
s
c
o
r
p
a
n

- The remotely sensed data has been shown to be an efficient means of assessing the condition of natural resources



Remote Sensing

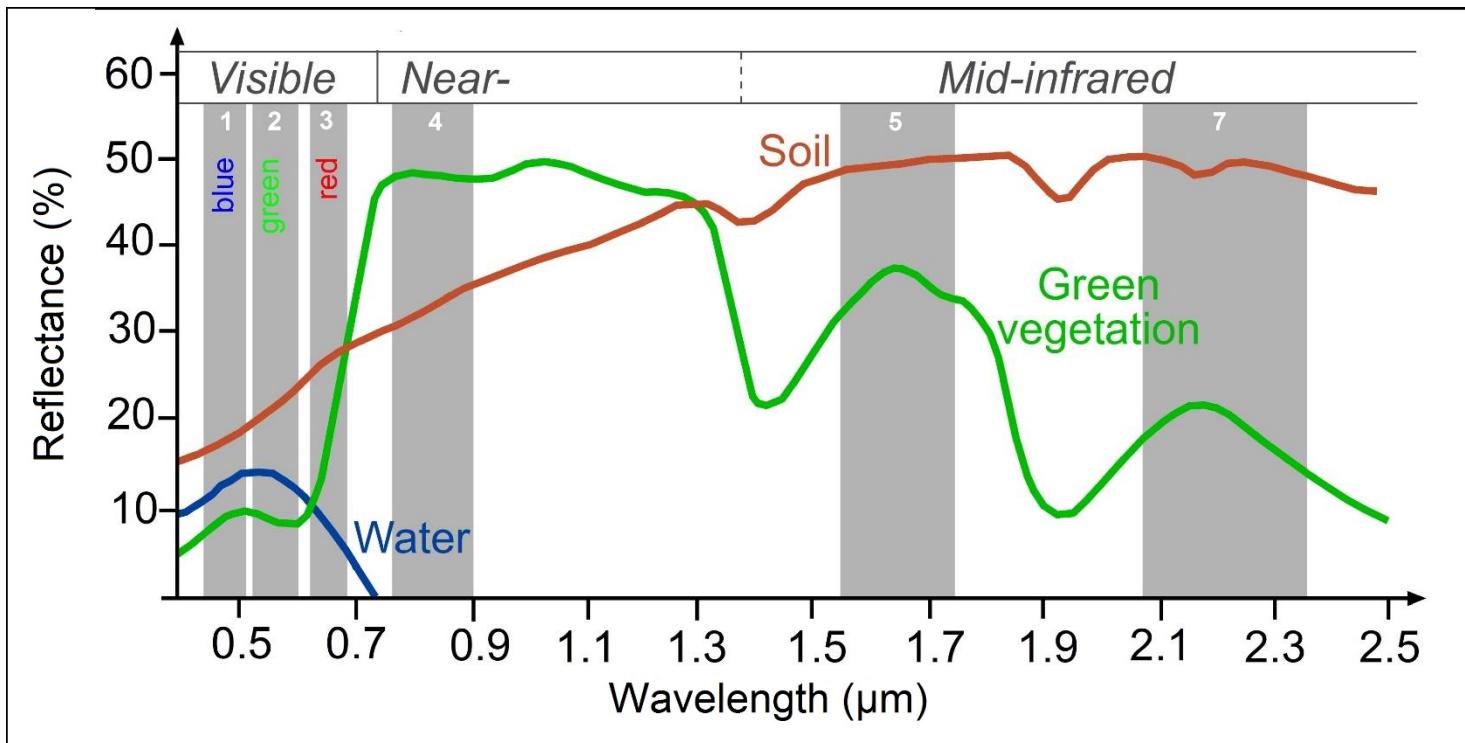
s
c
o
r
p
a
n



Spectral signatures of soil, vegetation and water.
Source: Siegmund, Menz 2005 with modifications

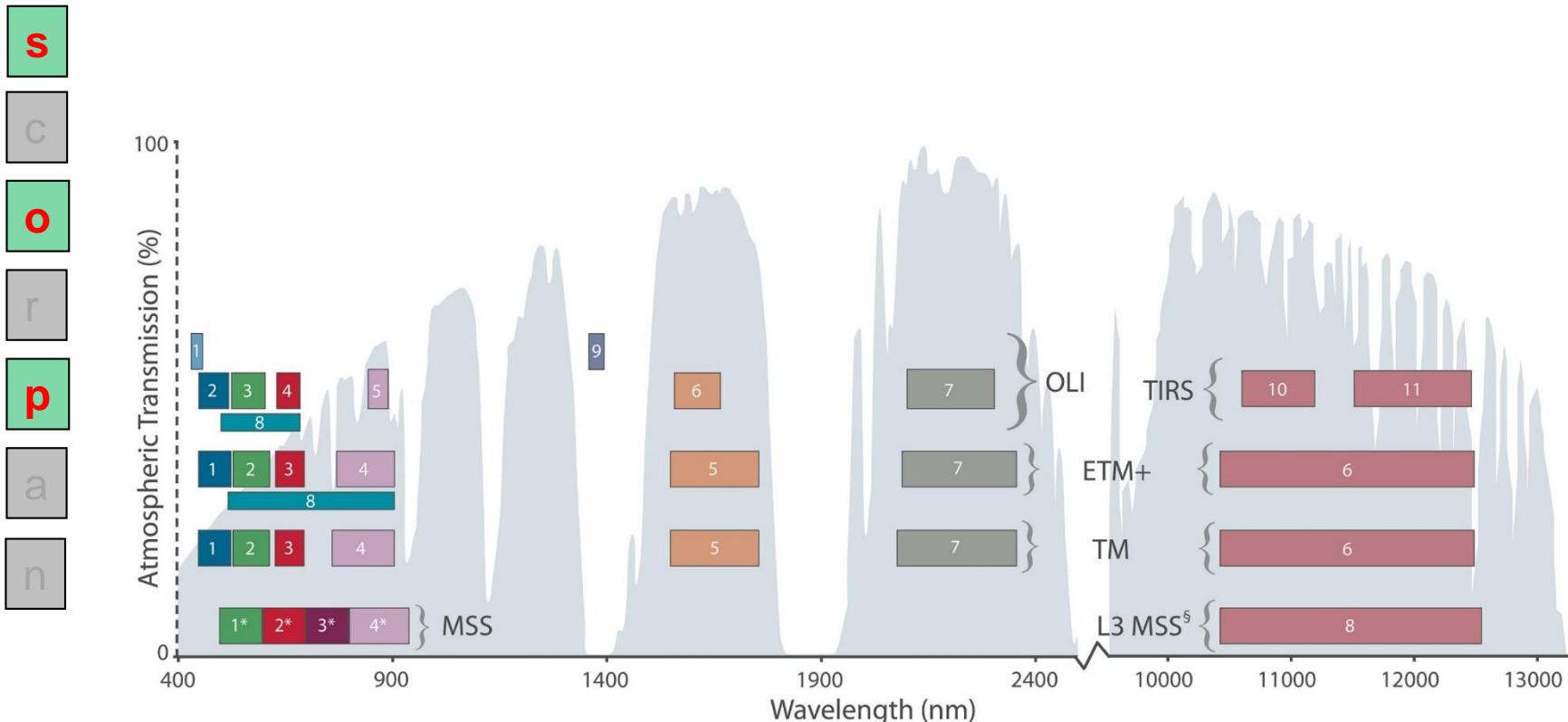
Remote Sensing

s
c
o
r
p
a
n



Spectral signatures of soil, vegetation and water, and spectral bands of LANDSAT 7.
Source: Siegmund, Menz 2005 with modifications

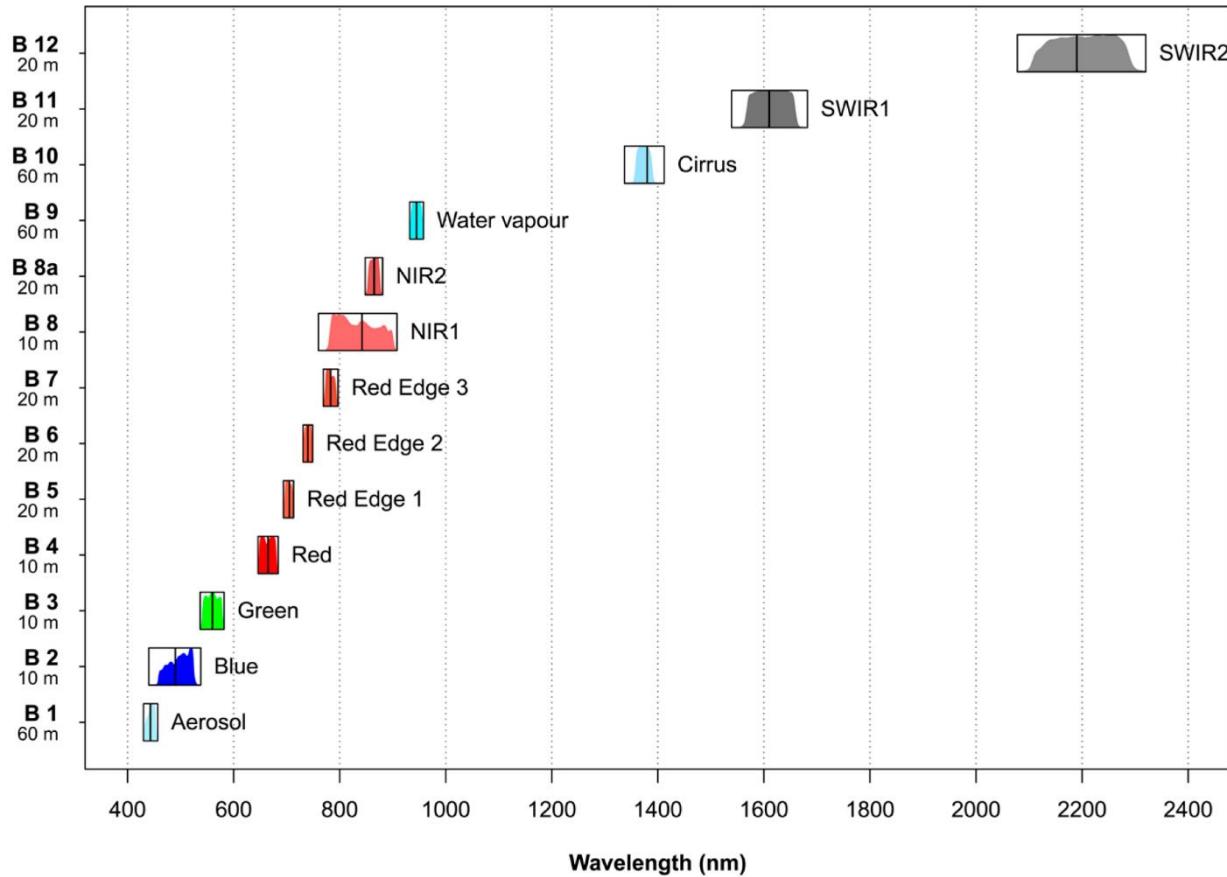
Remote Sensing: Landsat



Landsat sensors, spectral channels and band-passes, superimposed on atmospheric transmission percentage (grey background). MSS: Landsat-1 through -5; TM: Landsat-4 and -5; ETM+: Landsat-7; OLI and TIRS: Landsat-8. (Source: NASA/Landsat Legacy Project Team and American Society for Photogrammetry and Remote Sensing). Wulder et al., 2018

Remote Sensing: Sentinel 2

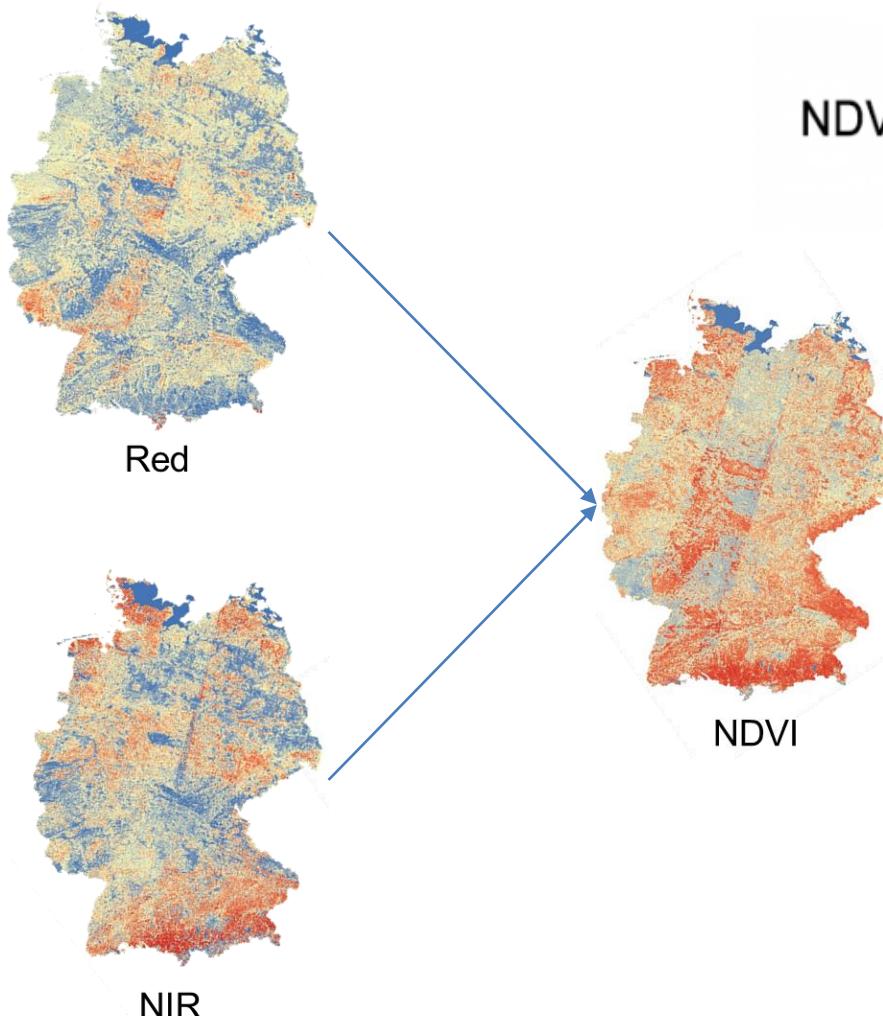
S
C
O
R
P
A
N



Characteristics of the Multi Spectral Instrument (MSI) on board Sentinel-2. The spectral response functions are shown in colors with the central wavelength in black. The band names and the corresponding spatial resolutions (in meters) are also indicated. Immitzer et al., 2016

Remote Sensing: band ratio

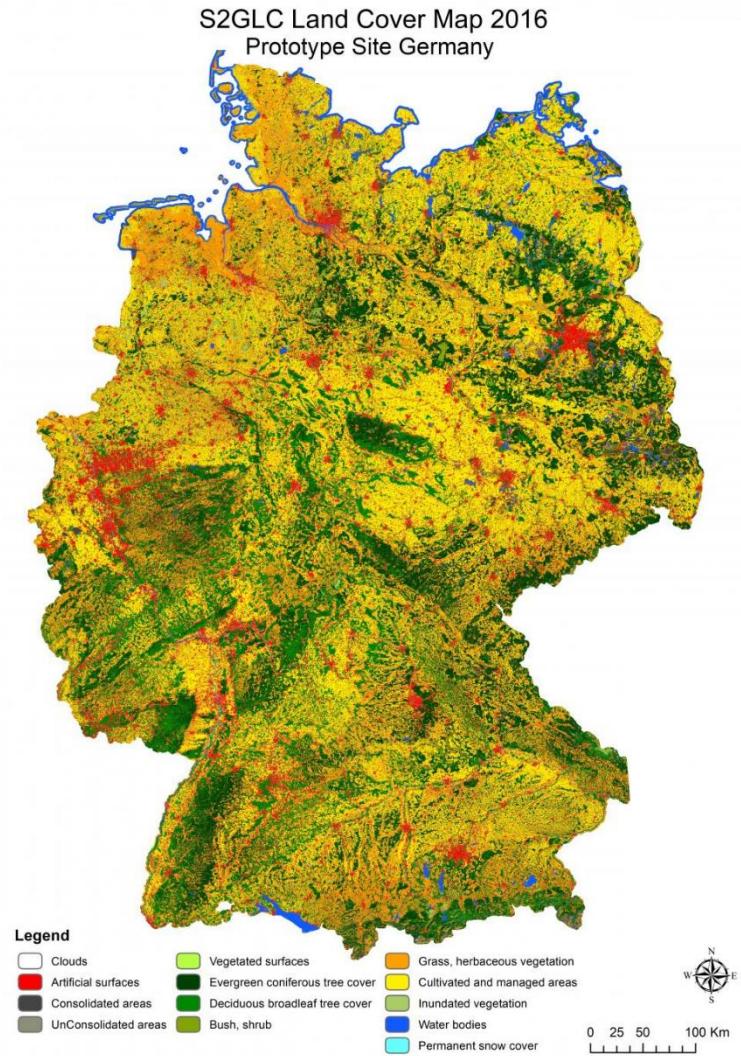
s
c
o
r
p
a
n



Red and NIR bands of Landsat-8 for Germany

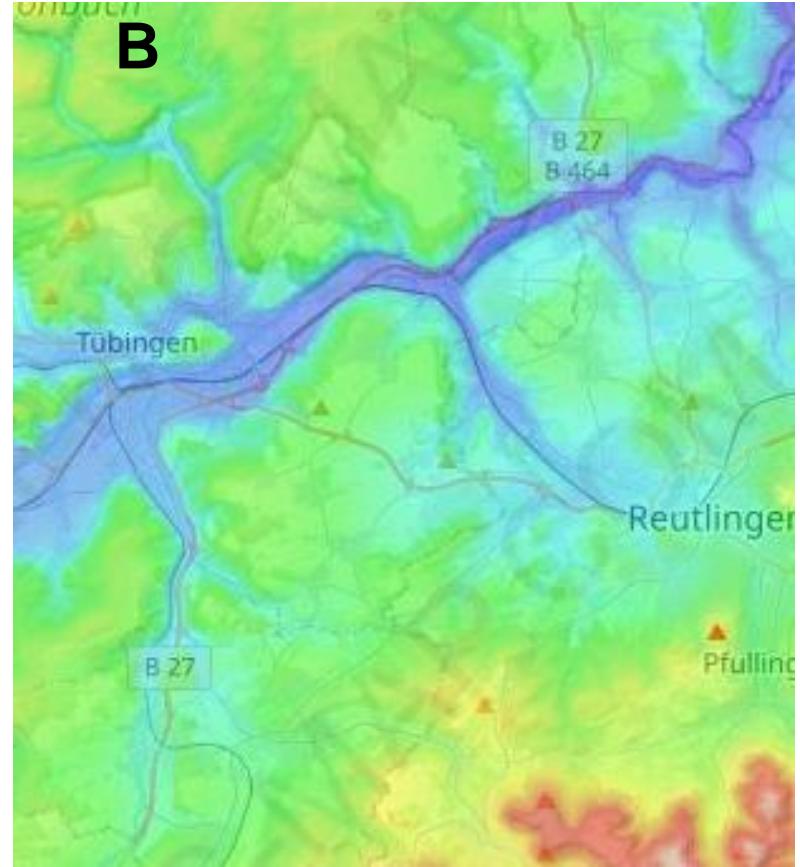
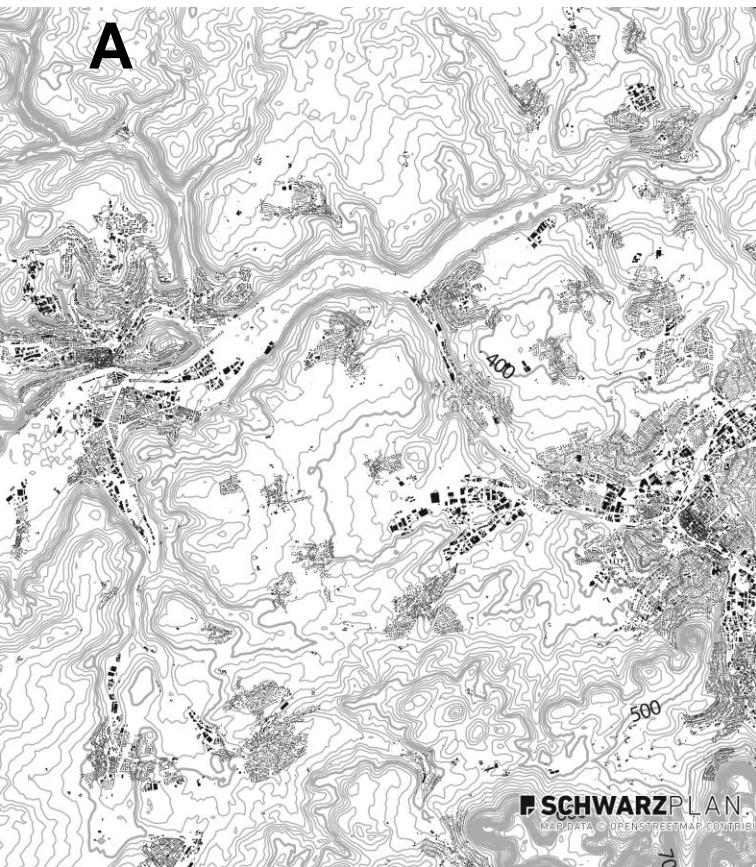
Remote Sensing: Land use/cover

S
C
O
R
P
A
N



Digital Elevation Model (DEM)

S
C
O
R
P
A
N

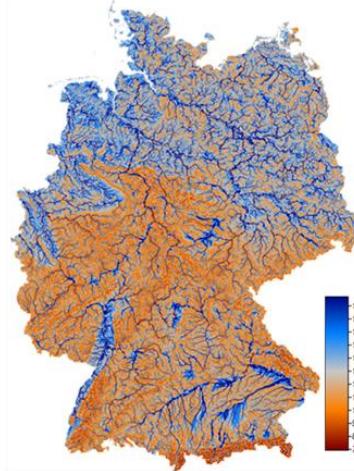


Elevation maps of Tuebingen, <https://en-gb.topographic-map.com/>

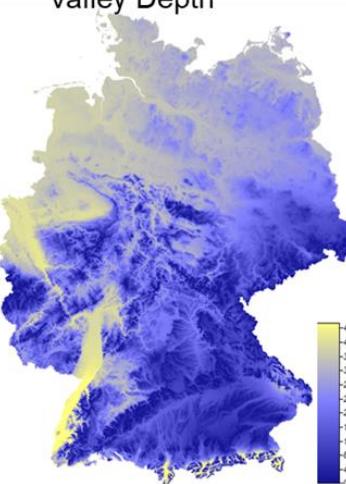
DEM: Terrain Attributes

S
C
O
r
P
A
N

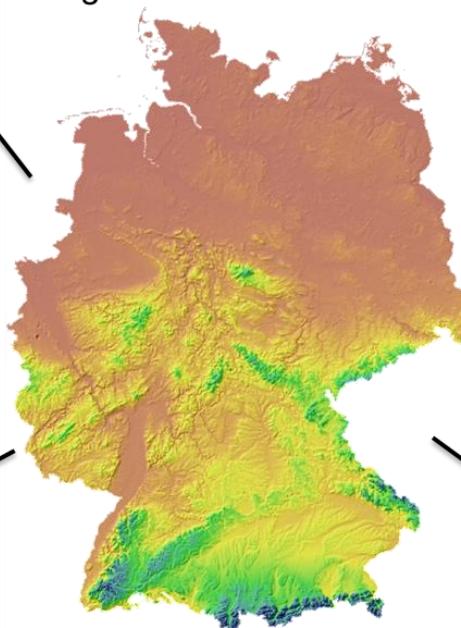
Topographic Wetness Index



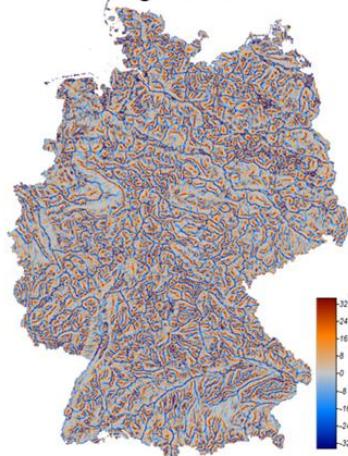
Valley Depth



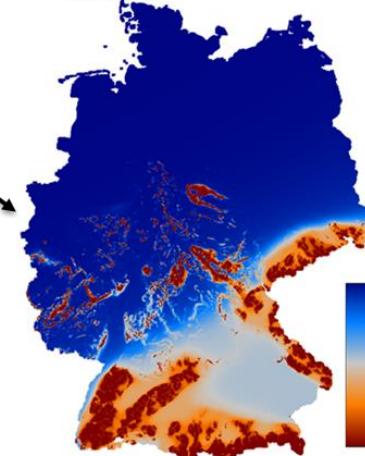
Digital Elevation Model



Convergence Index



MrVBF

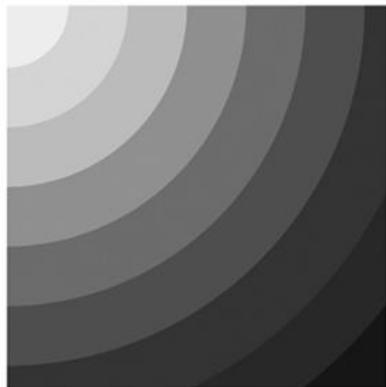


Basic Terrain Derivatives of Germany

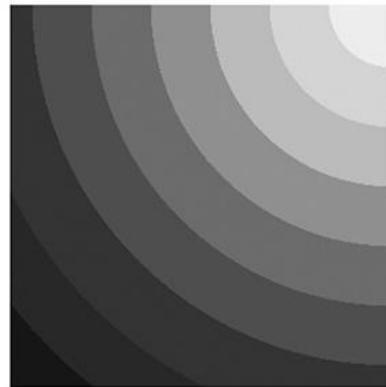
Distance layers

S
C
O
R
P
A
N

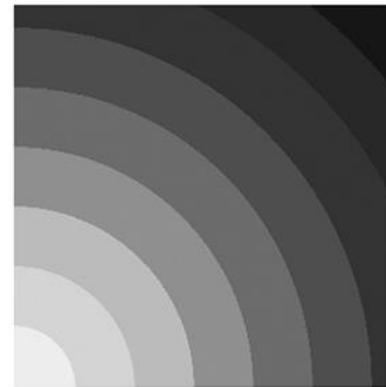
Upper left (C1)



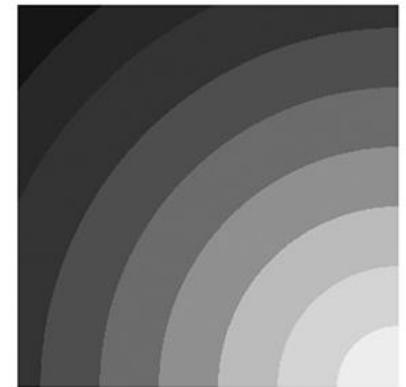
Upper right (C2)



Lower left (C3)



Lower right (C4)



Centre (CC)



European Journal of Soil Science, 2018

doi: 10.1111/ejss.12687

Spatial modelling with Euclidean distance fields and machine learning

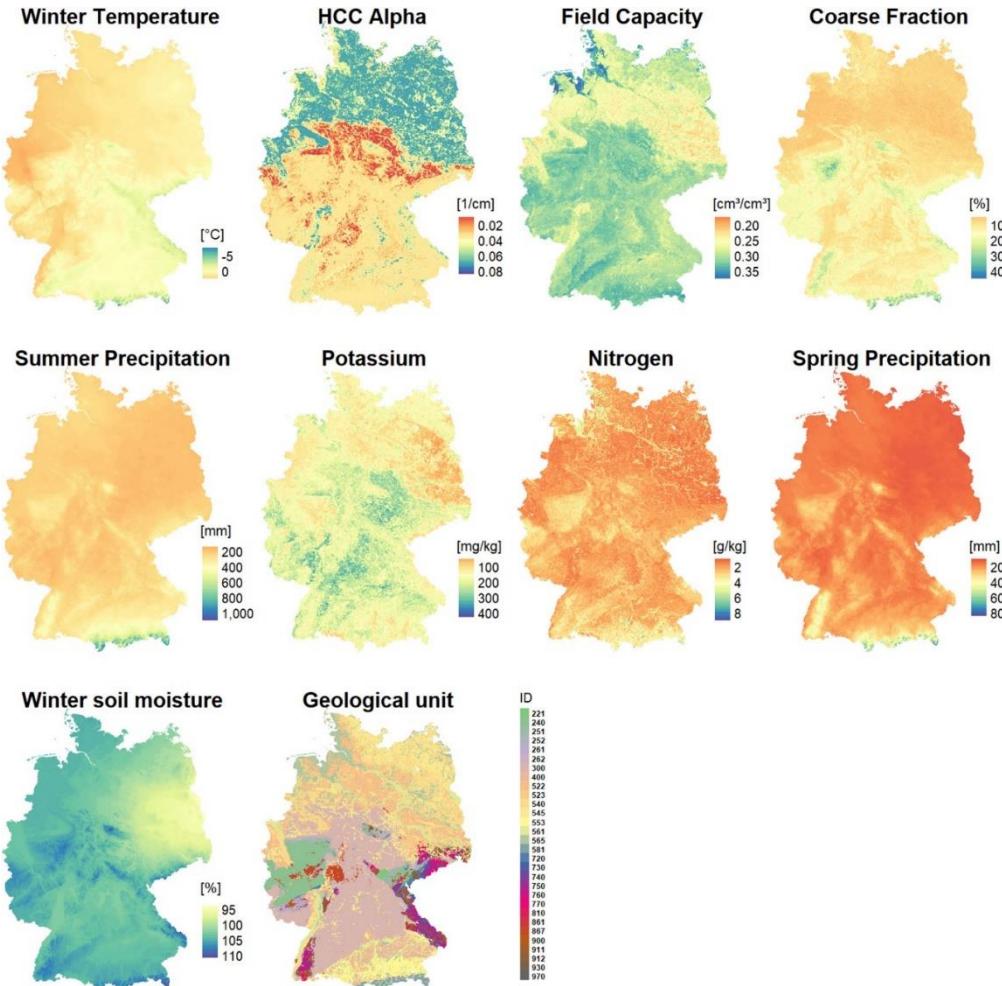
T. BEHRENS^a, K. SCHMIDT^a, R. A. VISCARRA ROSSEL^b, P. GRIES^a, T. SCHOLTEN^a & R. A. MACMILLAN^c

^aDepartment of Geosciences, Soil Science and Geomorphology, University of Tübingen, Rümelinstraße 19-23, 72074, Tübingen, Germany,

^bCSIRO Land and Water, GPO Box 1700, Canberra ACT 2601, Australia, and ^cLandMapper Environmental Solutions Inc., 7415 118 A Street NW, Edmonton, Alberta Canada

Other Covariates

s
c
o
r
p
a
n



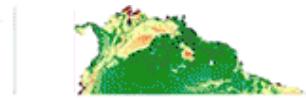
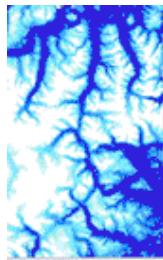
Other Geospatial Data of Germany (Petermann et al., 2020)

Where Can I Obtain the Covariates?

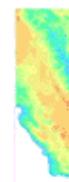
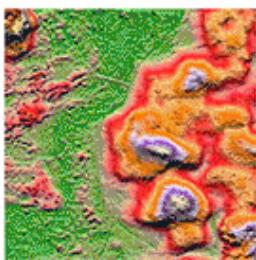
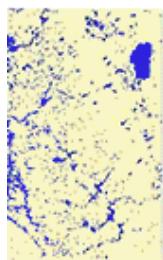
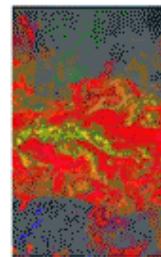
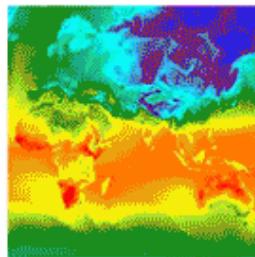
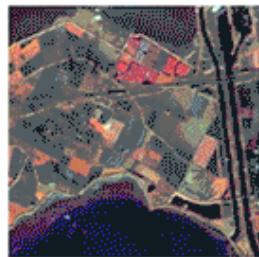
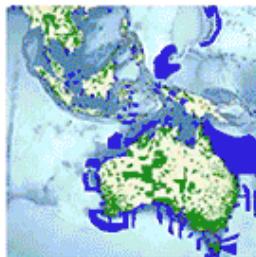
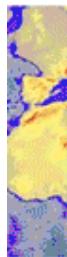
s
c
o
r
p
a
n



Google Earth Engine



A glimpse into the Earth Engine Data Catalog

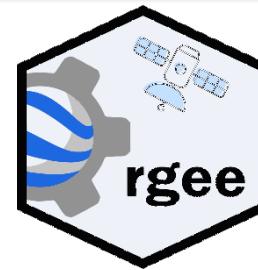


Where Can I Obtain the Covariates?

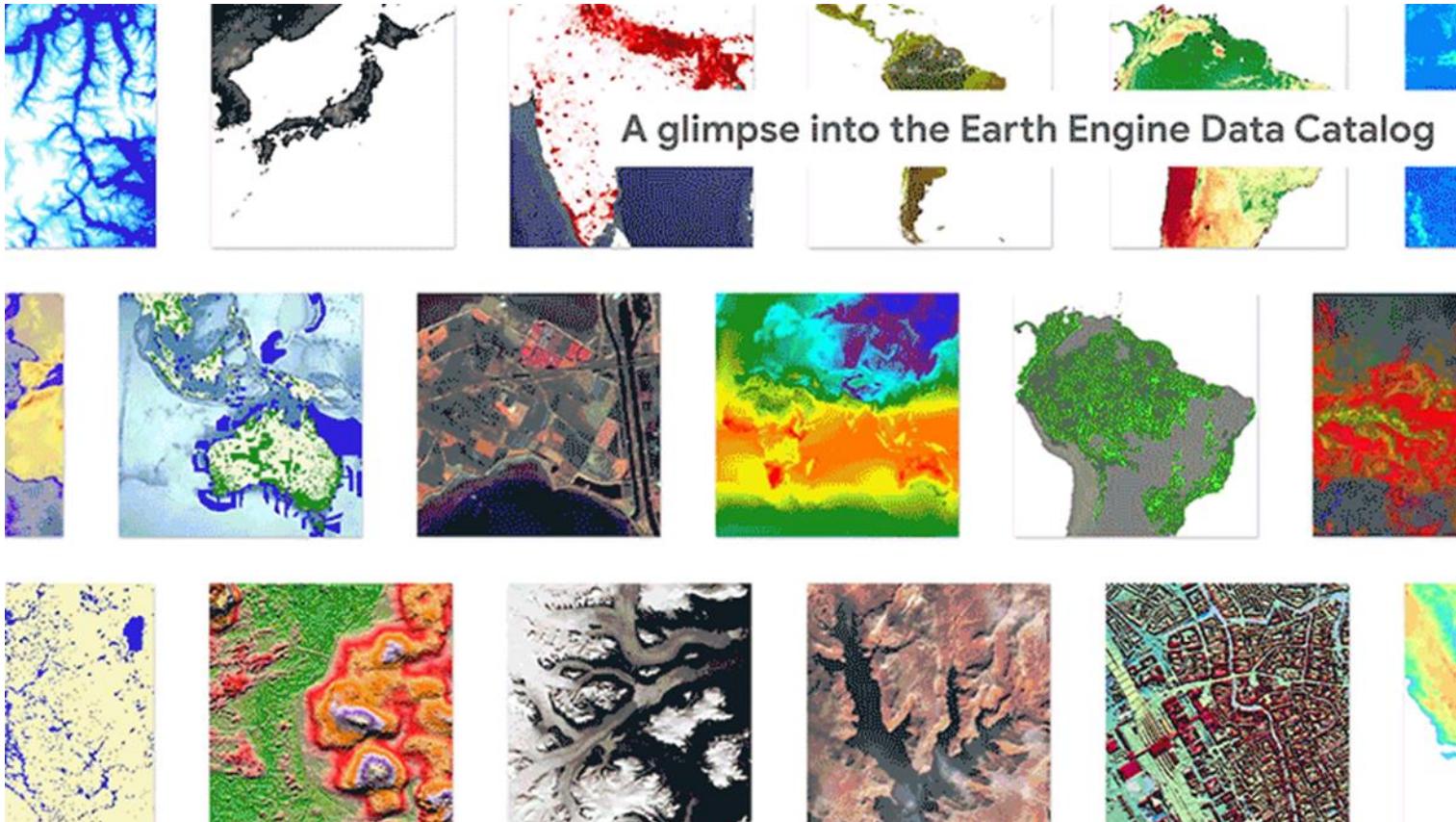
s
c
o
r
p
a
n



Google Earth Engine



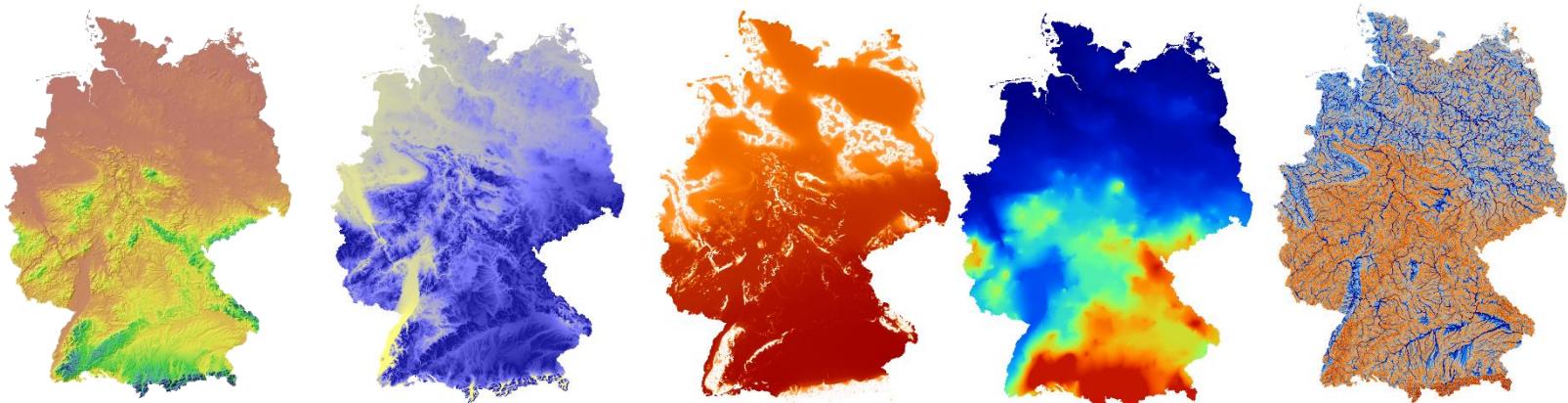
A glimpse into the Earth Engine Data Catalog



Choosing Covariates

s
c
o
r
p
a
n

- What factors influence soil formation?
 - Link to pedological knowledge
- What data do I have?
- Data should represent the multiple SORPAN covariates that influence soil development in the area

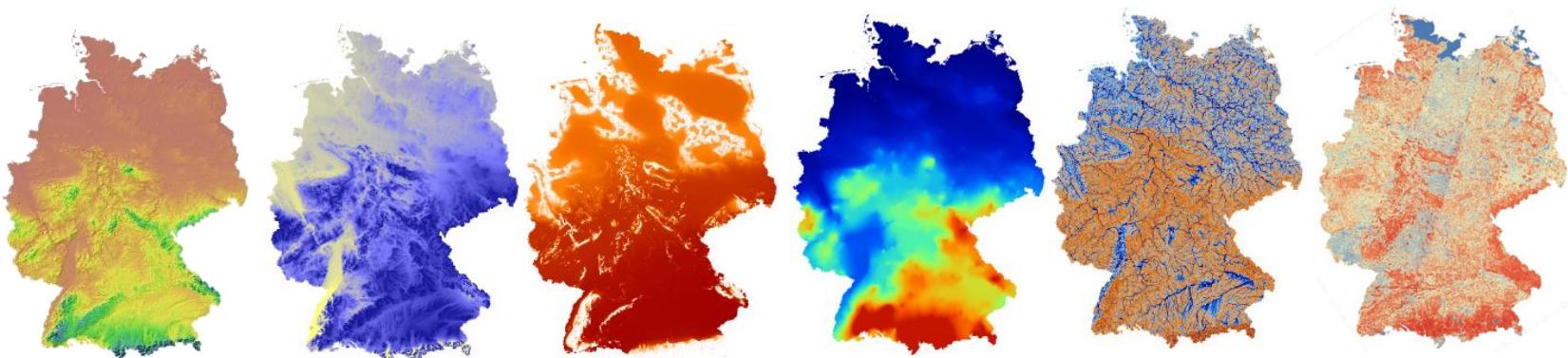


Technical and Practical Notes

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 multicollinearity problems,
4. Overlaying raster stacks and points

Technical and Practical Notes

1. *Converting polygon maps to rasters*

s

c

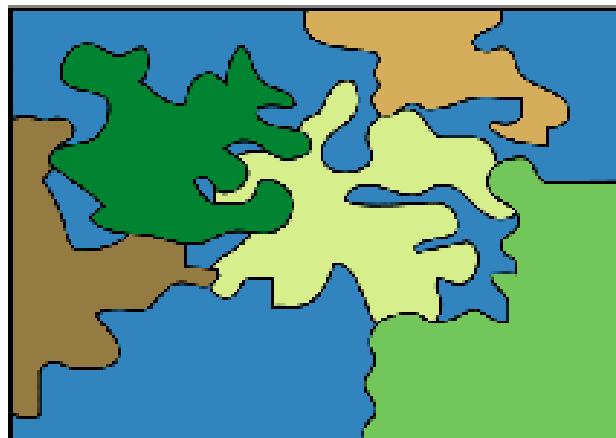
o

r

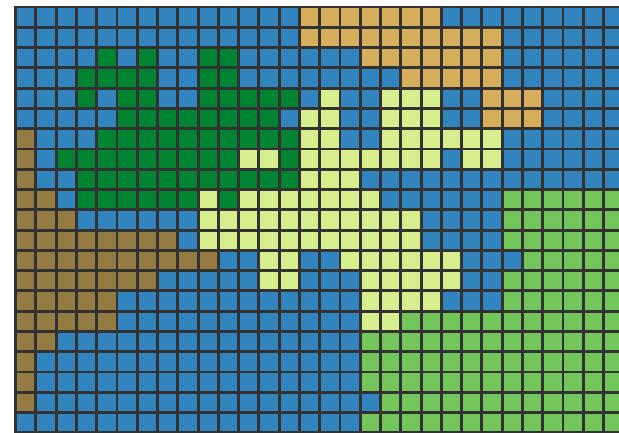
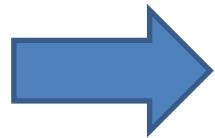
p

a

n



Polygon features



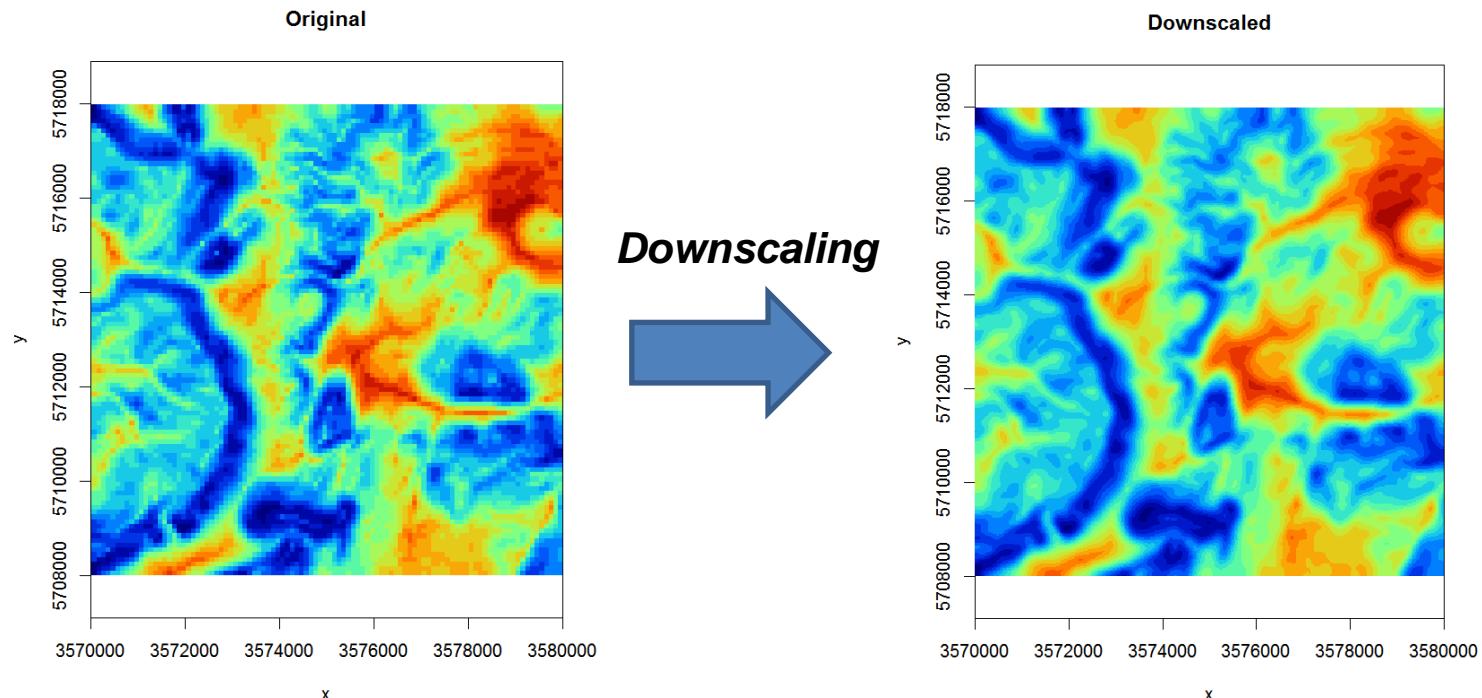
Raster polygon features

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

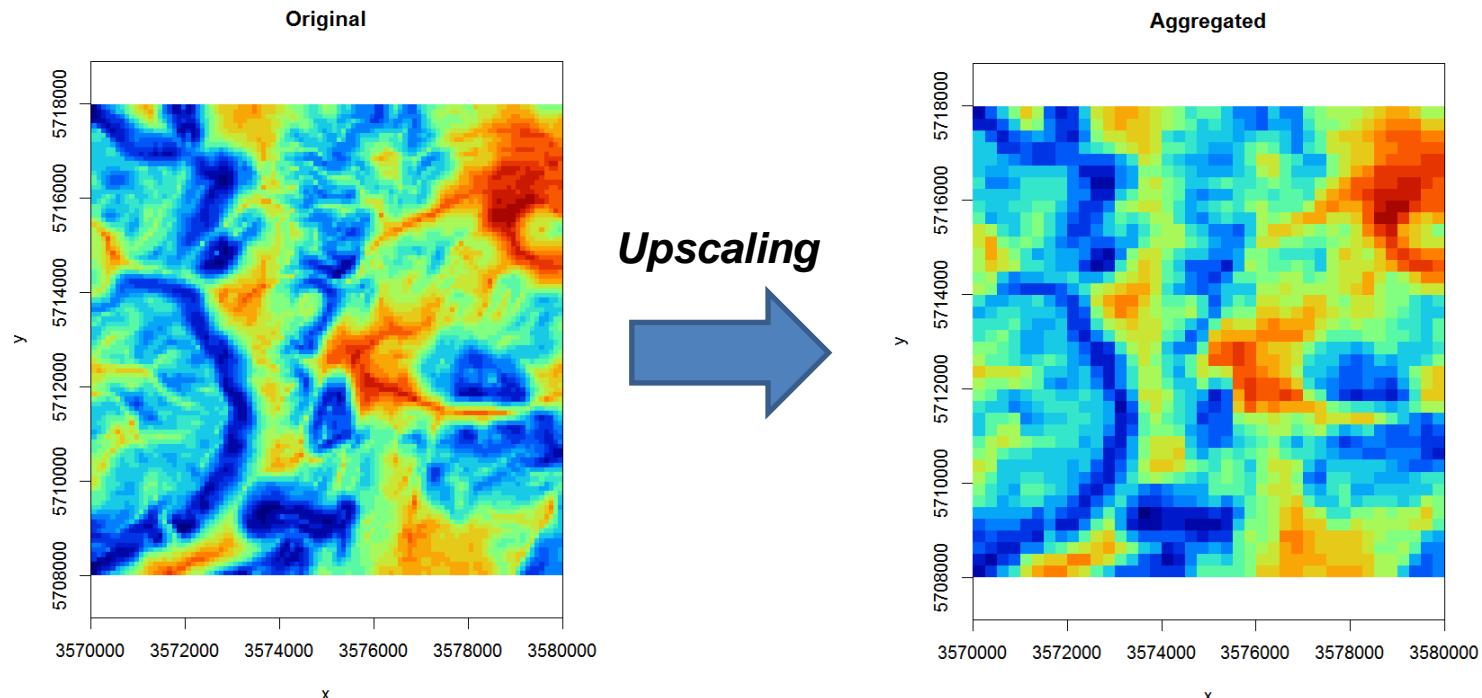


Technical and Practical Notes

s
c
o
r
p
a
n

2. Downscaling or upscaling rasters

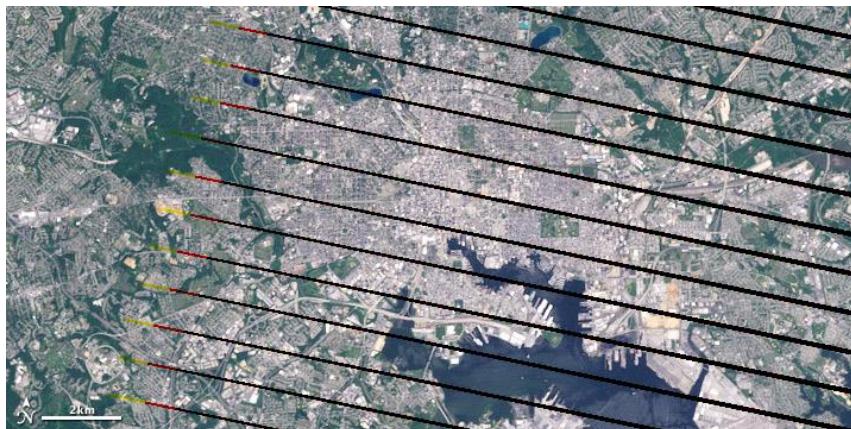
- To adjust the resolution of some covariates that have either too coarse or too fine a resolution compared to the target resolution
- The process of bringing raster layers to a common grid resolution is also known as resampling



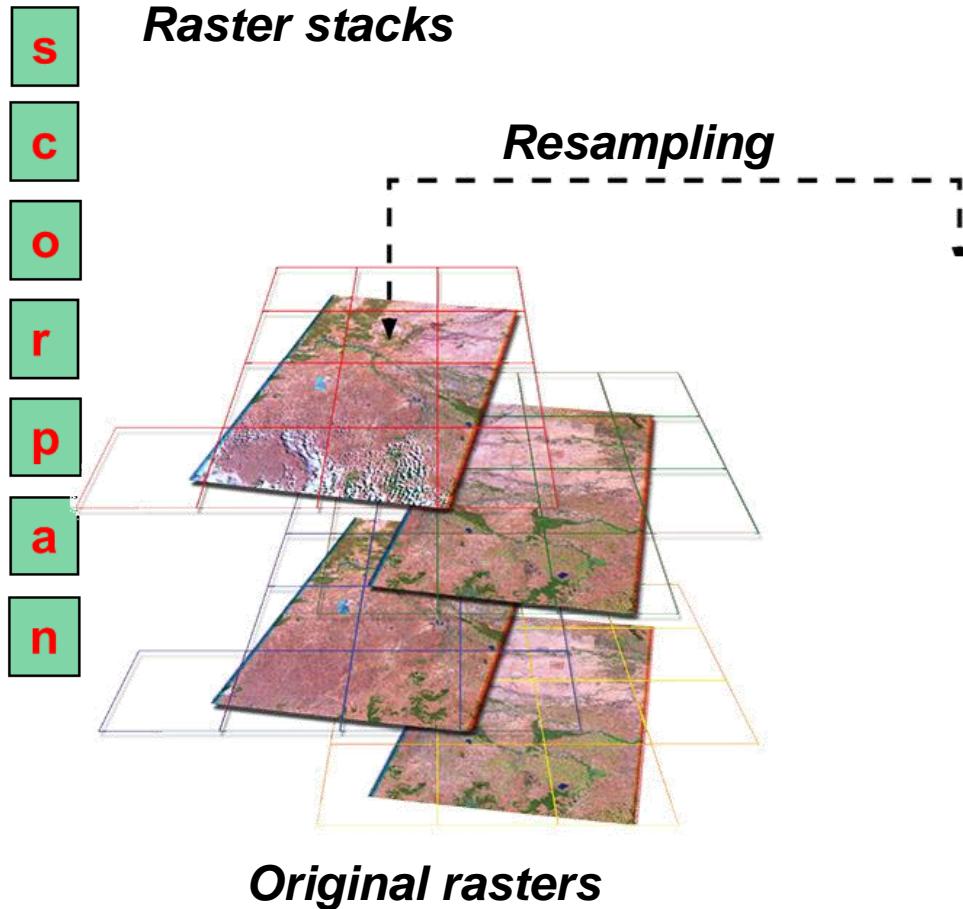
Technical and Practical Notes

3. *Filtering out missing pixels and artifacts*

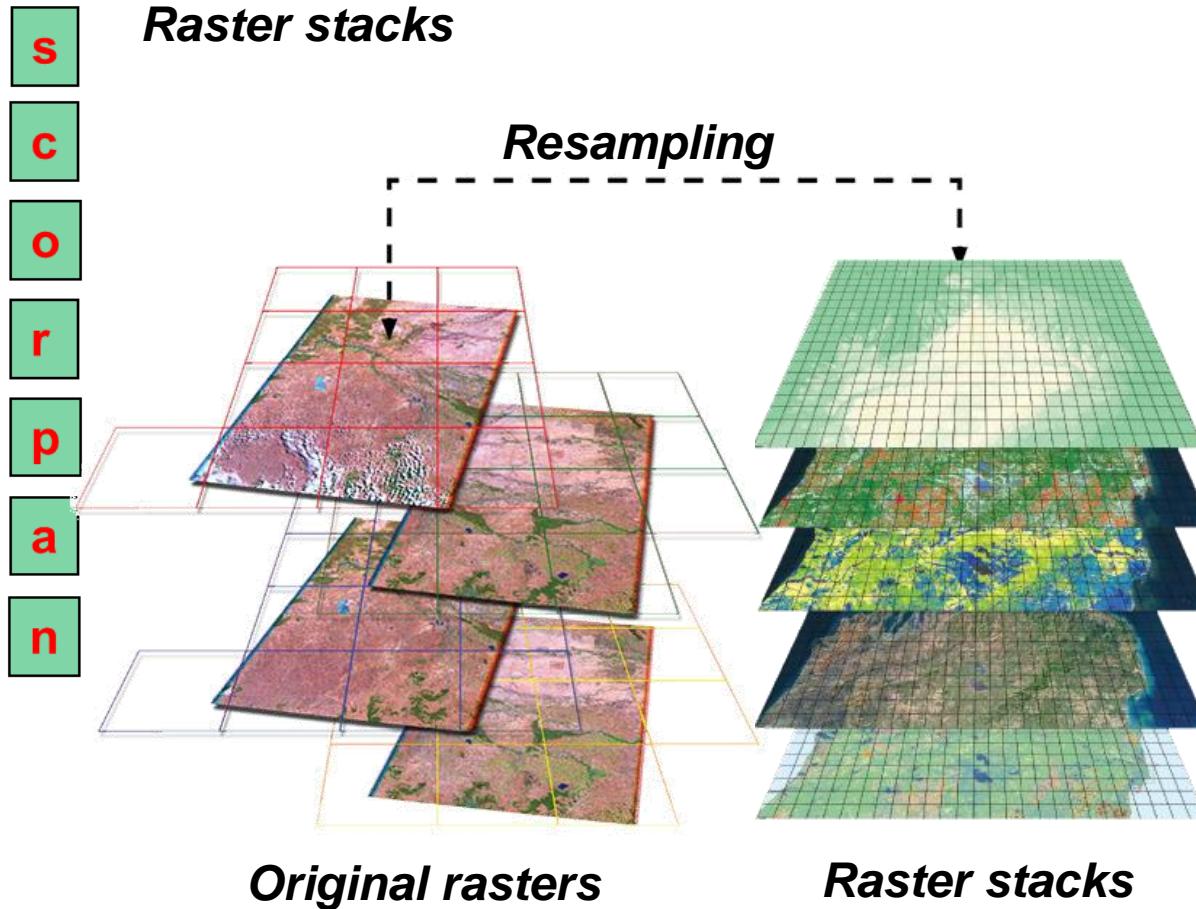
- could cause serious problems for producing soil maps as the missing pixels and artifacts would propagate to predictions: if only one layer in the raster stack misses values then predictive models might drop whole rows in the predictions even though data is available for 95% of rows.
- Missing pixels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



Technical and Practical Notes



Technical and Practical Notes

s

c

o

r

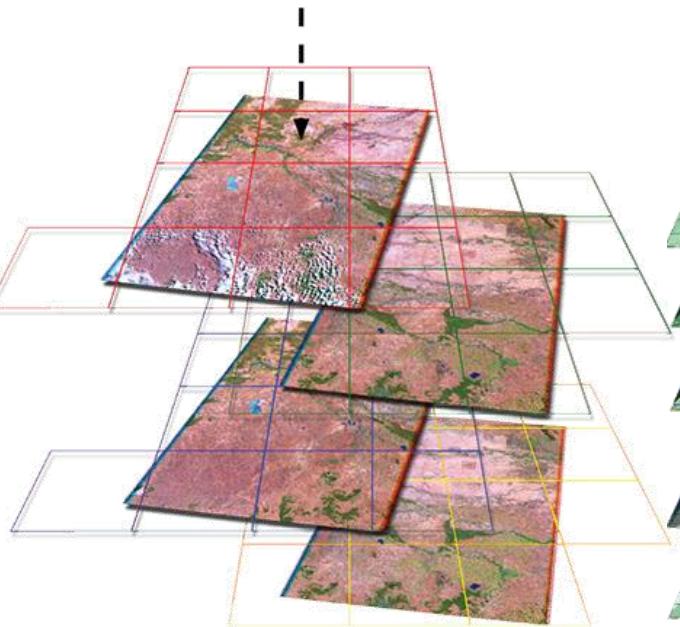
p

a

n

Raster stacks

Resampling



Original rasters

Raster stacks

Technical and Practical Notes

s

c

o

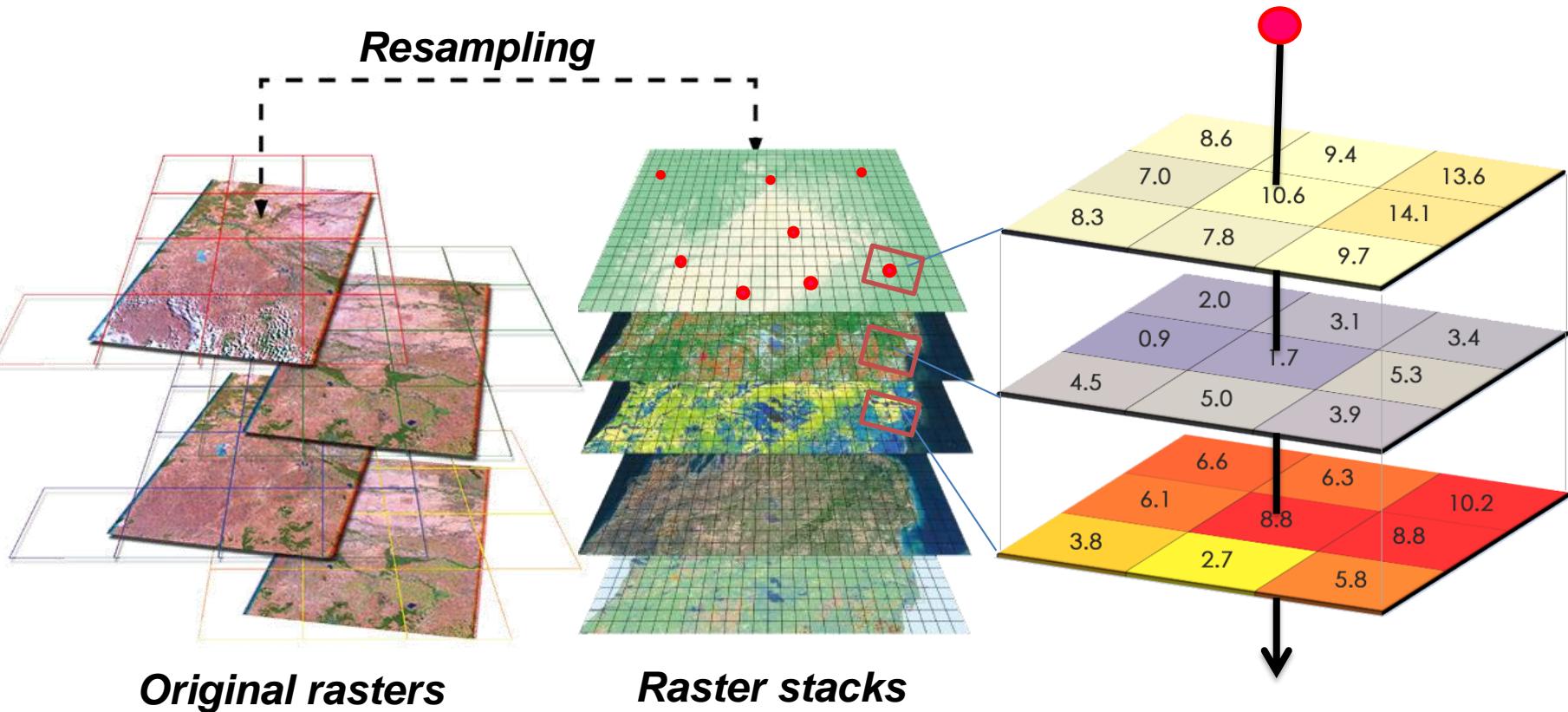
r

p

a

n

4. Overlaying raster stacks and points



Technical and Practical Notes

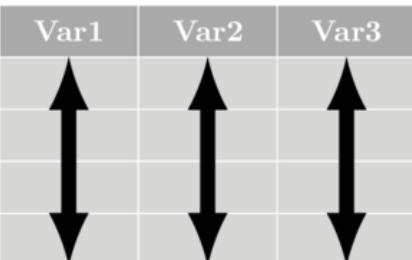
Geo-database
Tidy data
(*data.frame*)

Predictors
Covariates
Independent variables
Features

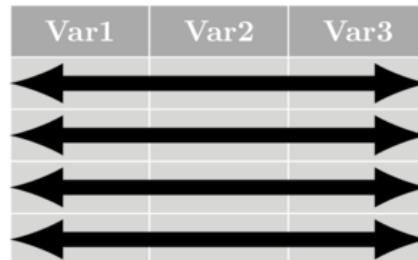


X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15	Class
-0.60	3.06	-5.77	0.69	1.46	1.22	1.94	0.61	-0.07	-2.02	-0.87	0.71	0.89	-0.91	0.21	a
0.61	1.61	0.21	2.74	0.46	1.29	0.13	0.55	-0.60	-0.53	0.70	-0.41	-0.62	1.17	-0.32	b
3.81	0.99	-1.21	0.30	0.11	-0.34	-0.29	0.15	-0.09	1.30	0.08	0.66	-0.44	0.45	-0.13	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

Variables in Columns

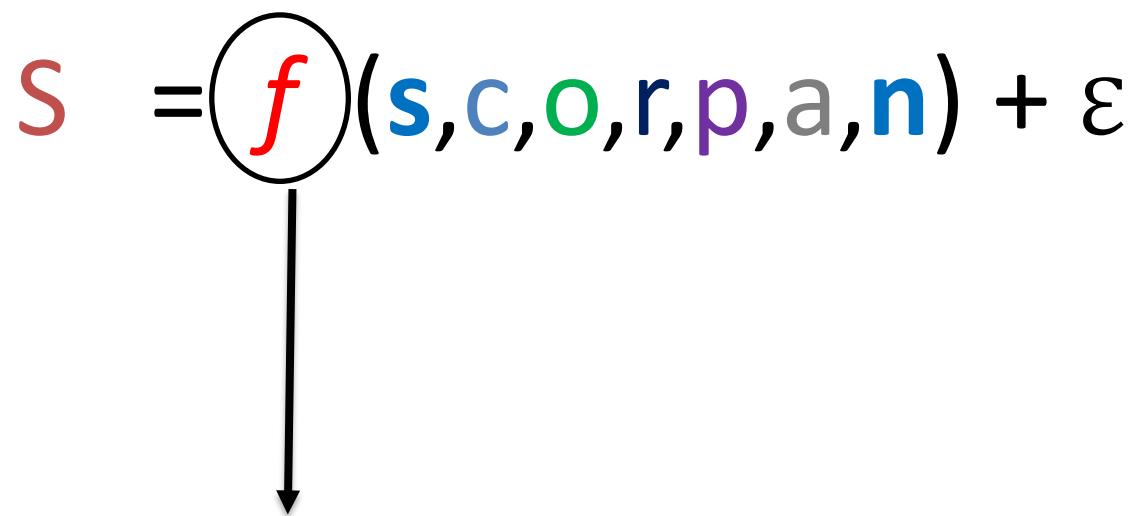


Observations in Rows



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

SCORPAN Model

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


Machine learning

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



Take A Break

5:00

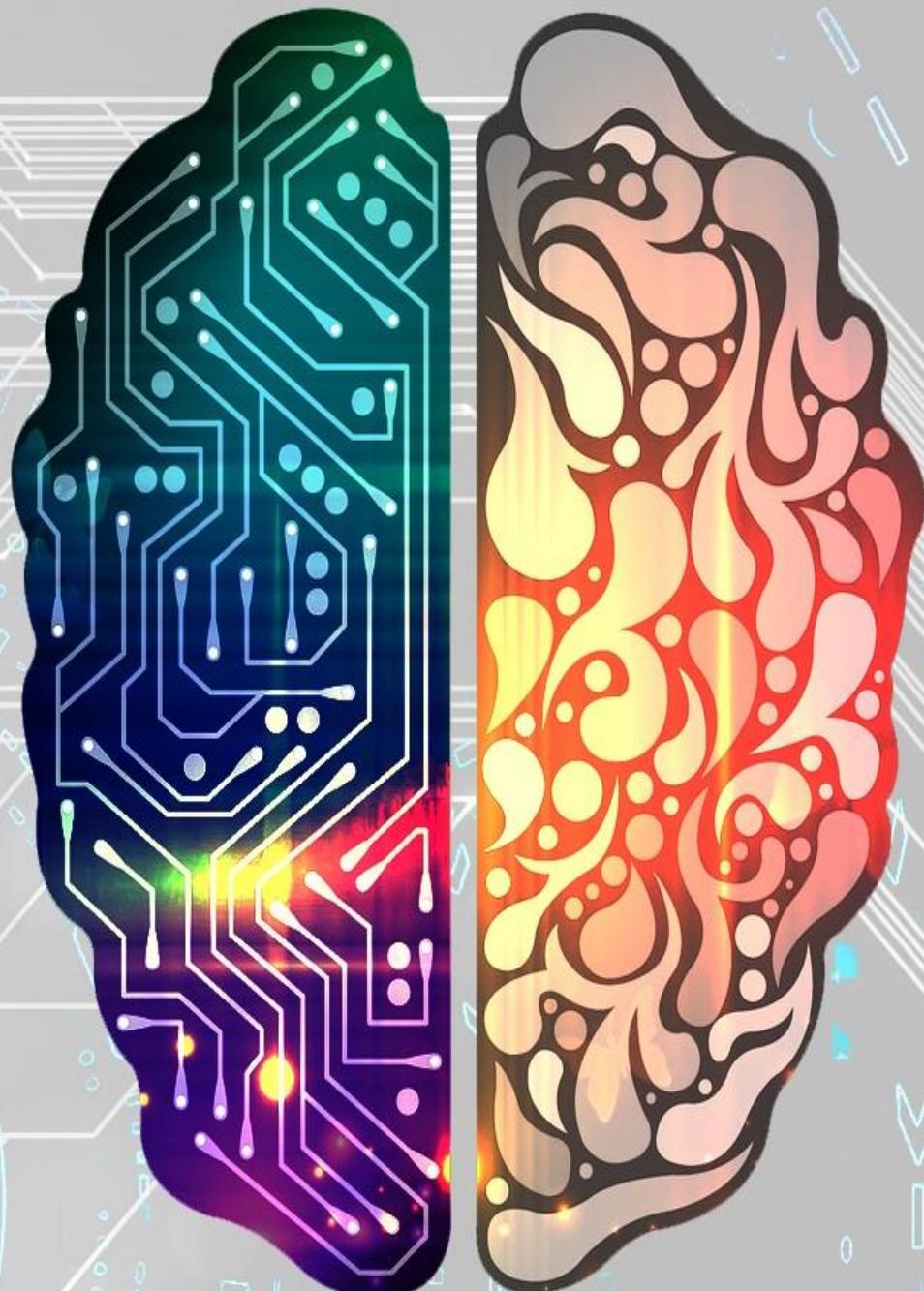
Sequence of DSM Steps



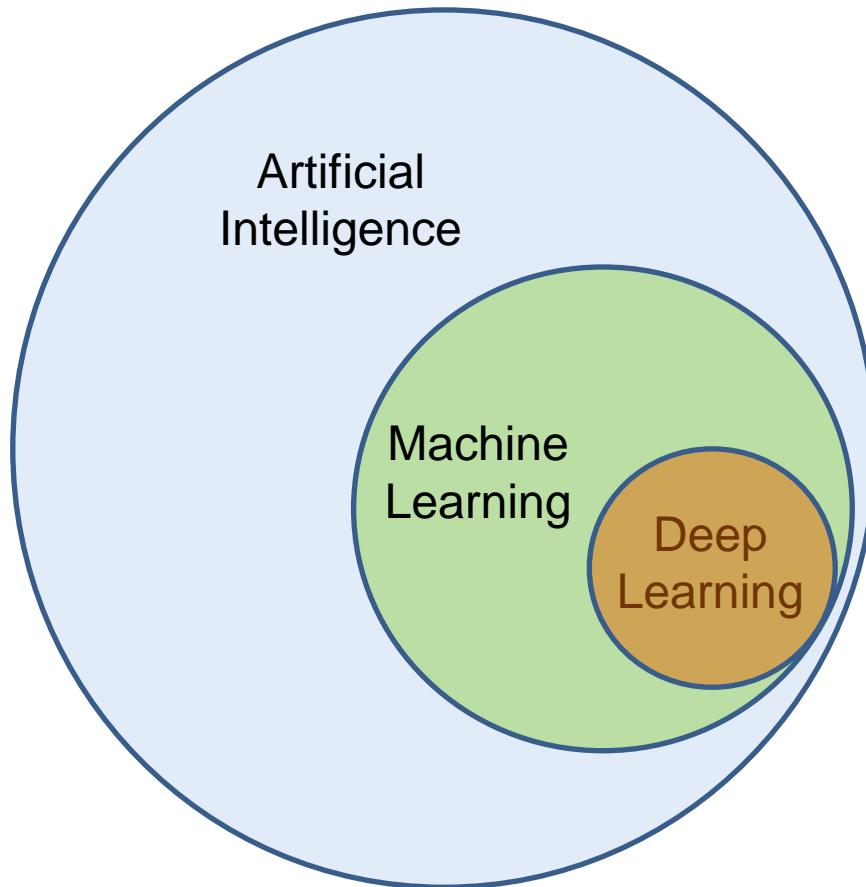
- 1 Environmental covariates, relevant as predictors of soil property/class, are derived from remote sensing, digital elevation, climatic datasets, ...
- 2 Soil samples are collected at the specified locations (e.g., Latin hypercube sampling) and soil property is measured in the laboratory.
- 3 Intersecting 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.

Introduction: some terms

MACHINE LEARNING



AI, ML, and DL

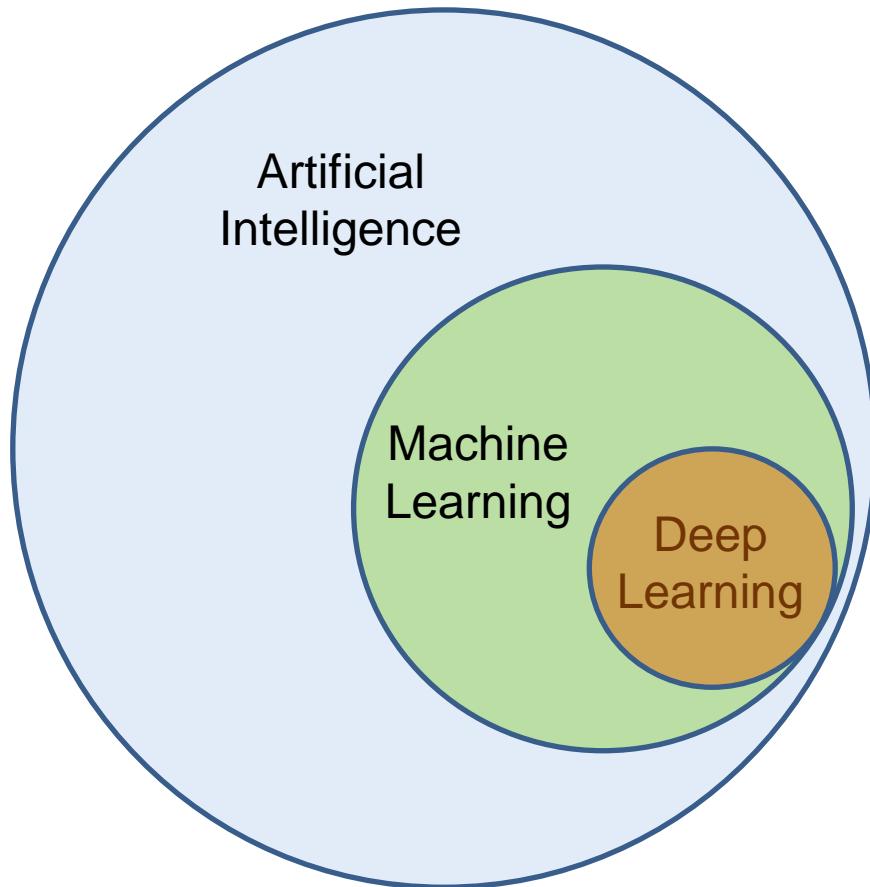


Artificial Intelligence
A technique which enables machines to mimic human behavior

Machine Learning
Subset of AI technique which use statistical methods to enable machines to improve with training and experience

Deep learning
Subset of ML technique where a system can train itself to perform tasks

AI, ML, and DL



Artificial Intelligence

A technique which enables machines to mimic human behavior



Machine Learning

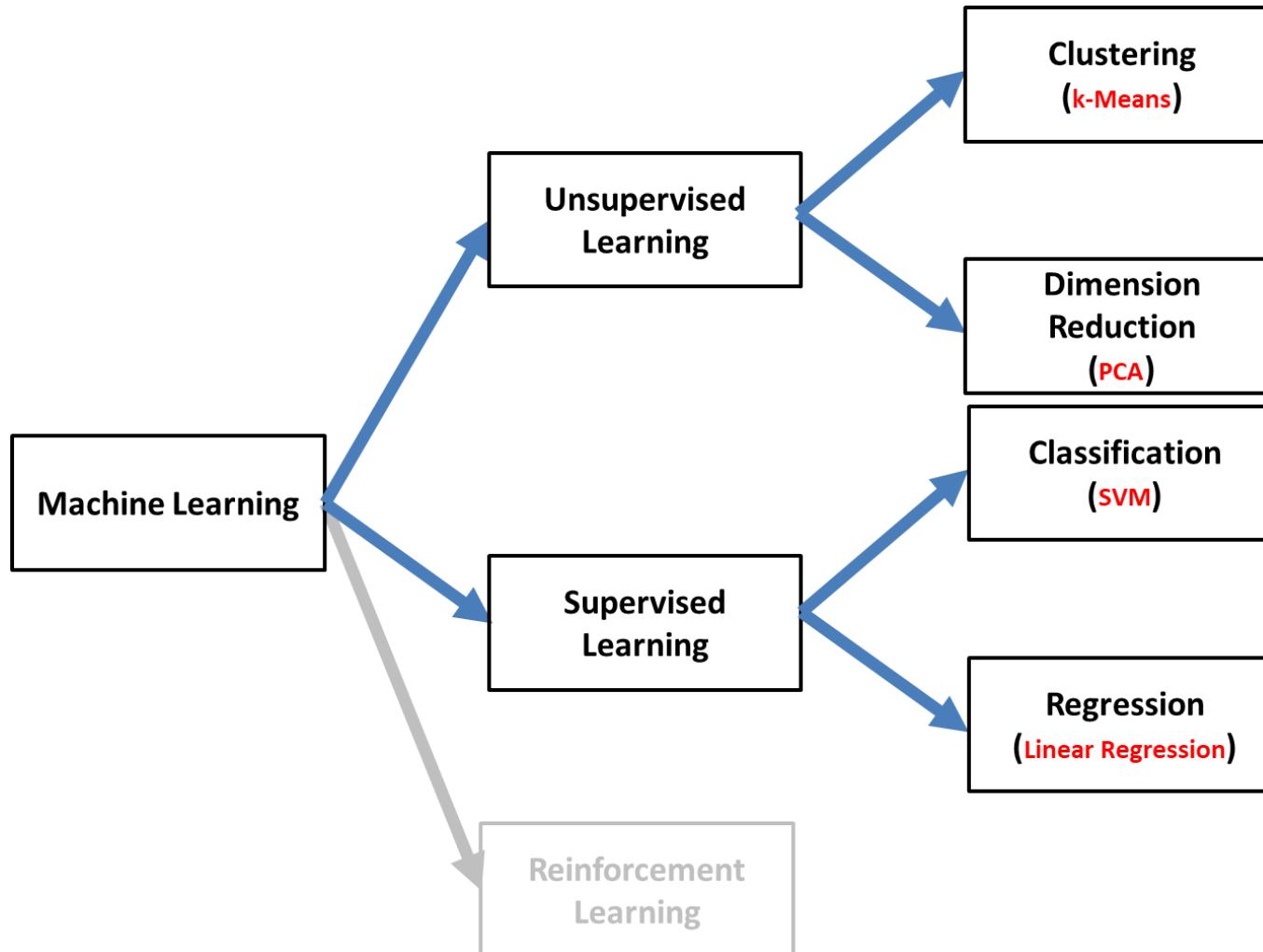
Subset of AI technique which use statistical methods to enable machines to improve with training and experience



Deep learning

Subset of ML technique where a system can train itself to perform tasks

Types of Machine Learning



Unsupervised Learning

Unsupervised learning

- No response, just “**covariates**”, (x_i)
- Algorithm identifies clusters
- No training data required
- e.g., clustering of satellite image by similar values

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

Advantages

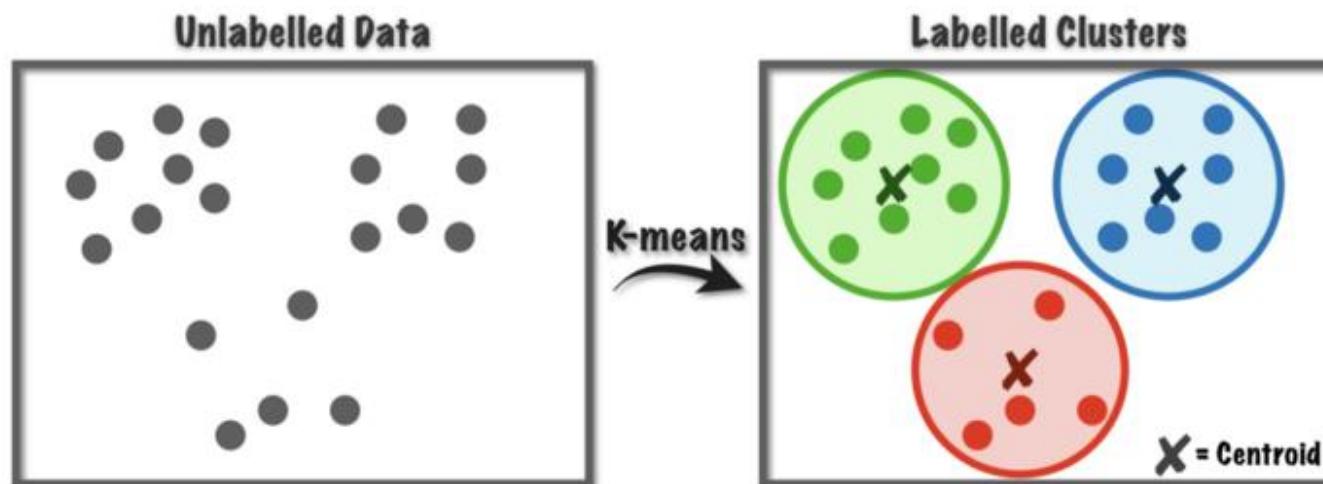
- Unsupervised classification is fairly quick and easy to run
- There is no extensive prior knowledge of area required

Disadvantages

- The classes do not always correspond to informational classes
- The user also has to spend time interpreting and label the classes following the classification

Clustering

- Clustering can be considered the most important unsupervised learning problem, it deals with finding a structure in a collection of **unlabeled data**.
- Clustering algorithms group the samples of a set such that two samples in the same cluster are **more similar** to one another than two samples from different clusters
- One option: small Euclidean Distance (squared)



A data set with clear cluster structure

k-means Clustering

Basic Algorithm:

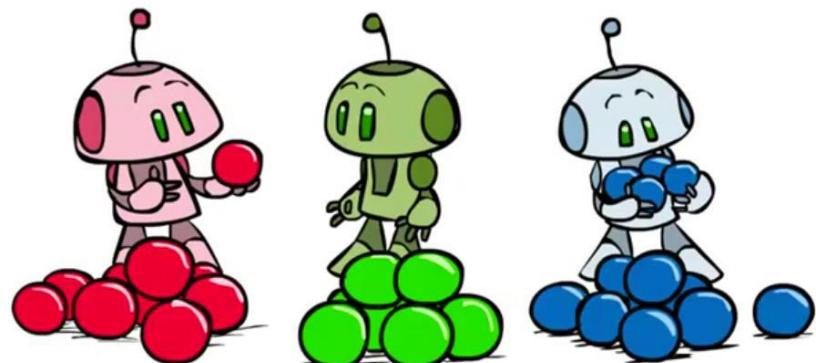
- Step 0: select K
- Step 1: randomly select k initial cluster centers
- Step 2: calculate distance from each object to each cluster centers.
- Step 3: assign each object to the closest cluster
- Step 4: compute the new centroid for each cluster

Iterate:

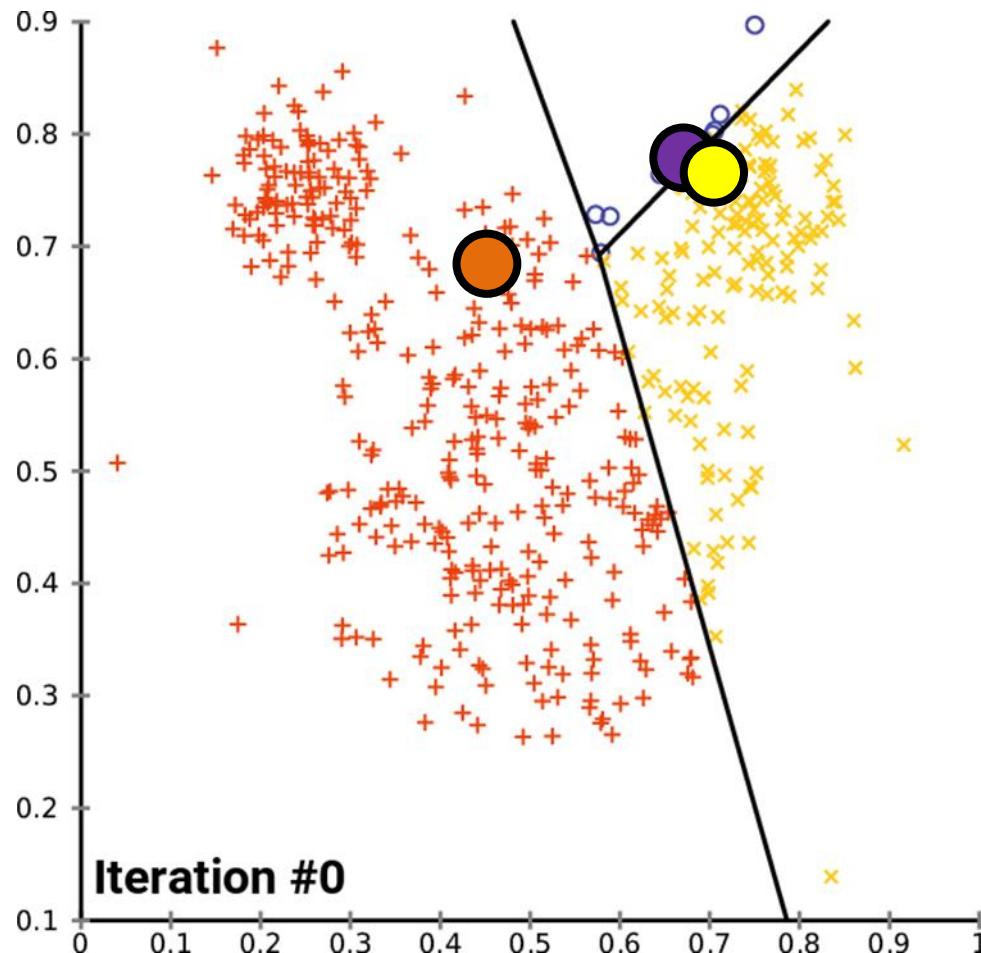
- calculate distance from objects to cluster centroids.
- assign objects to closest cluster
- recalculate new centroids

Stop based on convergence criteria

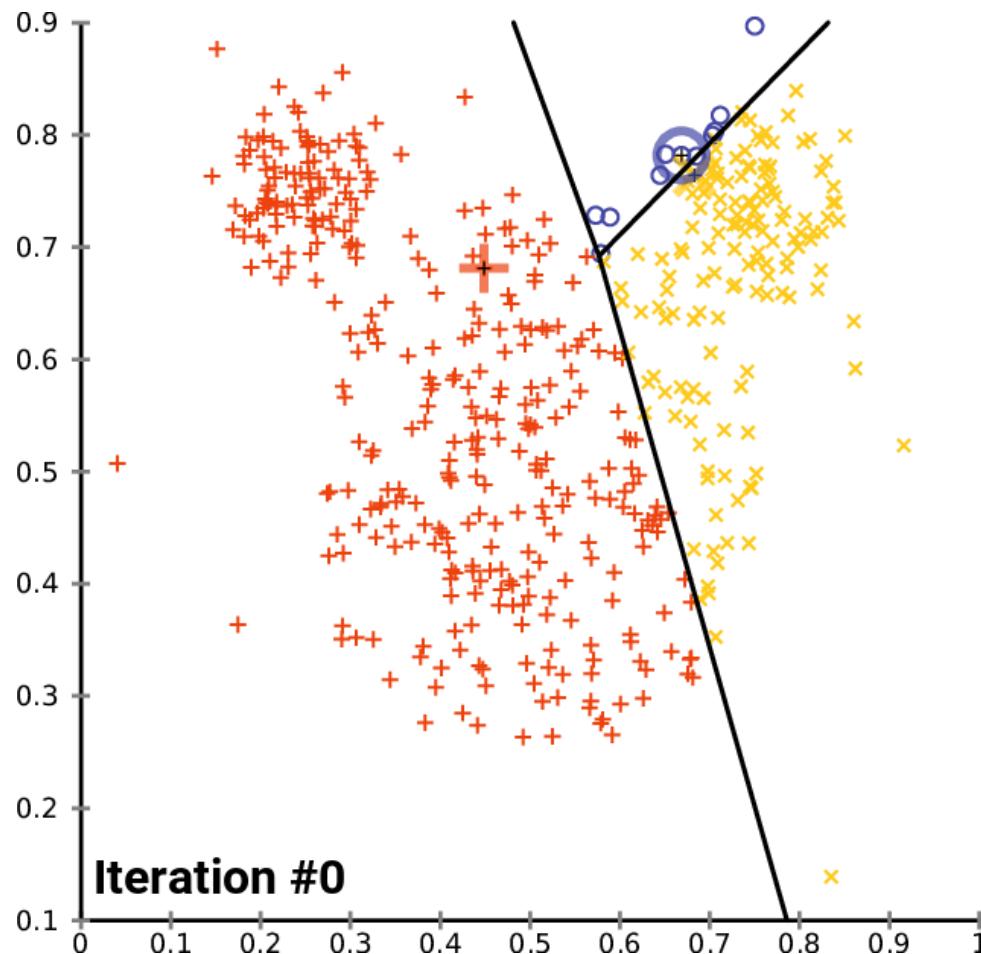
- no change in clusters
- max iterations



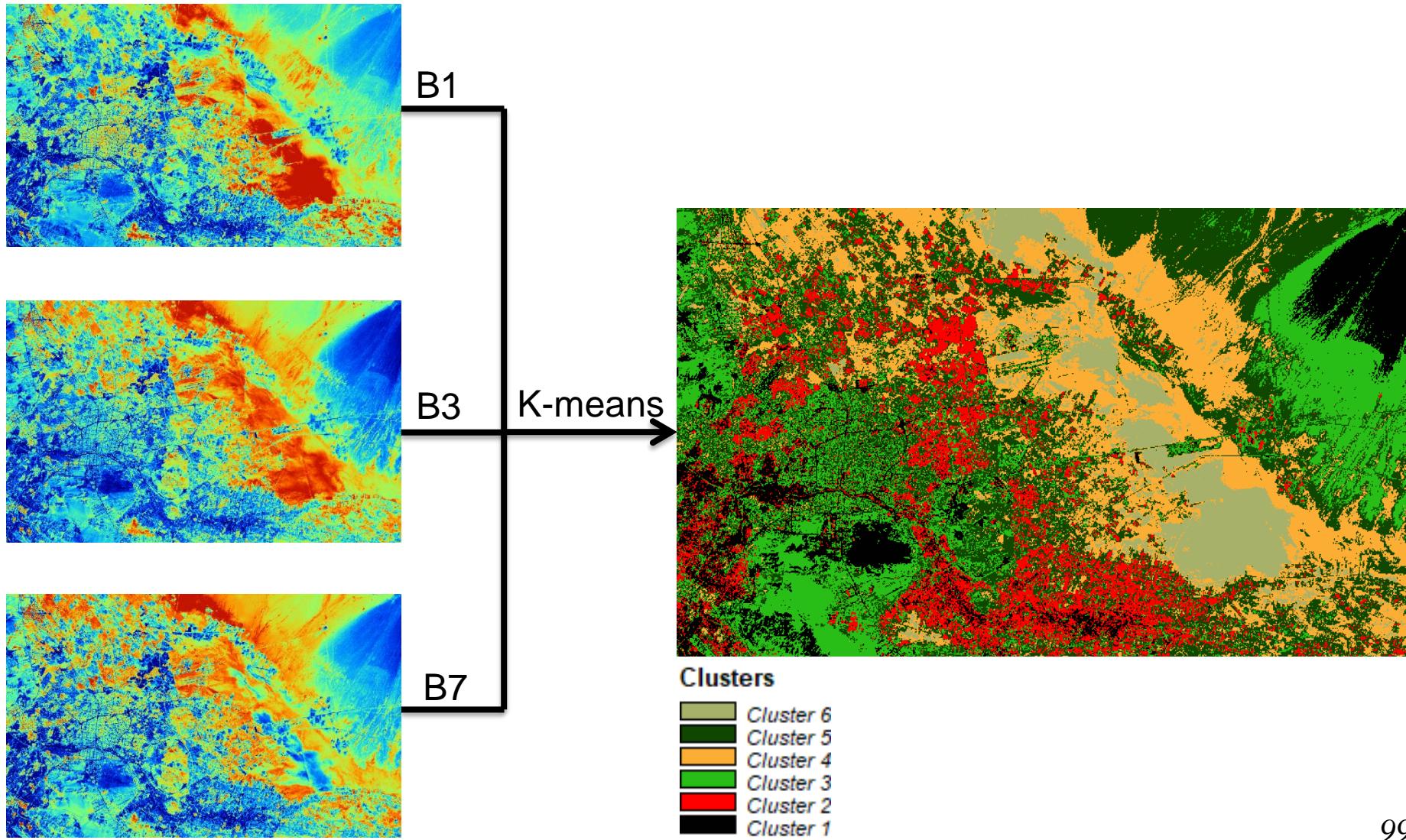
k-means Clustering



k-means Clustering



Clustering: example



Dimension Reduction

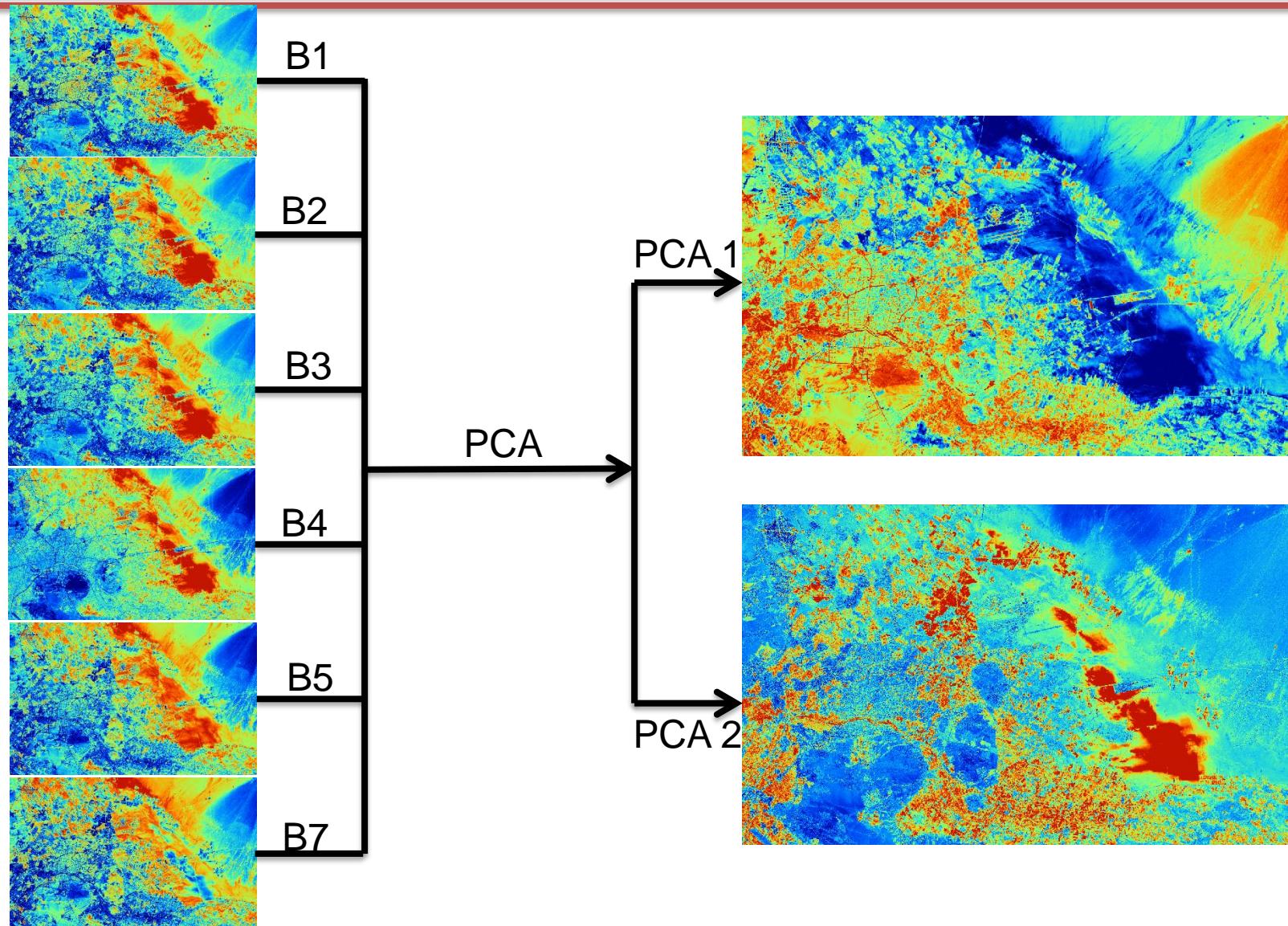
Principal Components Analysis Ideas (PCA)

- It is often necessary to reduce the dimension of a dataset, a large number of covariates
- Transform some large number of covariates into a smaller number of uncorrelated covariates called principal components (PCs).
- Developed to capture as much of the variation in data as possible (minimal loss of information)

■ See online tutorials such as

http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf

PCA: example



Supervised Learning

Supervised learning

- For each covariate value x_i there is also a response value 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

Cubist, Logistic Model Tree

Artificial Neural Networks,
Convolutional Neural
Networks

k Nearest Neighbors, Nearest
Shrunken Centroid

Tree-Based

**Linear-
Models**

**Model-
Trees**

**Expert-
Knowledge**

**Neural
Networks**

**Support
Vector
Machines**

**Distance-
Based
Learners**

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

Fuzzy Inference Systems,
Rule Induction Algorithms

Linear SVM, Polynomial SVM,
Radial Basis Function SVM



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

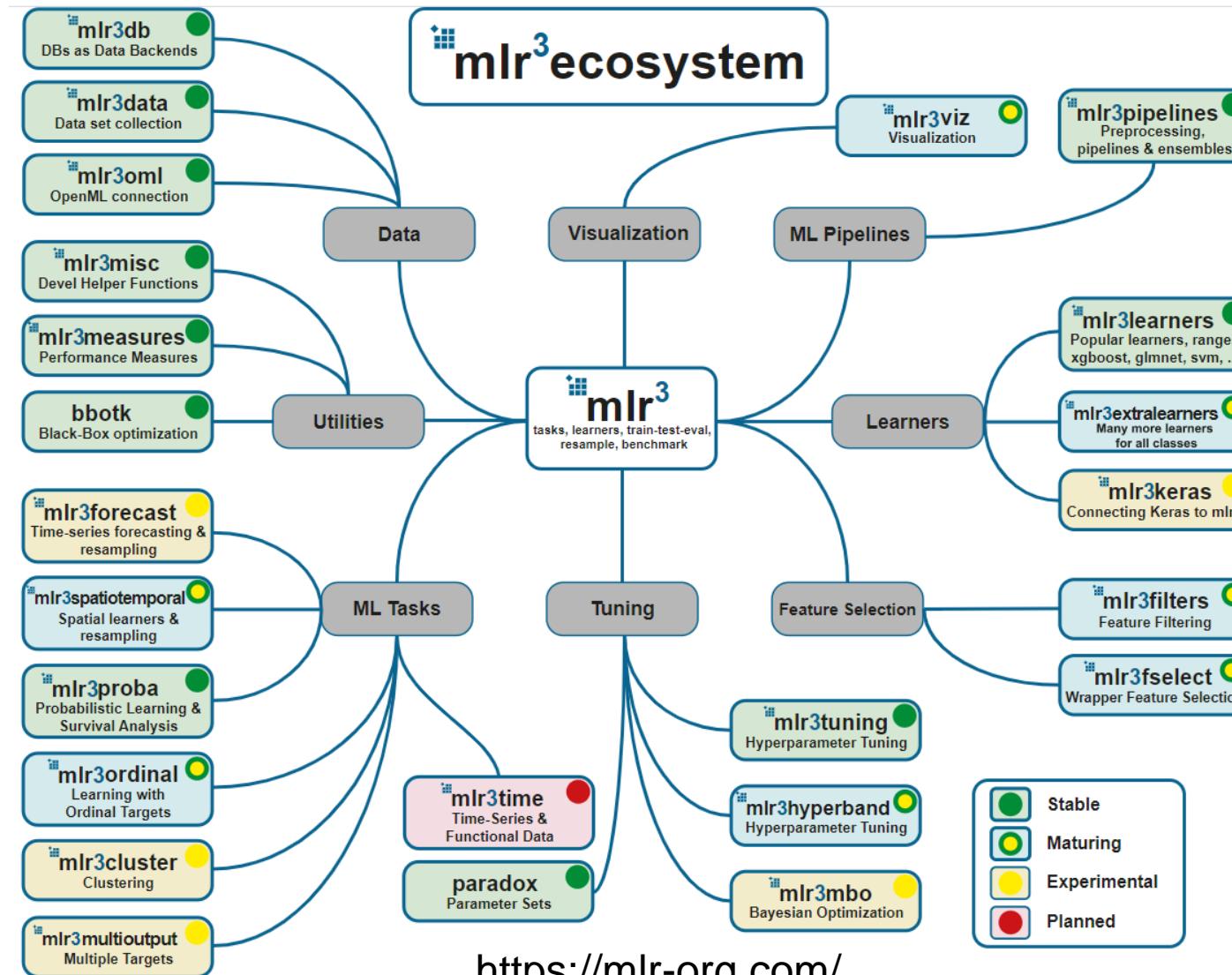
The models below are available in `train`. The code behind these protocols can be obtained using the function `getModelInfo` or by going to the [github repository](#).

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	amdaai		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

Available Models in R





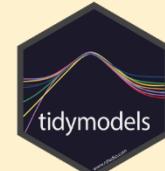
Available Models in R

Import



↓
Tidy → Transform

Visualize



Model



Share

- Whether you are just starting out today or have years of experience with modeling, tidymodels offers a consistent, flexible framework for your work.
- The tidymodels framework is a collection of packages for modeling and machine learning using tidyverse principles.

Things to Keep in Mind

- Each model predicts **one realization** of a DSM
- **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.
- Some models are more **computationally** efficient than others.
- Each model may produce drastically **different results** given the same inputs.
- Some models tend to **overfit**.

From Available Models

Decision
Tree

Random
Forest

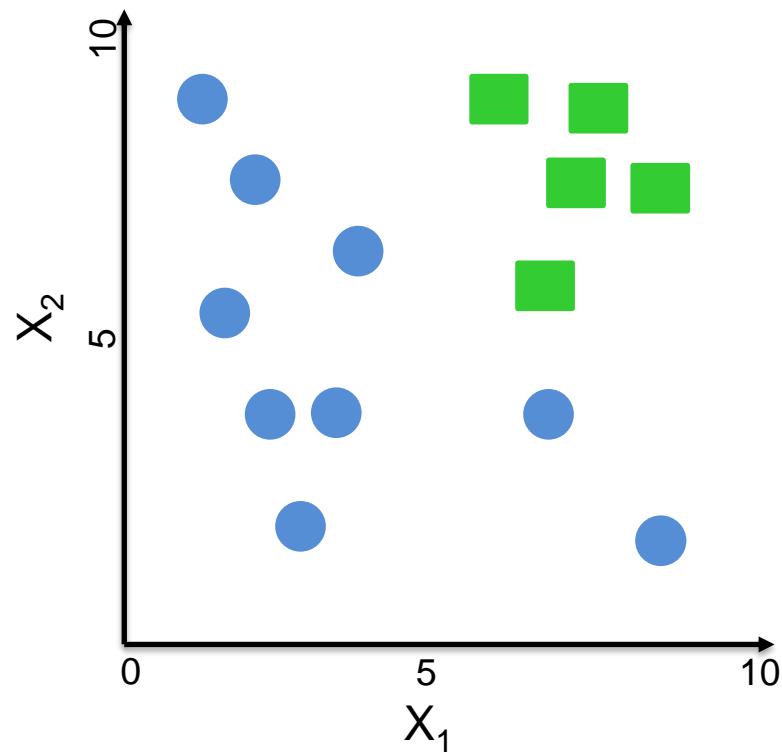
Decision Tree



Decision Tree: Definition

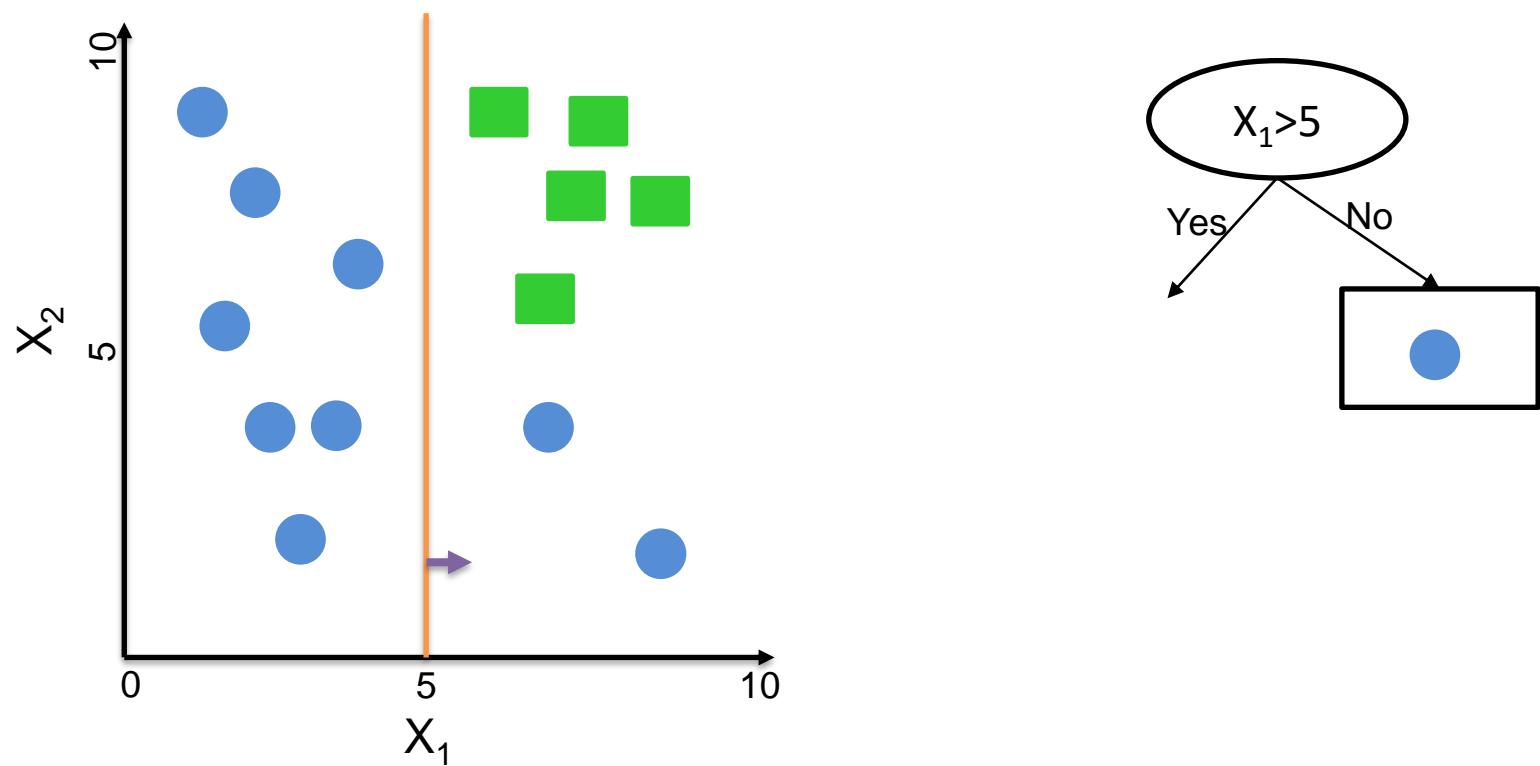
- Decision Tree is a supervised learning approach.
- It can be used for solving regression (e.g., clay content) and classification (e.g., soil types) problems.
- It partitions the data into subsets.
- The partitioning process starts with a binary split (Yes-No) and continues until no further splits can be made.
- It predicts the outcome by asking a set of **if-else** questions.
- It resembles a **tree-like graph, easy to understand**

Decision Tree: example I



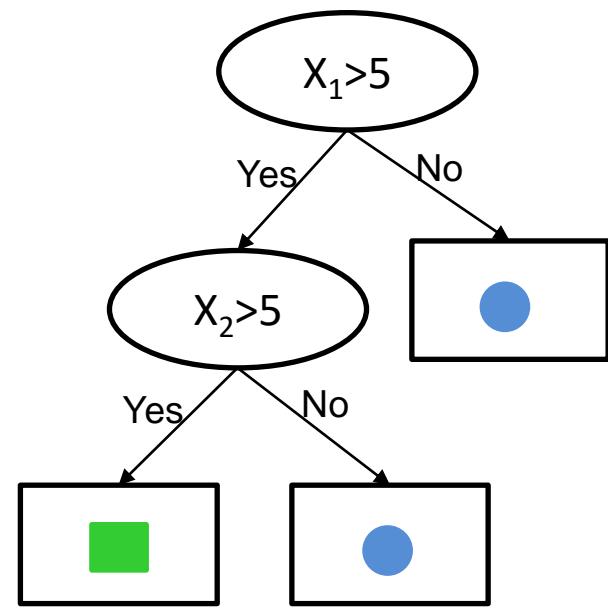
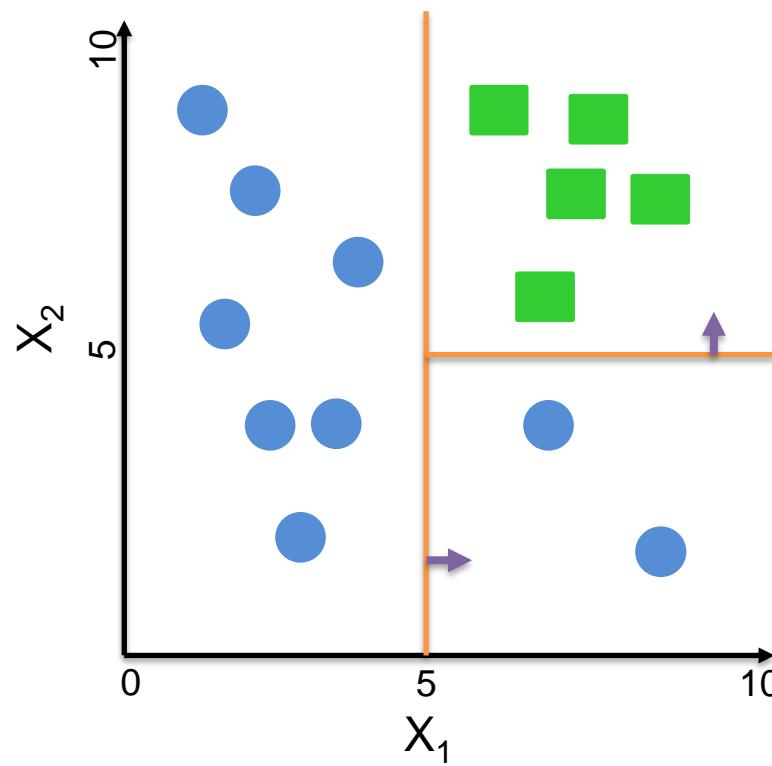
Example of a dataset and the corresponding decision tree. (Alpaydin, 2010)

Decision Tree: example I



Example of a dataset and the corresponding decision tree. (Alpaydin, 2010)

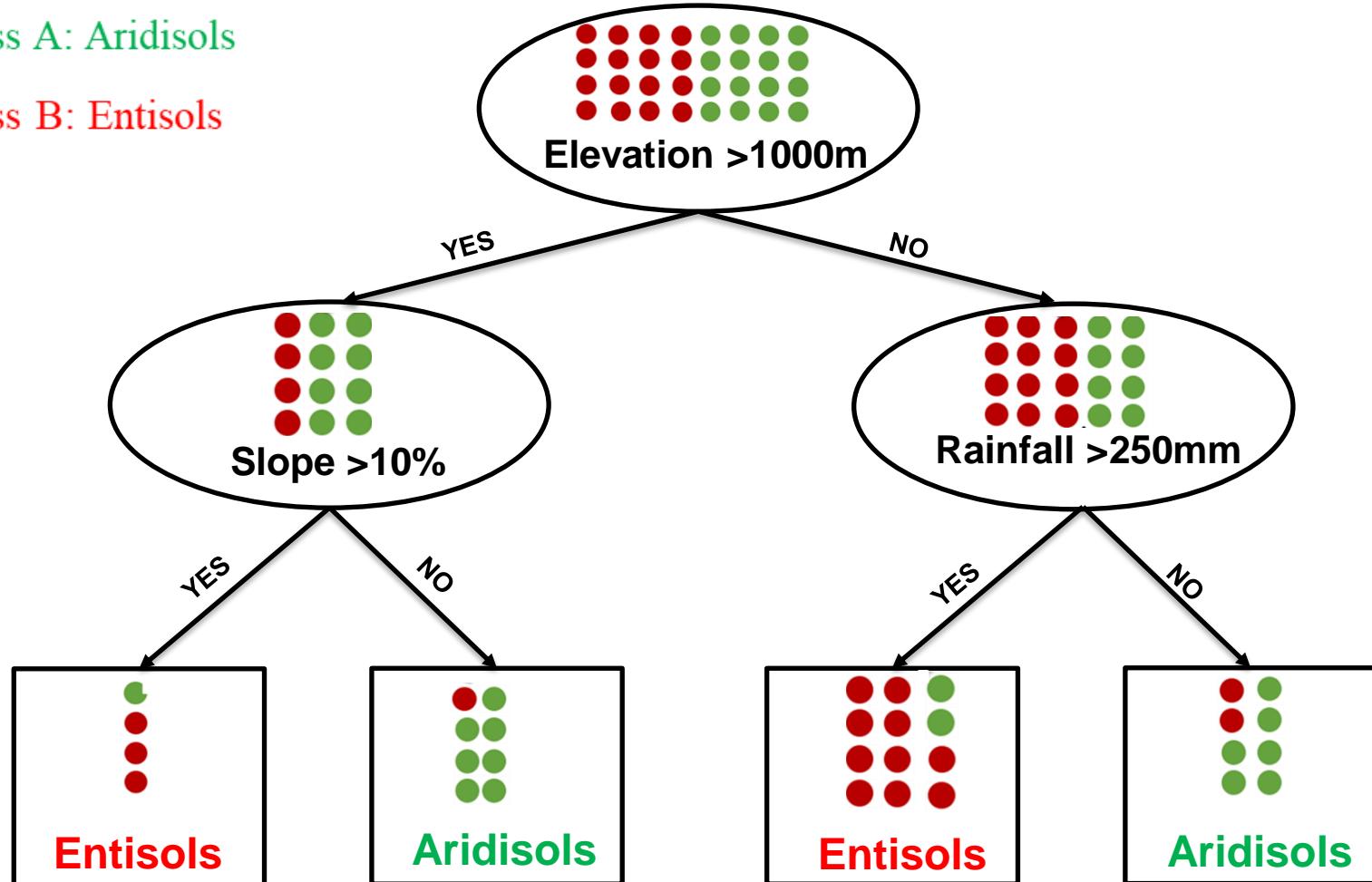
Decision Tree: example I



Example of a dataset and the corresponding decision tree. (Alpaydin, 2010)

Decision Tree: example II

- Class A: Aridisols
- Class B: Entisols

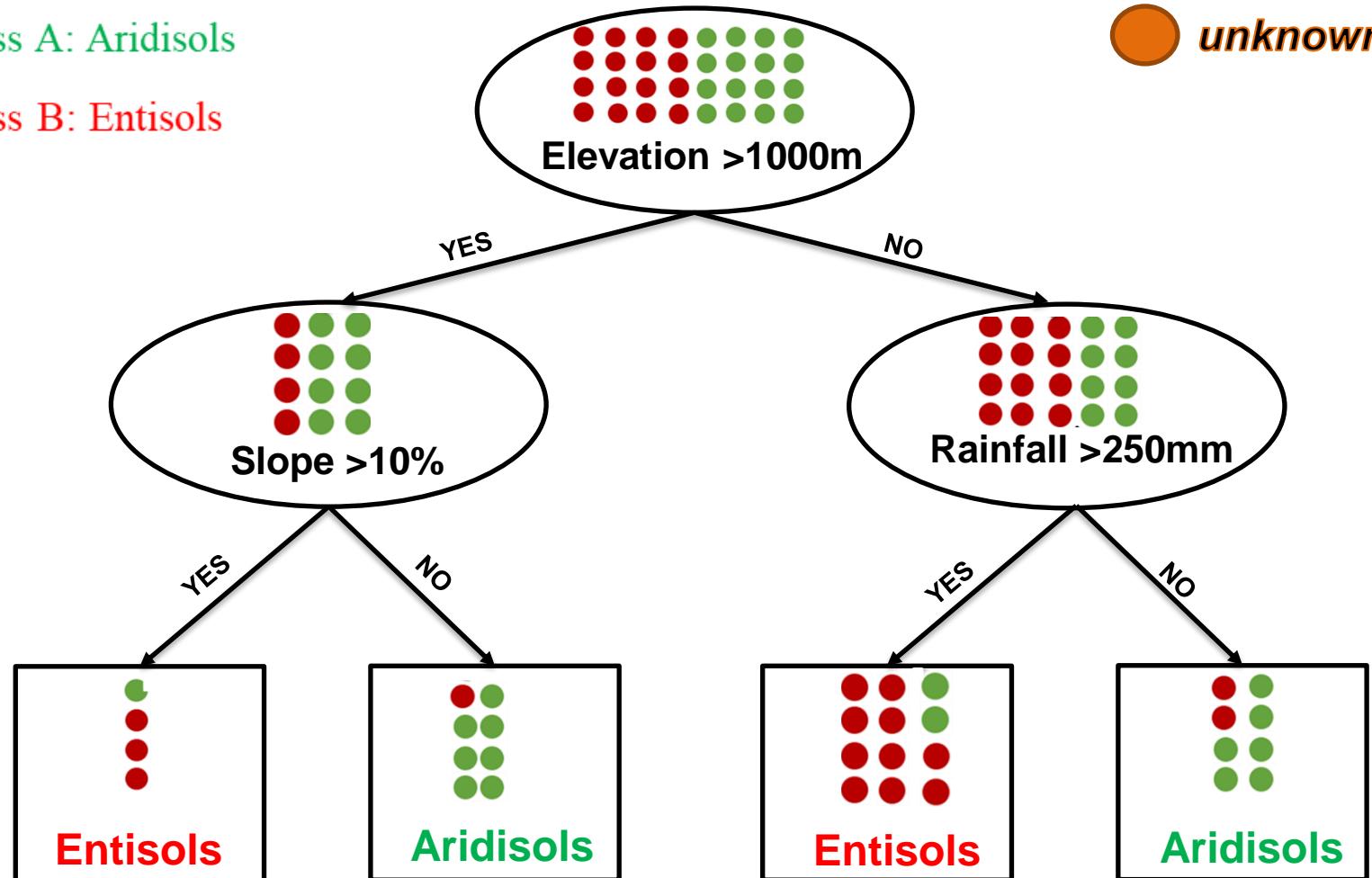


A decision tree model to predict Entisols and Aridisols

Decision Tree: example II

- Class A: Aridisols
- Class B: Entisols

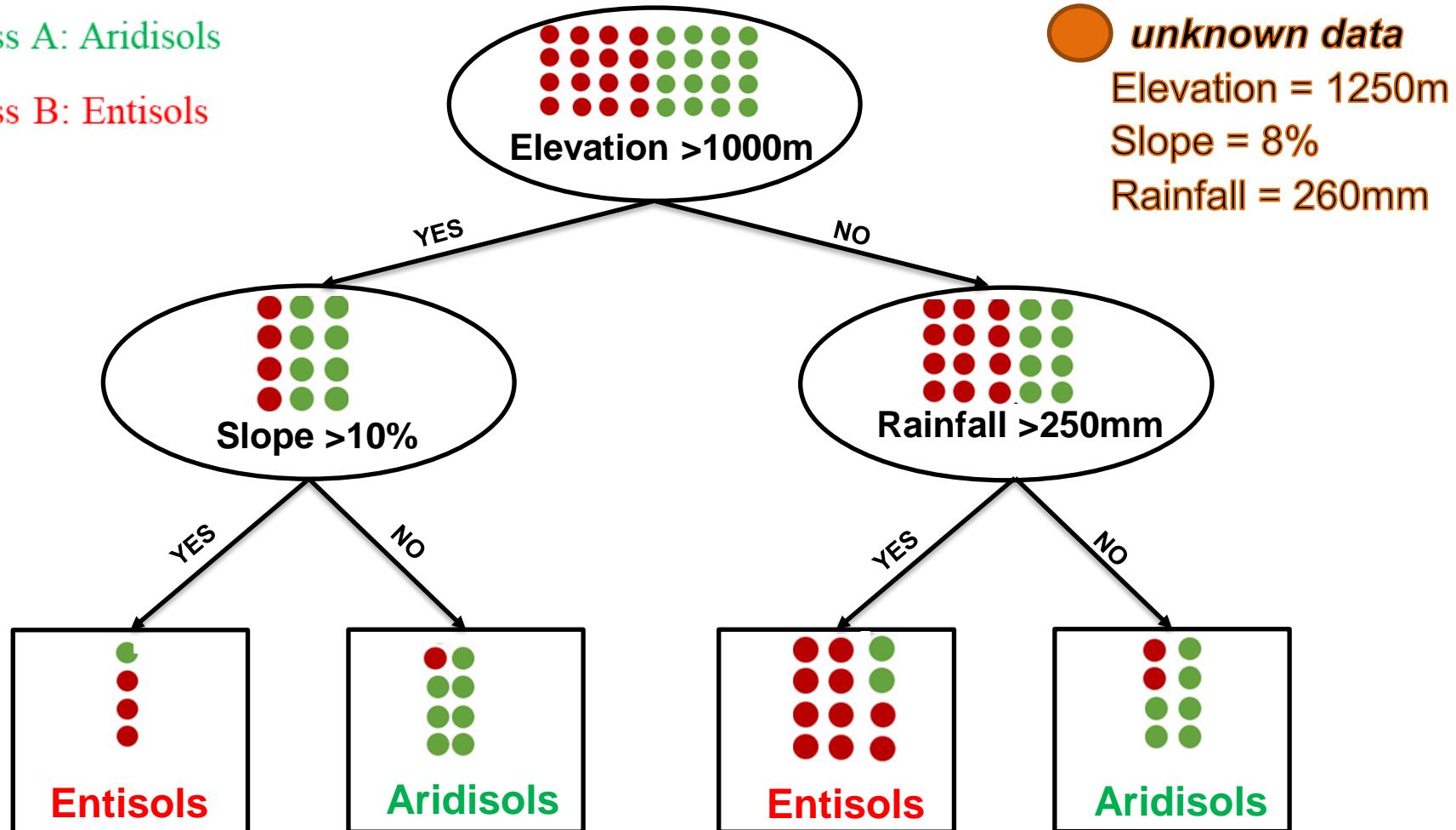
unknown data



A decision tree model to predict Entisols and Aridisols

Decision Tree: example II

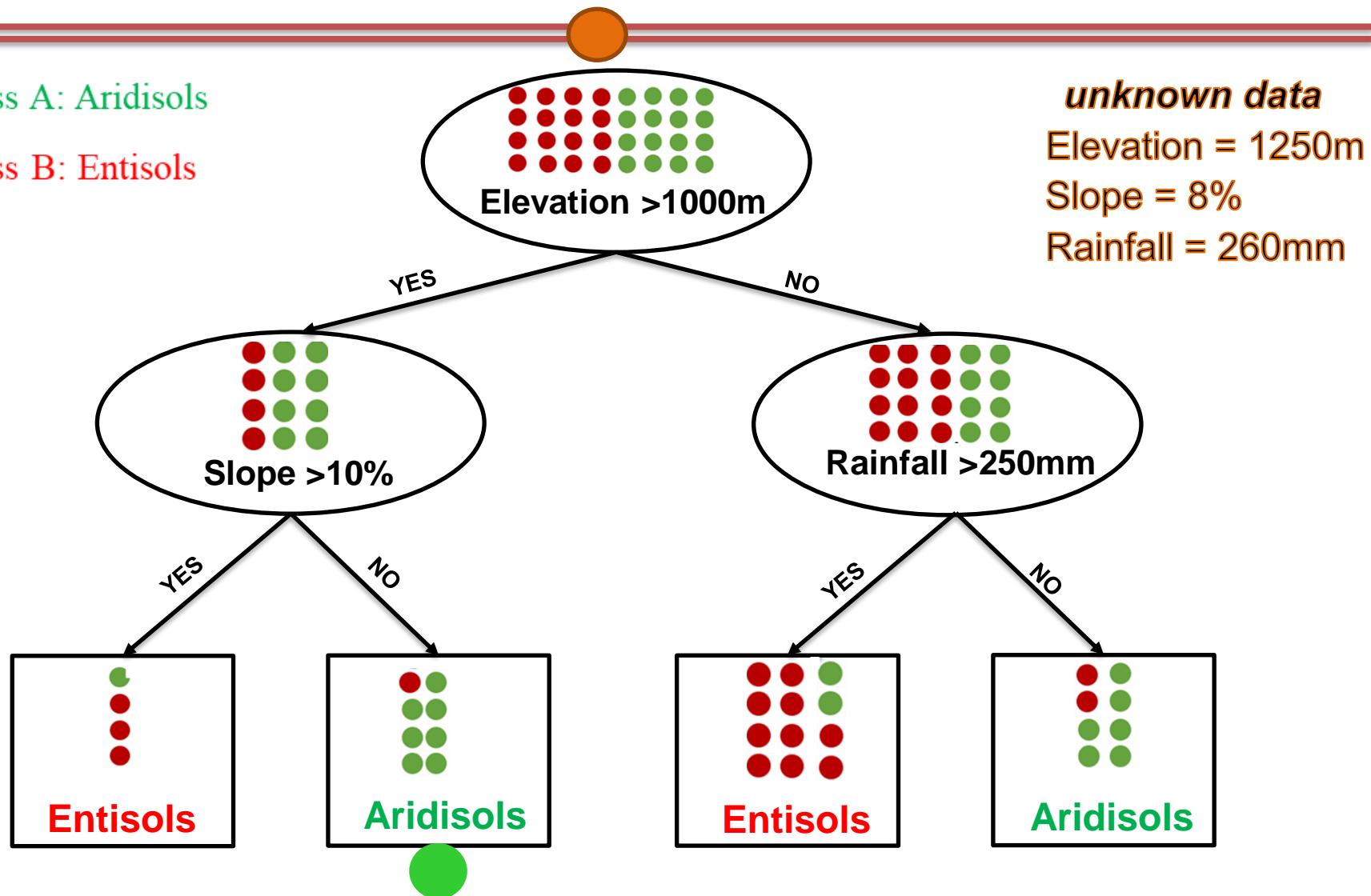
- Class A: Aridisols
- Class B: Entisols



A decision tree model to predict Entisols and Aridisols

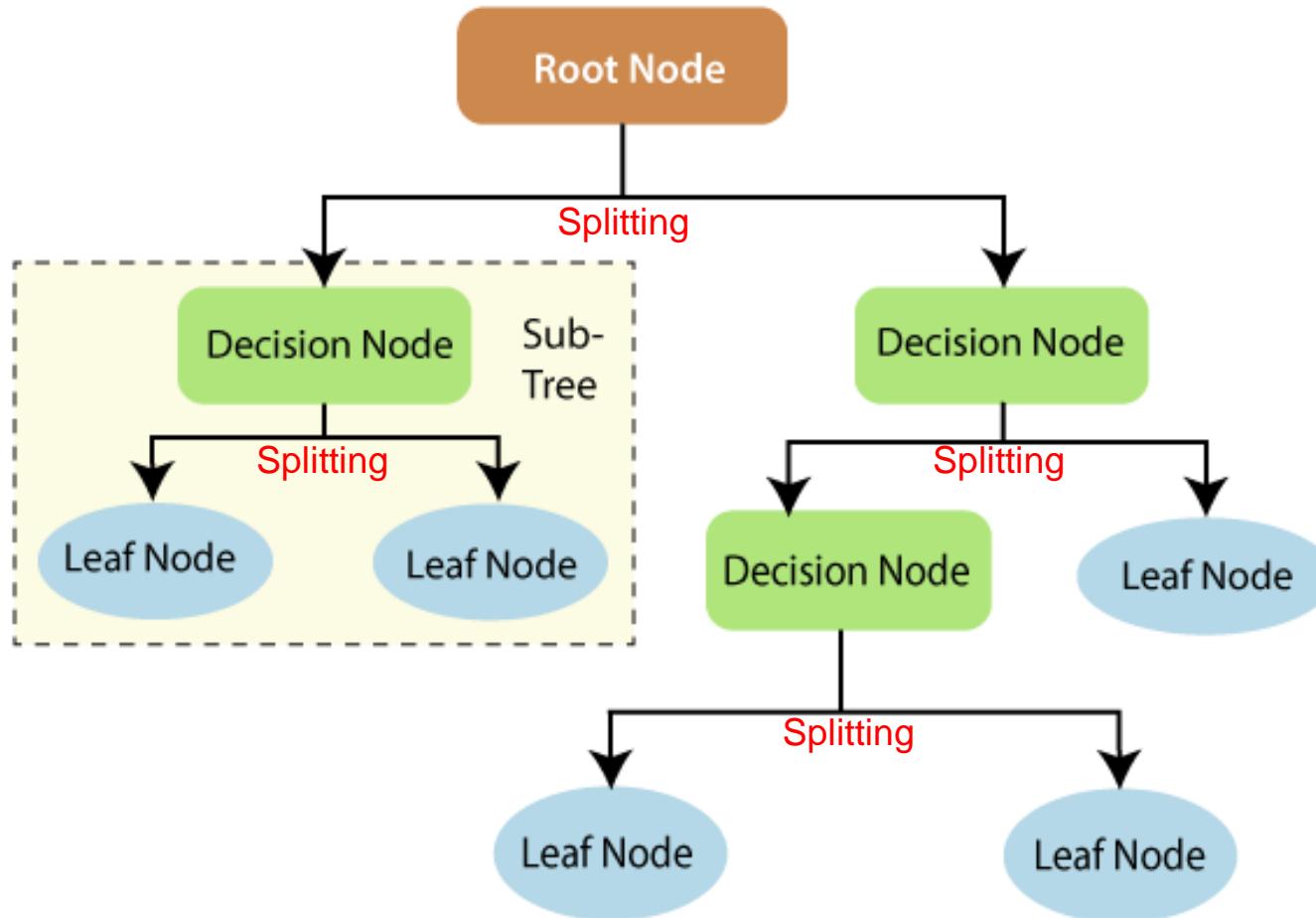
Decision Tree: example II

- Class A: Aridisols
- Class B: Entisols



A decision tree model to predict Entisols and Aridisols

Decision Tree: graphic



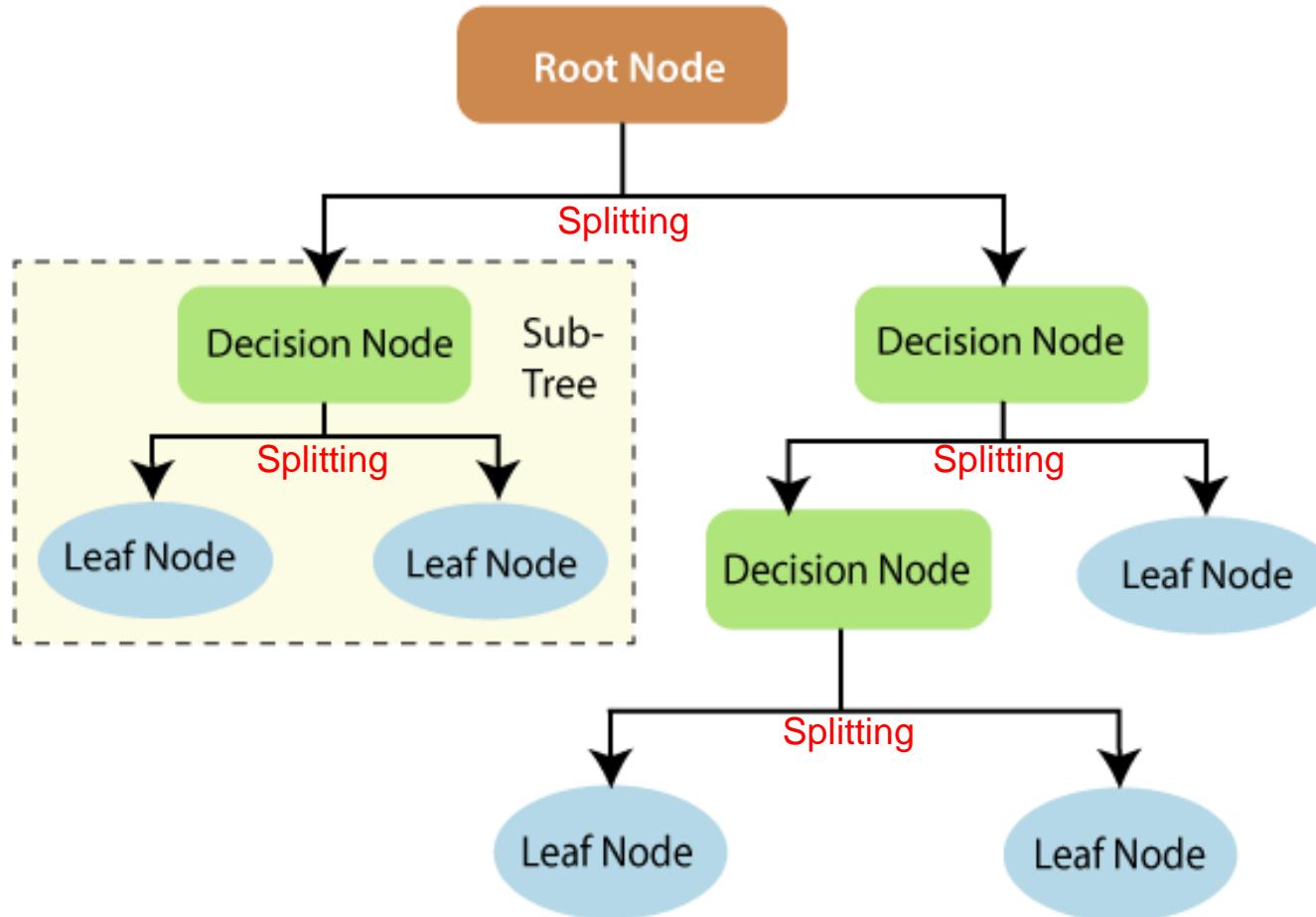
Decision Tree: Terminology

- **Root Node:** It represents the entire population or sample and this further gets divided into two or more homogeneous sets.
- **Decision Node:** When a sub-node splits into further sub-nodes, then it is called the decision node.
- **Leaf/Terminal Node:** Nodes do not split is called Leaf or Terminal node.
- **Parent and Child Node:** A node, which is divided into sub-nodes is called a parent node of sub-nodes whereas sub-nodes are the child of a parent node.
- **Splitting:** It is a process of dividing a node into two or more sub-nodes.
- **Pruning:** When we remove sub-nodes of a decision node, this process is called pruning. You can say the opposite process of splitting.
- **Branch/Sub-Tree:** A subsection of the entire tree is called branch or sub-tree.

Decision Tree: Terminology

- **Root Node:** It represents the entire population or sample and this further gets divided into two or more homogeneous sets.
- **Decision Node:** When a sub-node splits into further sub-nodes, then it is called the decision node.
- **Leaf/Terminal Node:** Nodes do not split is called Leaf or Terminal node.
- **Parent and Child Node:** A node, which is divided into sub-nodes is called a parent node of sub-nodes whereas sub-nodes are the child of a parent node.
- **Splitting:** It is a process of dividing a node into two or more sub-nodes.
- **Pruning:** When we remove sub-nodes of a decision node, this process is called pruning. You can say the opposite process of splitting.
- **Branch/Sub-Tree:** A subsection of the entire tree is called branch or sub-tree.

Splitting Issues

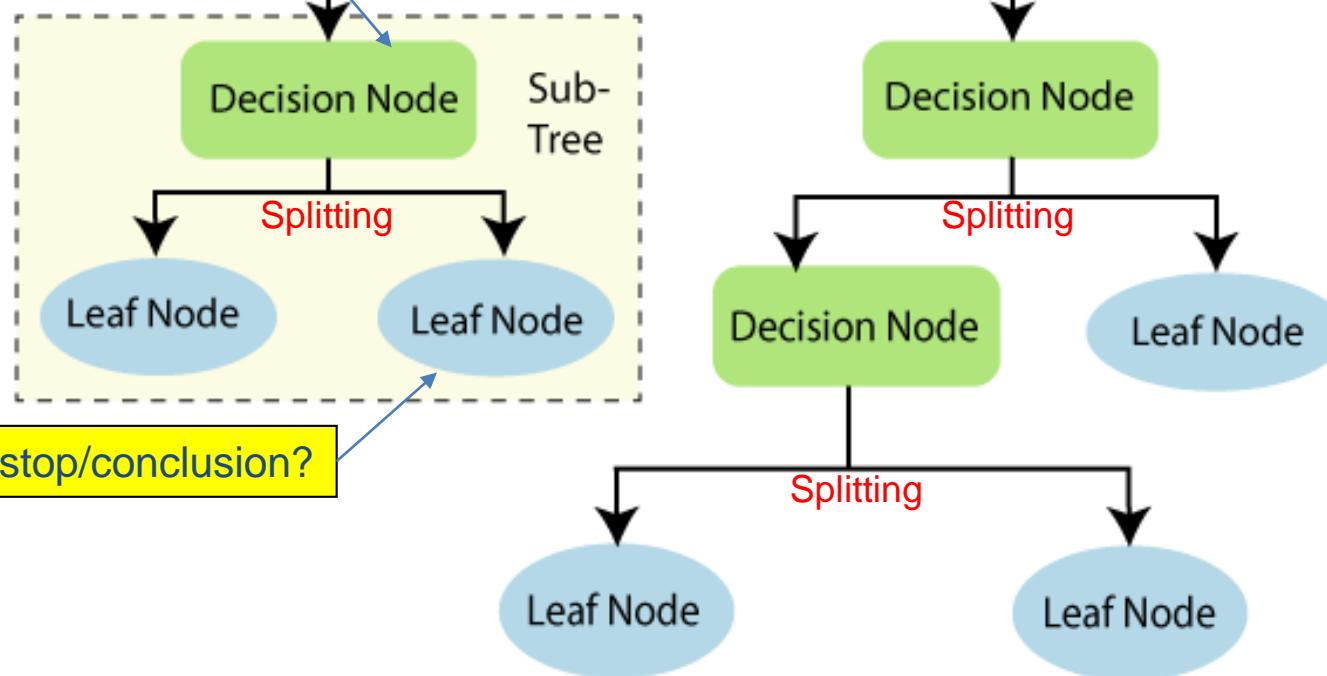


Splitting Issues

(1) Which to start? (root)



(2) Which node to proceed?



(3) When to stop/conclusion?

Splitting Issues

Issues

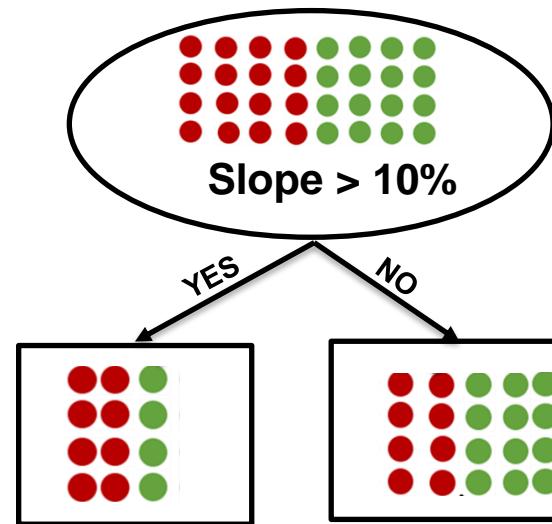
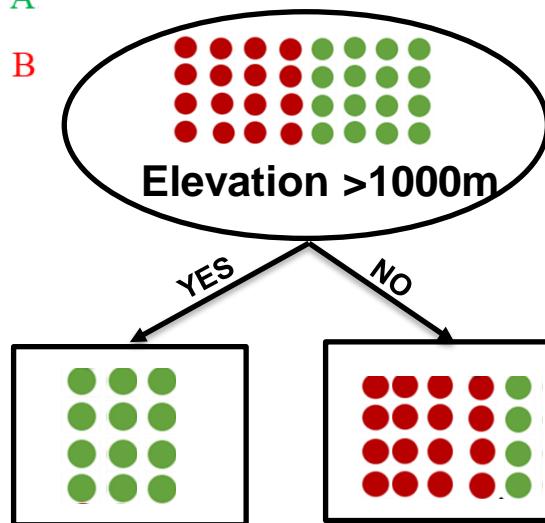
- How to determine the Best Split
- Determine when to stop splitting

How to determine the Best Split

- at each step, we look for **the best split** (Greedy Strategy).



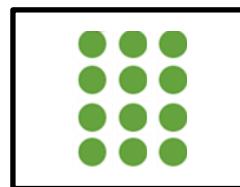
Class A
Class B



Which test split is the best?

How to determine the Best Split

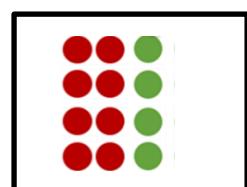
- Nodes with **homogeneous** class distribution are preferred
- Need a measure of node **impurity**:



● Class A
● Class B

12
0

Homogeneous,
Low degree of impurity



● Class A
● Class B

4
8

Non-homogeneous,
High degree of impurity

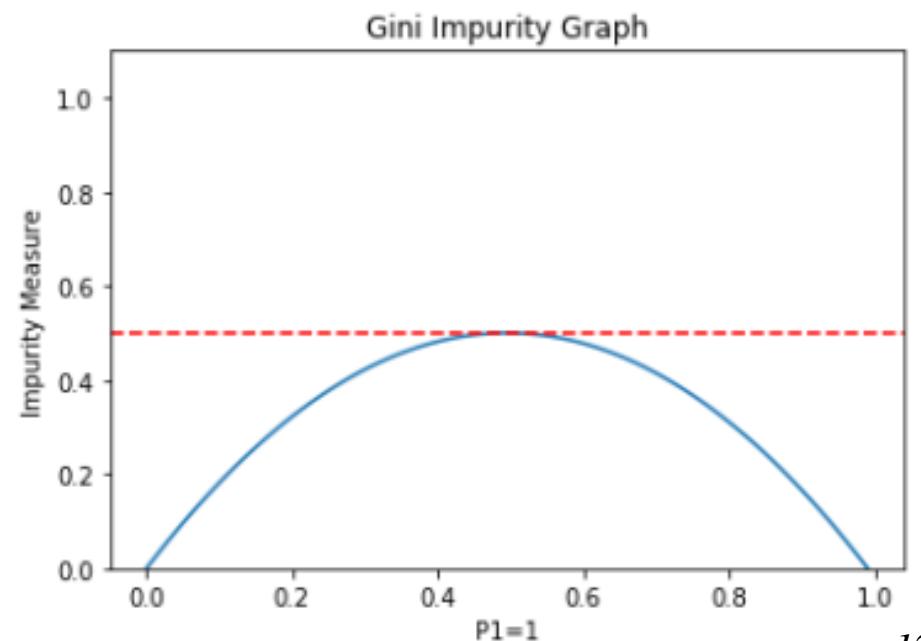
Measures of Node Impurity

- Gini Index or Gini ratio
- This measure provides a ranking to the attributes for splitting the data.
- It stores sum of squared probabilities of each class. We can formulate it as illustrated below.

$$Gini = 1 - \sum (P_i)^2$$

$i = 1$ to number of classes
 P = probabilities of each class

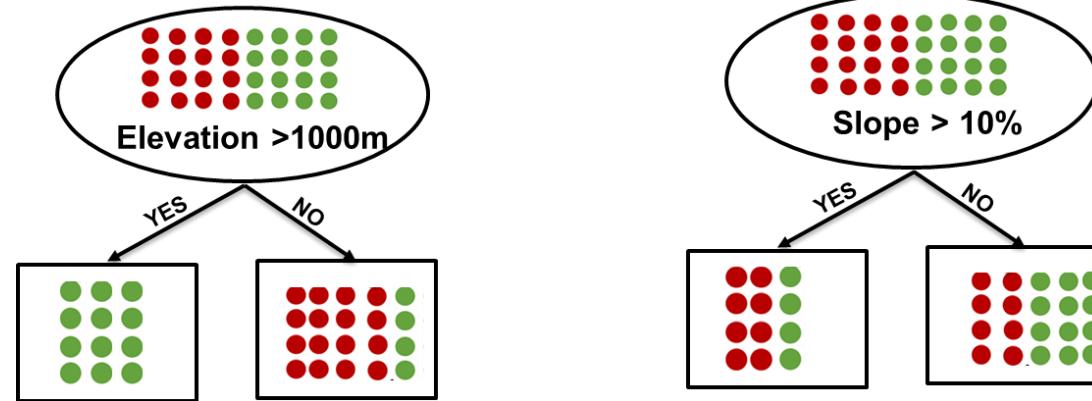
A Gini Impurity of **0** is the lowest.
It can only be achieved when everything is the same class



$$Gini = 1 - \sum(P_i)^2$$

Gini Index

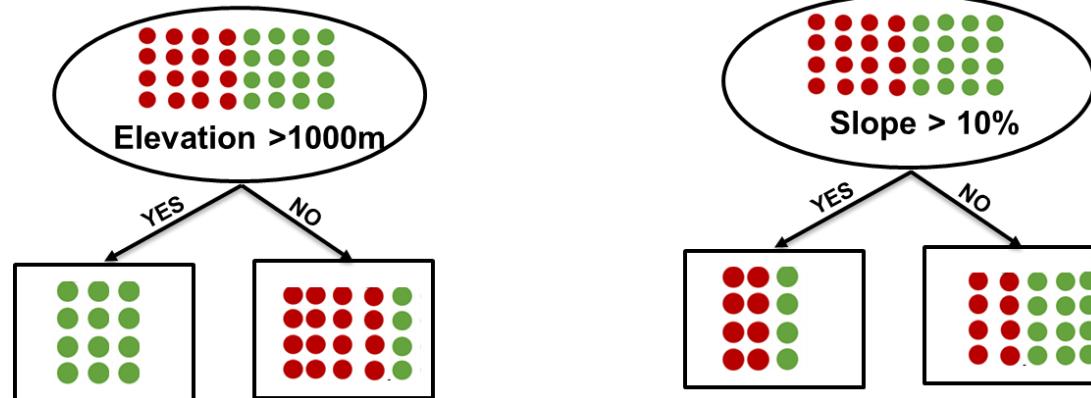
Class A
Class B



$$Gini = 1 - \sum(P_i)^2$$

Gini Index

Class A
Class B

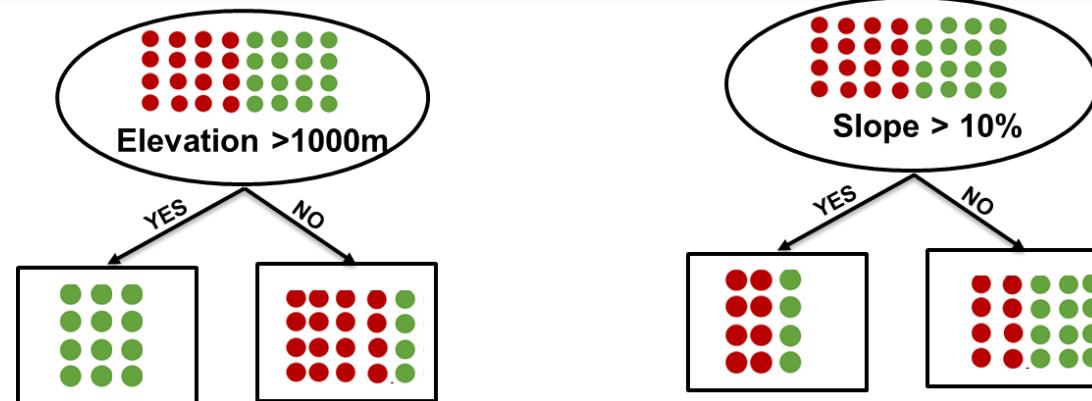


Elevation	A	B	P (A)	P (B)	Gini
>1000 m	12	0	12÷12=1	0÷12=0	1-[(1) ² +(0) ²]=0
<1000 m	4	16	4÷20=0.2	16÷20=0.8	1-[(0.2) ² +(0.8) ²]=0.32

$$Gini = 1 - \sum(P_i)^2$$

Gini Index

Class A
Class B



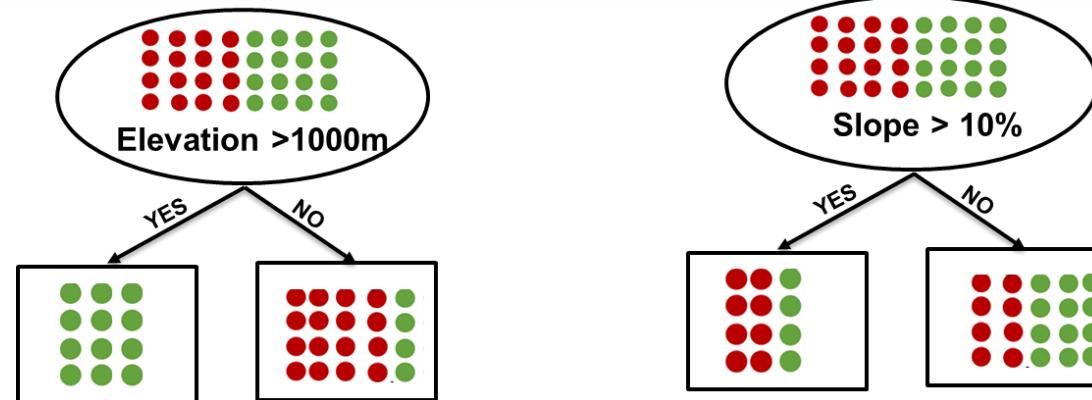
Elevation	A	B	P (A)	P (B)	Gini
>1000 m	12	0	12÷12=1	0÷12=0	1-[(1) ² +(0) ²]=0
<1000 m	4	16	4÷20=0.2	16÷20=0.8	1-[(0.2) ² +(0.8) ²]=0.32

$$Gini(Elevation) = \left(\frac{12}{32}\right) \times 0 + \left(\frac{20}{32}\right) \times 0.32 = 0.20$$

$$Gini = 1 - \sum(P_i)^2$$

Gini Index

Class A
Class B



Elevation	A	B	P (A)	P (B)	Gini
>1000 m	12	0	12÷12=1	0÷12=0	1-[(1) ² +(0) ²]=0
<1000 m	4	16	4÷20=0.2	16÷20=0.8	1-[(0.2) ² +(0.8) ²]=0.32

$$Gini(Elevation) = \left(\frac{12}{32}\right) \times 0 + \left(\frac{20}{32}\right) \times 0.32 = 0.20$$

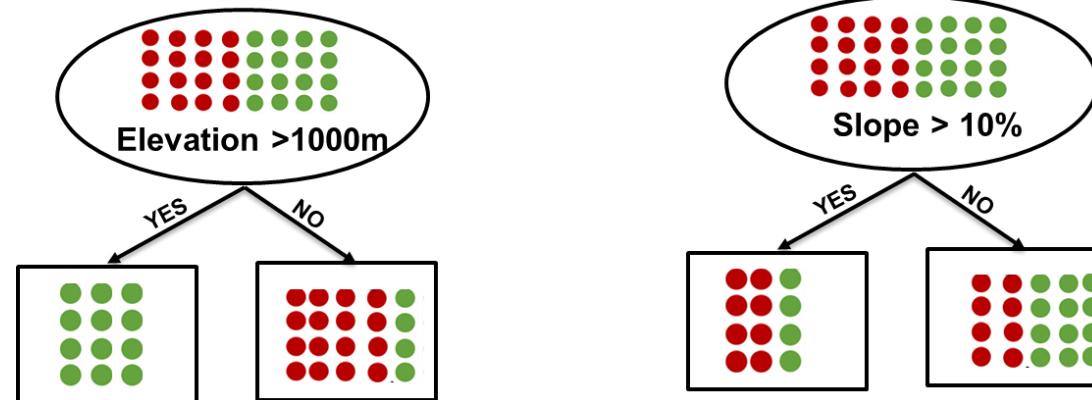
Slope	A	B	P (A)	P (B)	Gini
>10 %	4	8	4÷12=1	8÷12=0	1-[(0.22) ² +(0.66) ²]=0.44
<10 %	12	8	12÷20=0.2	8÷20=0.8	1-[(0.6) ² +(0.4) ²]=0.48

$$Gini(Slope) = \left(\frac{12}{32}\right) \times 0.44 + \left(\frac{20}{32}\right) \times 0.48 = 0.46$$

$$Gini = 1 - \sum(P_i)^2$$

Gini Index

Class A
Class B



Elevation	A	B	P (A)	P (B)	Gini
>1000 m	12	0	12÷12=1	0÷12=0	1-[1^2+0^2]=0
<1000 m	4	16	4÷20=0.2	16÷20=0.8	1-[0.2^2+(0.8)^2]=0.32

$$Gini(Elevation) = \left(\frac{12}{32}\right) \times 0 + \left(\frac{20}{32}\right) \times 0.32 = 0.20$$

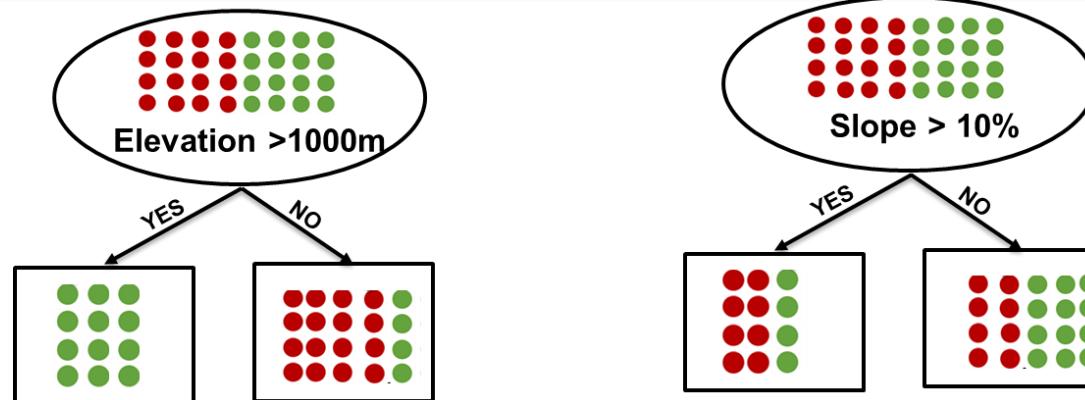
Slope	A	B	P (A)	P (B)	Gini
>10 %	4	8	4÷12=1	8÷12=0	1-[0.22^2+(0.66)^2]=0.44
<10 %	12	8	12÷20=0.2	8÷20=0.8	1-[0.6^2+(0.4)^2]=0.48

$$Gini(Slope) = \left(\frac{12}{32}\right) \times 0.44 + \left(\frac{20}{32}\right) \times 0.48 = 0.46$$

$$Gini = 1 - \sum(P_i)^2$$

Gini Index

Class A
Class B



Elevation	A	B	P (A)	P (B)	Gini
>1000 m	12	0	12÷12=1	0÷12=0	1-[(1) ² +(0) ²]=0
<1000 m	4	16	4÷20=0.2	16÷20=0.8	1-[(0.2) ² +(0.8) ²]=0.32



$$Gini(Elevation) = \left(\frac{12}{32}\right) \times 0 + \left(\frac{20}{32}\right) \times 0.32 = 0.20$$

Slope	A	B	P (A)	P (B)	Gini
>10 %	4	8	4÷12=1	8÷12=0	1-[(0.22) ² +(0.66) ²]=0.44
<10 %	12	8	12÷20=0.2	8÷20=0.8	1-[(0.6) ² +(0.4) ²]=0.48

$$Gini(Slope) = \left(\frac{12}{32}\right) \times 0.44 + \left(\frac{20}{32}\right) \times 0.48 = 0.46$$

Splitting Issues

Issues

- How to determine the Best Split
- Determine when to stop splitting

Determine when to stop splitting

- **Stopping Rule**
 - Typical stopping conditions for a node:
 - Stop if all instances belong to the same class
 - Stop if all the attribute values are the same
 - More restrictive conditions:
 - Stop if number of instances is less than some user-specified threshold
 - Stop if class distribution of instances are independent of the available features
 - Stop if expanding the current node does not improve impurity measures (e.g., Gini index).

Stop the algorithm before it becomes a fully-grown tree (**Pre-Pruning**)

Overfitting In Decision Trees

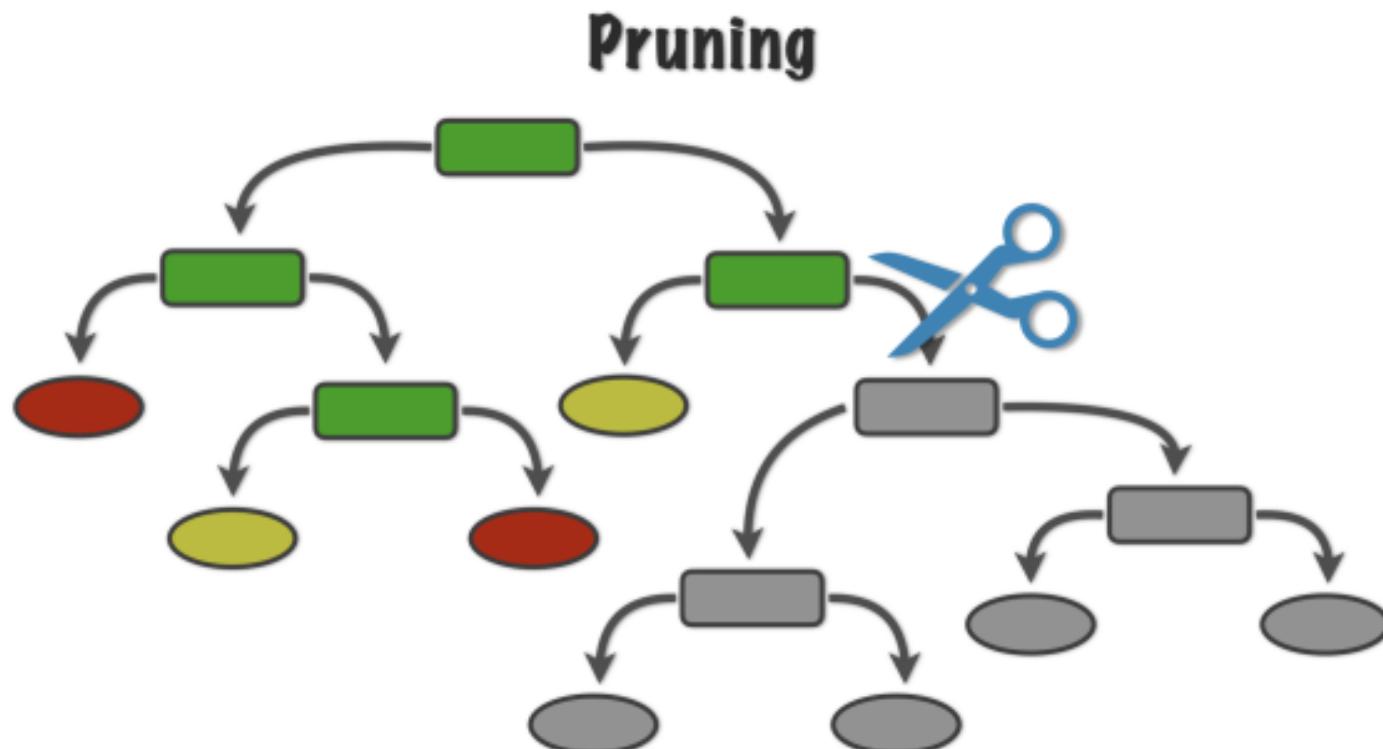
Overfitting happens when a decision tree tries to be as perfect as possible by increasing the depth of tests and thereby reduces the error.

- Post-pruning
 - Grow decision tree to its entirety
 - Trim the nodes of the decision tree in a bottom-up fashion
 - If generalization error improves after trimming, replace sub-tree by a leaf node.
 - Class label of leaf node is determined from majority class of instances in the sub-tree

Pruning

Comparing pre-pruning and post-pruning,

- Pre-pruning is faster but
- Post-pruning generally leads to more accurate trees.



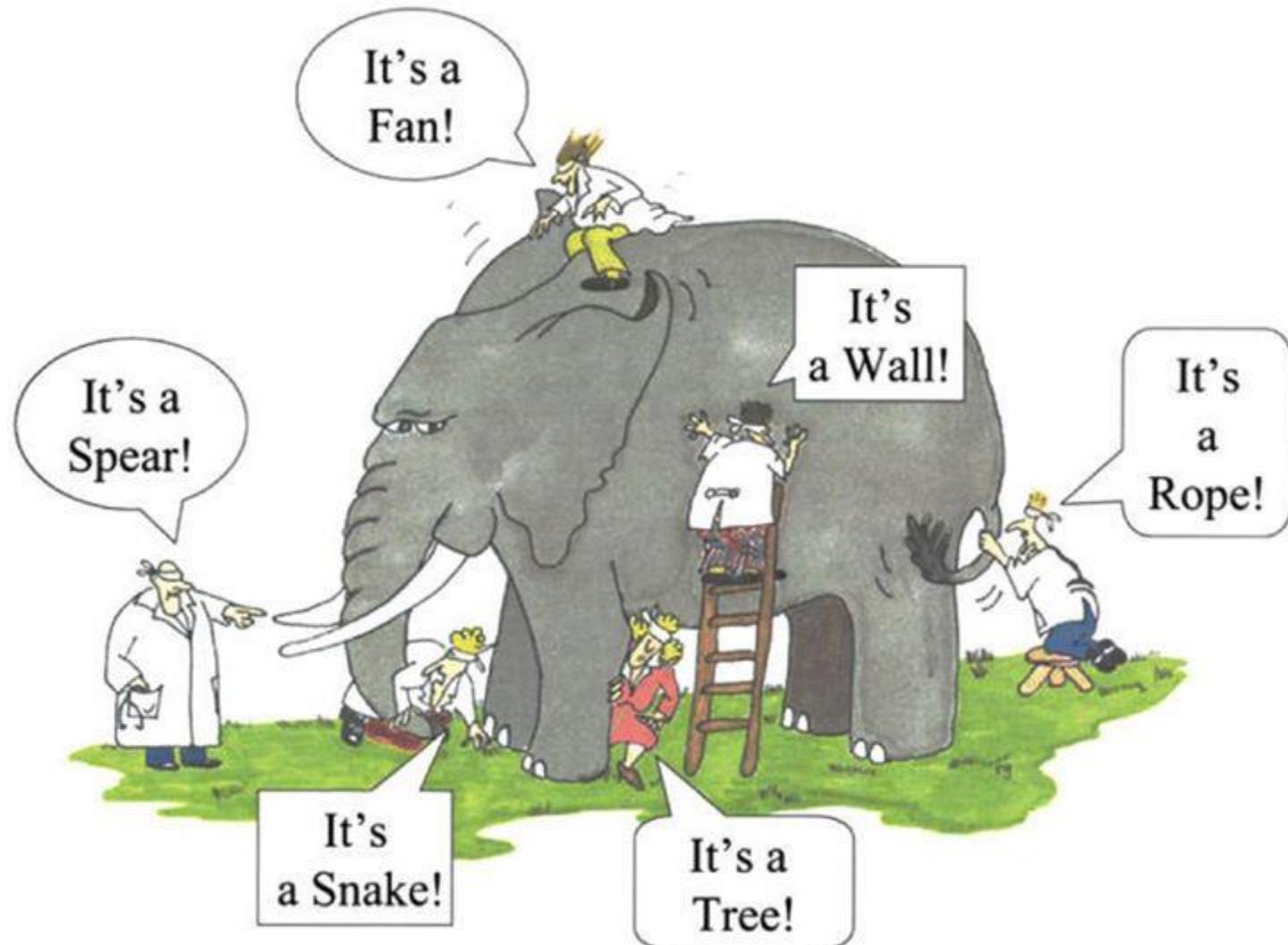
Advantages and Disadvantages

- **Advantages:**
 - Extremely fast at classifying unknown records.
 - Simple to understand and to interpret for small-sized trees
 - Trees can be visualized.
 - Accuracy comparable to other classification techniques for many simple data sets.
 - Excludes unimportant features.
 - Requires little data preparation. Other techniques often require data normalisation, dummy variables need to be created and blank values to be removed.
 - Able to handle both numerical and categorical data. Other techniques are usually specialised in analysing datasets that have only one type of variable.

Advantages and Disadvantages

- **Disadvantages:**
 - Easy to overfit.
 - Large trees can be difficult to interpret
 - Decision trees can be unstable because small variations in the data might result in a completely different tree being generated. This problem is mitigated by using decision trees within an **ensemble**.

Ensemble Methods



What Are Ensemble Methods?

- Ensemble learning is a machine learning paradigm where multiple models (e.g., multiple decision trees) (often called “weak learners”) are trained to solve the same problem (e.g., prediction of soil types) and combined to get better results.
- The main hypothesis is that when weak models are correctly combined we can obtain more accurate models.

Question

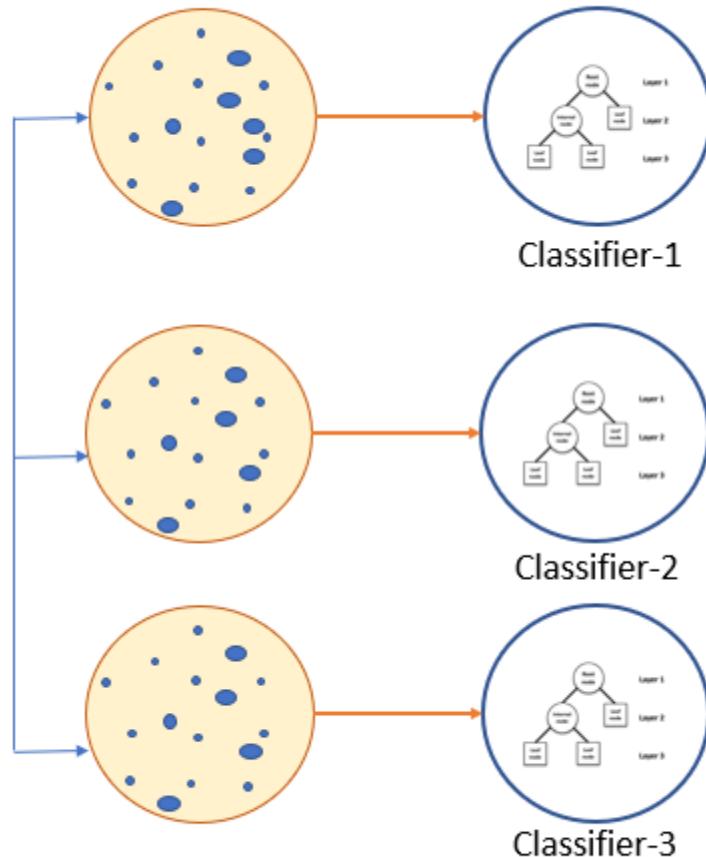
How to combine these models
(models do not all just learn the same)

Combining Weak Learners

- Bagging, Boosting, Stacking Models
 - bagging, that often considers **homogeneous** weak learners, learns them independently from each other in parallel and combines them following some kind of deterministic averaging process
 - boosting, that often considers **homogeneous** weak learners, learns them sequentially in a very adaptative way (a base model depends on the previous ones) and combines them following a deterministic strategy
 - stacking, that often considers **heterogeneous** weak learners, learns them in parallel and combines them by training a meta-model to output a prediction based on the different weak models predictions

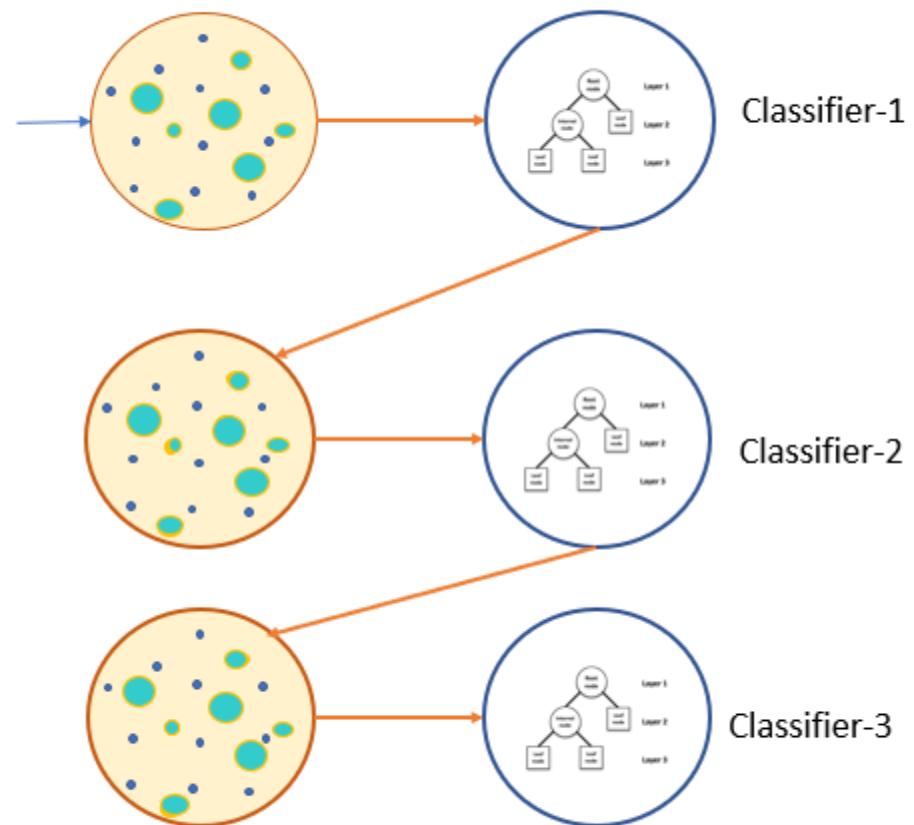
Bagging, Boosting, Stacking Models

Bagging



Parallel

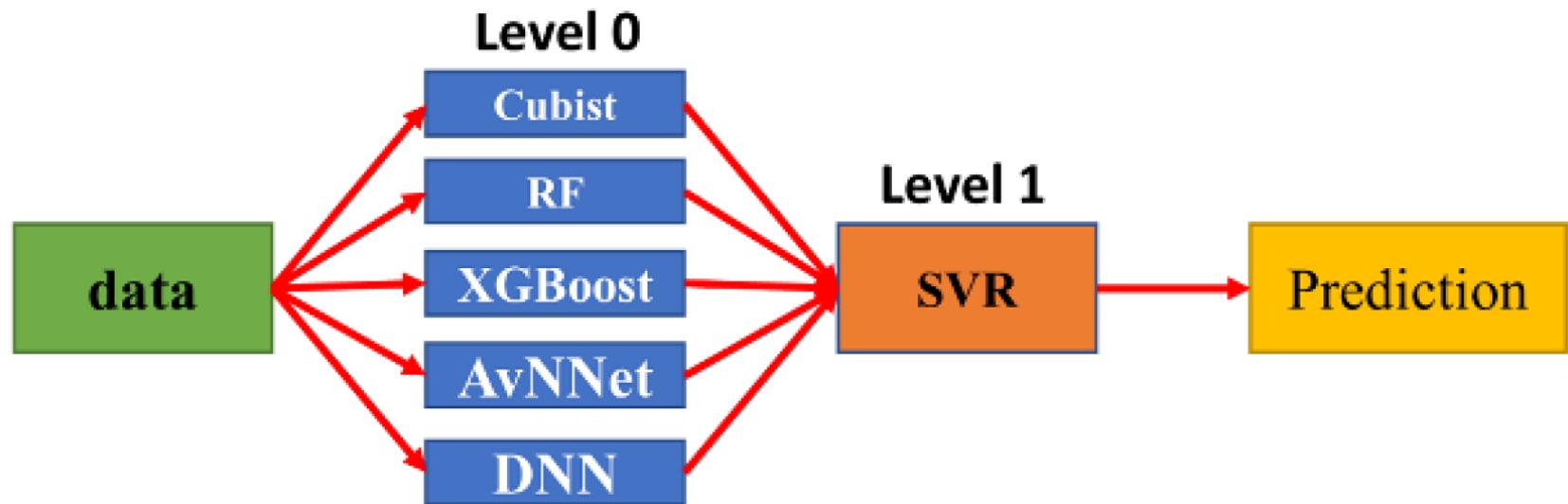
Boosting



Sequential

Bagging, Boosting, Stacking Models

- Stacking

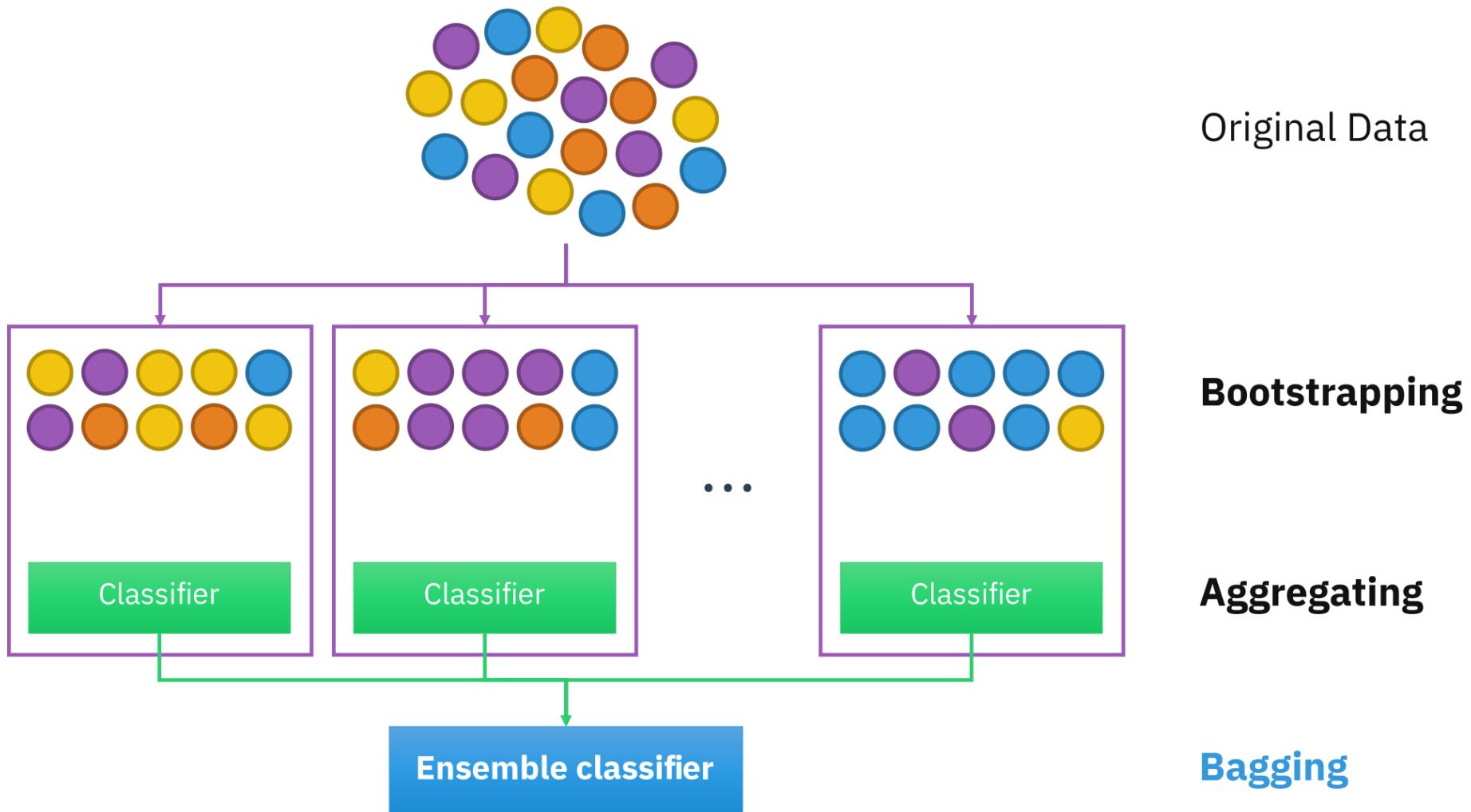


Bagging

- Bagging or Bootstrap aggregating (Breiman 1994):
- First, we **create multiple bootstrap samples** so that each new bootstrap sample will act as another (almost) independent dataset drawn from true distribution.
- Then, we can fit a weak learner (**decision tree**) for each of these samples
- Finally **aggregate them** (simple average for regression or simple majority vote for classification)

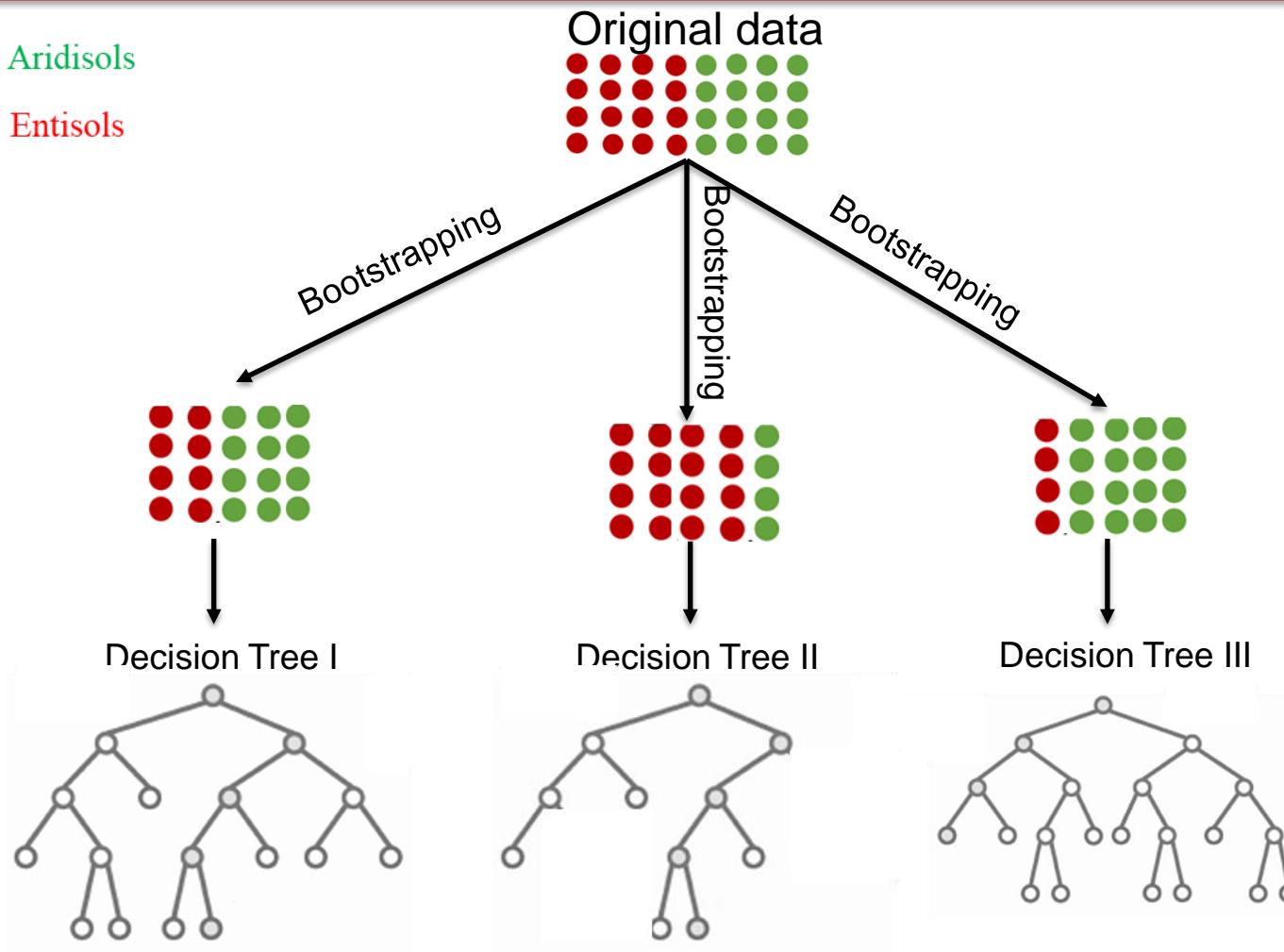
Bootstrap: a statistical technique, for generating samples of **size B** (called bootstrap samples) from an initial dataset of **size N** by randomly drawing with replacement B observations.

Bagging



Bagging: example I

- Class A: Aridisols
- Class B: Entisols



Bagging: example I

- Class A: Aridisols
- Class B: Entisols

Original data



unknown data
Elevation = 1250m
Slope = 8%
Rainfall = 260mm

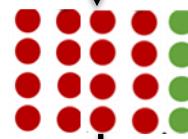
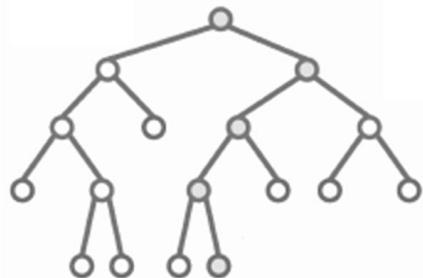
Bootstrapping

Bootstrapping

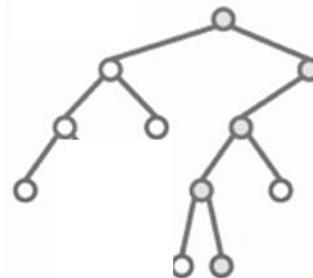
Bootstrapping



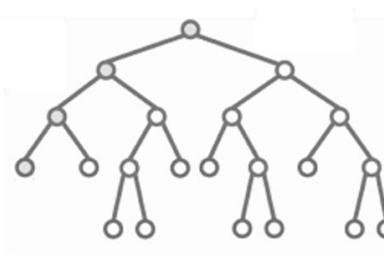
Decision Tree I



Decision Tree II

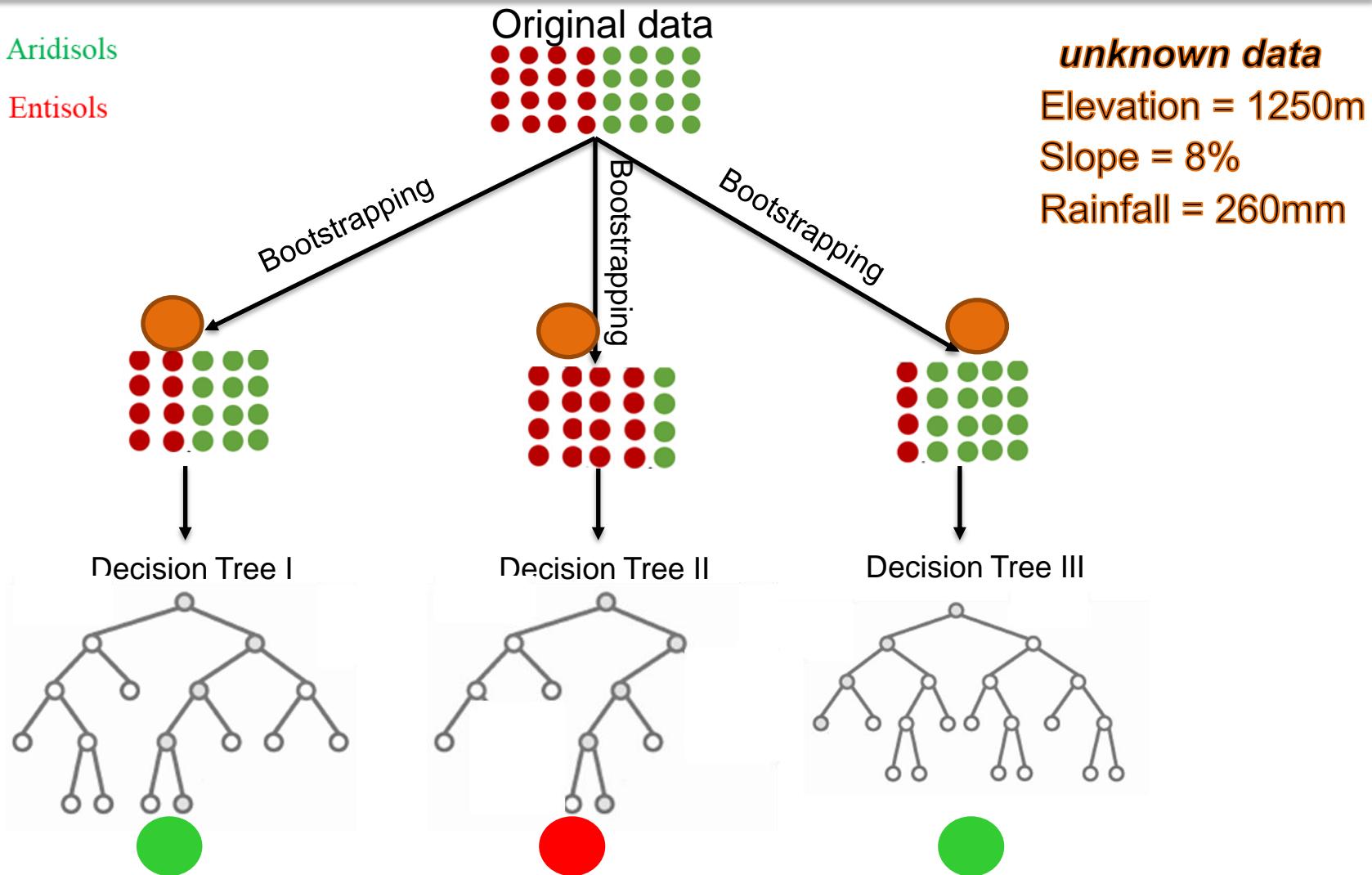


Decision Tree III



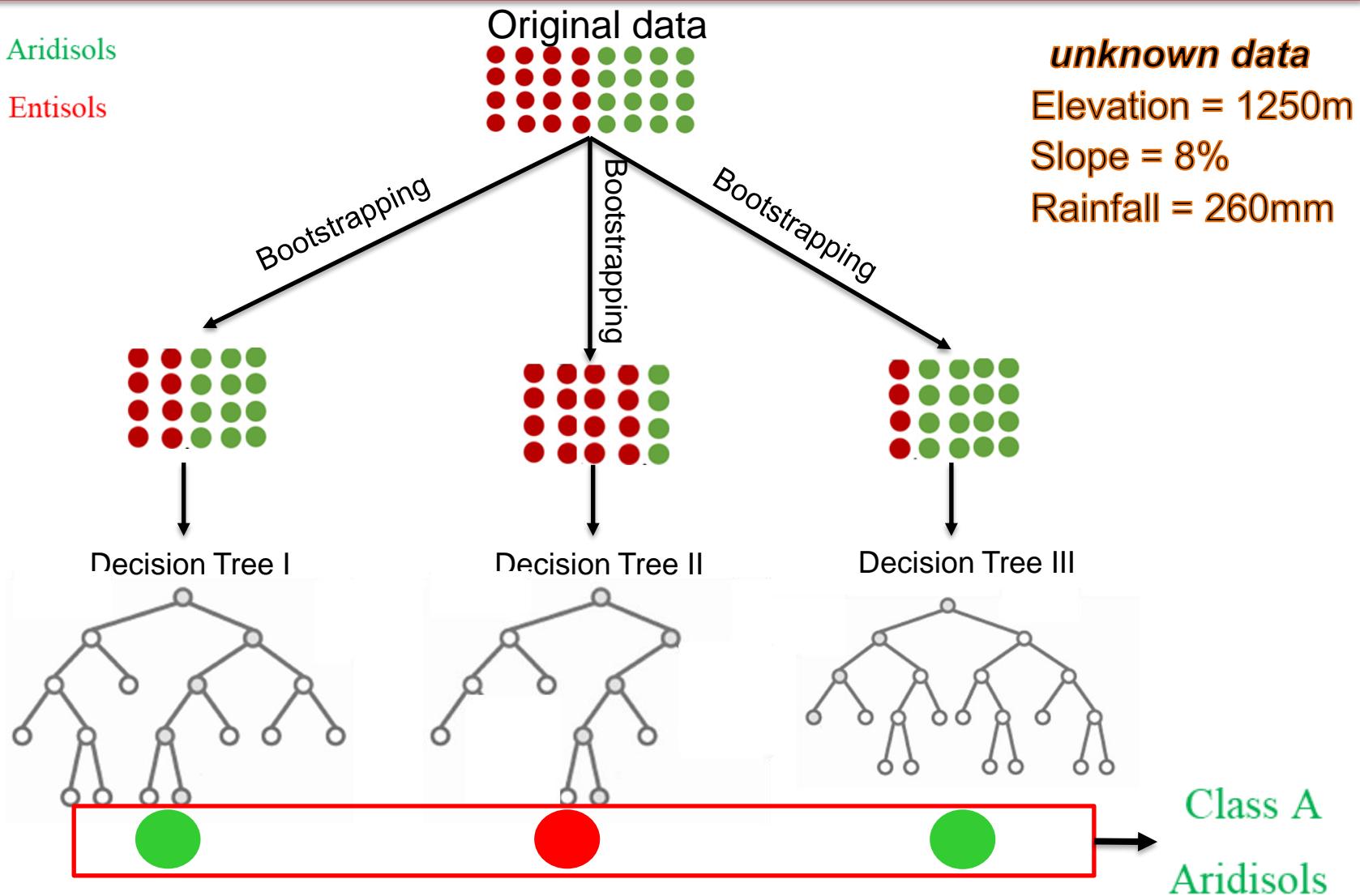
Bagging: example I

- Class A: Aridisols
- Class B: Entisols

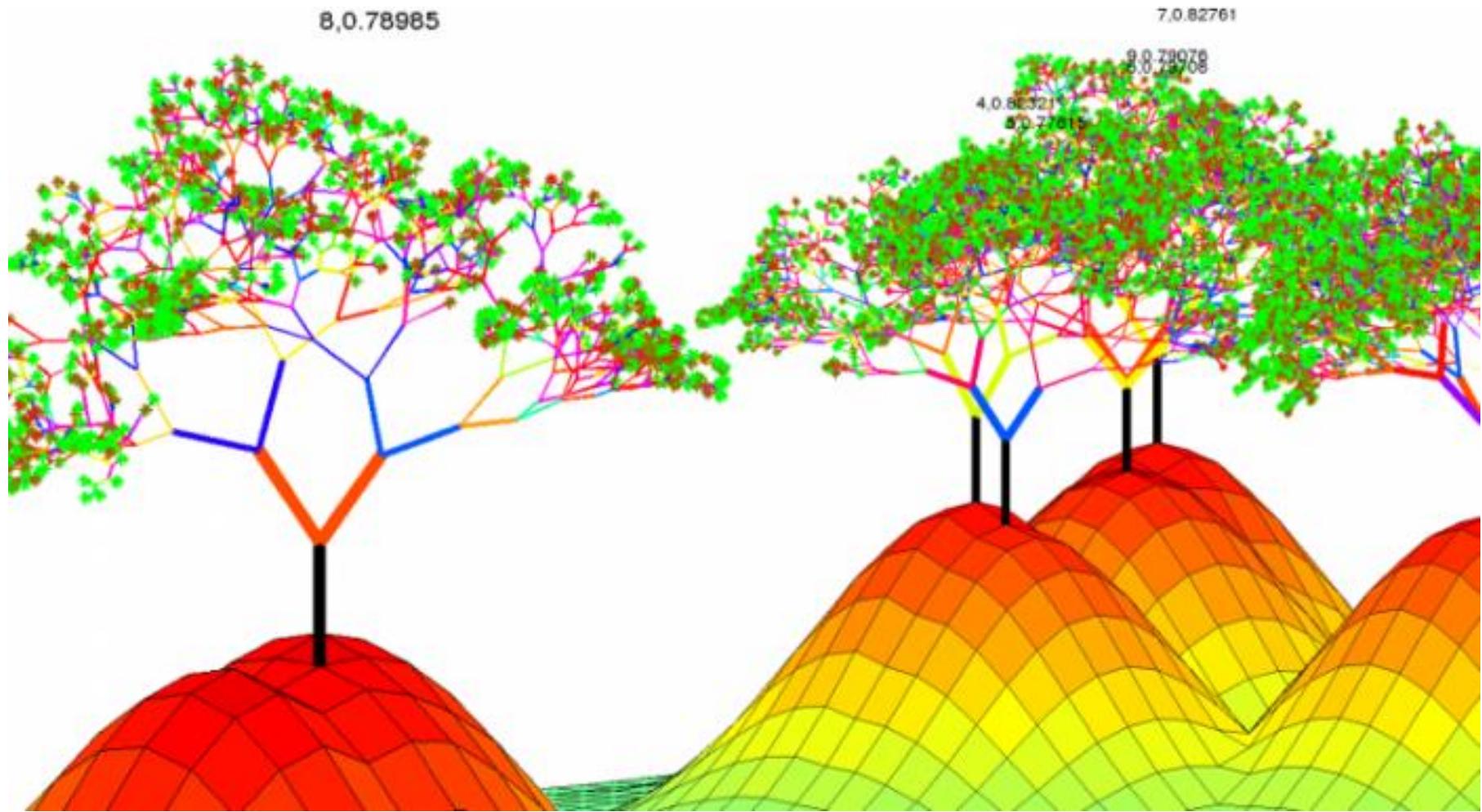


Bagging: example I

- Class A: Aridisols
- Class B: Entisols



Random Forest



Random Forest

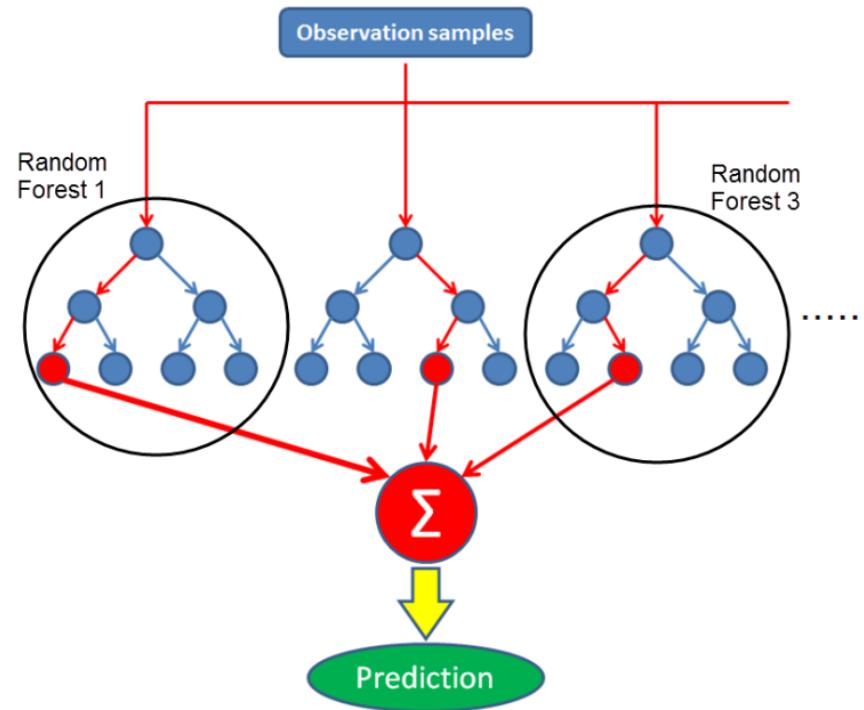
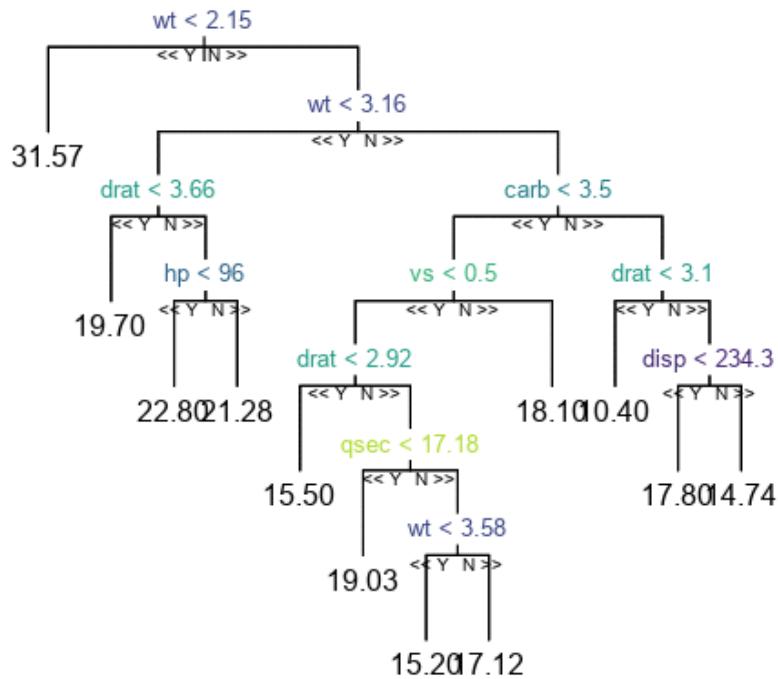
- **Random Forest:**
 - The random forest approach is a **bagging method** for classification and regression that operate by constructing multiple decision trees at training phase and outputting the class that is the mode of the classes or mean/average prediction of the individual trees.
 - Random forests also use another trick to make the multiple decision trees a bit less correlated with each others:
 - when growing each tree, instead of **only** sampling over the observations in the dataset to generate a bootstrap sample, we also sample over **covariates** and keep only a random subset of them to build the tree.

How Does It Work?

- Random Forest:
 - 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

Random Forest

Tree 1



Out Of Bag Accuracy Assessment

- **Random Forest:**
 - Random Forest comes with an internal accuracy assessment (based on cross validation).
 - Bootstrap sampling: the algorithm sets aside ~36% of the training data for each tree grown: out of bag data (OOB).
 - OOB data can be used to asses prediction accuracy:
 - Predict the data not in the bootstrap sample for each tree
 - Aggregate the OOB predictions: mean (continuous data), majority vote (categorical data).

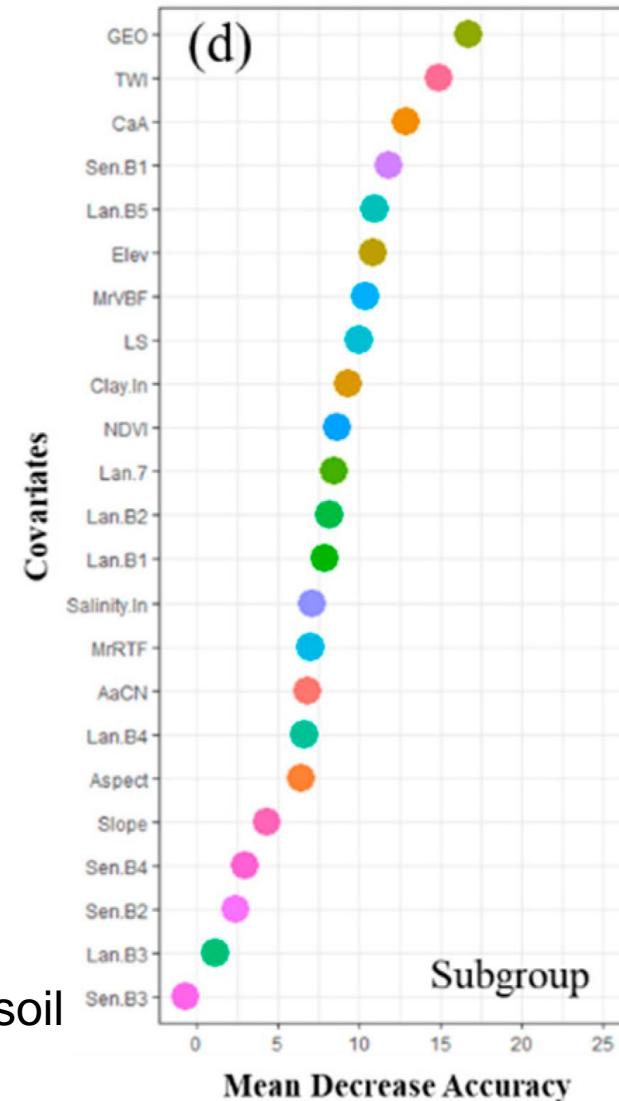
Variable Importance

- Random Forest:
 - Ensembles of trees are not easy to interpret.
 - Ensemble can reflect the potentially (complex) effect of a variable on the response
 - Variable importance plot: shows how much prediction error increases when the values of one predictor are permuted (break association with response) while all others are left unchanged.
 - Permuted variable is used together with other variables to predict the response, **prediction accuracy will decrease**.

Variable Importance Plot

Table 2. Environmental variables used as soil formation factors in the study area.

Definition	Code
Topographic attributes	
Slope	SLOPE
Aspect	ASPECT
LS factor	LS
Elevation	Elev
Catchments area	CaA
Catchment network base level	CNBL
Topographic Wetness Index	TWI
Altitude above channel network	AaCN
Multi-resolution Valley Bottom Flatness Index	MrVBF
Multi-resolution of ridge top flatness index	MrRTF
Remote sensing attributes	
Clay Index: (shortwave IR-1/shortwave IR-2)	Clay.In
Blue, green, red, near infrared, shortwave IR-1, shortwave IR-2,	Lan.B1-B7
Normalized Difference Vegetation Index: (Shortwave IR-1—Near infrared)/(Shortwave IR-1+ Near infrared)	NDVI
Salinity Index: (Red—Near infrared)/(Red + Near infrared)	Salinity.In
Sentinel Bands 1, 2, 3, and 4	Sen.B1-B4
Geomorphology map	
Hierarchical four level classification (geomorphic surfaces)	GEO

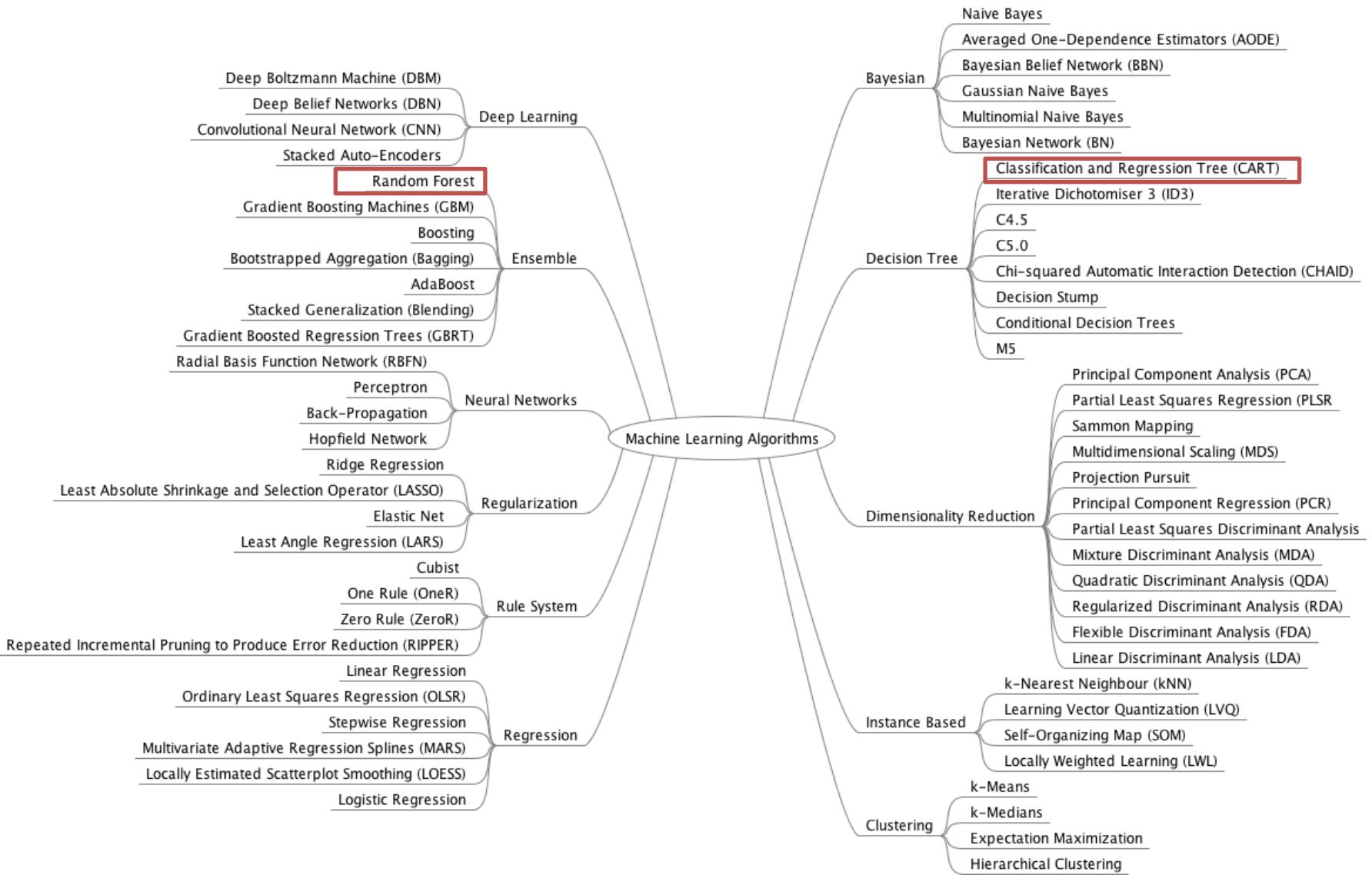


The importance of variables used for predicting soil classes in Subgroup (d).

Advantages and Disadvantages

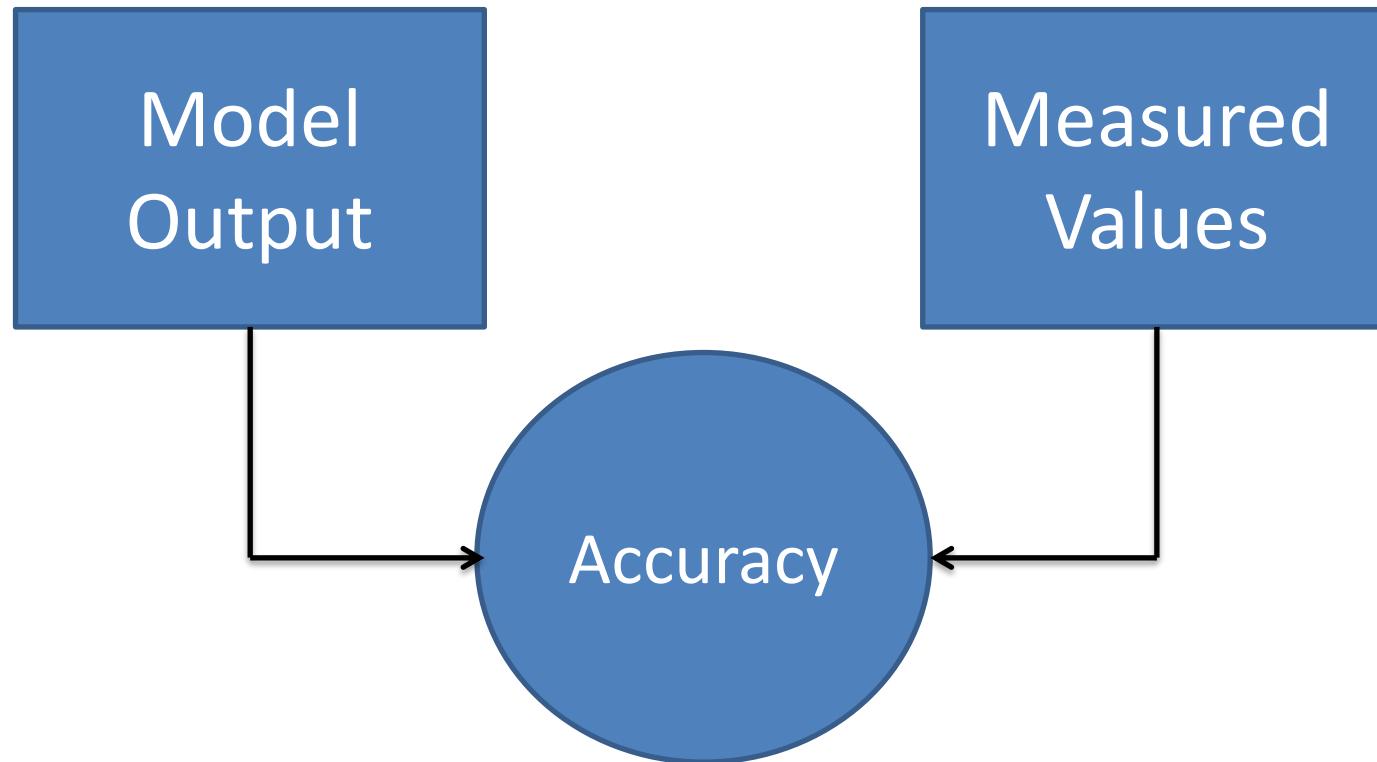
- **Advantages:**
 - Quite fast
 - Models interactions in the data / non linear relationships
 - Very good predictive power expected
 - Yields covariate importance, makes covariate selection possible
- **Disadvantages:**
 - Difficult to interpret

Different Models

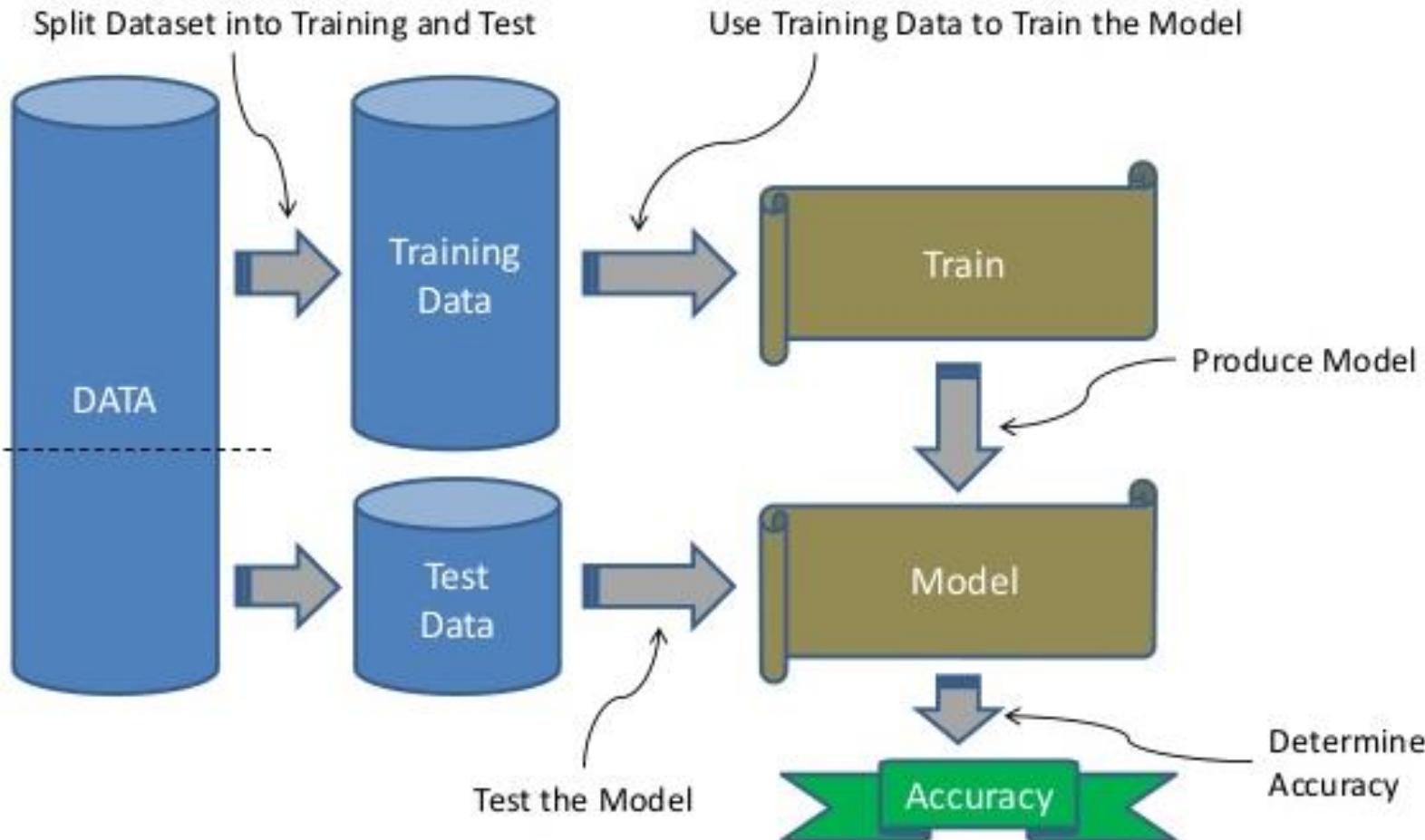


Accuracy Assessment of Models

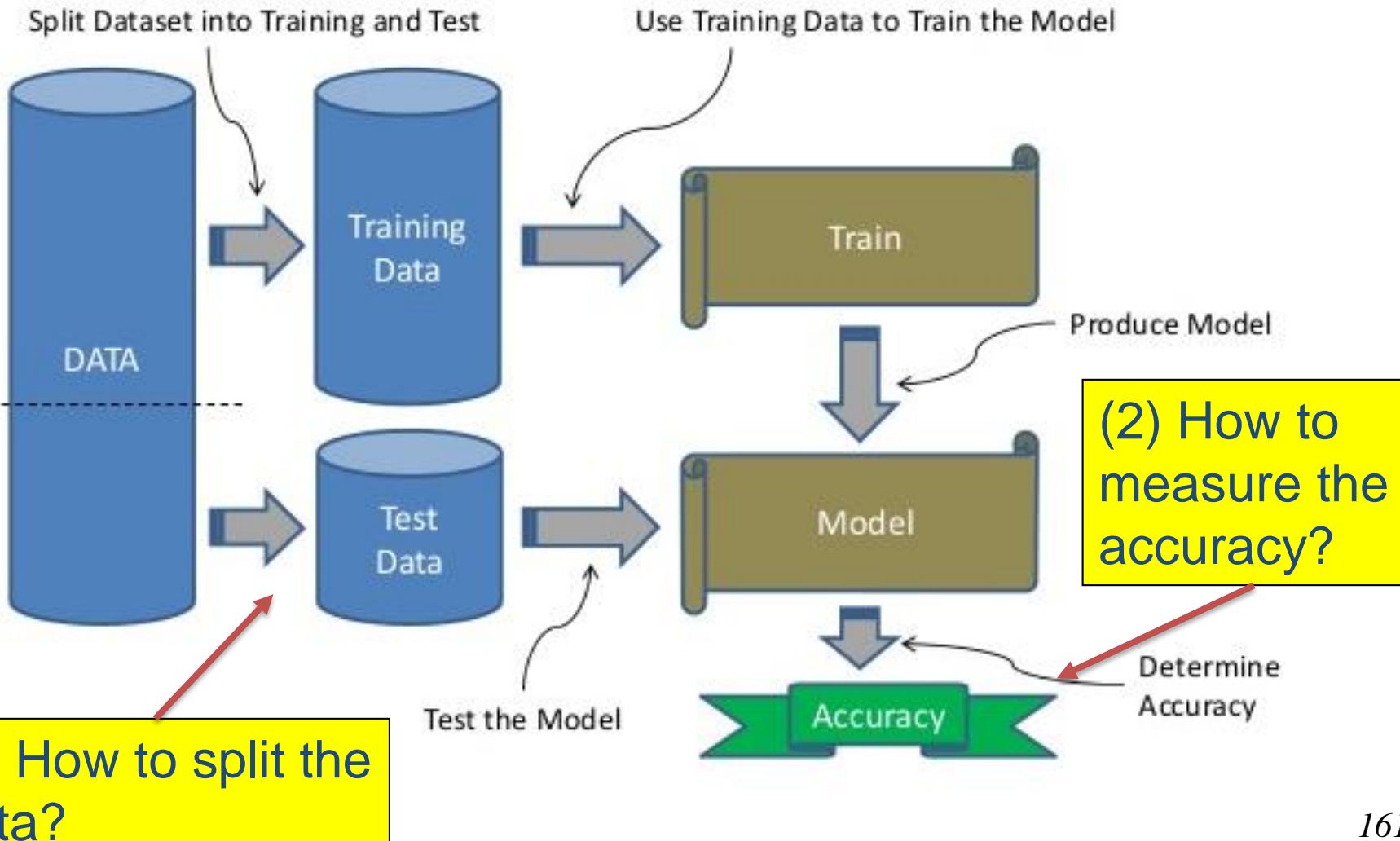
- Assess accuracy of a model output is one of the most important steps in digital soil mapping



Training and Testing the Models



Accuracy Assessment of Models



(1) How to split the data?

(2) How to measure the accuracy?

How to split the data?

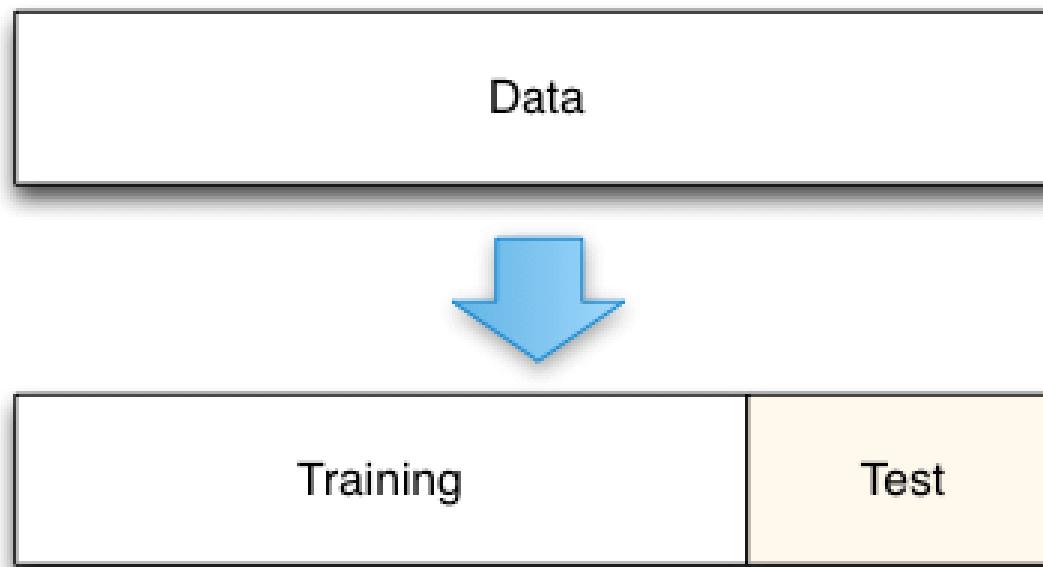
- Holdout method
- K-Folds Cross Validation
- Leave-one-out cross-validation
- Spatial K-Folds Cross Validation

Holdout method

- Holdout method:
 - In this approach we randomly split the complete data into training and test sets.
 - Then perform the model training on the training set and use the test set for validation purpose.
 - Ideally split the data into 70:30 or 80:20.
 - With this approach there is a possibility of high bias if we have limited data, because we would miss some information about the data which we have not used for training.
 - If our data is huge and our test sample and train sample has the same distribution then this approach is acceptable.

Holdout method

- Holdout method:



K-Folds Cross Validation

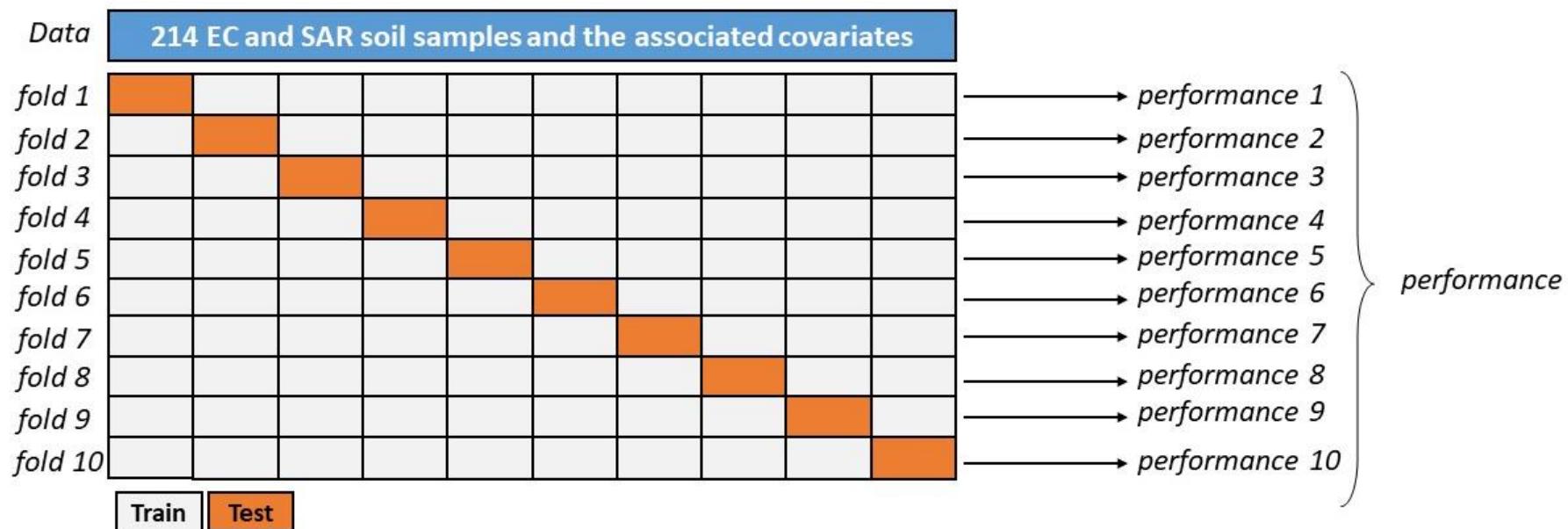
- K-Folds Cross Validation:
 - K-Folds technique is a popular method.
 - It generally results in a less biased model compare to other methods. Because it ensures that every observation from the original dataset has the chance of appearing in training and test set.
 - This is one among the best approach if we have a limited input data.

K-Folds Cross Validation

- K-Folds Cross Validation:
 1. Split the entire data randomly into K folds (value of K shouldn't be too small or too high, ideally we choose 5 to 10 depending on the data size).
 2. Then fit the model using the K-1 (K minus 1) folds and validate the model using the remaining Kth fold. Note down the scores/errors.
 3. Repeat this process until every K-fold serve as the test set. Then take the average of your recorded scores. That will be the performance metric for the model.

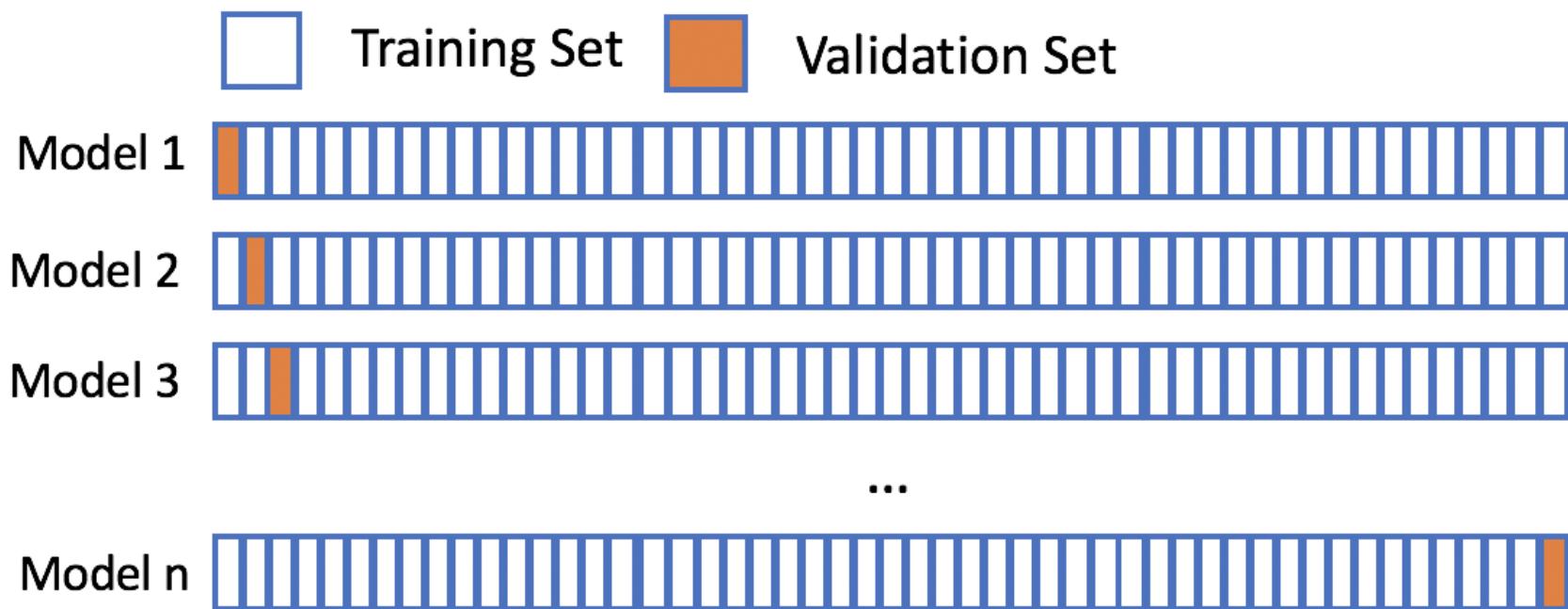
K-Folds Cross Validation

- K-Folds Cross Validation: $k = 10$



Leave-one-out cross-validation

- Leave-one-out cross-validation: $k=n$



Spatial K-Folds Cross Validation

- Spatial K-Folds Cross Validation:
- Spatial data exhibit a few properties that make it difficult to apply standard statistical methods to them.
- Spatial data exhibit spatial autocorrelation, where observations close to each other in space have related values
- K-fold CV, however, does not resample spatial groups of data. Instead, it uses random resampling, which produces training and validation sets whose points are distinct but come from overlapping spatial regions.
- Spatial cross-validation involves modifications of CV for spatial data. Instead of defining folds of points at random, folds are defined by spatial boundaries.

Spatial K-Folds Cross Validation

- Spatial K-Folds Cross Validation:

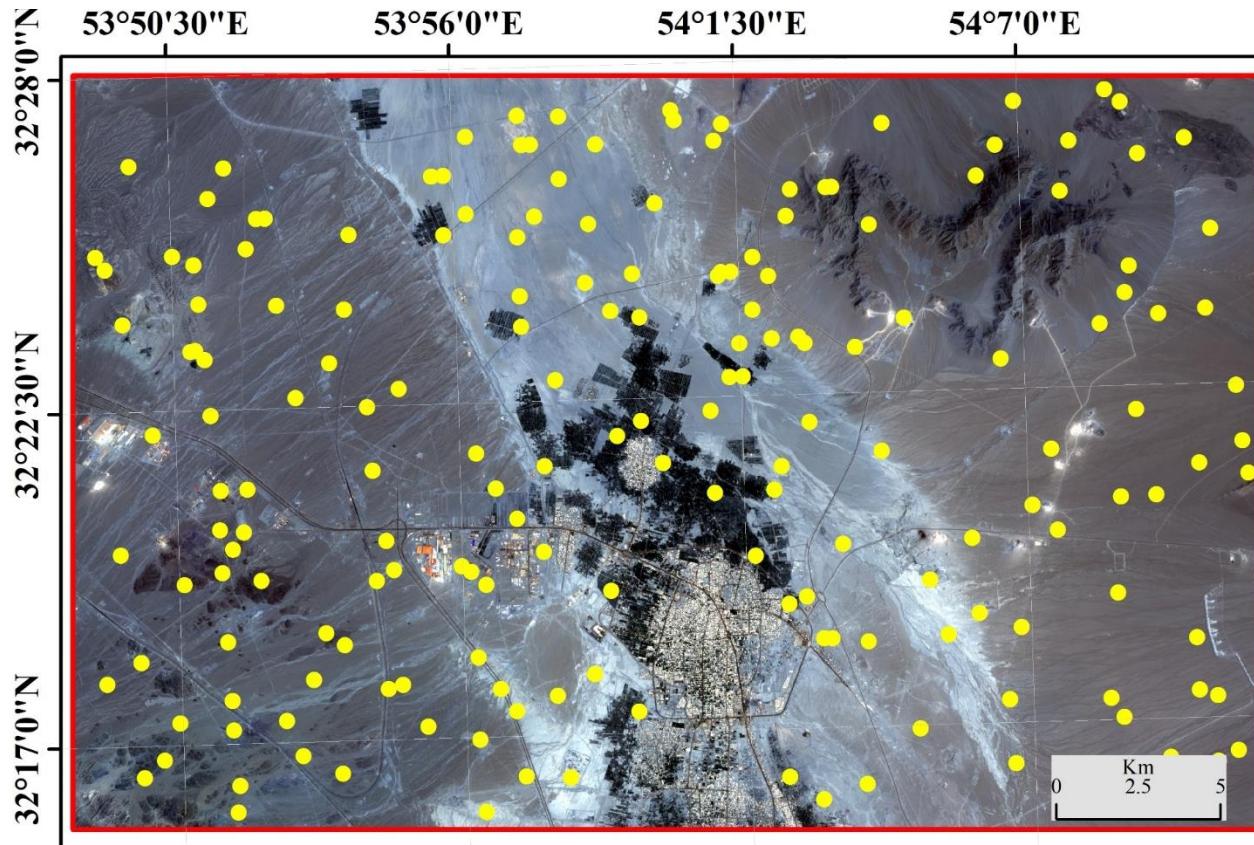


Illustration of the spatial blocking strategy. The numbers in the blocks are fold numbers, showing allocation of blocks to folds.

Spatial K-Folds Cross Validation

- Spatial K-Folds Cross Validation:

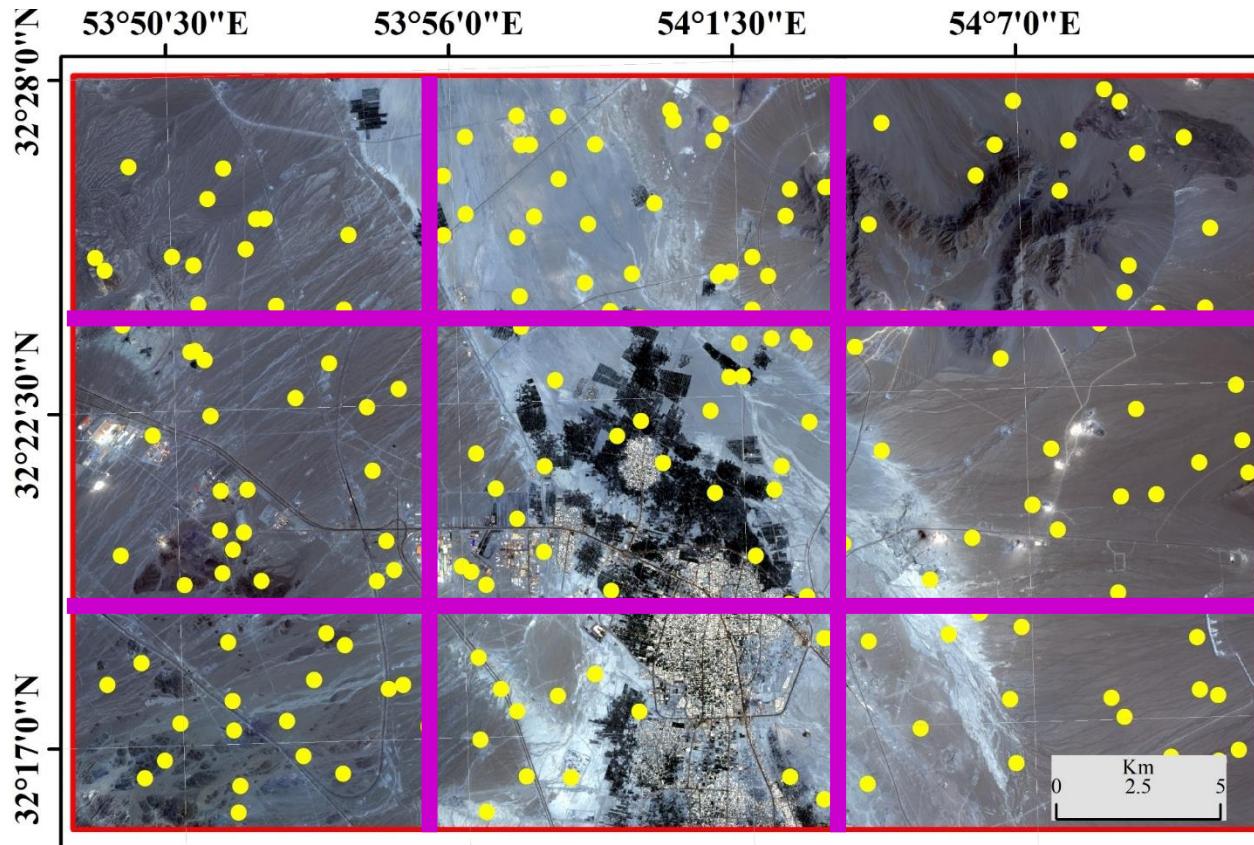


Illustration of the spatial blocking strategy. The numbers in the blocks are fold numbers, showing allocation of blocks to folds.

Spatial K-Folds Cross Validation

- Spatial K-Folds Cross Validation:

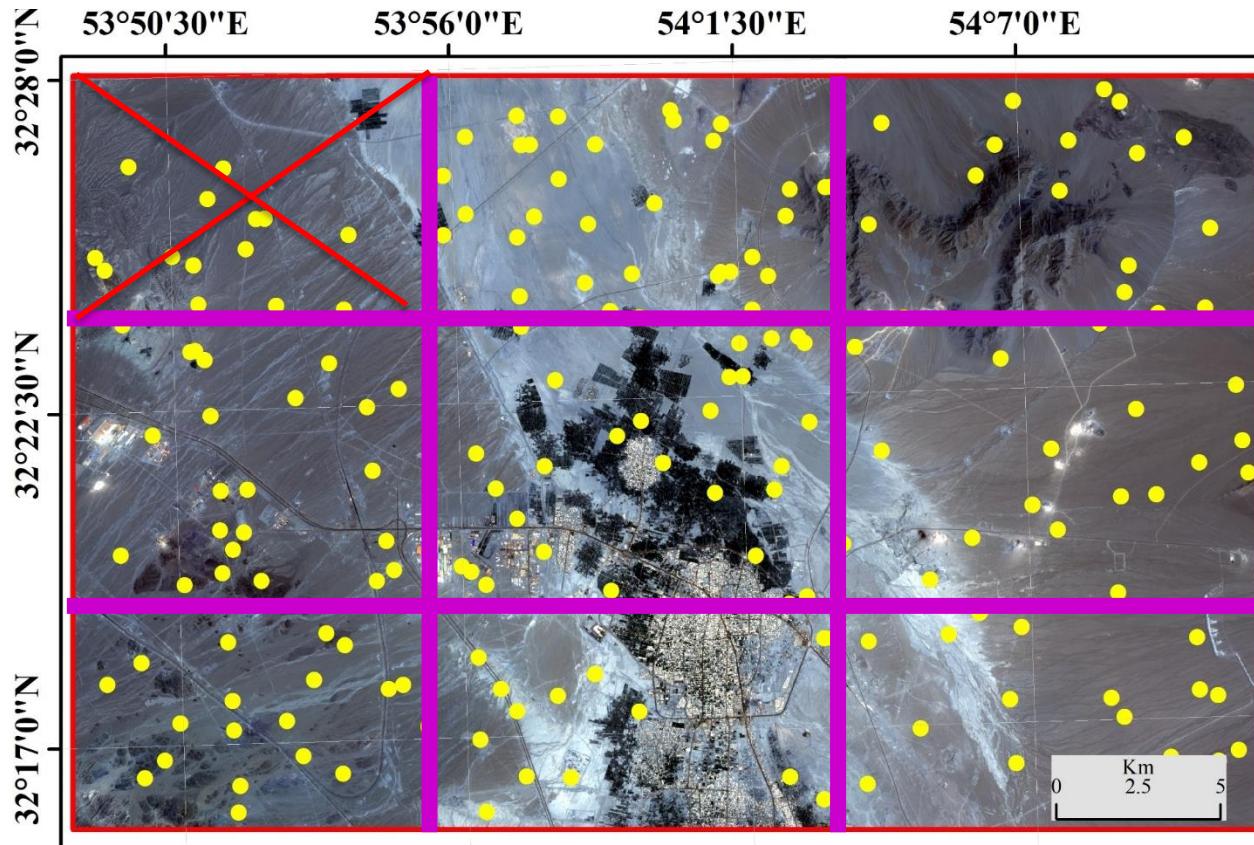


Illustration of the spatial blocking strategy. The numbers in the blocks are fold numbers, showing allocation of blocks to folds.

Spatial K-Folds Cross Validation

- Spatial K-Folds Cross Validation:

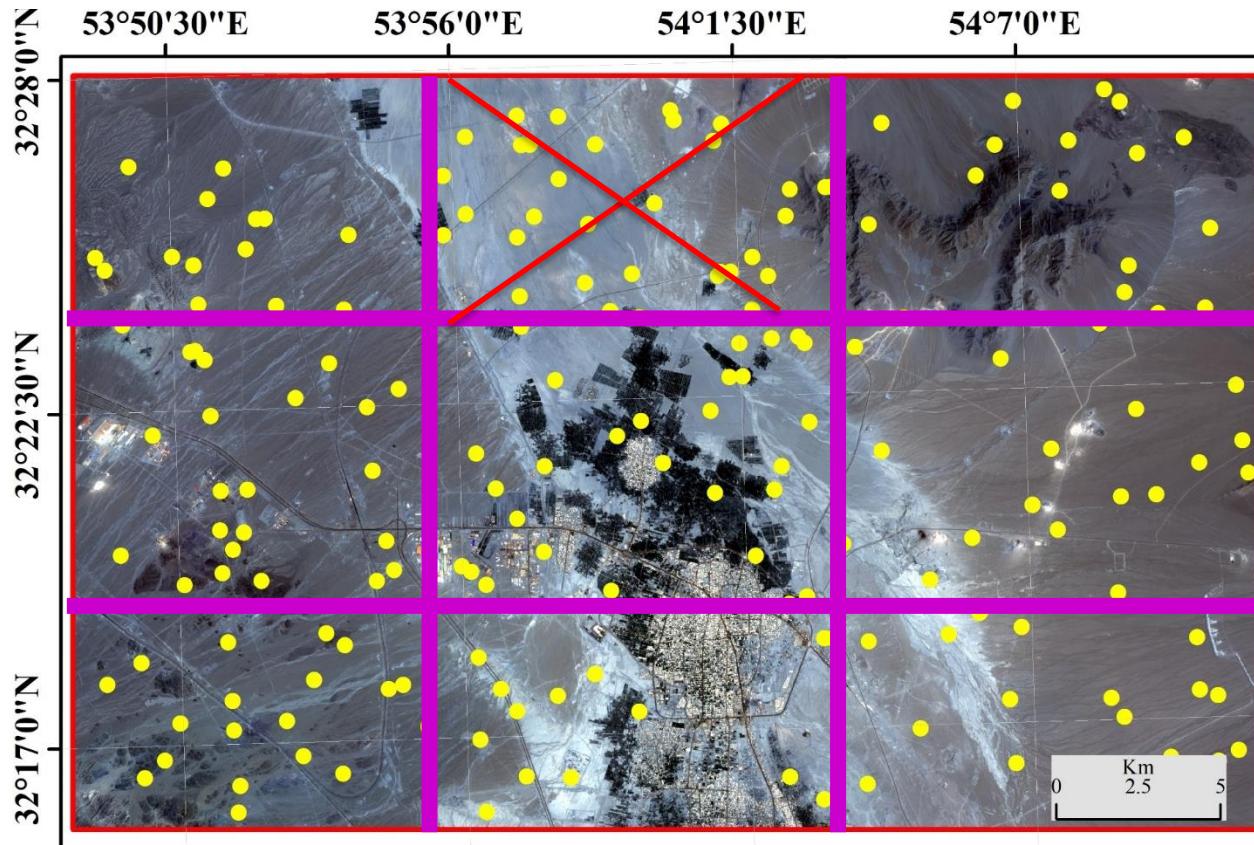


Illustration of the spatial blocking strategy. The numbers in the blocks are fold numbers, showing allocation of blocks to folds.

Spatial K-Folds Cross Validation

- Spatial K-Folds Cross Validation:

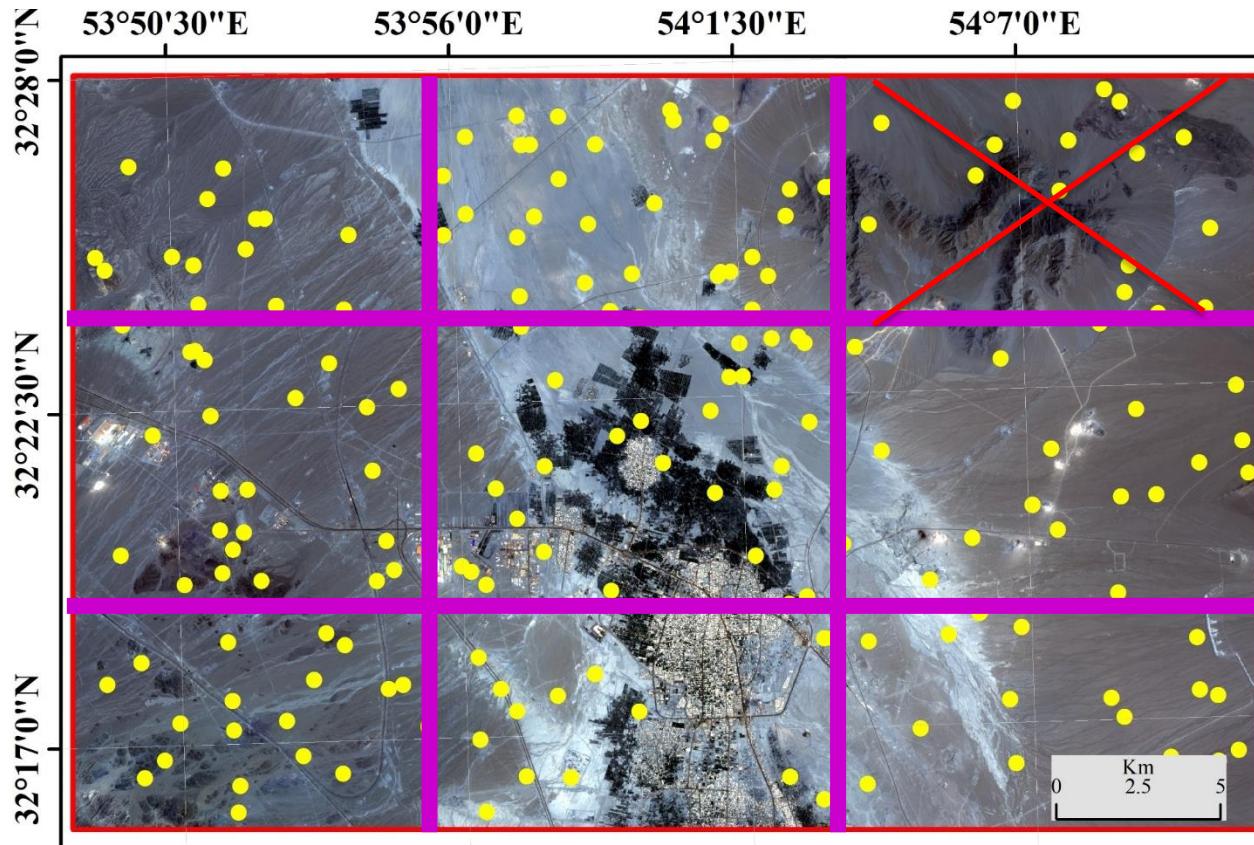


Illustration of the spatial blocking strategy. The numbers in the blocks are fold numbers, showing allocation of blocks to folds.

Spatial K-Folds Cross Validation

- Spatial K-Folds Cross Validation:

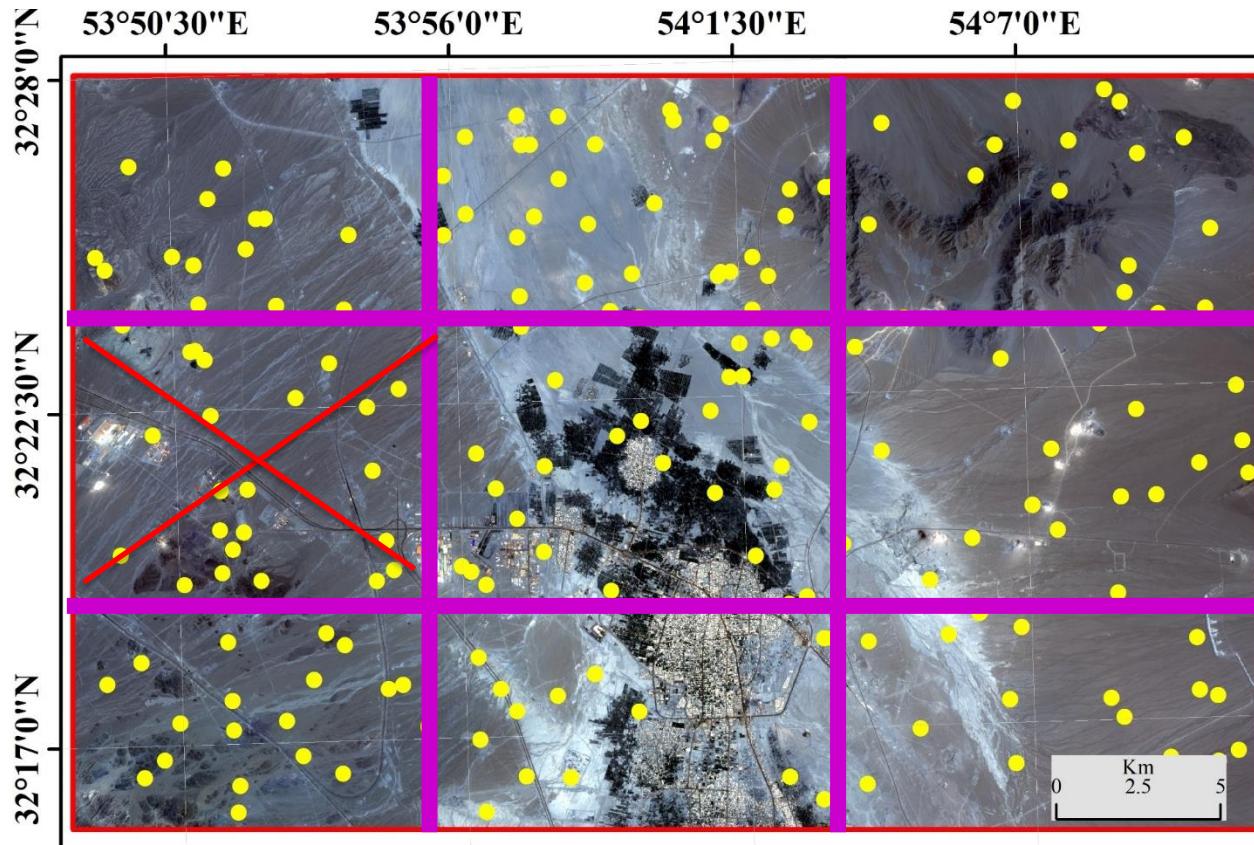
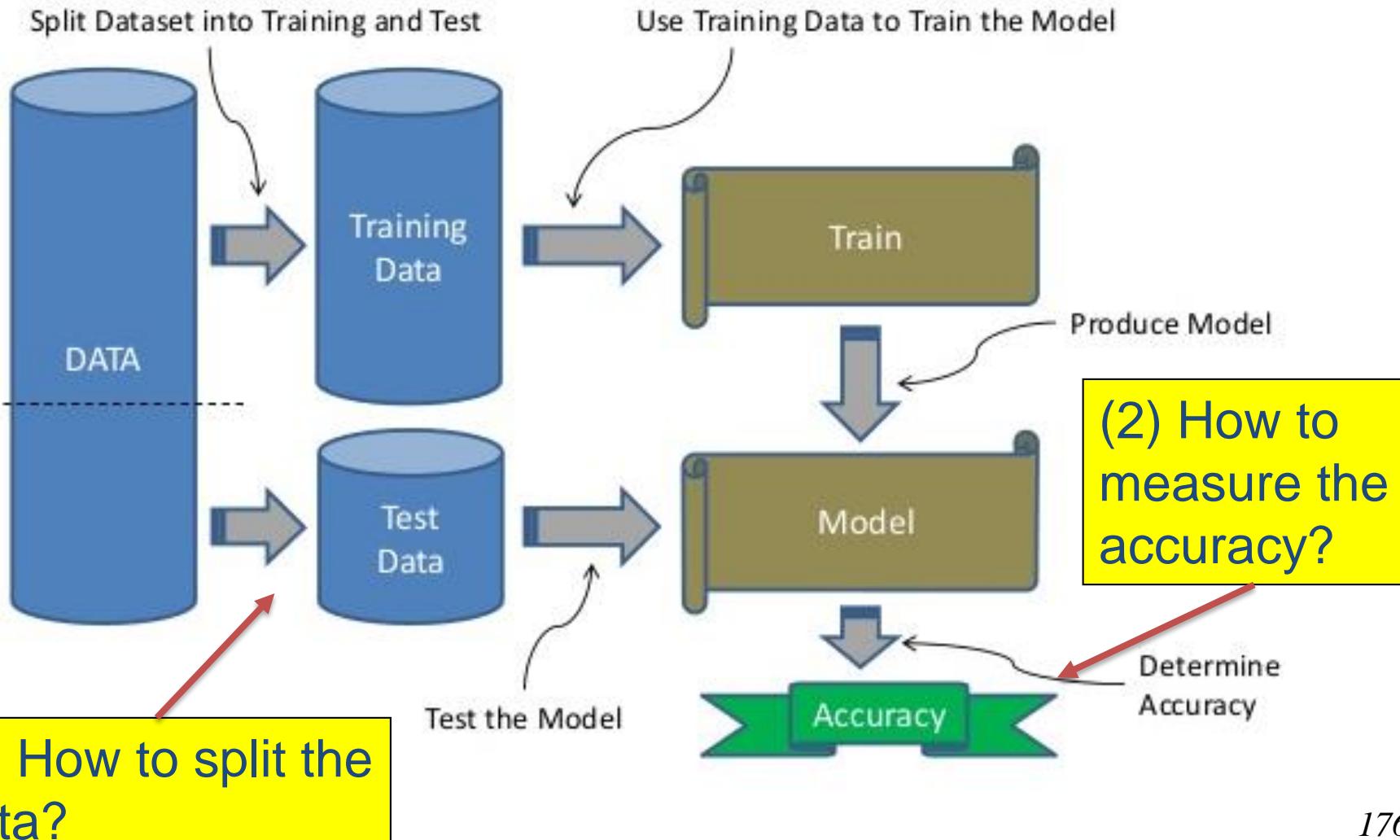


Illustration of the spatial blocking strategy. The numbers in the blocks are fold numbers, showing allocation of blocks to folds.

Accuracy Assessment of Models



How to measure the accuracy?

- Performance Metrics:

Regression	Classification
<ul style="list-style-type: none">• Mean Absolute Error (MAE)• Root Mean Squared Error (RMSE)• R-Squared and Adjusted R-Squared	<ul style="list-style-type: none">• Recall• Precision• F1-Score• Accuracy• Area Under the Curve (AUC)

Performance Metrics for Regression

- Mean Absolute Error:

$$MAE = \frac{1}{n} \sum_{\text{Sum of}} |y - \hat{y}|$$

Divide by the total number of data points

Predicted output value

Actual output value

The absolute value of the residual

- Root Mean Square Error:

$$RMSE = \sqrt{\sum_{i=1}^n \frac{(\hat{y}_i - y_i)^2}{n}}$$

Performance Metrics for Classification

- **Confusion Matrix:** is one of the most intuitive and easiest metrics used for finding the correctness and accuracy of the model.

		ACTUAL VALUES	
		POSITIVE	NEGATIVE
PREDICTED VALUES	POSITIVE	TP <i>correct</i>	FP
	NEGATIVE	FN	TN <i>correct</i>

where TP, TN, FP and FN are true positive, true negative, false positive and false negative, respectively

Performance Metrics for Classification

- Confusion Matrix:

- Class A: Aridisols
- Class B: Entisols

	Class A: Aridisoils	Class B: Entisols
Class A: Aridisoils		
Class B: Entisols		

Performance Metrics for Classification

- Confusion Matrix:

 Class A: Aridisols

 Class B: Entisols

Predicted Test data

Actual Test data

	Class A: Aridisols	Class B: Entisols
Class A: Aridisols		
Class B: Entisols		

Performance Metrics for Classification

- Confusion Matrix:

 Class A: Aridisols

 Class B: Entisols

Predicted Test data

     1 2 3 4 5

Actual Test data

     1 2 3 4 5

	Class A: Aridisols	Class B: Entisols
Class A: Aridisols		
Class B: Entisols		

Performance Metrics for Classification

- Confusion Matrix:

 Class A: Aridisols

 Class B: Entisols

Predicted Test data

 1  2  3  4  5

Actual Test data

1  2  3  4  5

	Class A: Aridisols	Class B: Entisols
Class A: Aridisols		
Class B: Entisols		

Performance Metrics for Classification

- Confusion Matrix:

- Class A: Aridisols
- Class B: Entisols

Predicted Test data



Actual Test data



	Class A: Aridisols	Class B: Entisols
Class A: Aridisols		
Class B: Entisols		

Performance Metrics for Classification

- Confusion Matrix:

 Class A: Aridisols

 Class B: Entisols

Predicted Test data

 1  2  3  4  5

Actual Test data

 1  2  3  4  5

	Class A: Aridisols	Class B: Entisols
Class A: Aridisols		
Class B: Entisols		

Performance Metrics for Classification

- Confusion Matrix:

 Class A: Aridisols

 Class B: Entisols

Predicted Test data

 1  2  3  4  5 

Actual Test data

 1  2  3  4  5 

	Class A: Aridisols	Class B: Entisols
Class A: Aridisols		
Class B: Entisols		

Performance Metrics for Classification

- Confusion Matrix:

 Class A: Aridisols

 Class B: Entisols

Predicted Test data

 1  2  3  4  5

Actual Test data

 1  2  3  4  5

	Class A: Aridisols	Class B: Entisols
Class A: Aridisols		
Class B: Entisols		

Performance Metrics for Classification

- **Overall Accuracy:** is a metric calculating the classifier overall accuracy

	Class A: Aridisoils	Class B: Entisols
Class A: Aridisoils		
Class B: Entisols		

$$OA = \left(\frac{\text{Correctly classified}}{\text{Total number of test data}} \right) * 100$$

$$OA = \left(\frac{2 + 1}{5} \right) * 100 = 60\%$$

Performance Metrics for Classification

- **Precision:** is the proportion of those predicted instances that are correctly classified

	Class A: Aridisols	Class B: Entisols
Class A: Aridisols		
Class B: Entisols		

$$Pr = \left(\frac{\text{Correctly classified for each class}}{\text{Total number of row data}} \right) * 100$$

$$Pr(\text{class } A) = \left(\frac{2}{2 + 1} \right) * 100 = 66\%$$

$$Pr(\text{class } B) = \left(\frac{1}{1 + 1} \right) * 100 = 50\%$$

Performance Metrics for Classification

- **Recall:** is the proportion of those instances that are correctly classified

	Class A: Aridisoils	Class B: Entisols
Class A: Aridisoils		
Class B: Entisols		

The diagram shows a 3x3 confusion matrix. The columns are labeled 'Class A: Aridisols' and 'Class B: Entisols'. The rows are labeled 'Class A: Aridisols' and 'Class B: Entisols'. The matrix contains black vertical bars representing correctly classified instances. Blue arrows point downwards from the top of each column, indicating the total number of instances in each class.

$$Re = \left(\frac{\text{Correctly classified for each class}}{\text{Total number of coloum data}} \right) * 100$$

$$Re(\text{class } A) = \left(\frac{2}{2 + 1} \right) * 100 = 66\%$$

$$Re(\text{class } B) = \left(\frac{1}{1 + 1} \right) * 100 = 50\%$$

Performance Metrics for Classification

- **F-score:** the F-score is the harmonic mean of precision and recall

$$F-score = \left(\frac{2 \times Precision \times Recall}{Precision + Recall} \right) * 100$$

Metric	Formula
Accuracy	$ACC = \frac{TP+TN}{TP+TN+FP+FN}$
Error rate	$ERR = \frac{FP+FN}{TP+TN+FP+FN}$
Precision	$PRC = \frac{TP}{TP+FP}$
Sensitivity	$SNS = \frac{TP}{TP+FN}$
Specificity	$SPC = \frac{TN}{TN+FP}$
ROC	$ROC = \frac{\sqrt{SNS^2+SPC^2}}{\sqrt{2}}$
F_1 score	$F_1 = 2 \frac{PRC \cdot SNS}{PRC + SNS}$
Geometric Mean	$GM = \sqrt{SNS \cdot SPC}$

Sequence of DSM Steps



- 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**.

DSM Examples using Machine Learning

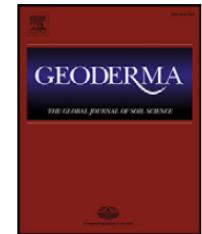
Geoderma 265 (2016) 62–77



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



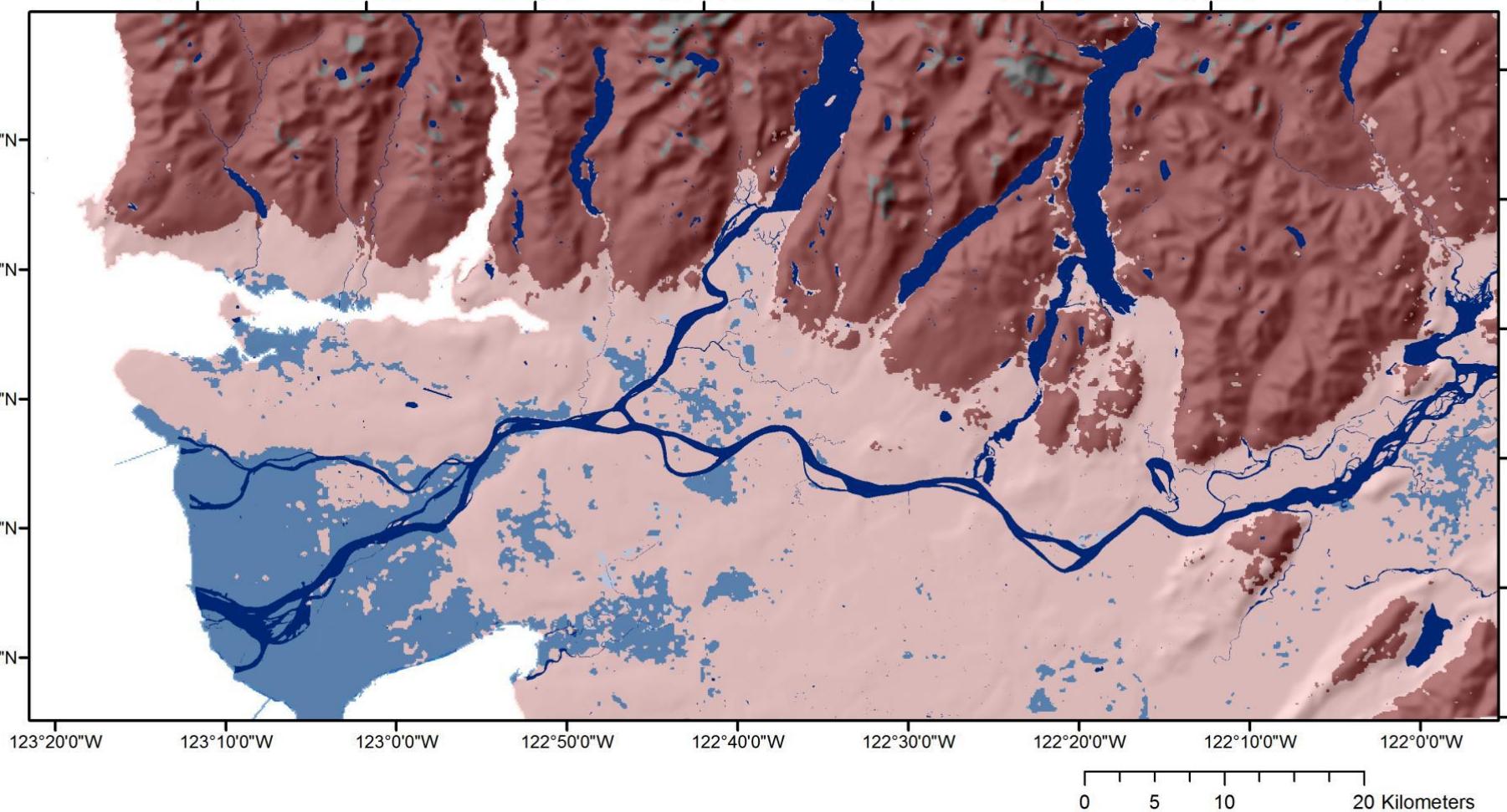
Brandon Heung ^{a,*}, Hung Chak Ho ^b, Jin Zhang ^a, Anders Knudby ^c, Chuck E. Bulmer ^d, Margaret G. Schmidt ^a

^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

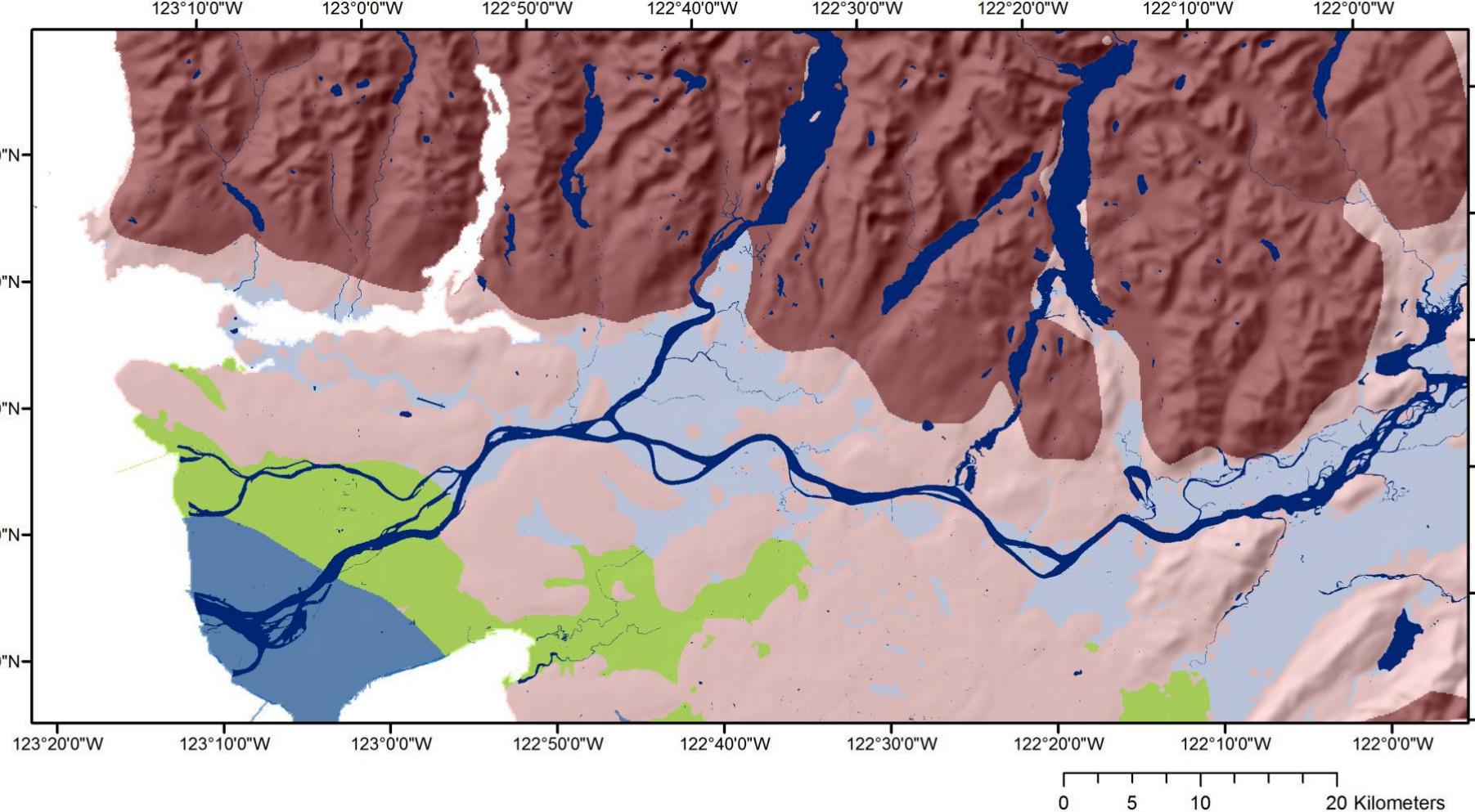
^c Department of Geography, University of Ottawa, 60 University, Ottawa, ON, K1N 6N5, Canada

^d British Columbia Ministry of Forests Lands and Natural Resources Operations, Natural Resource Sciences Section, Vernon, BC, V1B 2C7, Canada



Area-Weighted: Nearest Shrunken Centroid

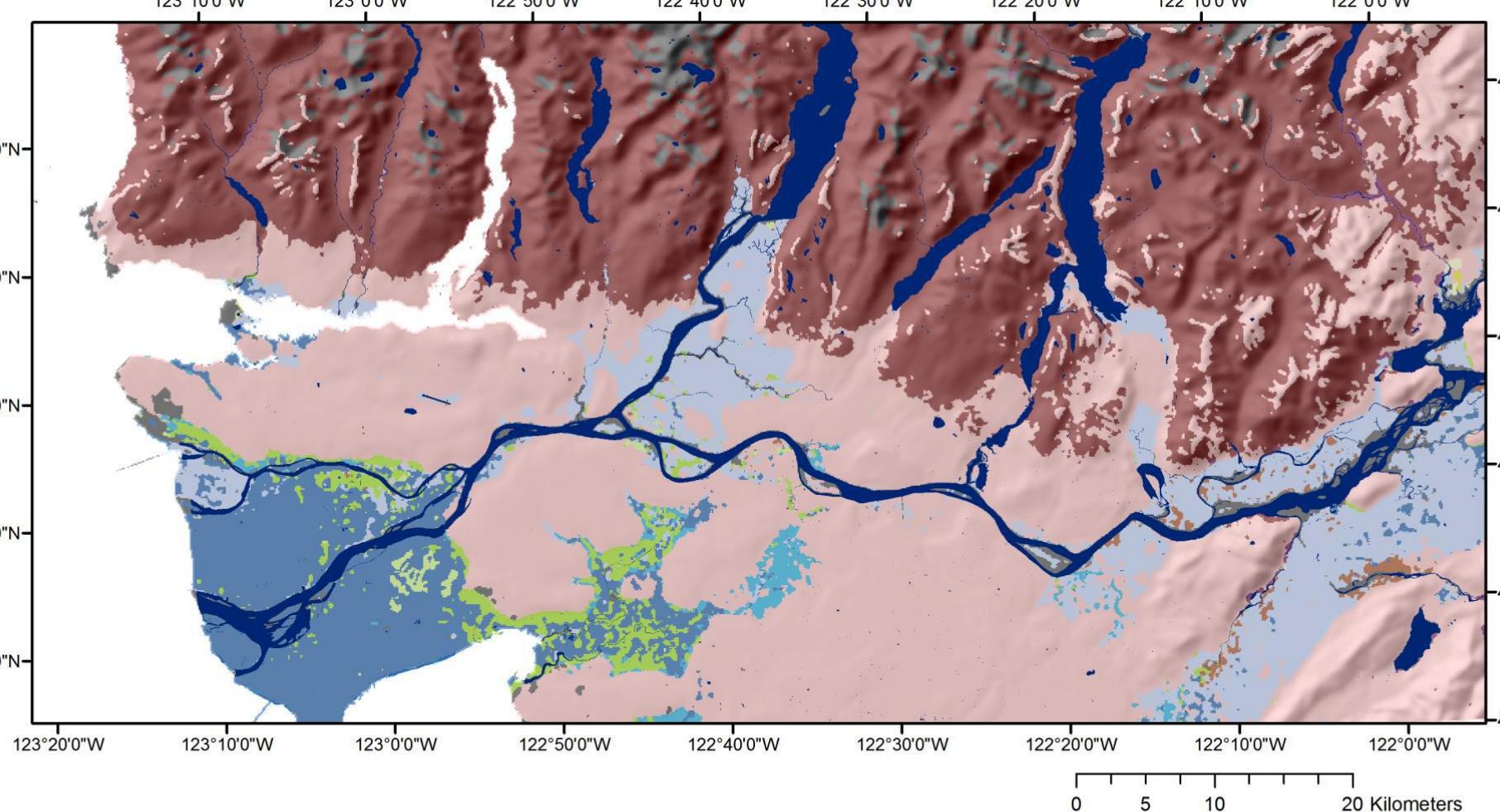
$C = 40\%$



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

$$C = 42\%$$

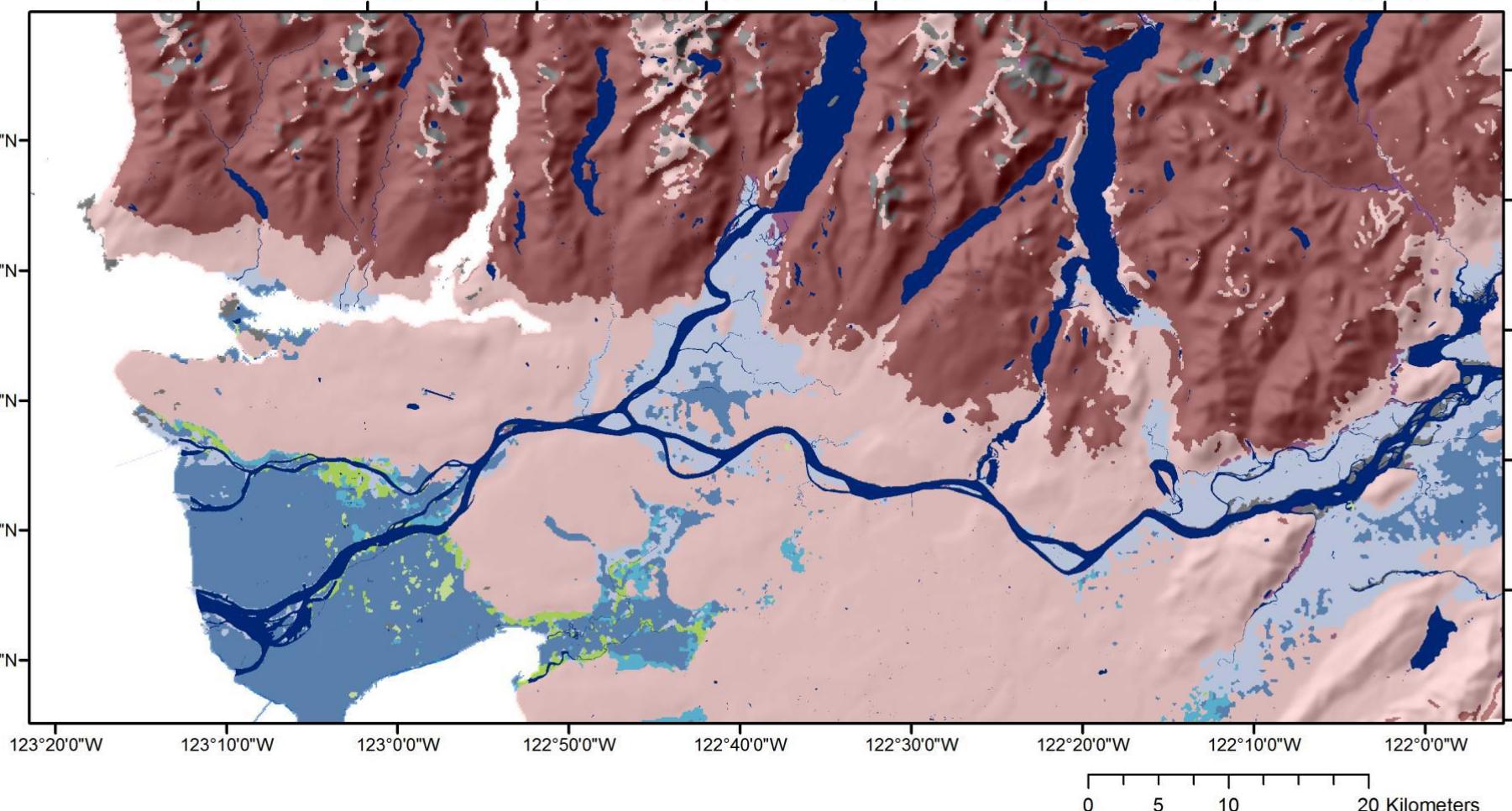
Area-Weighted: CART



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

$$C = 48\%$$

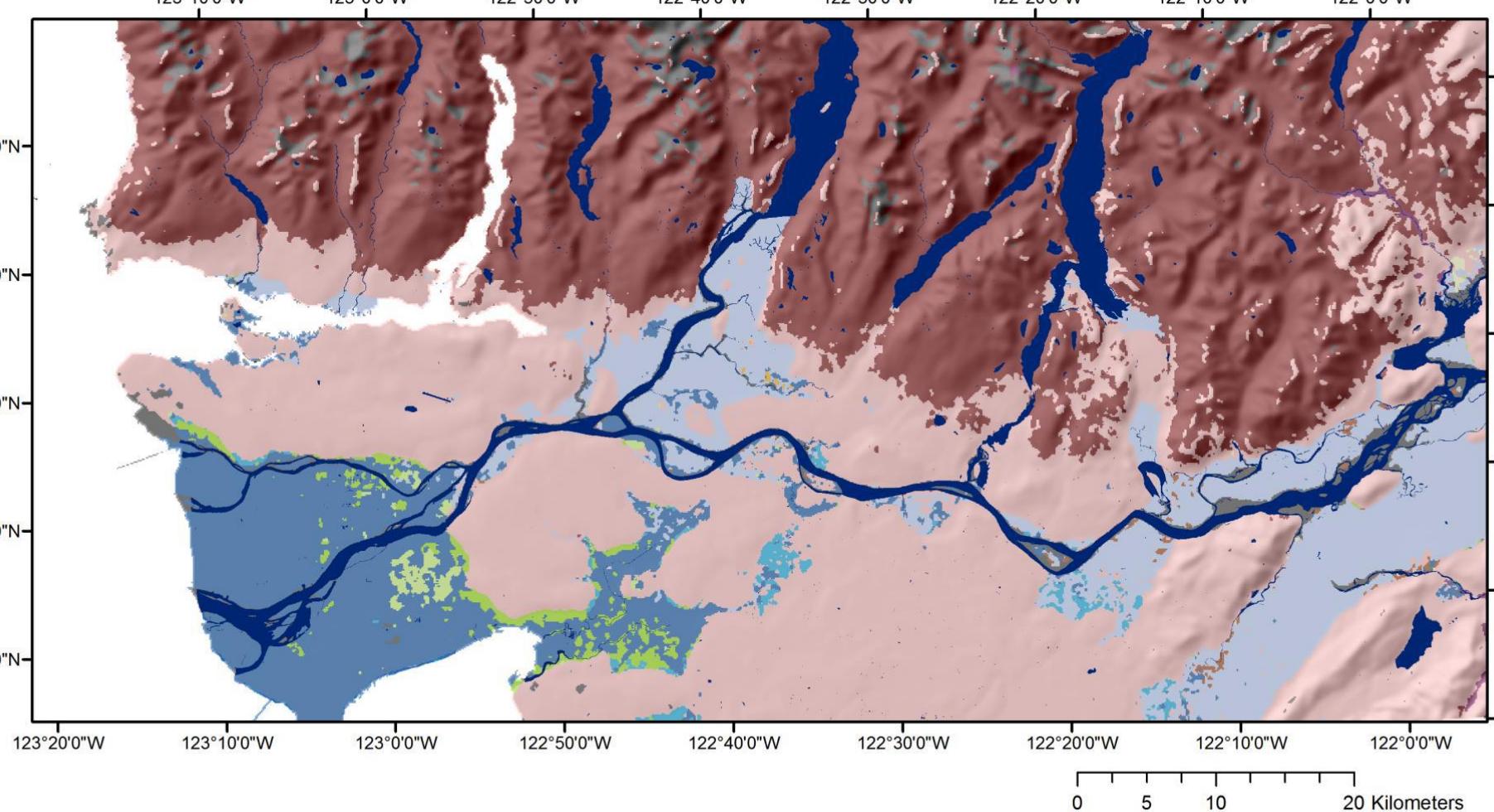
Area-Weighted: Multinomial Logistic Regression



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

$$C = 49\%$$

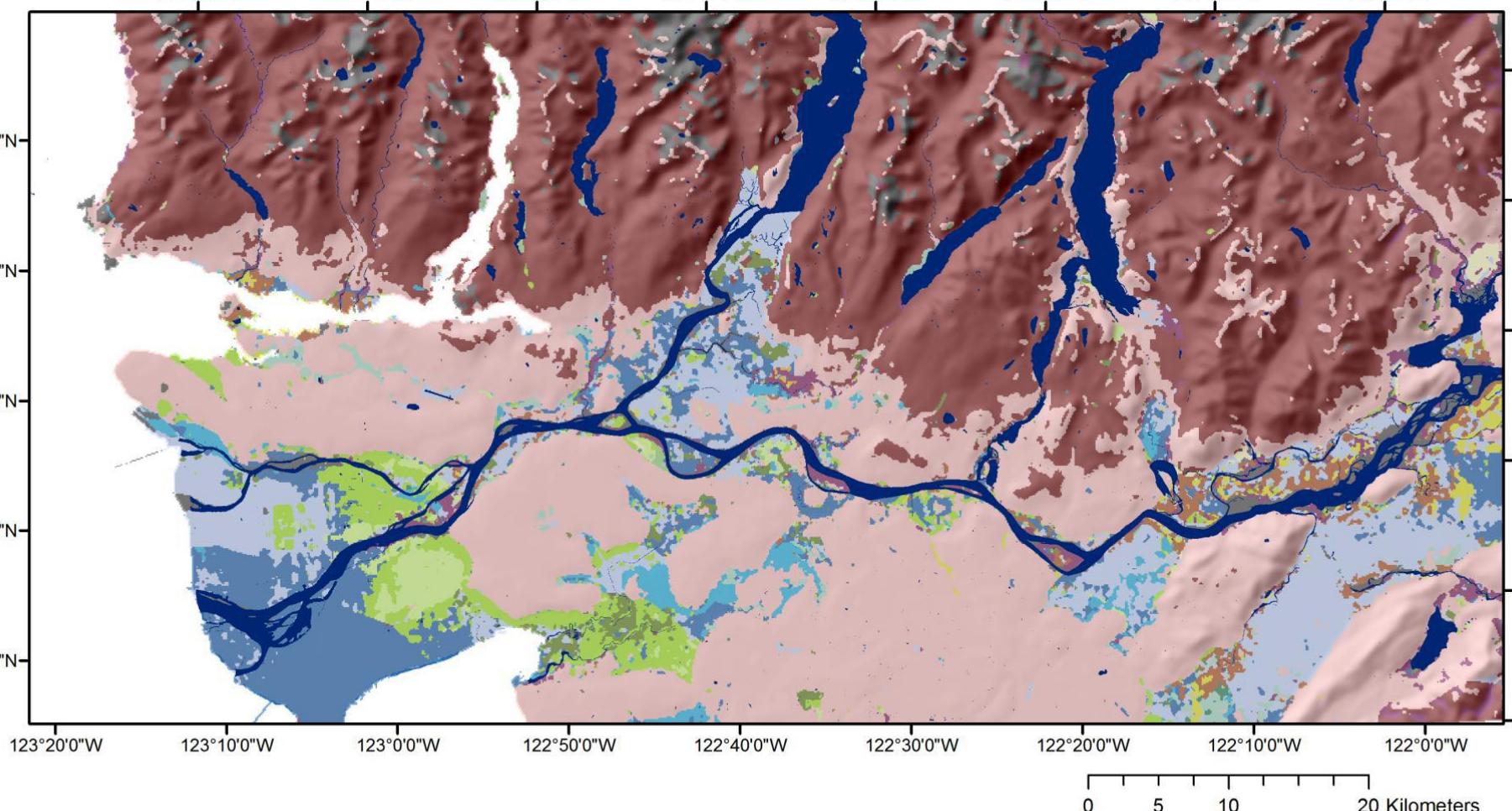
Area-Weighted: Artificial Neural Network



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

$C = 50\%$

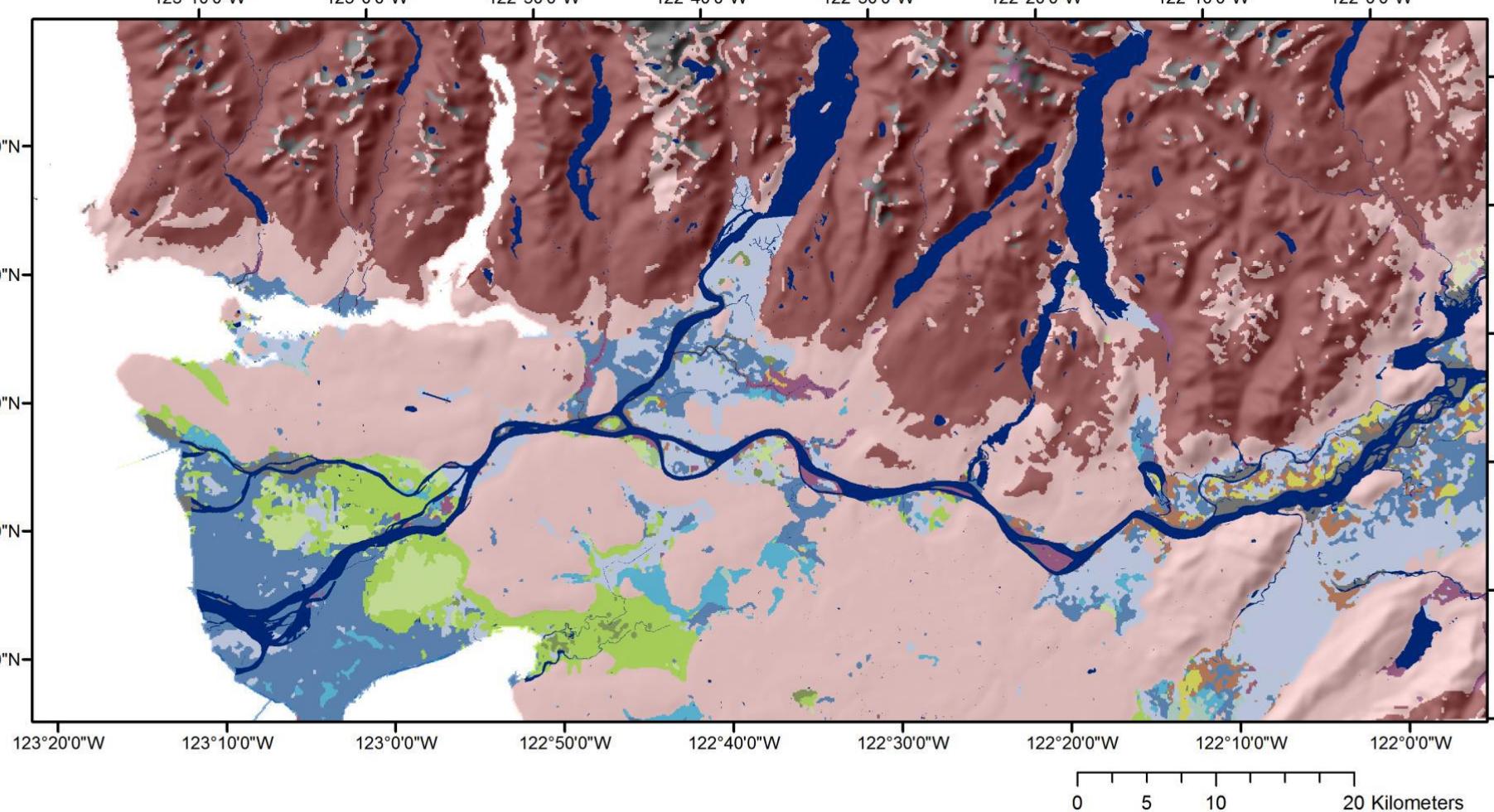
Area-Weighted: Support Vector Machine - Linear



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

$$C = 65\%$$

Area-Weighted: Logistic Model Tree



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

Area-Weighted: Random Forest

$$C = 66\%$$

Take-Home Message

- ML has great potential for spatial and spatio-temporal predictions
- Each model predicts one realization of a DSM
- R allows for easy model training
- Accuracy assessment is important
- Avoid overfitting by careful variable selection

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**.

Extend Soil Mapping to Soil Functions

Soil functions

Soils deliver ecosystem services that enable life on Earth



Food and Agriculture Organization of the United Nations

with the support of
Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

Provision of construction materials

Foundation for human infrastructure

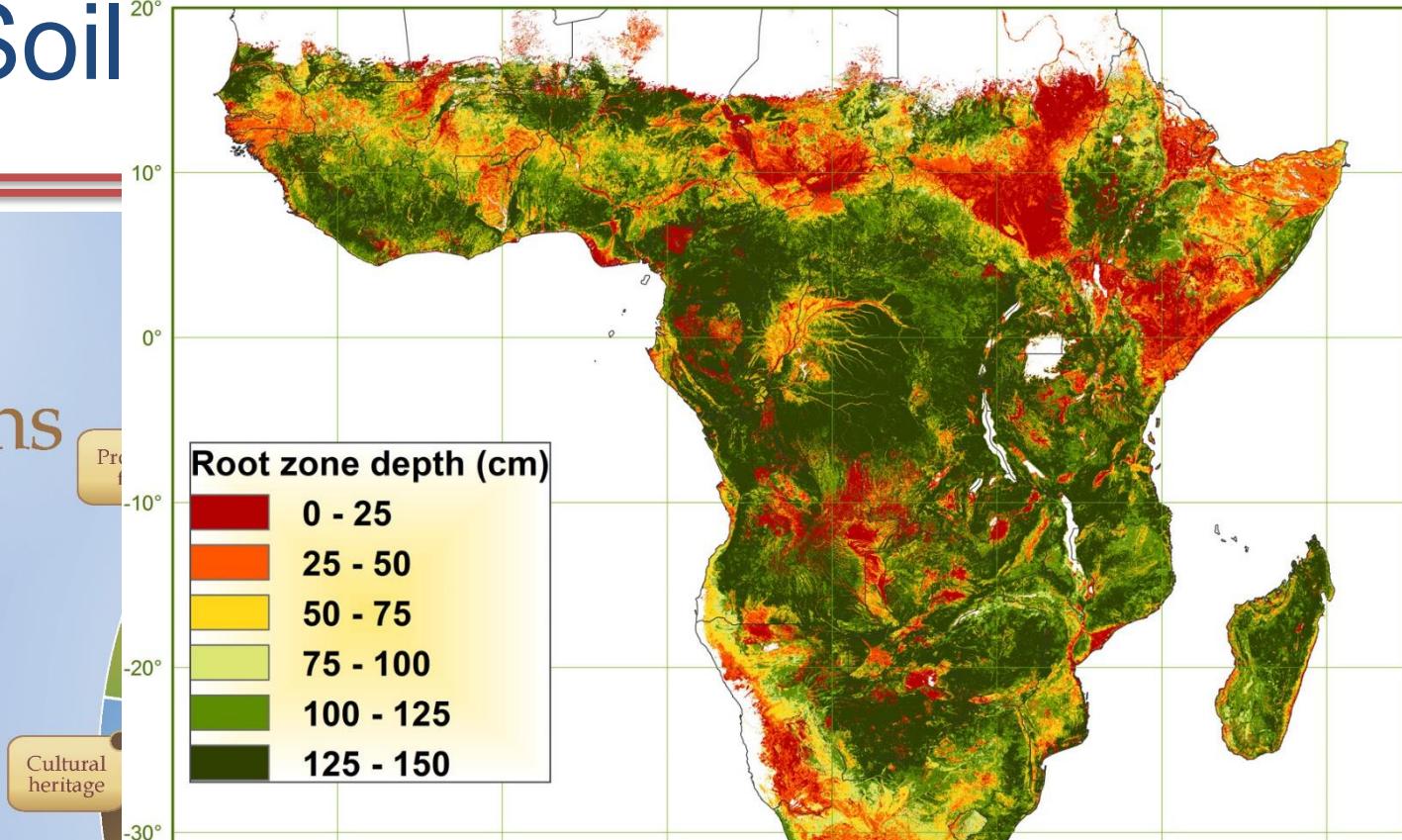
Source of pharmaceuticals and genetic resources



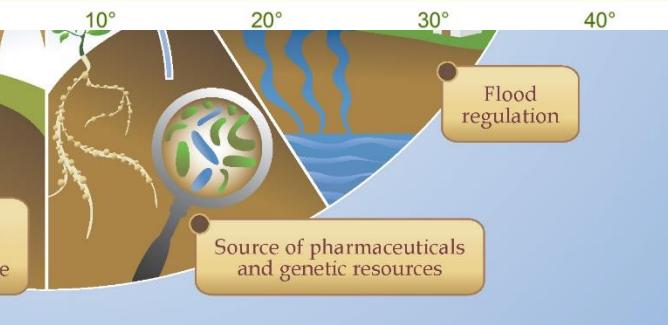
Extend Soil

Soil functions

Soils deliver ecosystem services that enable life on Earth



Geoderma
Volume 324, 15 August 2018, Pages 18-36

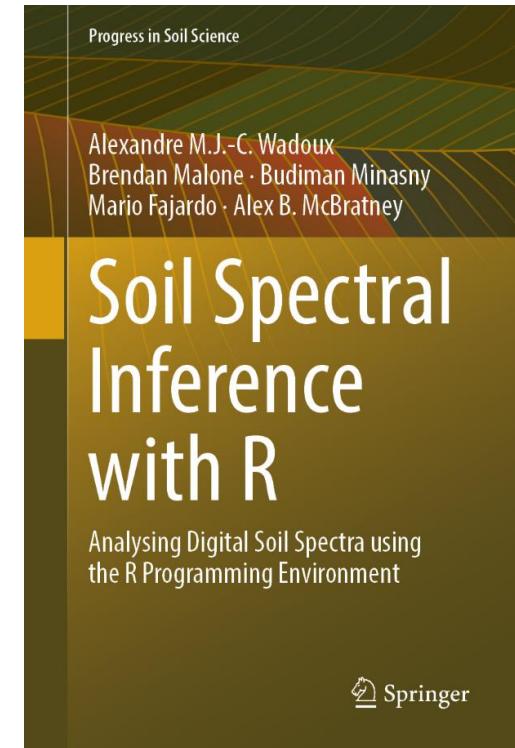
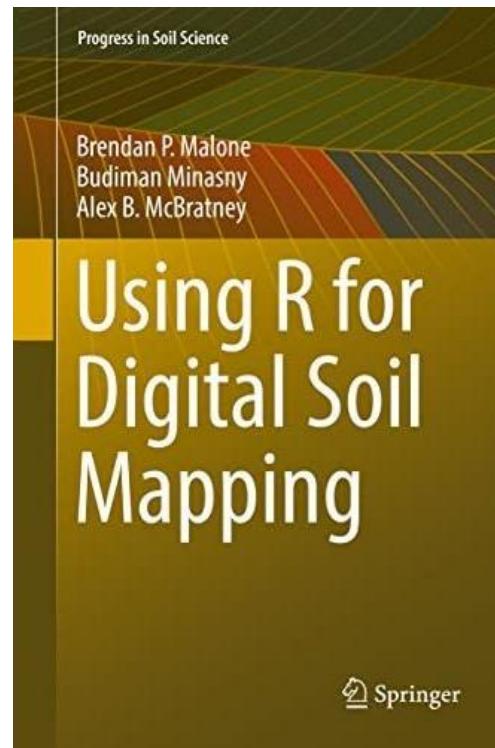
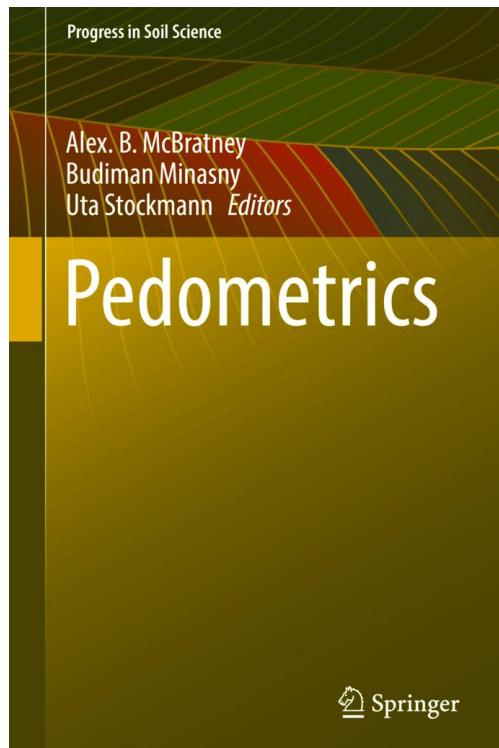


Mapping rootable depth and root zone plant-available water holding capacity of the soil of sub-Saharan Africa

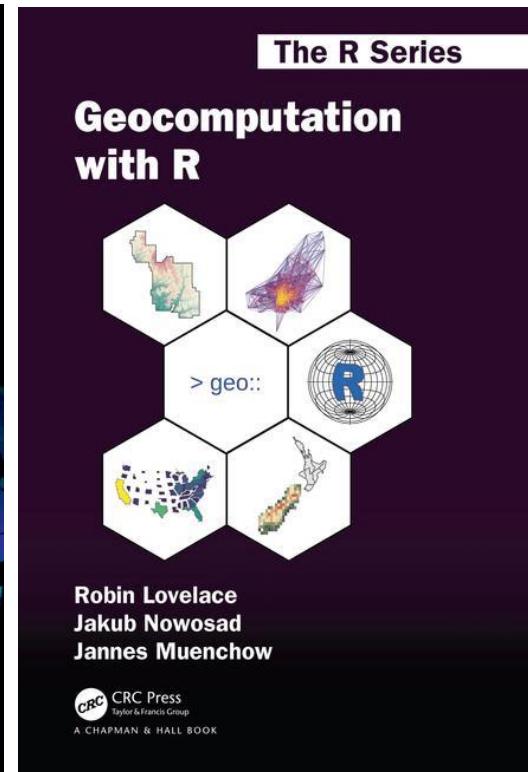
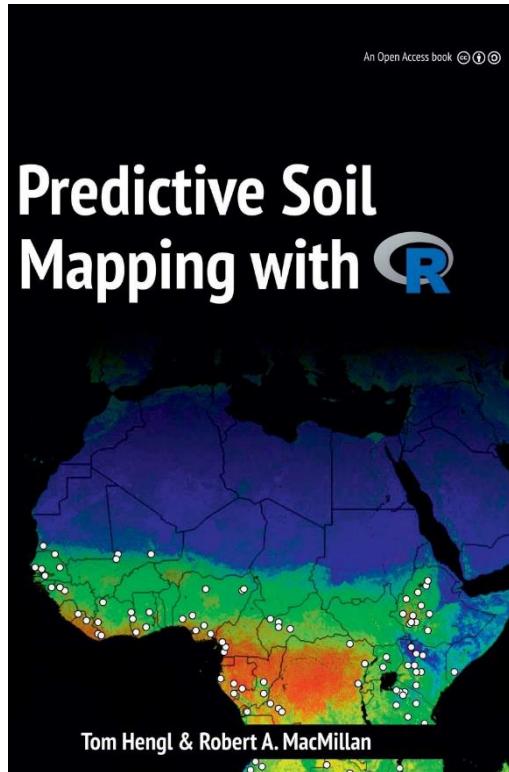
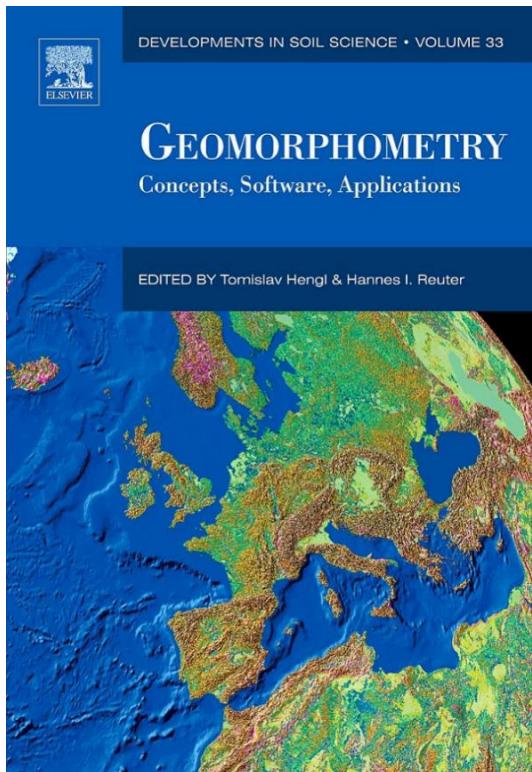
Johan G.B. Leenaars ^{a, b, c, d}, Lieven Claessens ^{b, e, f}, Gerard B.M. Heuvelink ^{b, c, d}, Tom Hengl ^b, María Ruiz-Pérez González ^b, Lenny G.J. van Bussel ^{a, f}, Nicolas Gulpert ^{b, h}, Haishun Yang ^b, Kenneth G. Cassman ^b

Show more ▾

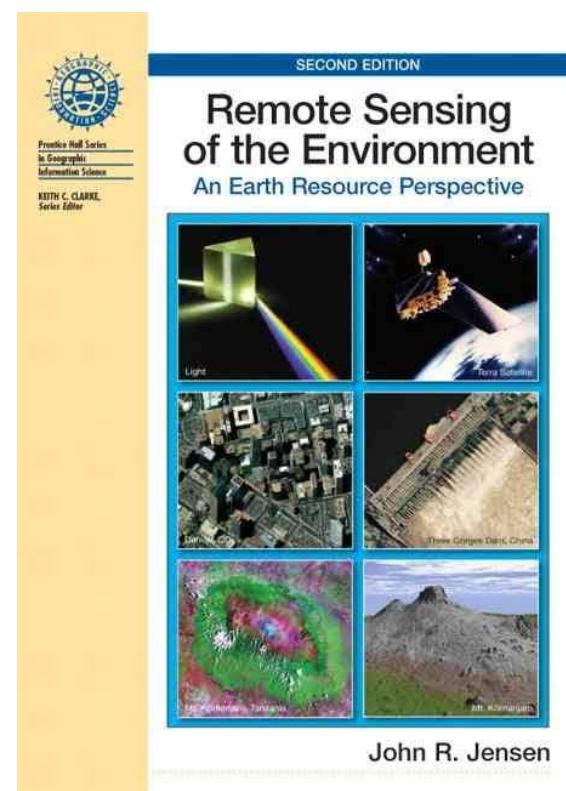
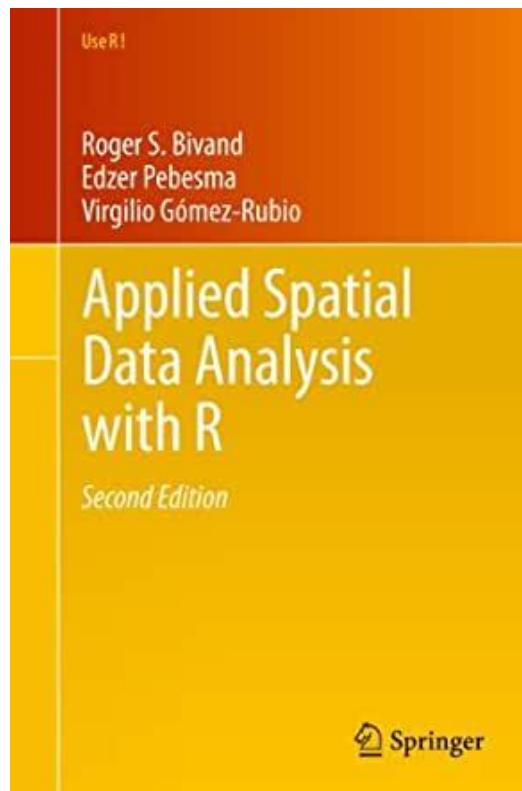
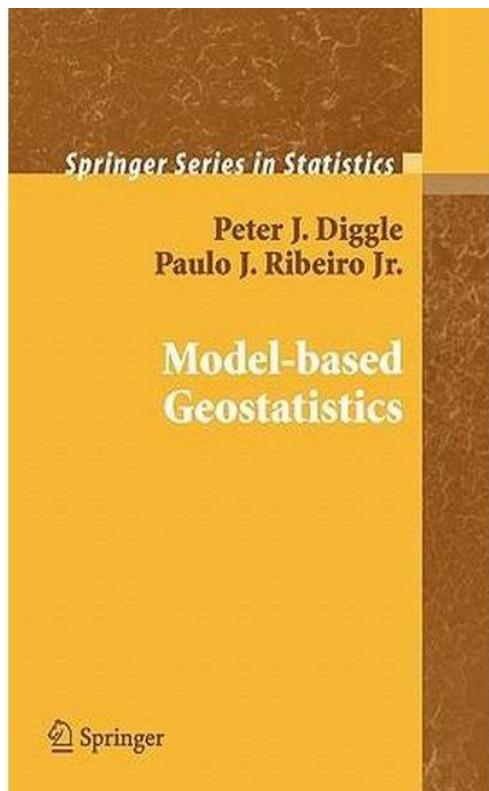
Useful Resources



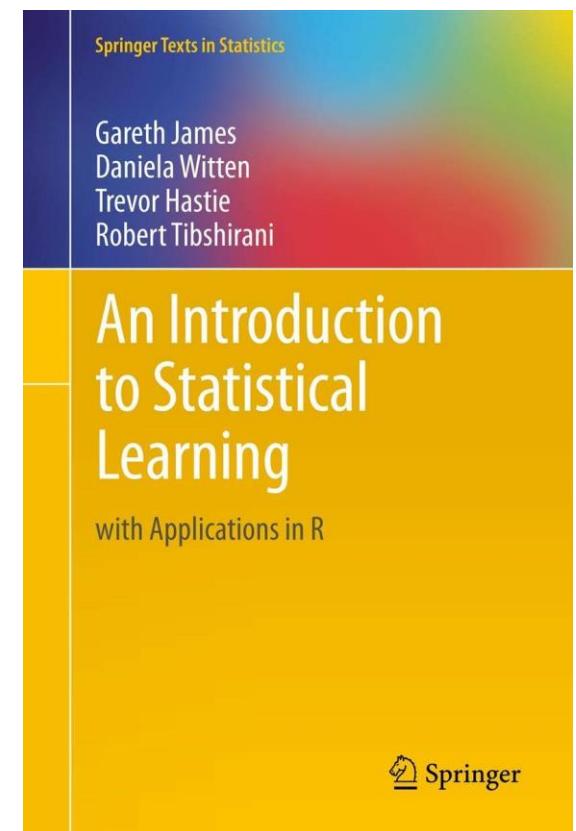
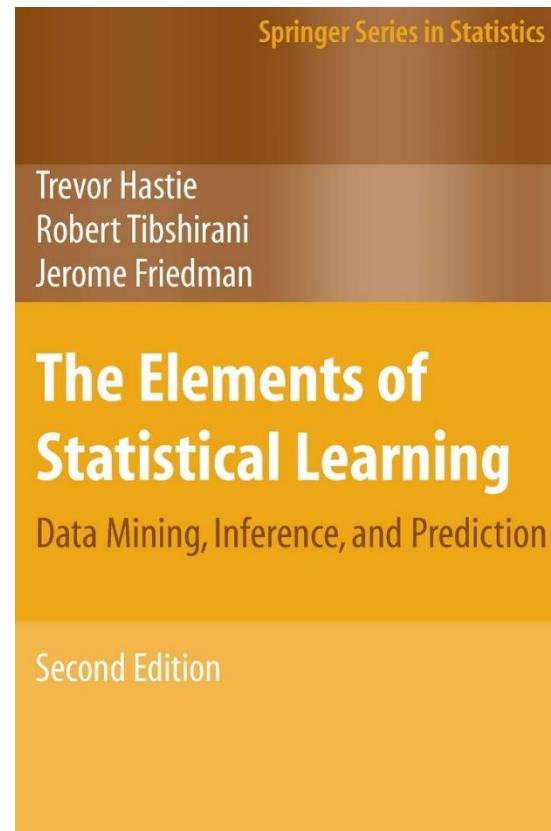
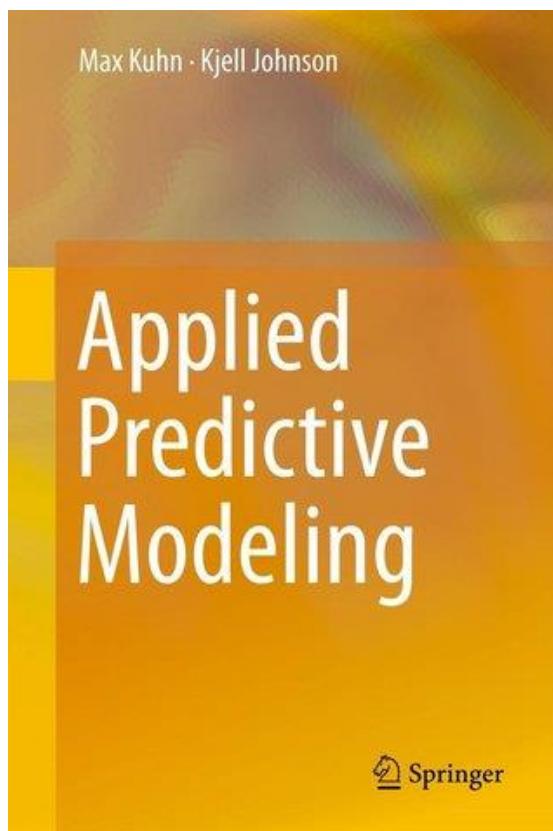
Useful Resources



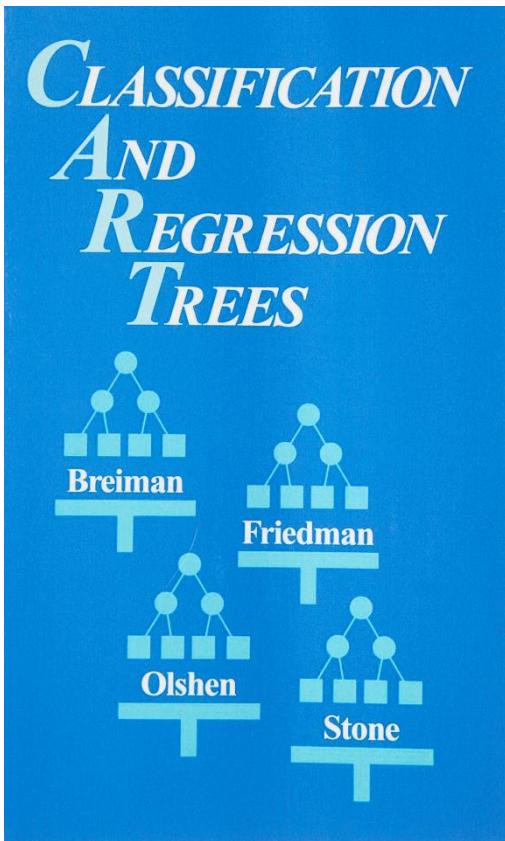
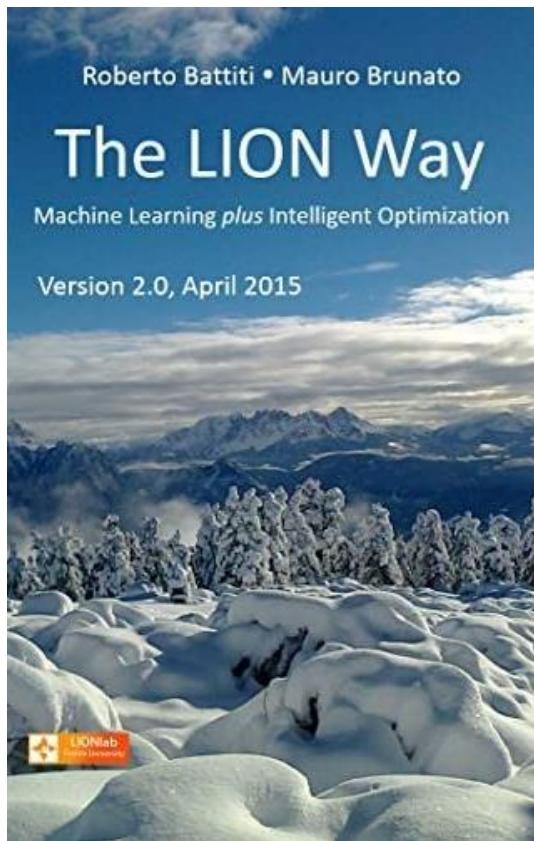
Useful Resources



Useful Resources



Useful Resources



Hothorn, Torsten. CRAN Task View: Machine Learning & Statistical Learning
<https://CRAN.R-project.org/view=MachineLearning>, 2020.

Useful Resources



<https://www.youtube.com/channel/UC6HFFFYiV4zEYJIQMIXemWA>

<https://www.youtube.com/channel/UCNi1XYjdXWF9eAjvG40KqWg>

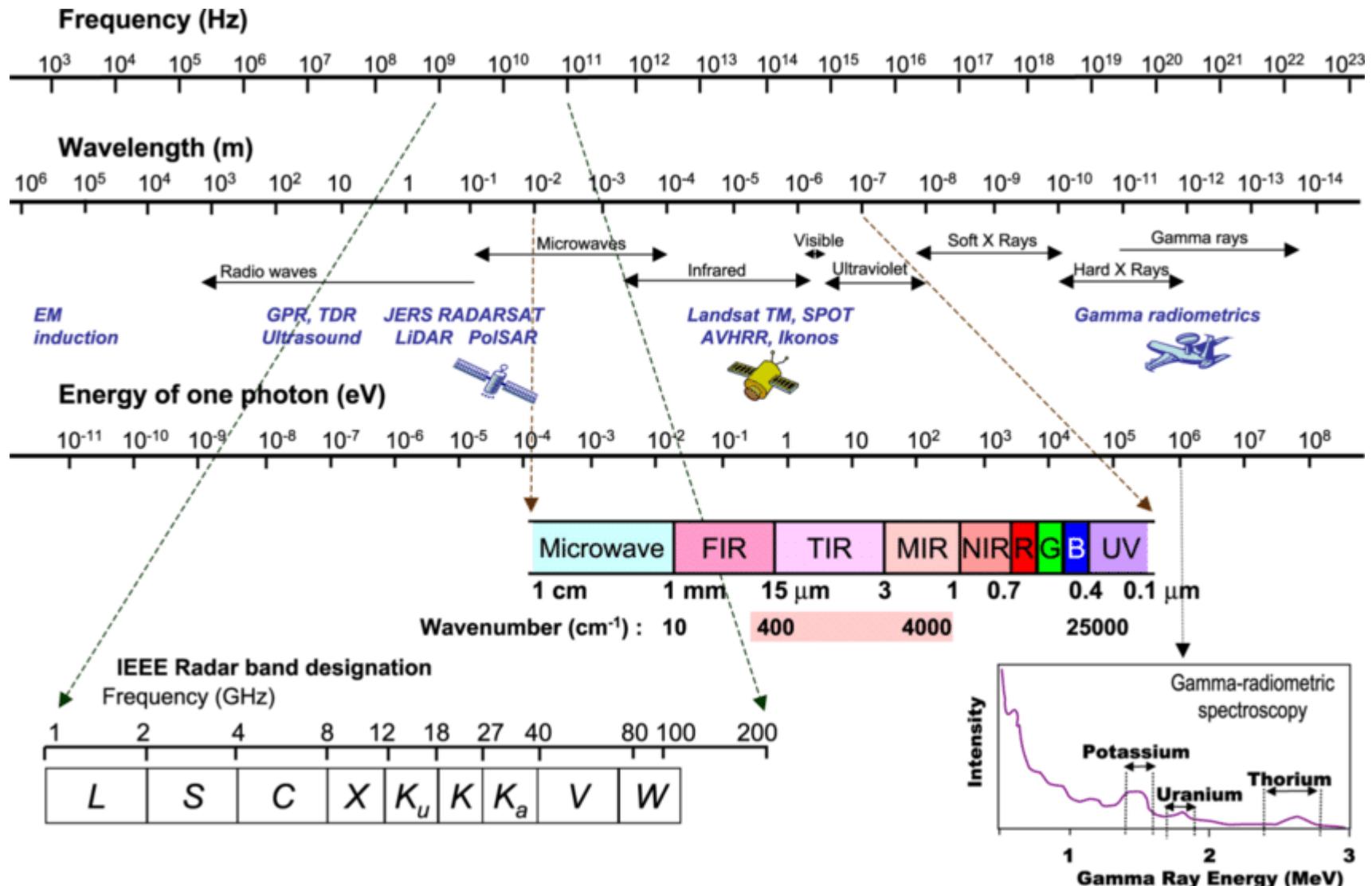
https://www.youtube.com/watch?v=PPLop4L2eGk&list=PLLssT5z_DsK-h9vYZkQkYNWcItqhIRJLN

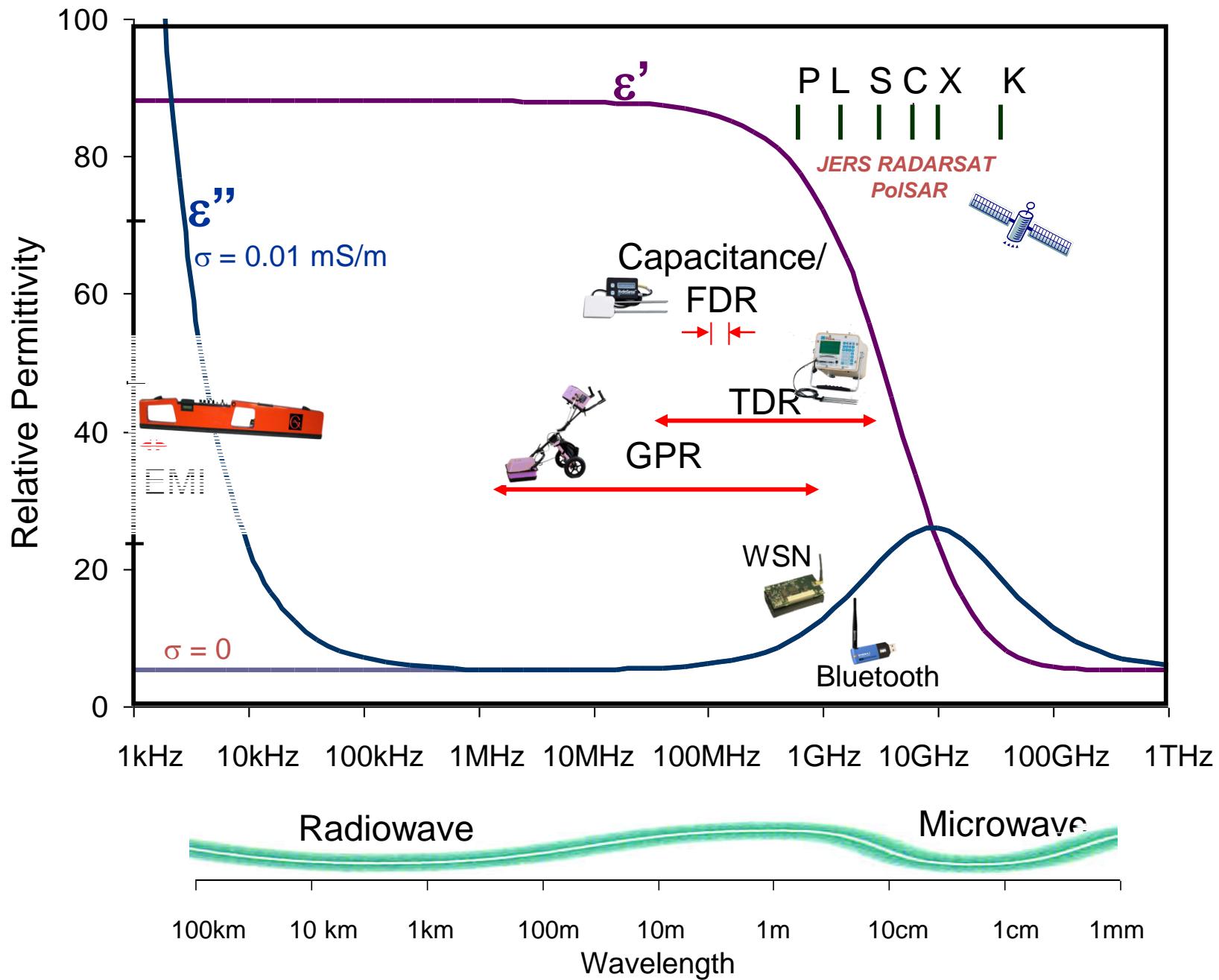
Thank You For Your Attention

Ruhollah Taghizadeh
Soil Science and Geomorphology

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

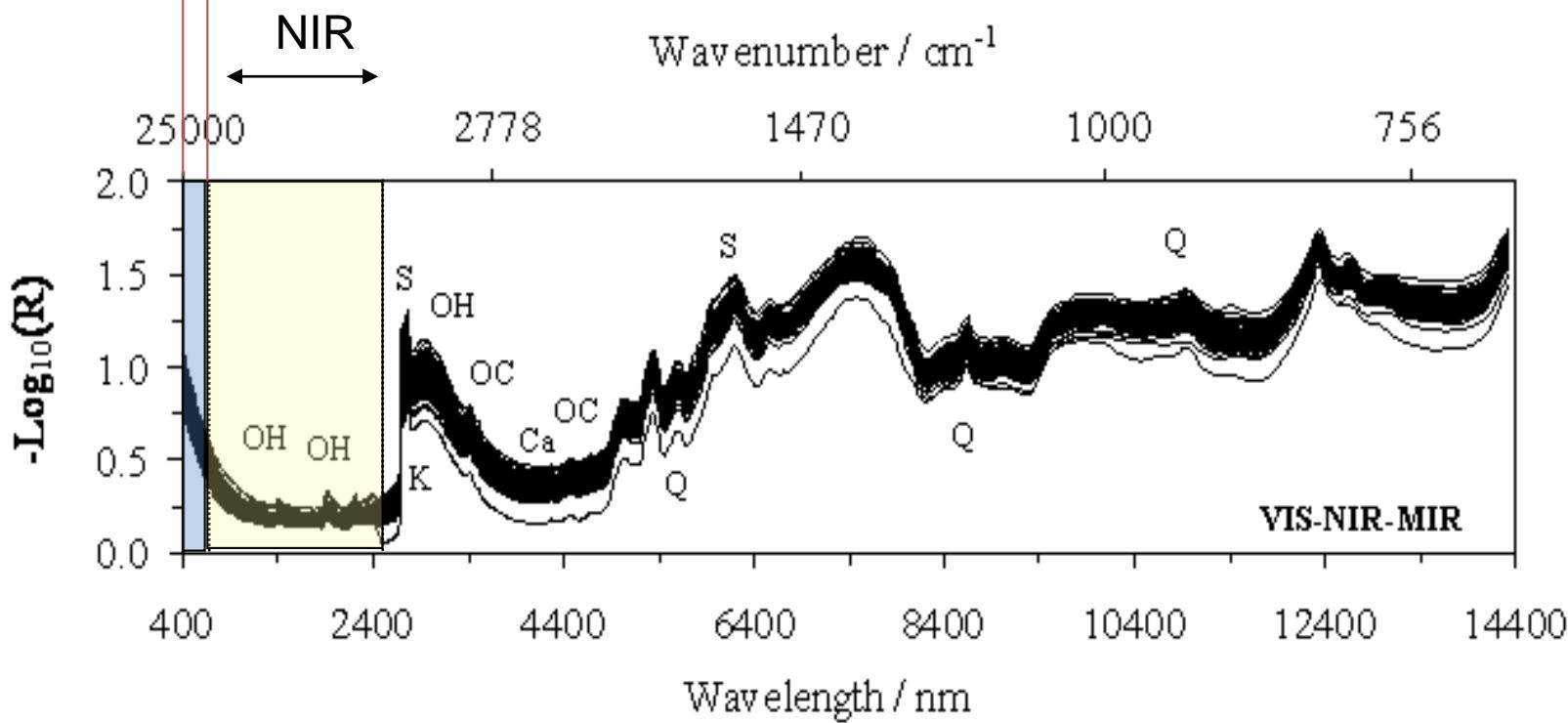
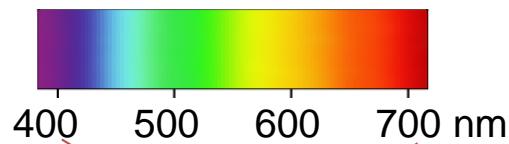
Electromagnetic Spectrum



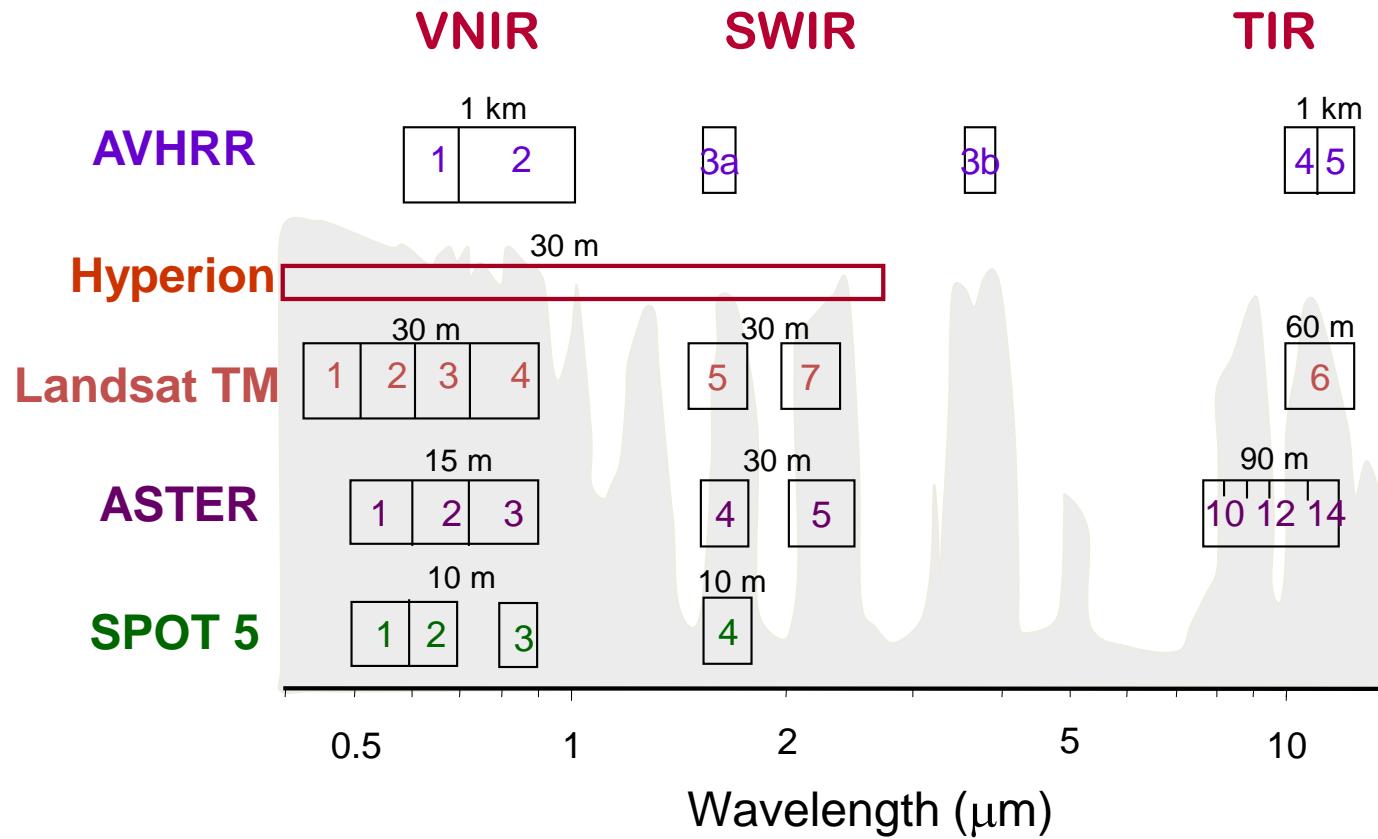


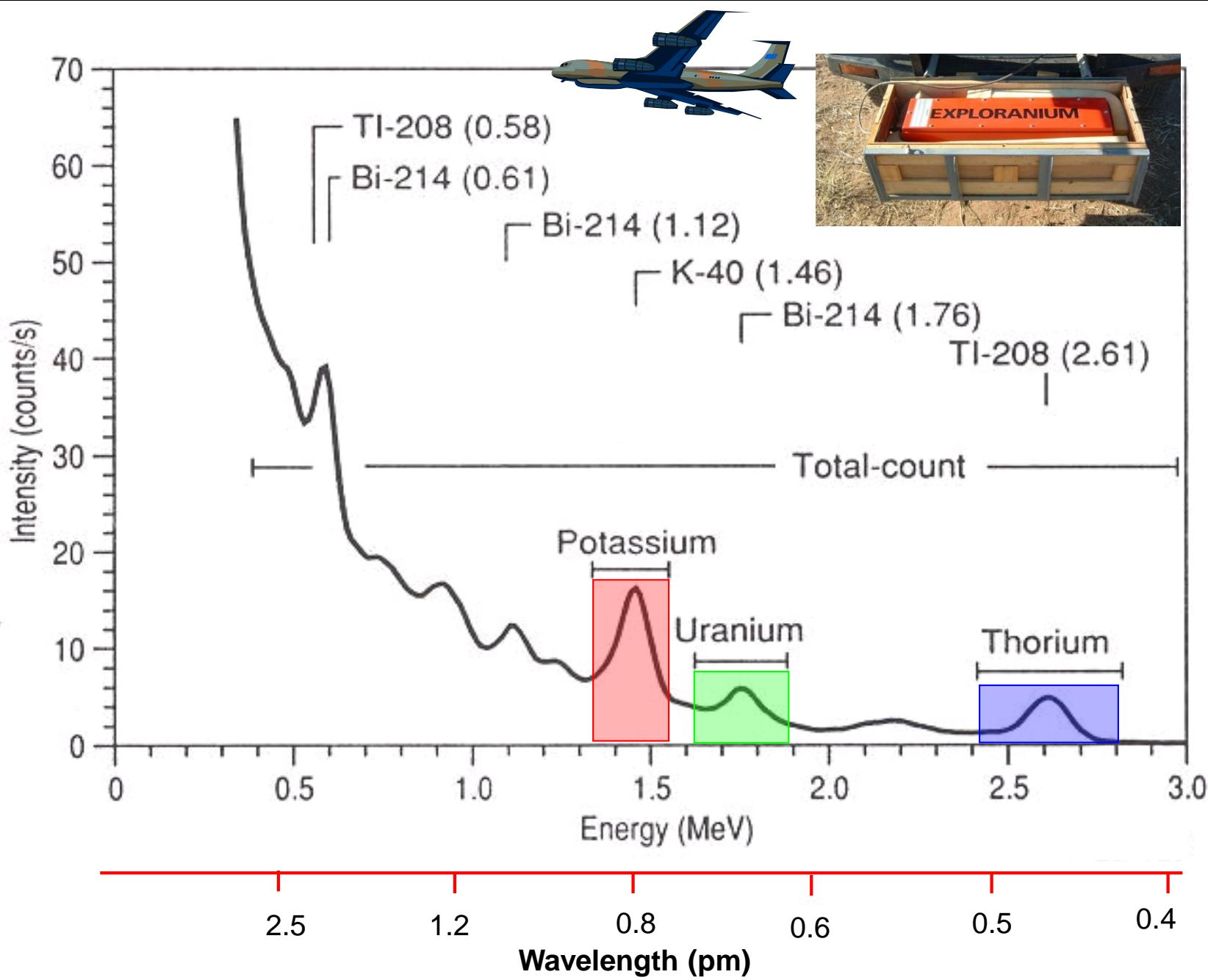


Vis, NIR, MIR



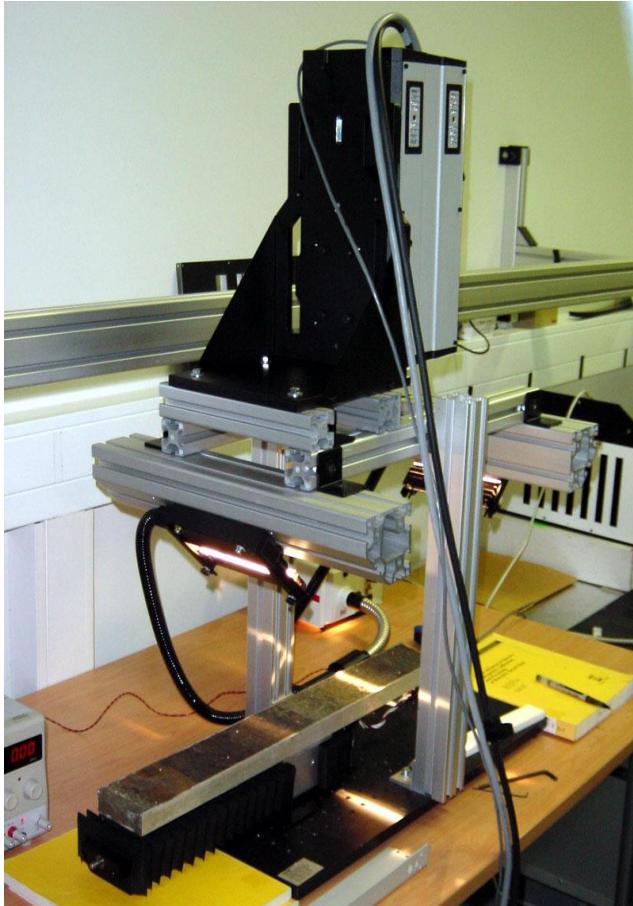
Remote Sensing





Infrared Spectroscopy





SPECTRAL CORE SCANNER

www.specim.fi

Spectral camera 200-2500 nm

Soil core is placed on a moving table in the Spectral Scanner system. The system scans spectral line images sequentially over the sample through movement of the table, providing full spectrum for each image pixel.

