Geomorphometry’2021, September 13–17, 2021, Perugia, Italy

Two case-based reasoning strategies of automatically selecting terrain covariates for geographical variable mapping

Cheng-Zhi QIN

State Key Laboratory of Resources and Environment Information System,
Institute of Geographic Sciences & Natural Resources Research, Chinese Academy of Sciences

2021-9-13

State Key Laboratory of Resources And Environmental Information System

Contents

1. Background

    - How to automatically select (terrain) covariates for building geographical variable-environment relationship?

2. Basic idea: Case-based reasoning

3. Two case-based reasoning strategies

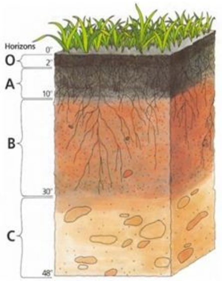
4. Experiment

5. Conclusion

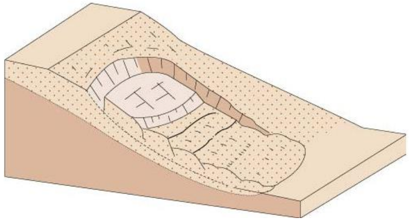
1. Background

- **Geographical variables:**

soil property



landslide susceptibility

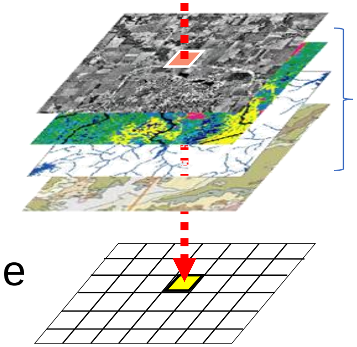


species habitat suitability

• • •

- **Geographical variable mapping** (GVM) through building geographical variable-environment relationship is widely used to obtain the spatial distribution information (often as a grid) of those geographical variables which are hard to acquire through direct observation (e.g., remote sensing).

Geographical variable = *f* (Covariates)



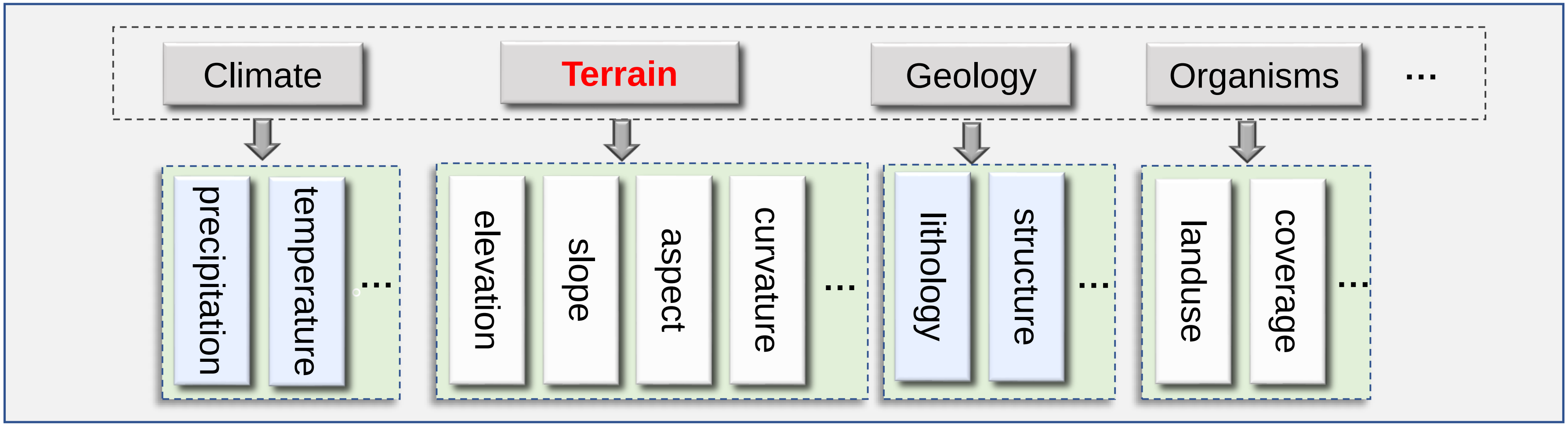
How to select proper covariates?
- a critical step (and hard for non-experts)

Zhu A-X, Lu G, Liu J, Qin C-Z, Zhou C. Spatial prediction based on Third Law of Geography. Annals of GIS, 2018, 24(4): 225-240.

Background

- Large number of potential (terrain) covariates for geographical variable mapping

(Ziadat, 2005; Zhu et al., 2010, Liu et al., 2013; Wiesmeier et al., 2014; Lecours et al., 2017)



Climate
Terrain
Geology
Organisms
…

precipitation
temperature
…

elevation
slope
aspect
curvature
…

lithology
structure
…

landuse
coverage
…

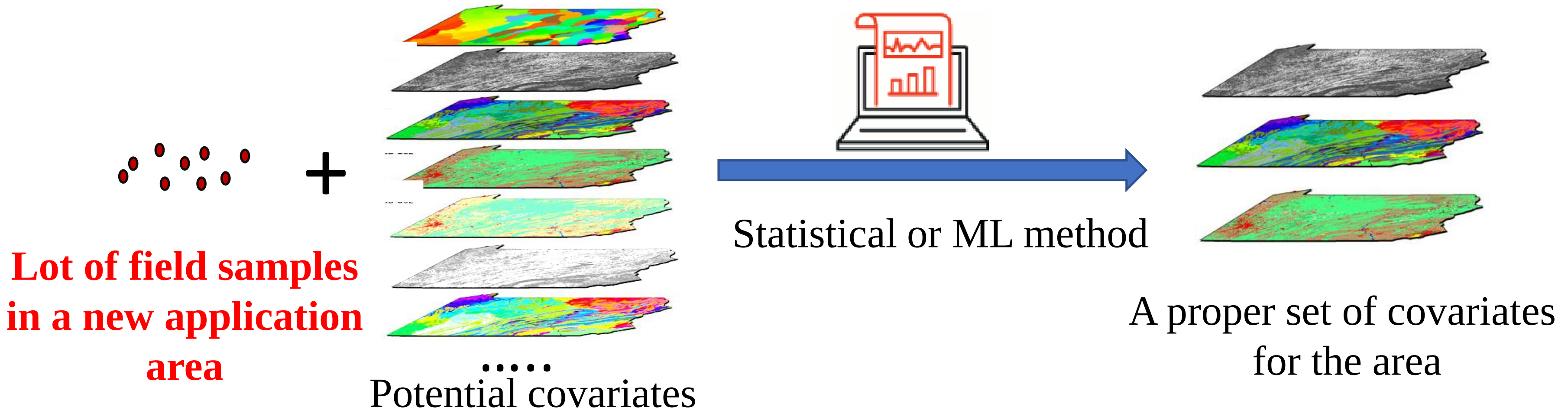
- Many tools exist for calculating covariates

ArcGIS
Grass
SAGA
LandSerf
TauDEM
Whitebox Geospatial Analysis Tools
Whitebox
…

However, lack clear guidance on which condition each potential covariate should be used in specific application contexts (target variable, study area characteristics, data availability, etc.) !

Existing methods of aiding users to select covariates for GVM

- By explicit, general rules (Lecours et al., 2017; Deng et al., 2007)
    - The related knowledge in many application domains are hard to form such explicit rules.
- Statistical (or machine learning) methods of selecting covariates
    - Filter: Pearson's correlation analysis (Lagacherie et al., 2013), moment correlation analysis (de Carvalho Junior et al., 2014), …
    - Wrapper: stepwise regression procedure (Zhu et al., 2015), recursive feature elimination (Shi et al., 2018)
    - Embedding: decision trees (Greve et al., 2012), cubist (Adhikari et al., 2014), random forests (Vaysse and Lagacherie, 2015)



Lot of field samples in a new application area

+

.....

Potential covariates

Statistical or ML method

A proper set of covariates for the area

When there exist few samples, statistical/ML methods often fail !

How to automatically select covariates when there are few samples ?

- Facts
  - Lots of practical applications conducted by domain experts have been published.

European Journal of Soil Science

Predictive soil mapping w

Pedology

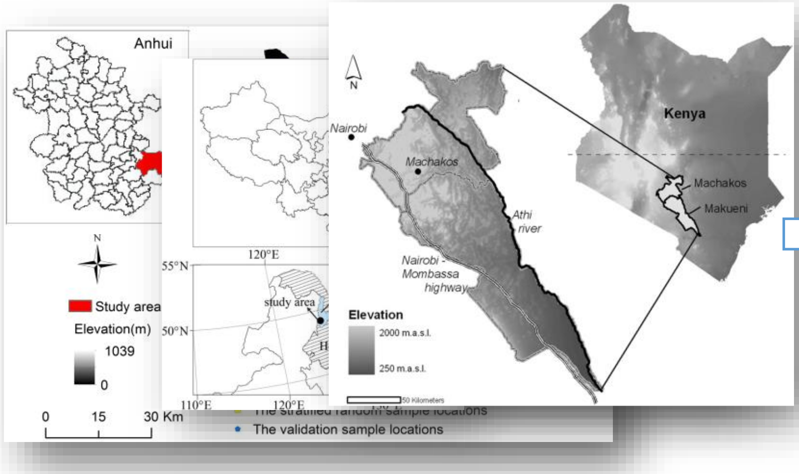
Evaluation of Integrative Hierarchical Stepwise Sampling for

Catena

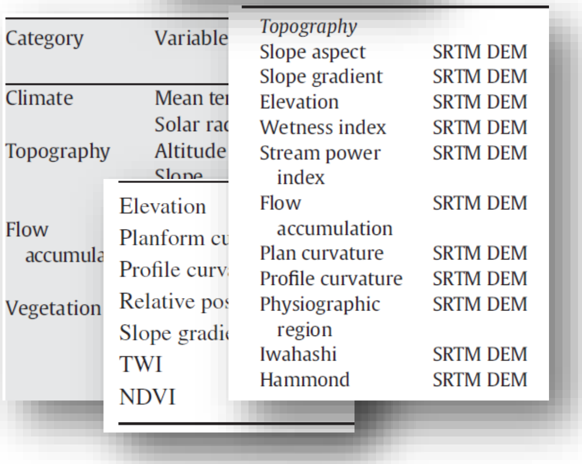
Small scale digital soil mapping in Southeastern Kenya

Geoderma

Construction of membership functions for predictive soil mapping under fuzzy logic



Domain experts



Category	Variable
Climate	Mean te Solar rac
Topography	Altitude Slope
Flow accumula	
Vegetation	

Elevation
Planform cu
Profile curv
Relative pos
Slope gradi
TWI
NDVI

Topography	
Slope aspect	SRTM DEM
Slope gradient	SRTM DEM
Elevation	SRTM DEM
Wetness index	SRTM DEM
Stream power index	SRTM DEM
Flow accumulation	SRTM DEM
Plan curvature	SRTM DEM
Profile curvature	SRTM DEM
Physiographic region	SRTM DEM
Iwahashi	SRTM DEM
Hammond	SRTM DEM

- Expert knowledge on selecting covariates (under specific application contexts) were implicitly contained in existing applications of geographic variable mapping.

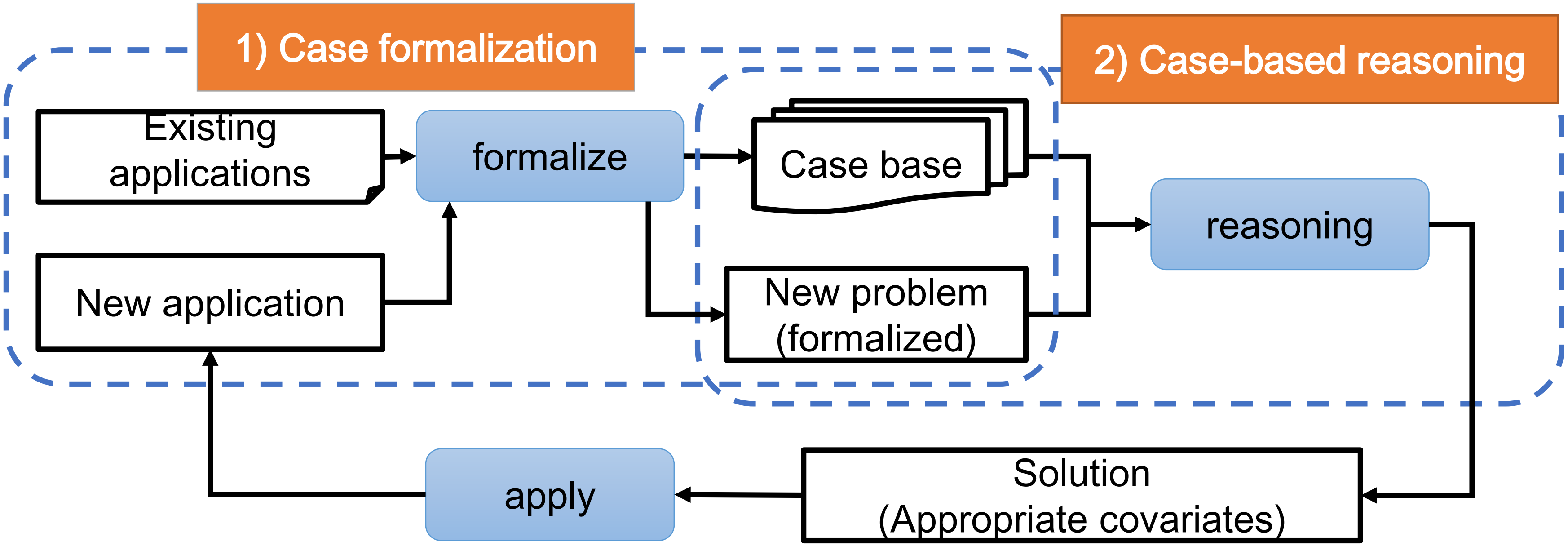
Research issue

How to use these implicit knowledge on selecting proper (terrain) covariates, which are contained in existing applications of geographical variable mapping?

2. Basic idea

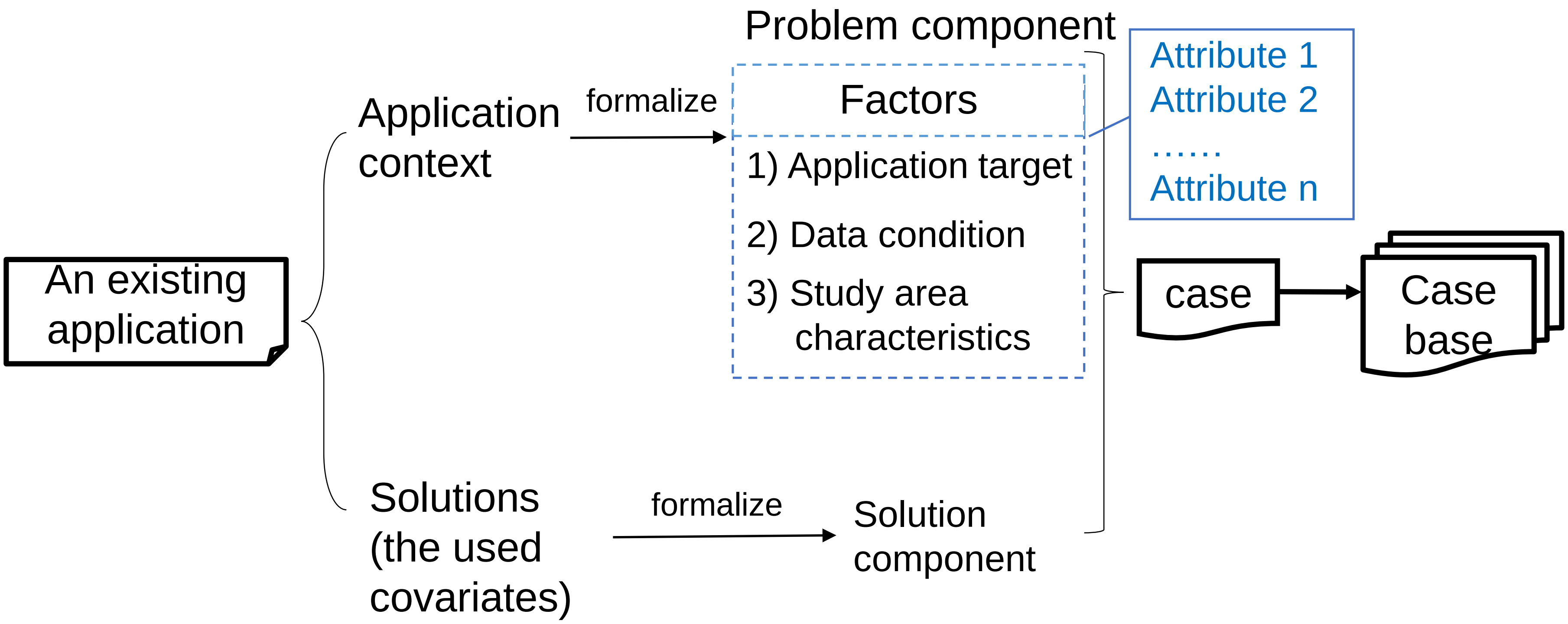
- **Cases:** a suitable way to formalize prior, non-systematic knowledge in the artificial intelligence domain (Kaster et al., 2005) :
  - **Problem component** -- describe application context information (Qin et al., 2016)
  - **Solution component**

- **Case-based reasoning:** find the existing case(s) which is/are similar to a new application, and then apply the solutions of the similar cases to the new application.



Case formalization

- **Problem component of cases**
  - Factors: describe the application context information
  - Attributes: quantify the factors, which can be directly used in case-based reasoning

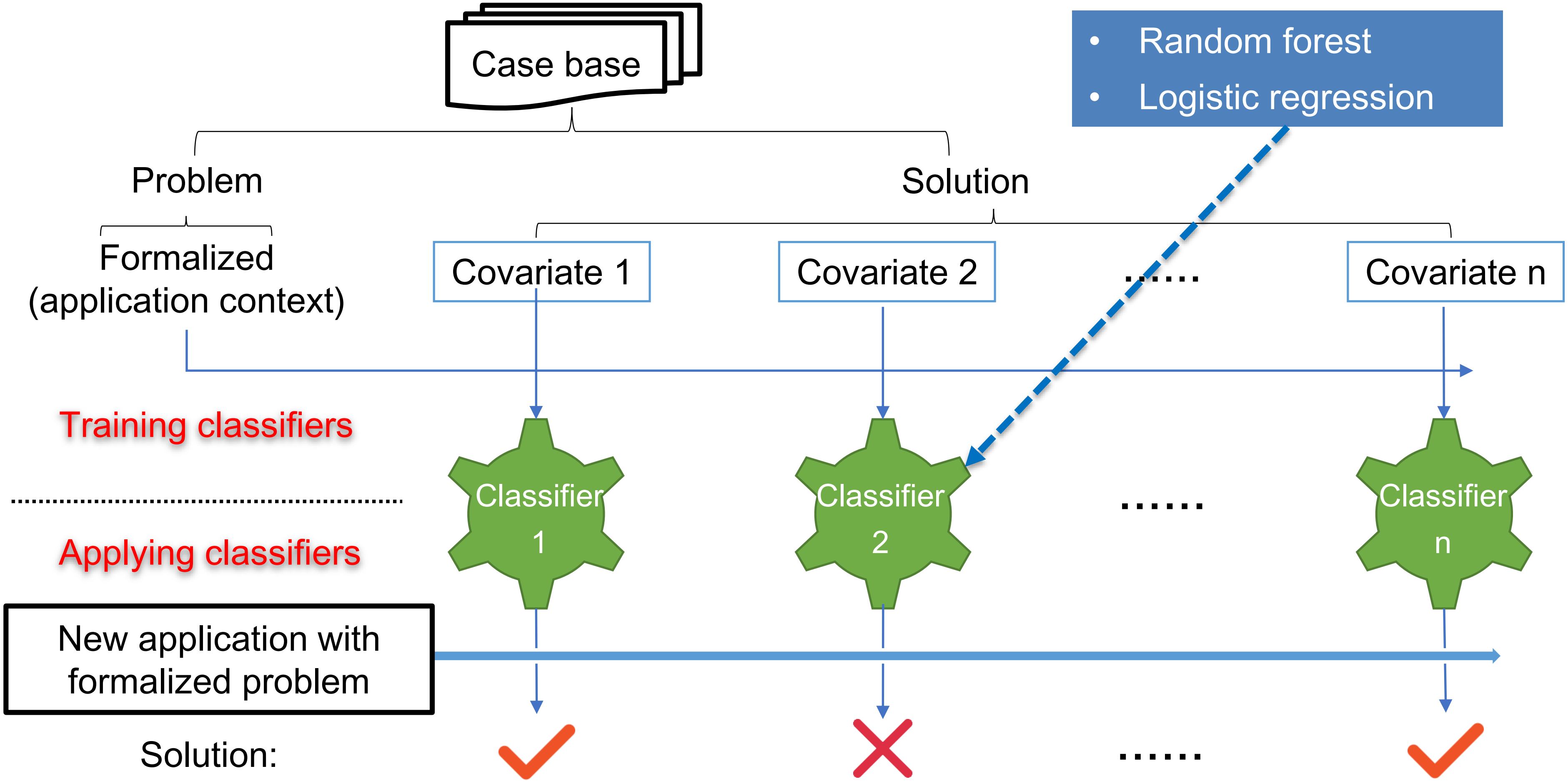


Qin C-Z, Wu X-W, Jiang J-C, Zhu A-X. Case-based knowledge formalization and reasoning method for digital terrain analysis - application to extracting drainage networks. *Hydrology and Earth System Sciences*, 2016, 20: 3379-3392.

3. Two case-based reasoning strategies for selecting proper covariates

1) The covariate-level binary classification strategy (or, the classification strategy)

For each covariate included in the case base: A binary classification problem



Case base

- Random forest
- Logistic regression

Problem

Solution

Formalized (application context)

Covariate 1

Covariate 2

……

Covariate n

Training classifiers

Applying classifiers

Classifier 1

Classifier 2

……

Classifier n

New application with formalized problem

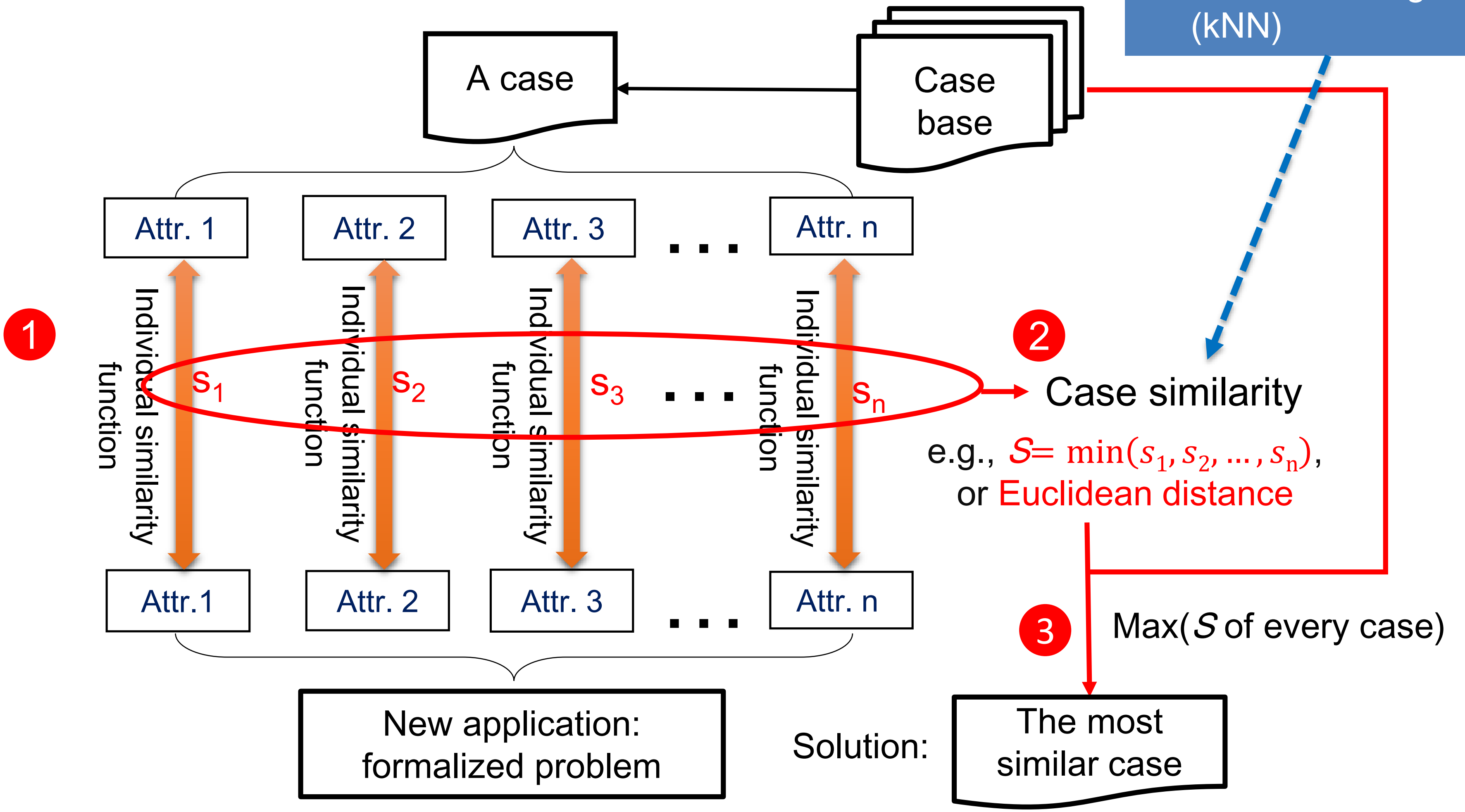
Solution: ✓ ✗ …… ✓

Liang P, Qin C-Z*, Zhu A-X, Hou Z-W, Fan N-Q, Wang Y-J. A case-based method of selecting covariates for digital soil mapping. *Journal of Integrative Agriculture*, 2020, 19(8): 2127-2136.

Two case-based reasoning strategies for selecting proper covariates

2) The most-similar-case strategy

- Minimum operator
- *k*-Nearest Neighbors (kNN)



e.g., $S = \min(s_1, s_2, \dots, s_n)$, or Euclidean distance

Liang P, Qin C-Z*, Zhu A-X, Zhu T-X, Hou Z-W, Fan N-Q, Wang Y-J. Using the most similar case method to automatically select environmental covariates for predictive mapping. *Earth Science Informatics*, 2020, 13(1): 39-53.

4. Experiments

- Classification strategy

+ different classifiers

  - Random forest (RF) method
  - Logistic regression (LG) method

vs.

- Most-similar-case strategy

+ different case similarity calculation

  - Minimum operator (MO) method
  - k-Nearest Neighbors (kNN) method

- Experiments: Taking digital soil mapping (DSM) as example
  - Terrain covariates have been predominantly used in DSM for building soil-environment relationship (McBratney et al., 2003)
  - When selecting terrain covariates, user needs to consider little beyond the study area characteristics (e.g., data availability)

1) Case formalization – e.g., digital soil mapping (DSM)

Case component	Case formalization		
	Factor group	Factor	Attribute
Problem (application context)	Mapping target	Mapping soil property	Soil property
		Mapping soil layer	Top (cm)
			Bottom (cm)
		Mapping resolution	Resolution (m)
	Mapping area characteristics	Mapping area size	Area size (km^2)
		Terrain condition	Total relief (m)
			SD(elev.) (m)
			Mean slope (°)
Solution	Terrain covariates used		

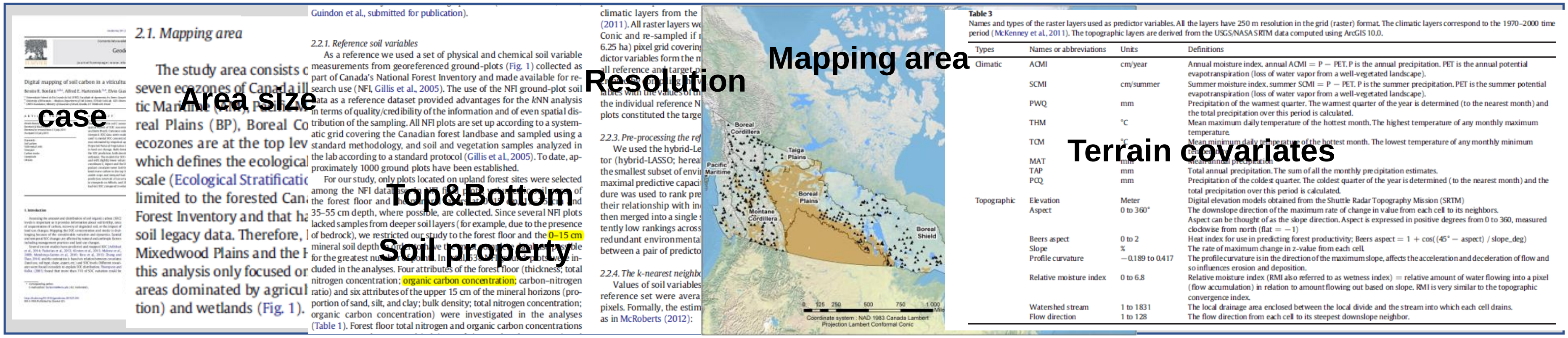
Liang P, Qin C-Z, Zhu A-X, Hou Z-W, Fan N-Q, Wang Y-J. A case-based method of selecting covariates for digital soil mapping. *Journal of Integrative Agriculture*, 2020, 19(8): 2127-2136.

2) Case base preparation

- Case formalization

Problem								Solution
Mapping target description				Mapping area characteristics				Terrain covariates
Soil property	Top	Bottom	Resolution	Area size	Total relief	SD(elev.)	Mean slope	

- Extract values for each attribute of a case


case
Area size
Resolution
Mapping area
Top&Bottom
Soil property
Terrain covariates
Boreal Cordillera
Taiga Plains
Pacific Maritime
Boreal Plains
Montane Cordillera
Boreal Shield
0 125 250 500 750 1 000 Miles
Coordinate system : NAD 1983 Canada Lambert
Projection Lambert Conformal Conic

2.1. Mapping area

The study area consists o
seven ecozones of Canada ill
tic Mar
real Plains (BP), Boreal Co
ecozones are at the top lev
which defines the ecological
scale (Ecological Stratificatio
limited to the forested Cana
Forest Inventory and that ha
soil legacy data. Therefore, l
Mixedwood Plains and the H
this analysis only focused on
areas dominated by agricul
tion) and wetlands (Fig. 1).

Guindon et al., submitted for publication).

2.2.1. Reference soil variables

As a reference we used a set of physical and chemical soil variable measurements from georeferenced ground-plots (Fig. 1) collected as part of Canada's National Forest Inventory and made available for research use (NFI, Gillis et al., 2005). The use of the NFI ground-plot soil data as a reference dataset provided advantages for the *k*NN analysis in terms of quality/credibility of the information and of even spatial distribution of the sampling. All NFI plots are set up according to a systematic grid covering the Canadian forest landbase and sampled using a standard methodology, and soil and vegetation samples analyzed in the lab according to a standard protocol (Gillis et al., 2005). To date, approximately 1000 ground plots have been established.

For our study, only plots located on upland forest sites were selected among the NFI data the forest floor and 35–55 cm depth, where possible, are collected. Since several NFI plots lacked samples from deeper soil layers (for example, due to the presence of bedrock), we restricted our study to the forest floor and the 0–15 cm mineral soil depth for the greatest nu cluded in the analyses. Four attributes of the forest floor (thickness; total nitrogen concentration; organic carbon concentration; carbon–nitrogen ratio) and six attributes of the upper 15 cm of the mineral horizons (proportion of sand, silt, and clay; bulk density; total nitrogen concentration; organic carbon concentration) were investigated in the analyses (Table 1). Forest floor total nitrogen and organic carbon concentrations

climatic layers from the
(2011). All raster layers we
Conic and re-sampled if r
6.25 ha) pixel grid covering
dictor variables form the m
all reference and target p
the individual reference N
plots constituted the targe

2.2.3. Pre-processing the ref

We used the hybrid-Le
tor (hybrid-LASSO; herea
the smallest subset of envi
maximal predictive capaci
dure was used to rank pre
their relationship with in
then merged into a single s
tently low rankings across
redundant environmental
between a pair of predicto

2.2.4. The k-nearest neighbo

Values of soil variables
reference set were avera
pixels. Formally, the estim
as in McRoberts (2012):

Table 3
Names and types of the raster layers used as predictor variables. All the layers have 250 m resolution in the grid (raster) format. The climatic layers correspond to the 1970–2000 time period (McKenney et al., 2011). The topographic layers are derived from the USGS/NASA SRTM data computed using ArcGIS 10.0.

Types	Names or abbreviations	Units	Definitions
Climatic	ACMI	cm/year	Annual moisture index. annual ACMI = P − PET. P is the annual precipitation. PET is the annual potential evapotranspiration (loss of water vapor from a well-vegetated landscape).
	SCMI	cm/summer	Summer moisture index. summer SCMI = P − PET. P is the summer precipitation. PET is the summer potential evapotranspiration (loss of water vapor from a well-vegetated landscape).
	PWQ	mm	Precipitation of the warmest quarter. The warmest quarter of the year is determined (to the nearest month) and the total precipitation over this period is calculated.
	THM	°C	Mean maximum daily temperature of the hottest month. The highest temperature of any monthly maximum temperature.
	TCM	°C	Mean minimum daily temperature of the hottest month. The lowest temperature of any monthly minimum temperature
	MAT	mm	Mean annual precipitation
	TAP	mm	Total annual precipitation. The sum of all the monthly precipitation estimates.
	PCQ	mm	Precipitation of the coldest quarter. The coldest quarter of the year is determined (to the nearest month) and the total precipitation over this period is calculated.
Topographic	Elevation	Meter	Digital elevation models obtained from the Shuttle Radar Topography Mission (SRTM)
	Aspect	0 to 360°	The downslope direction of the maximum rate of change in value from each cell to its neighbors. Aspect can be thought of as the slope direction. Aspect is expressed in positive degrees from 0 to 360, measured clockwise from north (flat = −1)
	Beers aspect	0 to 2	Heat index for use in predicting forest productivity; Beers aspect = 1 + cos((45° − aspect) / slope_deg)
	Slope	%	The rate of maximum change in z-value from each cell.
	Profile curvature	−0.189 to 0.417	The profile curvature is in the direction of the maximum slope, affects the acceleration and deceleration of flow and so influences erosion and deposition.
	Relative moisture index	0 to 6.8	Relative moisture index (RMI also referred to as wetness index) = relative amount of water flowing into a pixel (flow accumulation) in relation to amount flowing out based on slope. RMI is very similar to the topographic convergence index.
	Watershed stream	1 to 1831	The local drainage area enclosed between the local divide and the stream into which each cell drains.
	Flow direction	1 to 128	The flow direction from each cell to its steepest downslope neighbor.

Google Earth Engine

Mapping area characteristics

- Relief
- SD(elevation)
- Mean slope

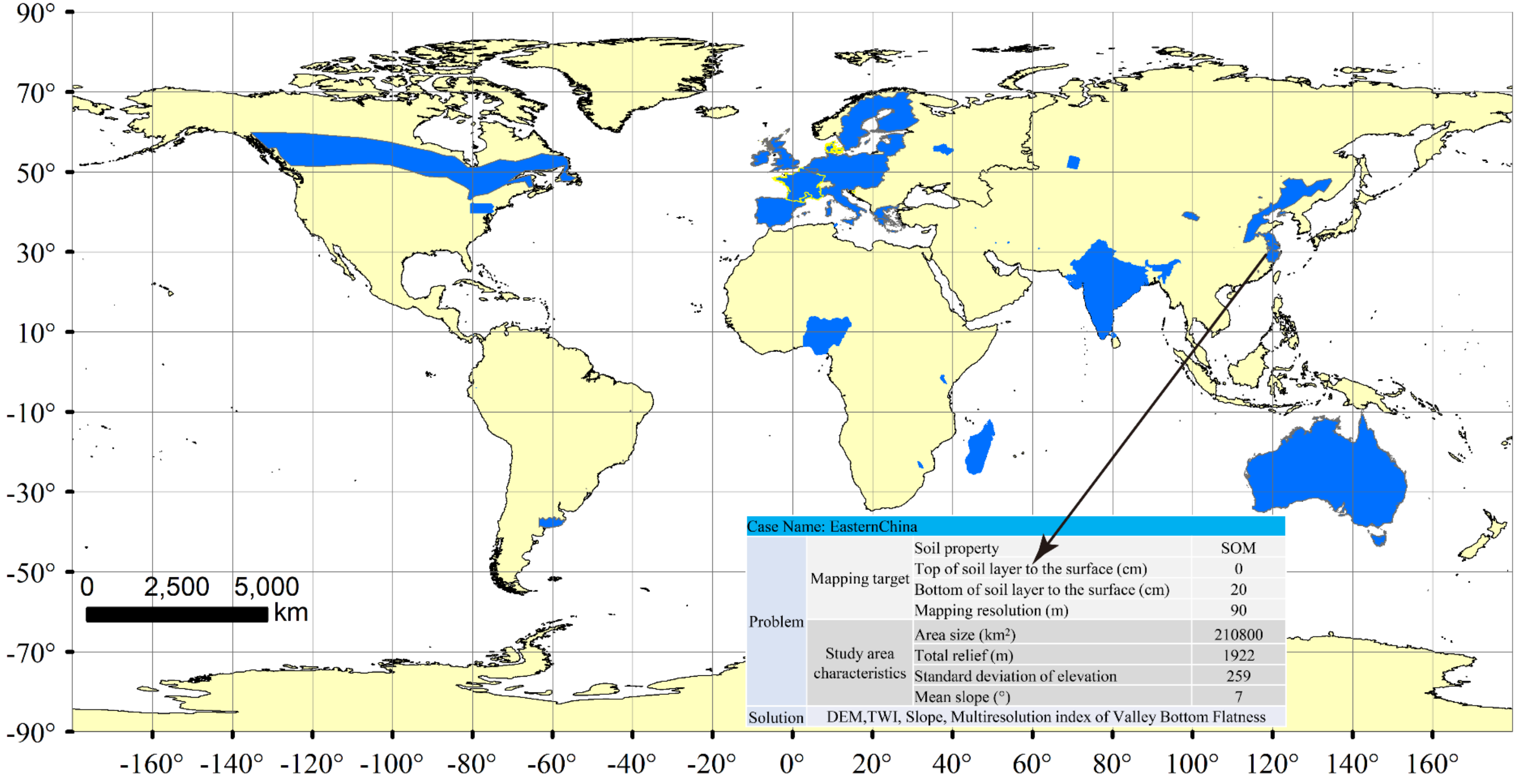
Covariates merging for

- Same covariate with different names
- Different covariates which have highly consistent effects from the perspective of DSM

Liang P, Qin C-Z*, Zhu A-X, Hou Z-W, Fan N-Q, Wang Y-J. A case-based method of selecting covariates for digital soil mapping. *Journal of Integrative Agriculture*, 2020, 19(8): 2127-2136.

Case base

- 191 cases collected from 56 papers in DSM-related journals (*Geoderma*, *European Journal of Soil Science*, *Soil Science Society of America Journal*, *Catena*, *Geoderma Regional*, *Plant and Soil*, *Science of the Total Environment*, *Ecological Indicators*, *Environmental Monitoring and Assessment*, *GIScience & Remote Sensing*, and *PLOS ONE*)
- A total of 38 terrain covariates used



Case Name: EasternChina			
Problem	Mapping target	Soil property	SOM
		Top of soil layer to the surface (cm)	0
		Bottom of soil layer to the surface (cm)	20
		Mapping resolution (m)	90
	Study area characteristics	Area size (km^2)	210800
		Total relief (m)	1922
		Standard deviation of elevation	259
		Mean slope (°)	7
Solution	DEM,TWI, Slope, Multiresolution index of Valley Bottom Flatness		

3) Cross-validation: a leave-one-out experiment

- Leave-one-out experiment:
  - 190 cases as the training set, the remaining 1 case as the new coming application.
  - Repeated 191 times

- Evaluation: How consistent between the covariates selected by each method and those originally used in the cases?

$$Recall = \frac{TP}{TP+FN} \qquad Precision = \frac{TP}{TP+FP} \qquad F1_{-score} = \frac{2 * (Precision * Recall)}{Precision + Recall}$$

(*TP*: True Positives; *FN*: False Negatives; *FP*: False Positives)

- ➢ *Recall* index: the ratio of covariates correctly selected by from a method to all covariates used in the original solution of the evaluation case.
- ➢ *Precision* index: the ratio of covariates correctly selected by a method to all covariates recommended by the method**.**
- ➢ *F1-score* index: The harmonic average of *Precision* and *Recall*

The larger of evaluation indices, the better performance of the proposed method

A reference method: “Novice”

- Novice method: pick those most often-used covariates (without considering the application context)
  - **Assumption:** the more frequently a covariate is used in the case base, the more popular that covariate is in the DSM domain.
  - **Preprocessing**: Sort the covariates according to the using frequency of each covariate used in the case base
  - **Usage**: Select the most frequently used covariates in the case base, according to the number of covariates used in the original solution of the validation case.

Liang P, Qin C-Z*, Zhu A-X, Hou Z-W, Fan N-Q, Wang Y-J. A case-based method of selecting covariates for digital soil mapping. *Journal of Integrative Agriculture*, 2020, 19(8): 2127-2136.

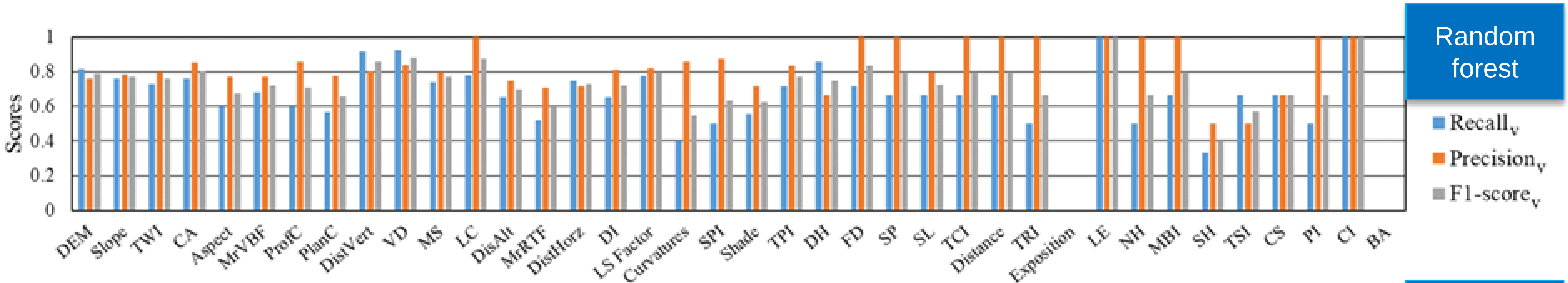
4) Experimental results and discussion

Strategy	Method	Evaluation index	Mean	Median	Max	Min	Std.
Covariate-level binary classification	Random forest	Recall	**0.644**	0.667	1	0	0.38
		Precision	**0.704**	1	1	0	0.391
		F1-score	**0.624**	0.667	1	0	0.362
	Logistic regression	Recall	0.414	0.333	1	0	0.350
		Precision	0.546	0.6	1	0	0.407
		F1-score	0.332	0.4	1	0	0.275
Most-similar-case	Minimum Operator	Recall	0.587	0.6	1	0	0.396
		Precision	0.589	0.6	1	0	0.396
		F1-score	0.552	0.571	1	0	0.372
	kNN	Recall	0.568	0.6	1	0	0.4
		Precision	0.577	0.6	1	0	0.404
		F1-score	0.532	0.545	1	0	0.376
Novice method		Recall / Precision / F1-score	**0.474**	0.5	1	0	0.321

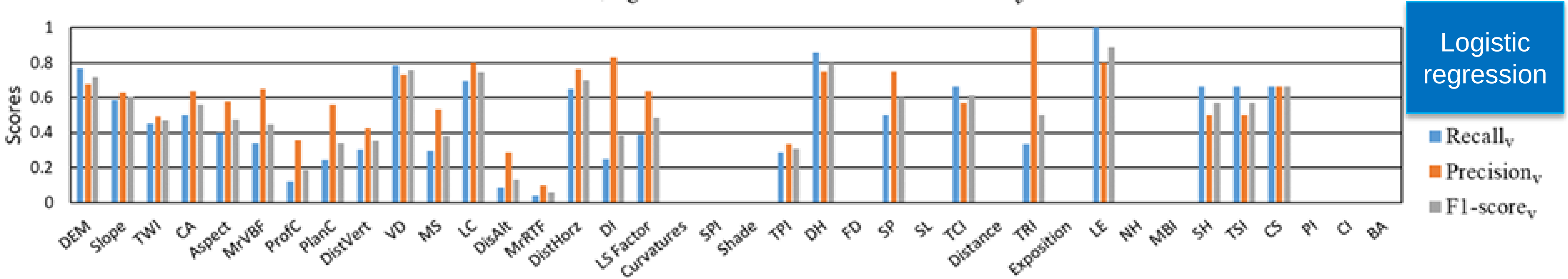
- Compared with the novice method, the RF method and two most-similar-case methods (MO and kNN) improved 24~35% consistency between the selected covariates and the original solution in the evaluation cases.

Discussion: performance of the classification strategy

- Random forest showed advantage, when current case base is highly imbalanced (80% covariates used in less than 40 among 191 cases.

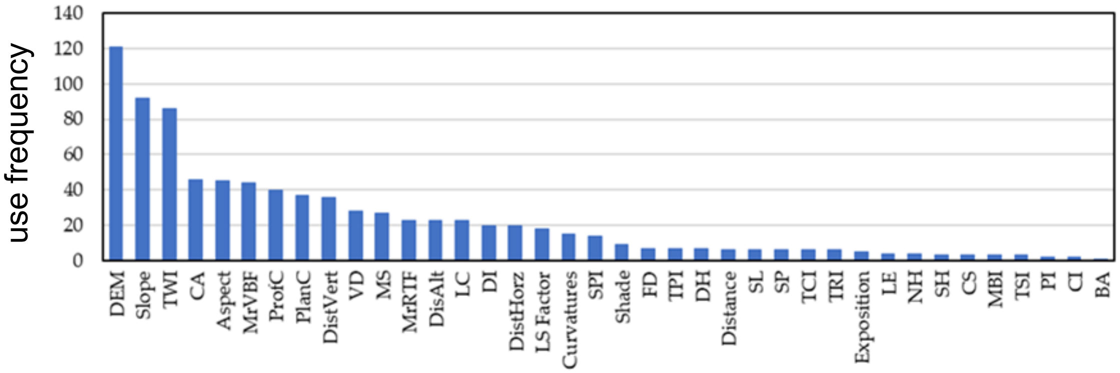


Random forest



Logistic regression

Covariate (use frequency: high → low)



Discussion: performance of the most-similar-case strategy

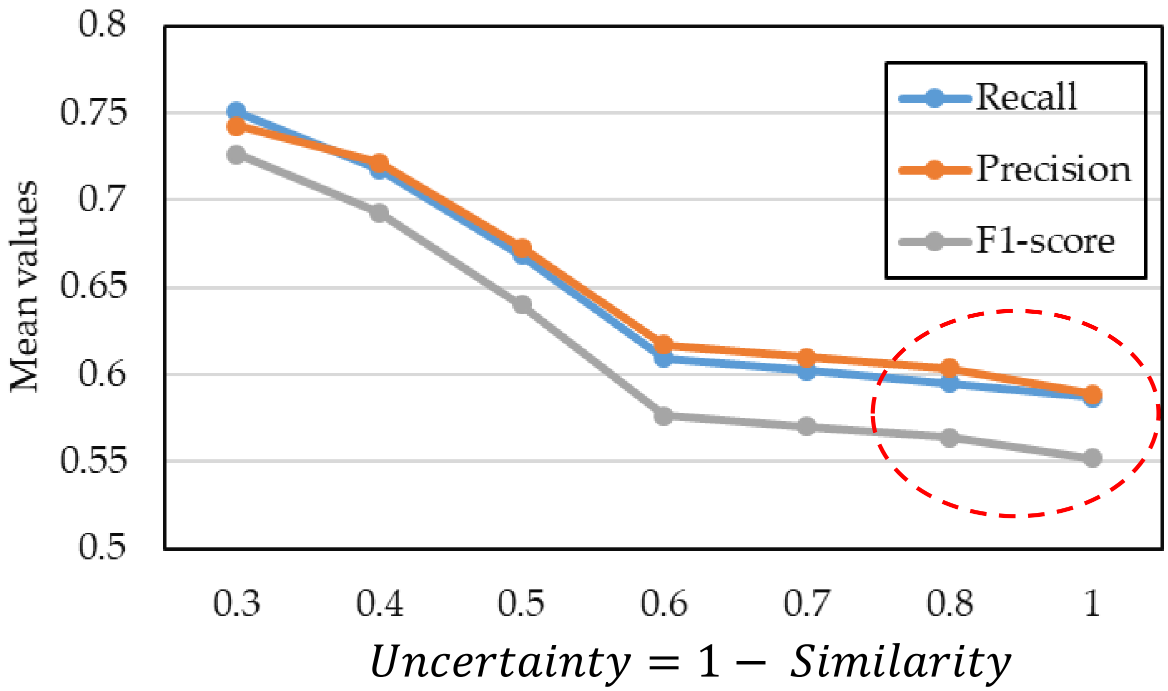
- Relationship between the evaluation indices and the case similarity from the MO method

Index intervals	**Eval. indices**	**S∈[0.8,1]**	**S∈[0.7,0.8)**	**S∈[0.6,0.7)**	**S∈[0.5,0.6)**	**S∈[0,0.5)**	**Total count**
[0.9,1]	Recall	**40**	**14**	12	7	4	**77**
	Precision	**39**	**14**	14	8	2	**77**
	F1-score	**37**	**11**	9	3	0	**60**
[0.7,0.9)	Recall	4	3	1	2	0	10
	Precision	4	4	1	1	1	11
	F1-score	4	6	1	1	0	12
[0.6,0.7)	Recall	6	1	0	2	2	11
	Precision	7	1	0	3	2	13
	F1-score	10	1	3	6	2	22
[0.5,0.6)	Recall	9	3	3	4	5	24
	Precision	8	1	4	4	4	21
	F1-score	6	3	5	5	2	21
[0.3,0.5)	Recall	5	0	4	5	0	14
	Precision	5	2	1	4	4	16
	F1-score	8	1	1	2	8	20
[0,0.3)	Recall	9	2	7	11	26	55
	Precision	10	1	7	11	24	53
	F1-score	10	1	8	12	25	56

For most of the evaluation cases, the results from the method were good

(i.e., high evaluation index value; consistent results as the original solutions of the evaluation cases)

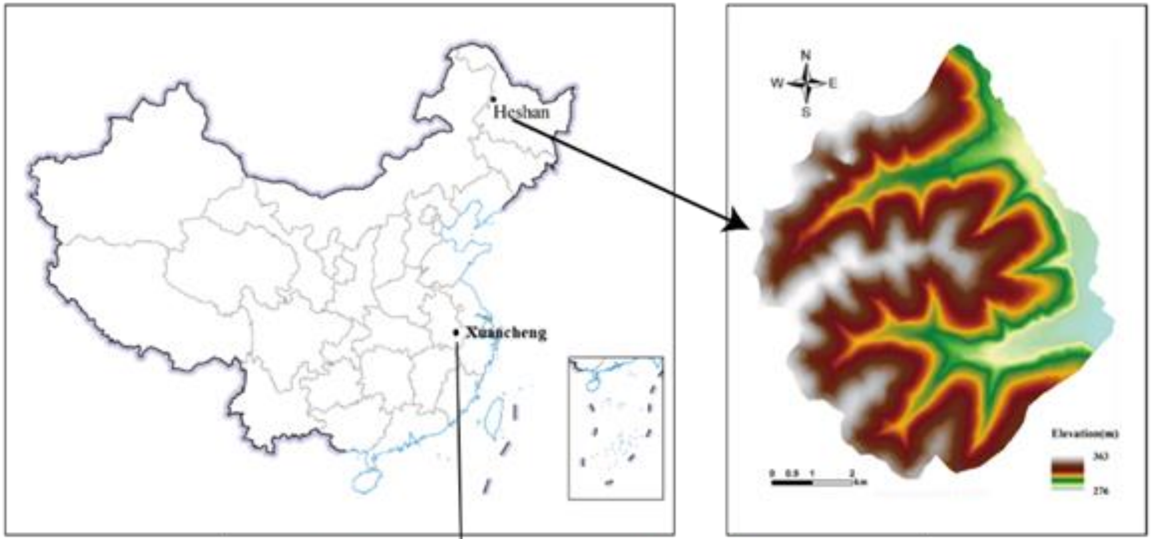
Discussion: performance of the most-similar-case strategy



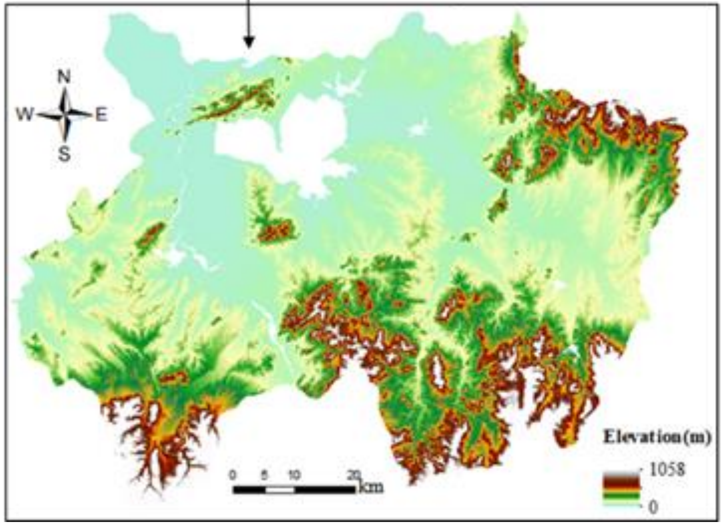
- The minimum operator method performed reasonably
  - The lower uncertainty (i.e., the higher the case similarity), the more consistent are the predicted covariates with the original solution of the evaluation case.
  - High uncertainty means there is no similar cases in the case base, which lowers the performance of the method under test. -- Size of case base does matter!

5) Practical DSM applications (with soil samples)

- Evaluate the mapping accuracy with the covariates selected by different methods

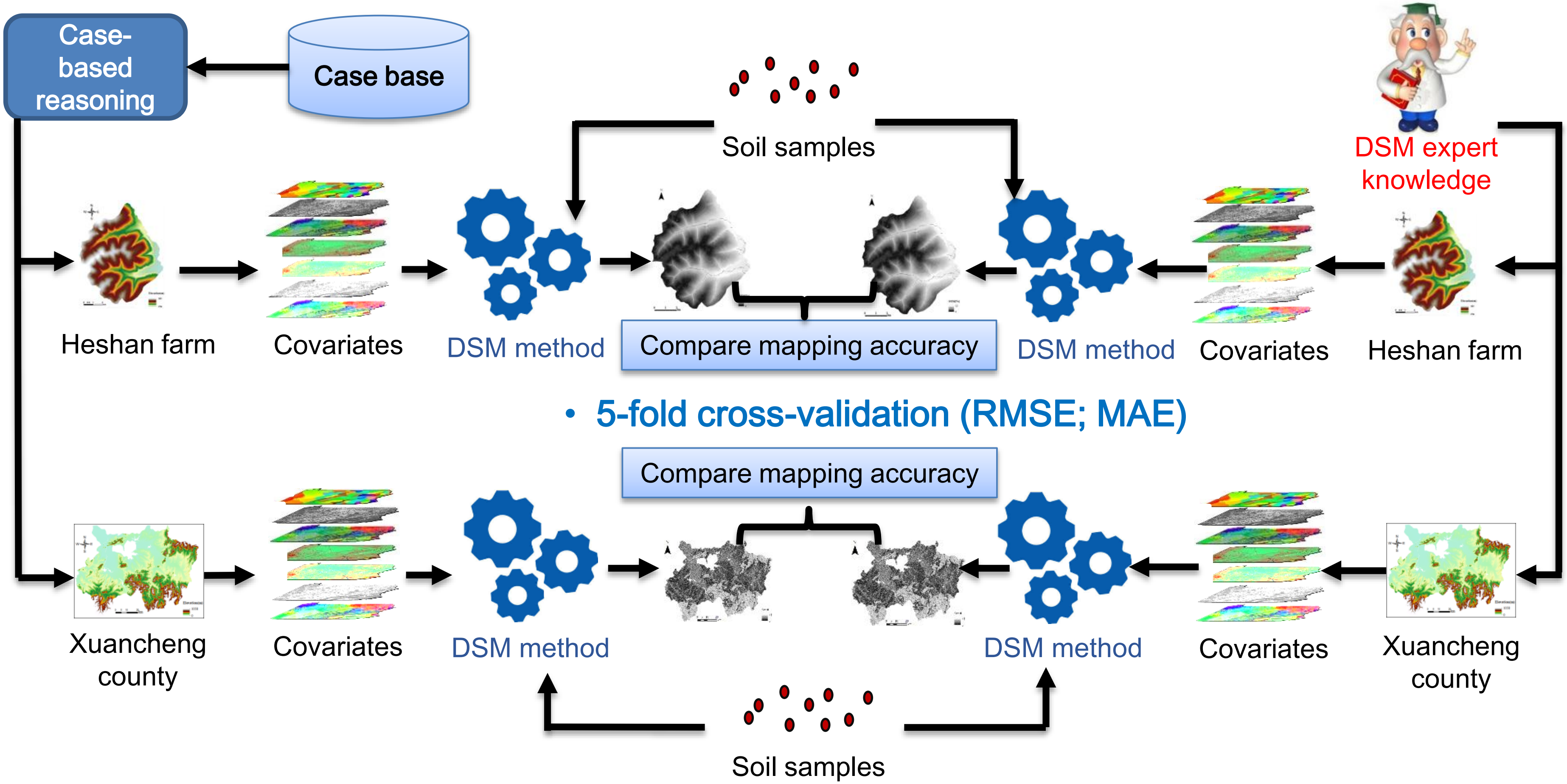


(1) Heshan farm:
- Low-relief
- 60 km^2
- Soil organic matter (%) in topsoil layer
- 83 soil samples



(2) Xuancheng county:
- Complex terrain conditions
- 5900 km^2
- Sand content (%) in topsoil layer
- 295 soil samples

Practical DSM applications (with soil samples)



- 5-fold cross-validation (RMSE; MAE)

- Digital soil mapping method:
  - individual predictive soil mapping (iPSM) (Zhu et al., 2015);
  - Random forest mapping

- DSM expert knowledge:
  - Heshan farm (Zhu et al., *EJSS*, 2015);
  - Xuancheng county (Yang et al., *SSSAJ*, 2016)

Results: Heshan farm (SOM in topsoil layer)

- Covariates selected by different methods

Covariate	Expert choice (Zhu et al., 2015)	Most-similar-case strategy		Covariate-level binary classification strategy	
		Minimum operator	kNN	RF	Logistic regression
Aspect		●			
DEM	●		●	●	
LS-Factor		●			
Plan Curvature	●	●			
Profile Curvature	●	●			
Slope	●	●	●	●	●
TWI	●	●	●	●	●
Catchment Area		●			
Relative position index (RPI)	●				
Recall		**0.67**	0.5	0.5	0.33
Precision		0.57	**1**	**1**	**1**
F1-score		0.61	**0.67**	**0.67**	0.5

- Mapping accuracy with the covariates selected by different methods
  - RMSE, MAE: about 3%~15% larger than those from expert choice.

DSM method	Evaluation index	Expert choice	Most-similar-case strategy		Covariate-level classification strategy	
			Minimum operator	kNN	RF	Logistic regression
iPSM	MAE	0.86	0.91	0.87	0.87	0.90
	RMSE	1.23	**1.28**	1.24	1.24	1.27
Random forest mapping	MAE	0.91	0.939	0.969	0.969	0.997
	RMSE	1.250	1.278	1.399	1.399	**1.469**

Results: Xuancheng county (Sand content in topsoil layer)

- Covariates selected by different methods

Covariate	Expert choice (Yang et al., 2016)	Most-similar-case strategy		Covariate-level binary classification strategy	
		Minimum operator	kNN	RF	Logistic regression
Aspect		●			
Curvature			●		
DEM			●	●	●
Landform		●			
LS-Factor		●			
MRRTF			●		
MRVBF			●		
Plan Curvature	●	●			
Profile Curvature	●	●	●		
Slope	●	●	●	●	●
Catchment Area		●			
TWI	●	●	●	●	●
Aspect		●			
Recall		**1**	0.75	0.5	0.5
Precision		0.5	0.43	**0.67**	**0.67**
F1-score		**0.67**	0.55	0.57	0.57

- Mapping accuracy with the covariates selected by different methods
  - RMSE, MAE: about 0.%~3% difference with those from expert choice.

DSM method	Evaluation index	Expert choice	Most-similar-case strategy		Covariate-level classification strategy	
			Minimum operator	kNN	RF	Logistic regression
iPSM	MAE	15.19	15.41	15.177	15.294	15.294
	RMSE	18.82	**19.193**	18.664	18.776	18.776
Random forest mapping	MAE	15.262	15.403	15.356	15.934	15.934
	RMSE	18.934	18.976	19.30	**19.572**	**19.572**

Mapping accuracies with the automatically-selected covariates were acceptable, while no one method performed the best at all times.

5. Conclusion

- Research issue: How to use those implicit knowledge contained in existing applications to automatically select proper (terrain) covariates for building geographical variable-environment relationship for geographical variable mapping?
- Case-based reasoning: Two strategies
  - ➢ The covariate-level binary classification strategy & the most-similar-case strategy
    - Preliminary evaluation showed the reasonableness of case-based reasoning.
    - The classification strategy is sensitive to the classification method and the imbalanced case base. Random forest method performed the best, while the logistic regression method also adopting the classification strategy performed the worst.
    - Performance of methods with the most-similar-case strategy are comparatively stable.
- Potential: Intelligent modeling
  - use those implicit, non-systematic, empirical knowledge on geographic modeling to help users (especially non-experts with few mapping knowledge) to automatically build application-context-specific model (not only covariate-selecting).
- Future work ...
  - Size of case base does matter!
  - Other domains of geographical variable mapping
  - Integrate into modeling tools

Thank You for Your Attention !

Liang P, Qin C-Z*, Zhu A-X. Comparison on two case-based reasoning strategies of automatically selecting terrain covariates for digital soil mapping. *Transactions in GIS*, 2021. doi:10.1111/TGIS.12831.

Qin C-Z, Liang P, Zhu A-X. A case-based classification strategy of automatically selecting terrain covariates for modeling geographic variable-environment relationship. In: M Alvioli, I Marchesini, L Melelli, P Guth, eds., *Proceedings of the Geomorphometry 2020 Conference*, p. 33-36. (extended abstract)

秦承志 (QIN Cheng-Zhi)

Email: qincz@lreis.ac.cn

Webpage: http://people.ucas.ac.cn/~qincz?language=en