Food and Agriculture
Organization of the
United Nations

GLOBAL SYMPOSIUM ON
**SOIL INFORMATION
AND DATA**

MEASURE
MONITOR
MANAGE

Advances In DSM For Global And Continental Applications: Innovative Covariates, Model Applicability And Spatial Uncertainty Assessment

Laura Poggio - ISRIC

September 25-28, 2024
Nanjing, China

GLOBAL SOIL PARTNERSHIP

Advances In DSM For Global And Continental Applications: Innovative Covariates, Model Applicability And Spatial Uncertainty Assessment

Laura Poggio, Niels Batjes, Bas Kempen, Giulio Genova, Fenny van Egmond, David Rossiter, Rik van den Bosch

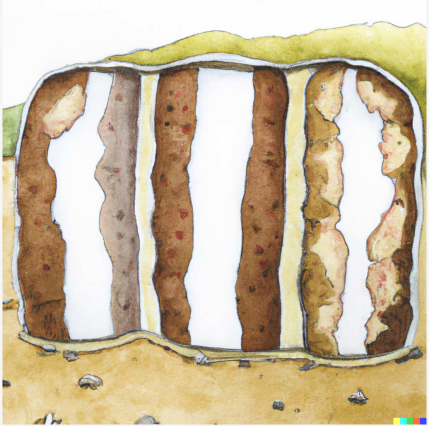
Outline

- **Models and their inputs**

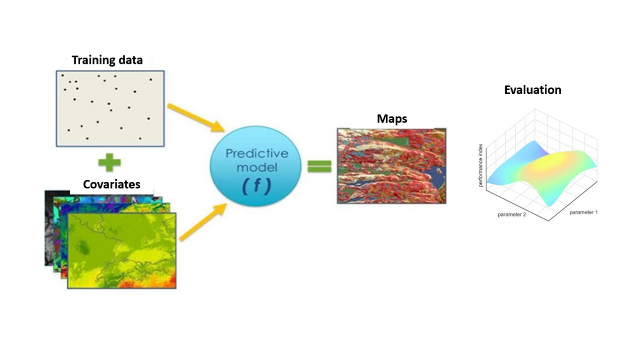
  Observations, covariates and models.
- **Maps and their evaluation**

  Maps and maps and some numbers.

  Uncertainty.
- **Concluding remarks**



main ingredients of DSM



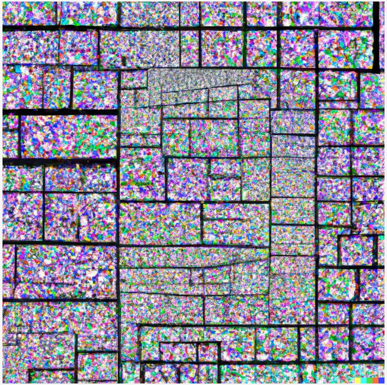
- Training data
- Covariates
- (Spatial) Predictive model
- Maps and their evaluation

what for?

The most useful map for the given application at the appropriate scale.

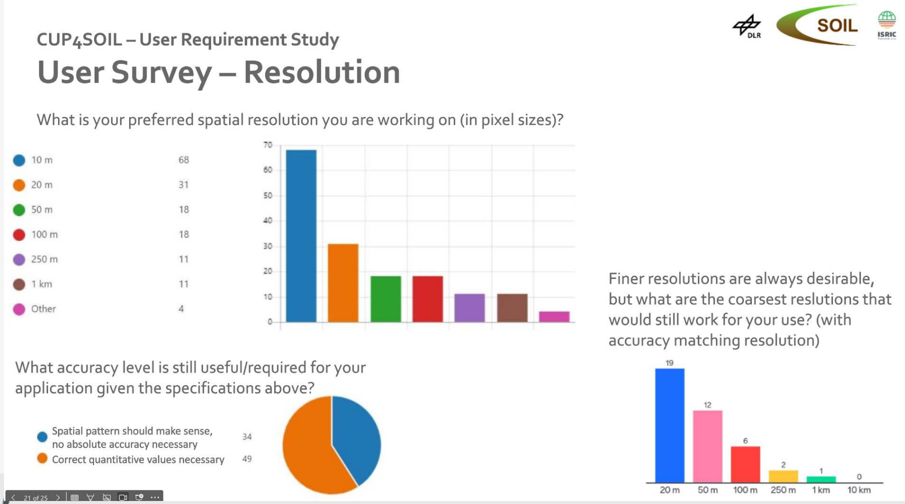
Not necessarily the most accurate.

DSM - critical evaluation?



- All training data available even if of dubious quality
- All covariates that can be handled
- Most complex model or ensembles of models
- Simple point wise evaluation

DSM - What users want? an example



CUP4SOIL – User Requirement Study

User Survey – Resolution

What is your preferred spatial resolution you are working on (in pixel sizes)?

● 10 m	68
● 20 m	31
● 50 m	18
● 100 m	18
● 250 m	11
● 1 km	11
● Other	4

What accuracy level is still useful/required for your application given the specifications above?

● Spatial pattern should make sense, no absolute accuracy necessary	34
● Correct quantitative values necessary	49

Finer resolutions are always desirable, but what are the coarsest reslutions that would still work for your use? (with accuracy matching resolution)

Models and their inputs

Observations, covariates and models.

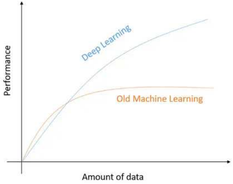
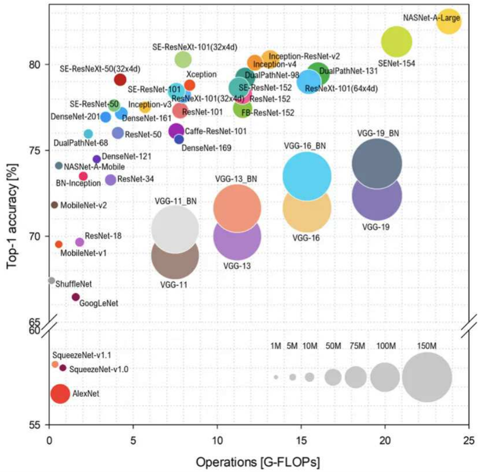


Figure 7. The performance of deep learning with respect to the amount of data.

Alom et al, Electronics, 2019



Bianco et al, IEEE Access, 2018

Everything but the kitchen sink?

KNOTS
Backpacking
YOU DO KNOW "EVERYTHING BUT THE KITCHEN SINK" IS AN EXPRESSION, NOT A CHALLENGE RIGHT?!
HEE, HEE, OH YEAH! WAIT, WHAT?
TREATS! TREATS! TREATS!
KAMP KNOTS
RLDiesslin © 2011

- Are all these nicely available covariates necessary?
- Are all these nicely available covariates *useful*?
- How many data points available?

Innovative covariates

Example France

SCMaP products

Mean Surface Reflectance

- Sentinel-2
- L2A reflectance (MAJA processed)
- 2018 – 2022

SOIL
Paris
Nancy

ID	mtry	ntree	MEC	RMSE
1	27	200	0.38	67.41
2	19	200	0.35	68.68
3	7	200	0.24	74.42
4	21	200	0.37	67.62
5	27	200	0.38	67.49
6	27	200	0.37	67.58
7	26	200	0.38	67.47

Maps and their evaluation

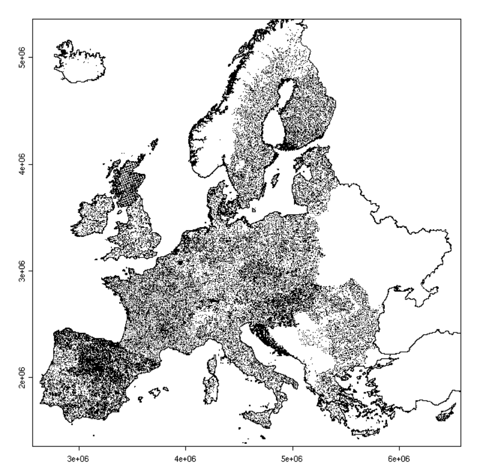
Maps and maps and some numbers.

Evaluating DSM products

- validation statistics, point-wise usually by cross-validation.
- So we know how well DSM can reproduce these known points.
- map users are looking for areas in the soil landscape.
- evaluate DSM products by their spatial patterns, i.e., how well they reproduce the soil landscape.

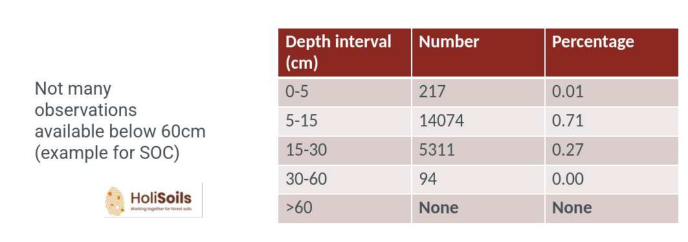
Europe as example of data rich continent

Europe as example of data rich continent



About 94K input observations for various properties : SOC,SIC,pH,nitrogen,sand,silt,clay, bulk density, coarse fragments

Europe as example of data rich continent



Not many observations available below 60cm (example for SOC)

HoliSoils

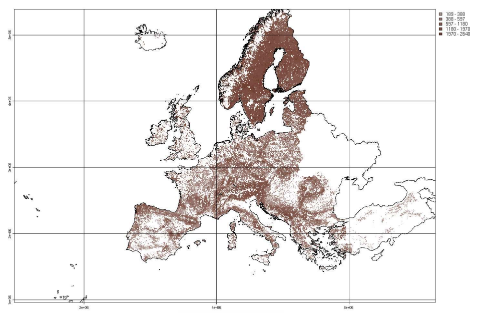
Working together for forest soils

Depth interval (cm)	Number	Percentage
0-5	217	0.01
5-15	14074	0.71
15-30	5311	0.27
30-60	94	0.00
>60	**None**	**None**

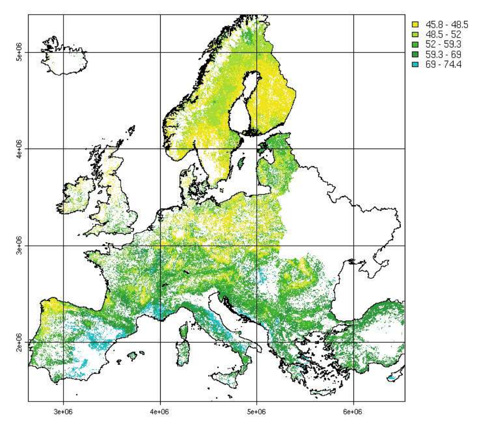
Depth distribution ...

Some maps

SOC

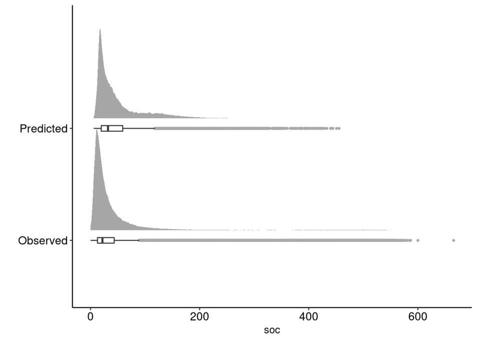


pH - H2O

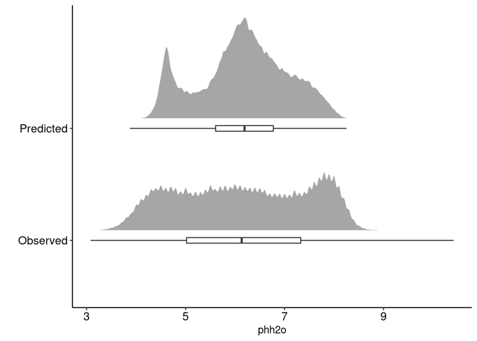


Some cross-validation

SOC

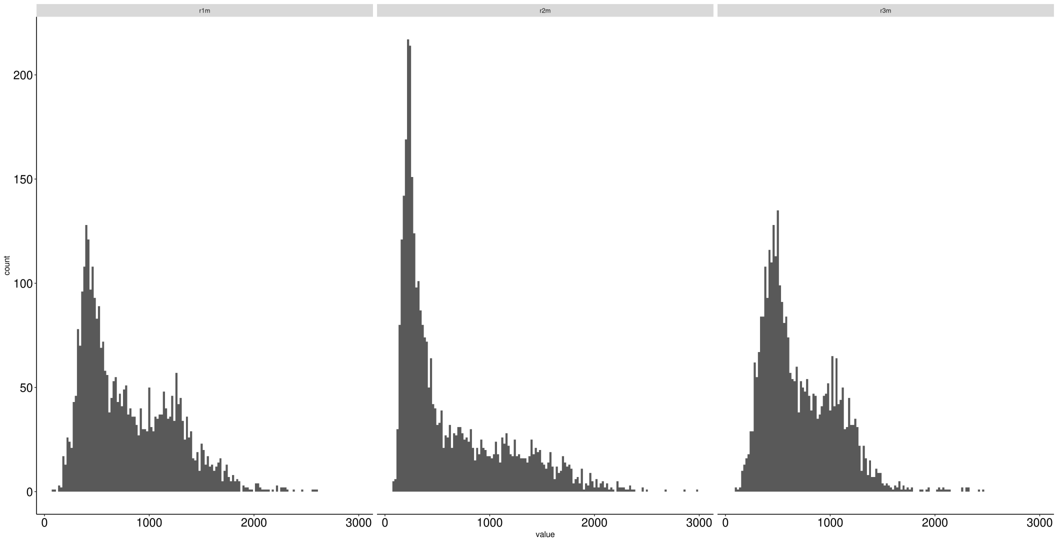


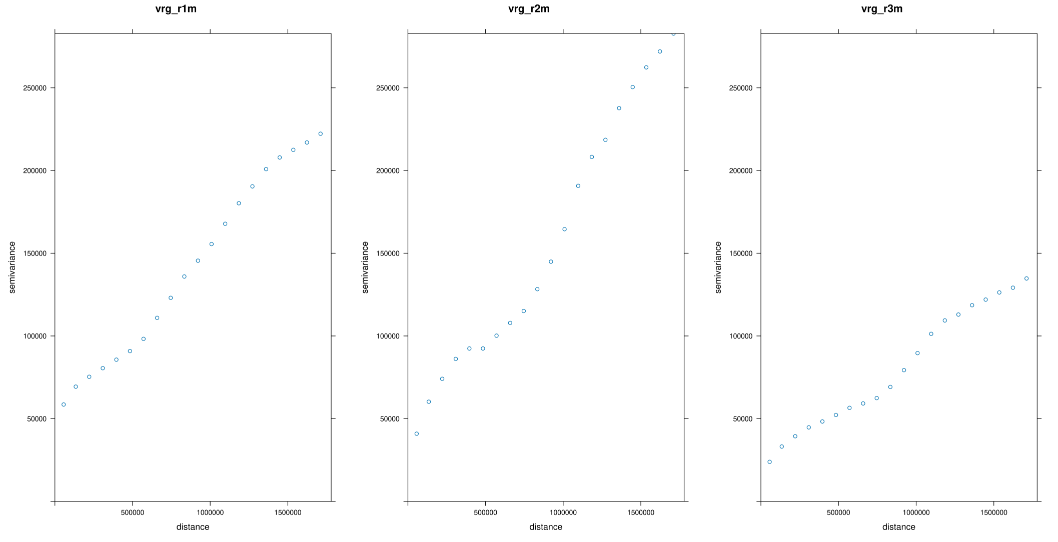
pH - H2O

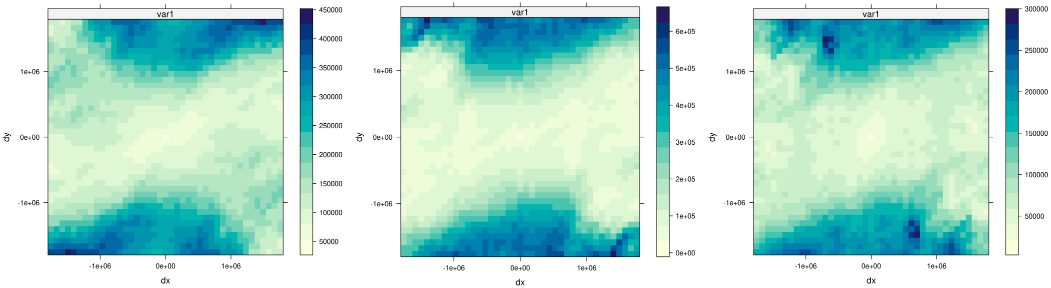


Some point accuracy

map	Resolution.m.	model	Covariates	rmse	mae	mec	ccc	mtry
r1	100	Europe	63	69.20	34.69	0.34	0.50	14
r2	100	Europe	193	73.64	34.84	0.28	0.45	28
r3	20	Europe	196	67.41	33.35	0.38	0.53	27


r1m
r2m
r3m
count
200
150
100
50
0
0
1000
2000
3000
value


vrg_r1m
vrg_r2m
vrg_r3m
semivariance
250000
200000
150000
100000
50000
distance
500000
1000000
1500000


var1
1e+06
0e+00
-1e+06
dy
-1e+06
0e+00
1e+06
dx
450000
400000
350000
300000
250000
200000
150000
100000
50000
var1
1e+06
0e+00
-1e+06
dy
-1e+06
0e+00
1e+06
dx
6e+05
5e+05
4e+05
3e+05
2e+05
1e+05
0e+00
var1
1e+06
0e+00
-1e+06
dy
-1e+06
0e+00
1e+06
dx
300000
250000
200000
150000
100000
50000

Evaluating DSM products

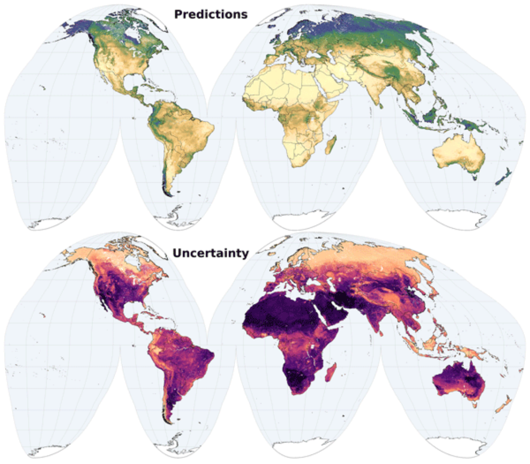
Which is the most suitable for the user case?

- the most accurate?
- the one with most reliable landscape link?
- the one created for the scale of the application?
- the most complex?

Maps and their evaluation Uncertainty.

Evaluating DSM products: uncertainty

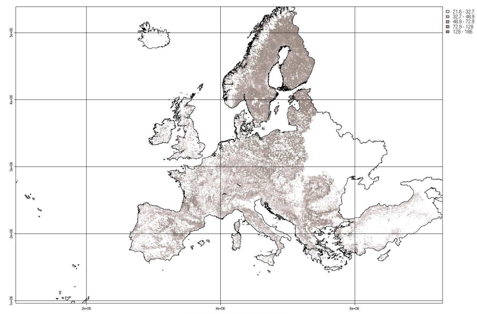
- Spatial uncertainty assessment is a key aspect of DSM products
- Most often represented with inter-quantile range $Q95 - Q05$ or ratio $(Q95 - Q05)/Q50$
- Many components in uncertainty.



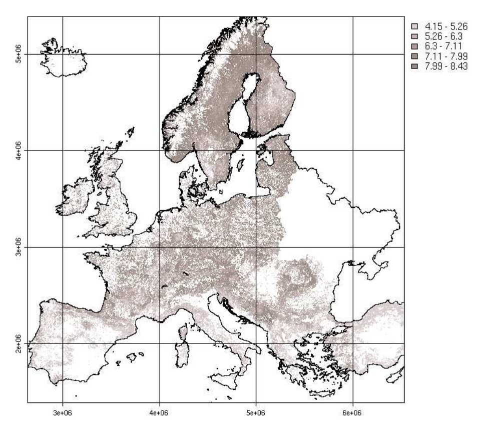
Poggio et al, 2021

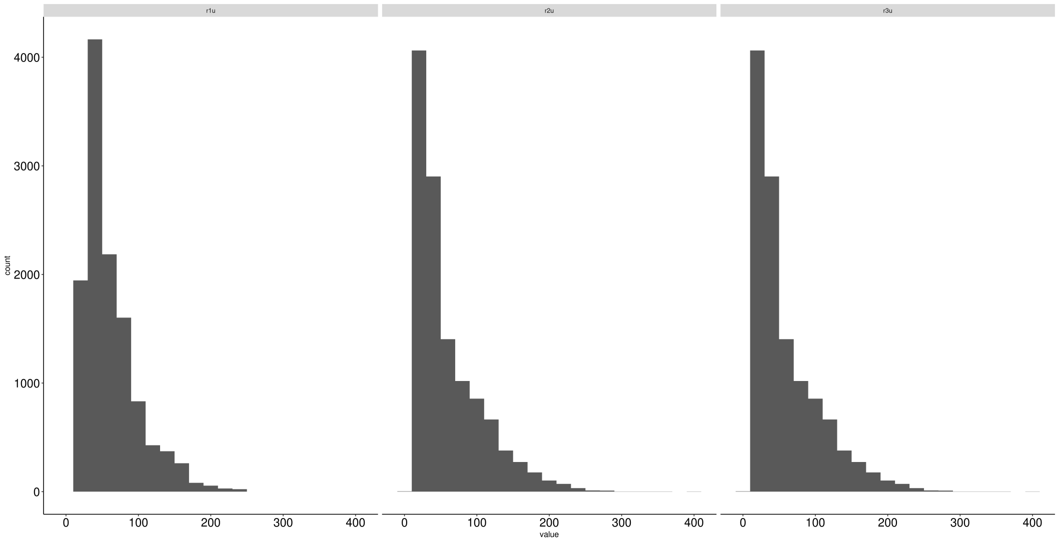
Some uncertainty maps

SOC



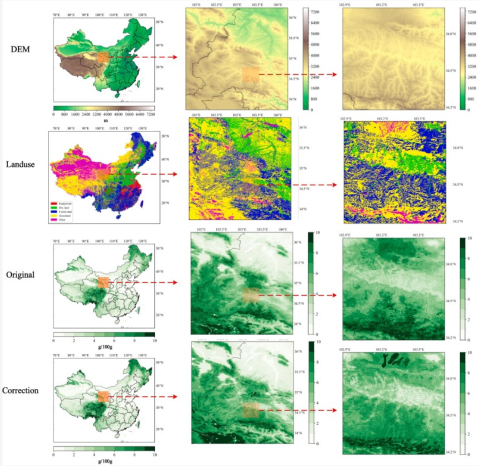
pH - H2O




r1u
r2u
r3u
count
4000
3000
2000
1000
0
0
100
200
300
400
value

Evaluating DSM products: uncertainty

- Positional accuracy
- Laboratory measurements
- Model
- Covariates
- ...



Shi et al, 2024

Evaluating DSM products: uncertainty

Area of Applicability

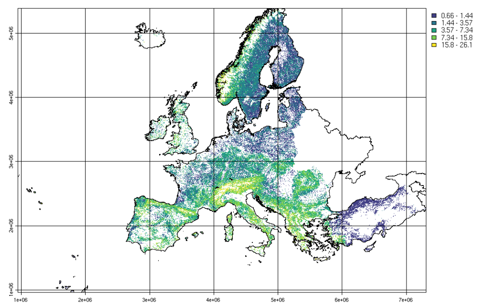
- Are the sampled locations representative of the covariates space?
- Do we have enough observations?
- Do we have too many covariates?

Area of applicability (AOA)

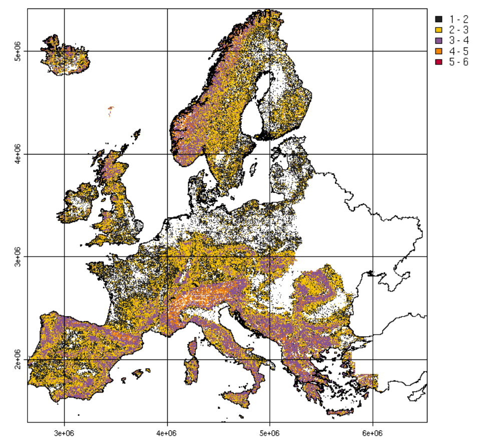
- *area of applicability* (AOA) of (spatial) prediction models (Meyer, Pebesma, 2021)
- The AOA is defined as the area where we enabled the model to learn about relationships based on the training data, and where the estimated cross-validation performance holds.
- A dissimilarity index (DI) is calculated based on distances to the training data in the multidimensional predictor variable space.
- variables are weighted by the model-derived importance scores prior to distance calculation.
- The AOA is then derived by applying a threshold based on the DI observed in the training data using cross-validation.

AOA/DI maps

DI example

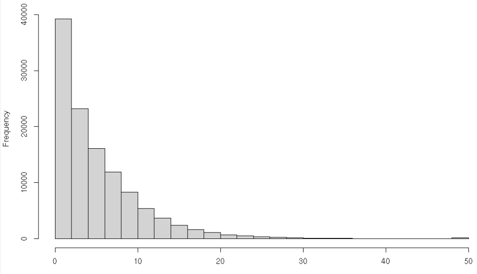


DI (reclassed percentiles)

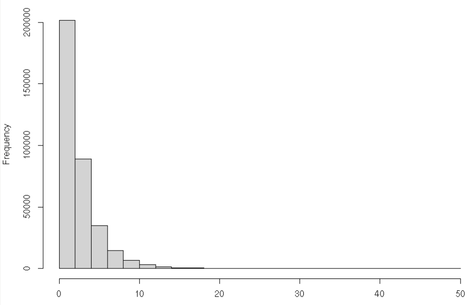


AOA/DI values

DI histogram (property one)



DI histogram (property two)

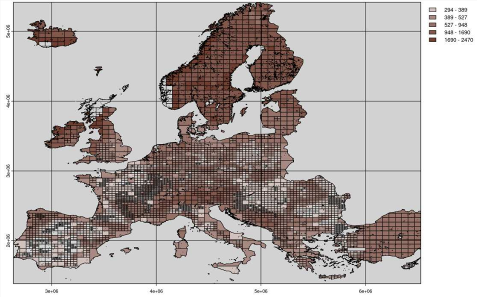


Quadmap

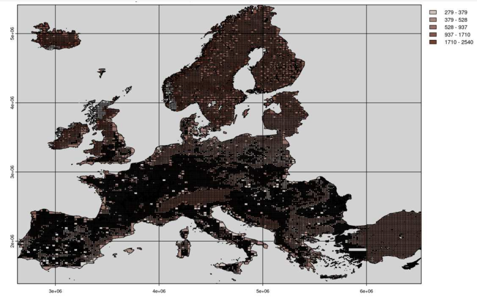
- QuadMap: Variable resolution maps to better represent spatial uncertainty (Padarian, McBratney, 2023)
- a method to create a single map with the uncertainty encoded as the pixel size.
- based on quadtree algorithm recursively partitioning the map into quadrants until the uncertainty criteria are fulfilled
- Using different uncertainty thresholds can yield dramatically different maps
- The selection of the final target will depend on the application

Quadmap

Low uncertainty threshold



High uncertainty threshold



Quadmap stats

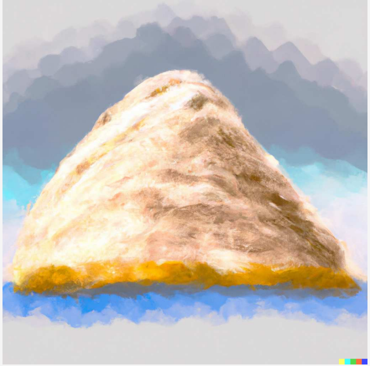
Percentage of superpixels of different size for three uncertainty thresholds

size	low	medium	high
1	17	24	45
<10	12	40	49
<100	25	32	6
>100	46	4	0

- size 1 correspond to the resolution of the map, size 10 to a superpixel encompassing 10 map pixels
- not many areas support a resolution of 100m

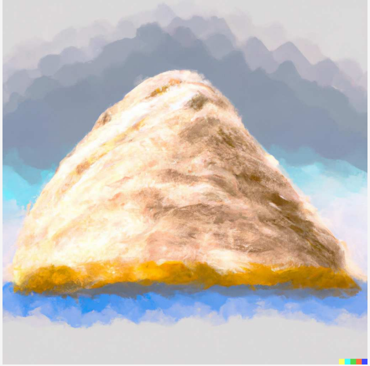
Concluding remarks.

Concluding remarks



- Every statistical model ever created in the history of the human race is subjective
- There are endless possibilities of how data and methods can be combined.
- Uncertainty and error propagation are a crucial issue.

Concluding remarks



- DSM is a crucial tool for integration of soil and land management domains
- Expert knowledge (domain, users, stakeholders) is fundamental to evaluate DSM products

Many thanks to colleagues at ISRIC - World Soil Information and around the World

ISRIC

World Soil Information

laura.poggio@isric.org

www.isric.org