NPTEL ONLINE CERTIFICATION COURSES

Machine Learning for Soil and Crop Management

Prof. Somsubhra Chakraborty
Agricultural and Food Engineering Department
Indian Institute of Technology Kharagpur

Week 12: DIGITAL SOIL MAPPING WITH CATEGORICAL VARIABLES

LECTURE 56

CONCEPTS COVERED

- **REGRESSION KRIGING**

KEYWORDS

- Regression Kriging
- Variogram
- Interpolation
- Cubist
- goof

Regression Kriging

- The Best Linear Unbiased Predictor of spatial data
- Matheron (1969) proposed that a value of a target variable at some location can be modelled as a sum of the deterministic and stochastic components:

$$Z(\mathbf{s}) = m(\mathbf{s}) + \varepsilon'(\mathbf{s}) + \varepsilon''$$

We know that both deterministic and stochastic components of spatial variation can be modelled separately. By combining the two approaches, we obtain:

$$\begin{aligned}\hat{z}(\mathbf{s}_0) &= \hat{m}(\mathbf{s}_0) + \hat{e}(\mathbf{s}_0) \\ &= \sum_{k=0}^{p} \hat{\beta}_k \cdot q_k(\mathbf{s}_0) + \sum_{i=1}^{n} \lambda_i \cdot e(\mathbf{s}_i)\end{aligned}$$

where

$\hat{m}(\mathbf{s}_0)$ = fitted deterministic part

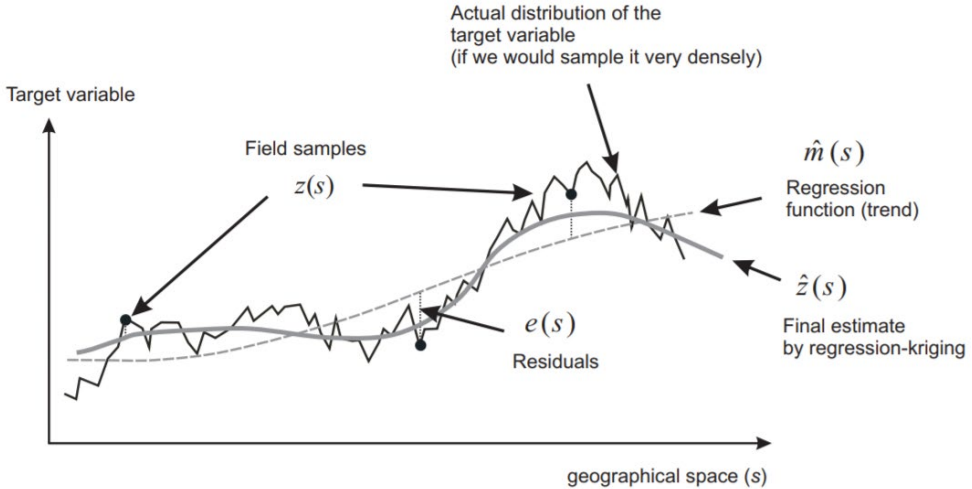
$\hat{e}(\mathbf{s}_0)$ = interpolated residual

$\hat{\beta}_k$ =estimated deterministic model coefficients

λi =kriging weights determined by the spatial dependence structure of the residual and where e(si) is the residual at location si

Regression Kriging (Hybrid Approach)

The deterministic model essentially "detrends" the data, leaving behind the residuals for which we need to investigate whether there is additional spatial structure which could be added to the regression model predictions. These residuals are the random component of the *SCORPAN + e model*.



A schematic example of regression-kriging: fitting a vertical cross-section with assumed distribution of an environmental variable in horizontal space.

REFERENCE

Malone, B. P., Minasny, B., & McBratney, A. B. (2017). *Using R for digital soil mapping* (Vol. 35). Cham, Switzerland: Springer International Publishing.


Thank
you

NPTEL ONLINE CERTIFICATION COURSES

Machine Learning for Soil and Crop Management

Prof. Somsubhra Chakraborty
Agricultural and Food Engineering Department
Indian Institute of Technology Kharagpur

Week 12: DIGITAL SOIL MAPPING WITH CATEGORICAL VARIABLES

LECTURE 57

CONCEPTS COVERED

- Categorical Modelling in DSM
- Kappa coefficient and other accuracy measures
- Multinomial logistic model

KEYWORDS

- Categorical model
- Kappa coefficient
- User’s accuracy
- Producer’s accuracy
- Terron

IMPORTANT QUALITY MEASURES

1. Overall accuracy
2. User's accuracy
3. Producer's accuracy
4. Kappa coefficient of agreement

RECALL: ACCURACY OF INDIVIDUAL CLASSES

- Accuracy of individual classes can be computed in a similar manner as that of overall accuracy.
- However, there is a choice of dividing the number of correct predictions for each class by either the totals (observations or predictions) in the corresponding columns or rows respectively.
- Traditionally, the total number of correct predictions of a class is divided by the total number of observations of that class (i.e. the column sum).

RECALL: ACCURACY OF INDIVIDUAL CLASSES

- This accuracy measure indicates the probability of an observation being correctly classified and is really a measure of omission error, or the "producer's accuracy".
- This is because the producer of the model is interested in how well a certain class can be predicted.

RECALL: ACCURACY OF INDIVIDUAL CLASSES

- Alternatively, if the total number of correct predictions of a class is divided by the total number of predictions that were predicted in that category, then this result is a measure of commission error, or "user's accuracy".
- This measure is indicative of the probability that a prediction on the map actually represents that particular category on the ground or in the field.

RECALL: KAPPA COEFFICIENT

- The Kappa coefficient is another statistical measure of the fidelity between observations and predictions of a classification.
- The calculation is based on the difference between how much agreement is actually present ("observed" agreement) compared to how much agreement would be expected to be present by chance alone ("expected" agreement).
- The observed agreement is simply the overall accuracy percentage.

RECALL: KAPPA COEFFICIENT

- We may also want to know how different the observed agreement is from the expected agreement.
- The Kappa coefficient is a measure of this difference, standardized to lie on a -1 to 1 scale, where 1 is perfect agreement, 0 is exactly what would be expected by chance, and negative values indicate agreement less than chance, i.e., potential systematic disagreement between observations and predictions. The Kappa coefficient is defined as:

$$K = \frac{p_o - p_e}{1 - P_e}$$

where *po* is the overall or observed accuracy, and *pe* is the expected accuracy, where:

$$p_e = \sum_{i=1}^{n} \left(\frac{colSum_i}{TO}\right) \times \left(\frac{rowSum_i}{TO}\right)$$

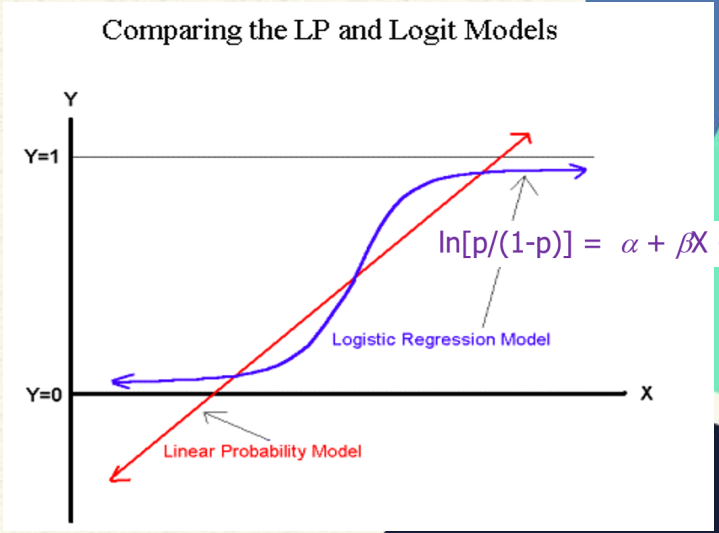
TO is the total number of observations and *n* is the number of classes

RECALL: KAPPA COEFFICIENT

- < Less than chance agreement.
- 0.01–0.20 Slight agreement.
- 0.21–0.40 Fair agreement.
- 0.41–0.60 Moderate agreement.
- 0.61–0.80 Substantial agreement.
- 0.80–0.99 Almost perfect agreement

MULTINOMIAL LOGISTIC REGRESSION

- Used to model nominal outcome variables, in which the log odds of the outcomes are modeled as a linear combination of the predictor variables.
- Since dealing with categorical variables, it is necessary that logistic regression take the natural logarithm of the odds (log-odds) to create a continuous criterion.
- The logit of success is then fit to the predictors using regression analysis.



REFERENCE

Malone, B. P., Minasny, B., & McBratney, A. B. (2017). *Using R for digital soil mapping* (Vol. 35). Cham, Switzerland: Springer International Publishing.


Thank
you

NPTEL ONLINE CERTIFICATION COURSES

Machine Learning for Soil and Crop Management

Prof. Somsubhra Chakraborty
Agricultural and Food Engineering Department
Indian Institute of Technology Kharagpur

Week 12: DIGITAL SOIL MAPPING WITH CATEGORICAL VARIABLES

LECTURE 58

CONCEPTS COVERED

- C5 decision tree
- Random Forest classification model

KEYWORDS

- C5
- Random Forest
- rasterVis
- Terron
- goofcat

RANDOM FOREST: RECALL

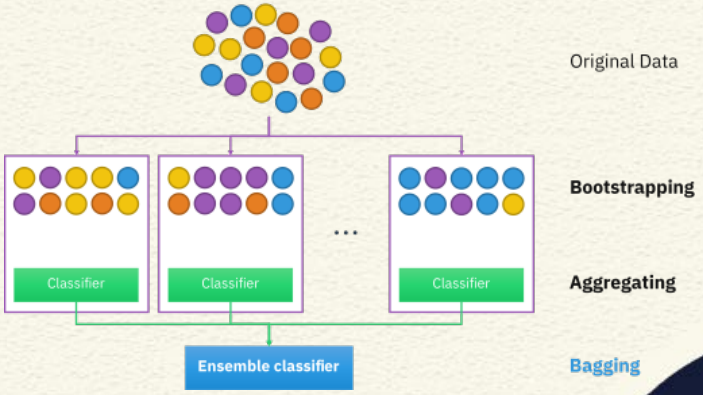
- Forms lots of decision trees (regression/classification) with random selection of samples/observations and random selection of features/variables
- Provides the class of dependent variable based on many trees
- Random Trees
- Many random trees= Random forest

WHY RANDOM FOREST IS WIDELY ACCEPTED?

- Most of tree can provide correct prediction of class for most part of the data
- The trees are making mistakes at different places
- Strong fundamental concept: a large number of relatively uncorrelated models (trees) operating as a committee will outperform any of the individual constituent models.

RECALL: BAGGING

- Reduces over-fitting of the model.
- Handles higher dimensionality data very well.
- Maintains accuracy for missing data.



Source: Sirakorn, https://commons.wikimedia.org/wiki/File:Ensemble_Bagging.svg (CC BY-SA 4.0)

HOW RF IS EXECUTED?

Each tree is constructed using the following algorithm:

1. Let the number of training cases be *N*, and the number of variables/features in the classifier be *M*.
2. Let us assume that *m* of input variables to be used to determine the decision at a node of the tree; *m* should be much less than *M*.
3. Choose a training set for this tree by choosing *n* times with replacement from all *N* available training cases (i.e. take a bootstrap sample). Use the rest of the cases to estimate the error of the tree, by predicting their classes.

HOW RF IS EXECUTED?

4. For each node of the tree, randomly choose *m* variables on which to base the decision at that node. Calculate the best split based on these *m* variables in the training set.

5. Each tree is fully grown and not pruned (as may be done in constructing a normal tree classifier).

For prediction a new sample is pushed down the tree. It is assigned the label of the training sample in the terminal node it ends up in. This procedure is iterated over all trees in the ensemble, and the average vote of all trees is reported as random forest prediction.

- **Bagging: training each individual learner on different bootstrapped subsets of the data and then averaging the predictions**

RF ALGORITHM- OOB

- Out of bag (OOB) score validates the RF model. Below is a simple example of how is it calculated followed by a description of how it is different from normal validation score and where it is advantageous.
- For the description of OOB score calculation, let's assume there are five DTs in the random forest ensemble labeled from 1 to 5. For simplicity, suppose we have a simple original training data set as below.

Weather	Temp	Humidity	Wind	Harvest Crop
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Sunny	Hot	High	Weak	Yes
Windy	Cold	Low	Weak	Yes

RF ALGORITHM- OOB

- Let the first bootstrap sample is made of the first three rows of this data set as shown in the red box below. This bootstrap sample will be used as the training data for the DT "1" in the RF model.

Bootstrap

Weather	Temp	Humidity	Wind	Harvest Crop
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Sunny	Hot	High	Weak	Yes
Windy	Cold	Low	Weak	Yes

RF ALGORITHM- OOB

Weather	Temp	Humidity	Wind	Harvest Crop
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Sunny	Hot	High	Weak	Yes
Windy	Cold	Low	Weak	Yes

OOB for DT1

RF ALGORITHM- OOB

After the DTs models have been trained, this leftover row or the OOB sample will be given as unseen data to the DT 1. The DT 1 will predict the outcome of this row. Let DT 1 predicts this row correctly as "YES". Similarly, this row will be passed through all the DTs that did not contain this row in their bootstrap training data. Let's assume that apart from DT 1, DT 3 and DT 4 also did not have this row in their bootstrap training data. The predictions of this row by DT 1, 3, 4 are summarized in the table below.

DT	Prediction (Harvest Crop)
1	Yes
3	No
4	Yes
Majority Vote: Yes	

RF ALGORITHM- OOB

- Note that the final prediction of this row by majority vote is a correct prediction since originally in the "Harvest Crop" column of this row is also a "YES".
- Similarly, each of the OOB sample rows is passed through every DT that did not contain the OOB sample row in its bootstrap training data and a majority prediction is noted for each row.
- Finally, the OOB score is computed as the number of correctly predicted rows from the out of bag sample.

DT	Prediction (Harvest Crop)
1	Yes
3	No
4	Yes
Majority Vote: Yes	

REFERENCE

Malone, B. P., Minasny, B., & McBratney, A. B. (2017). *Using R for digital soil mapping* (Vol. 35). Cham, Switzerland: Springer International Publishing.


Thank
you

NPTEL ONLINE CERTIFICATION COURSES

Machine Learning for Soil and Crop Management

Prof. Somsubhra Chakraborty
Agricultural and Food Engineering Department
Indian Institute of Technology Kharagpur

Week 12: DIGITAL SOIL MAPPING WITH CATEGORICAL VARIABLES

LECTURE 59

CONCEPTS COVERED

- **Combined Continuous and Categorical Models**

KEYWORDS

- Combined model
- Soil horizon
- Ap horizon
- Quantile regression forest
- nnet

WHY COMBINE MODELS?

- To gain some insights into a digital soil mapping approach that uses a combination of both continuous and categorical attribute modeling

- Mapping of occurrence and thickness of soil profiles

MODEL PERFORMANCE

Table 1: Selected model validation diagnostics returned for each horizon class and associated depth model.

	Presence/Absence of Horizon			Depth of Horizon		
Horizon	**Overall Accuracy**	**User's Accuracy**	**Kappa Statistic**	**Concordance**	**RMSE**	**PICP**
A1	87%	Pres = 89% Abs = 54%	0.19	0.05	10	46%
A2	87%	Pres = 100% Abs = 87%	0.04	0.10	12	42%
AP	86%	Pres = 50% Abs = 88%	0.15	0.00	12	53%
B1	91%	Pres = 0% Abs = 91%	0	0.16	12	45%
B21	97%	Pres = 97% Abs = 0%	0	0.05	17	41%
B22	73%	Pres = 73% Abs = 34%	0	0.10	14	41%
B23	78%	Pres = 0% Abs = 78%	0	0.04	12	45%
B24	97%	Pres = 0% Abs = 97%	0	0.00	22	46%
BC	74%	Pres = 68% Abs = 75%	0.20	0.06	18	29%
C	95%	Pres = 0% Abs = 95%	0	0	NA	68%

REFERENCE

Malone, B. P., Minasny, B., & McBratney, A. B. (2017). *Using R for digital soil mapping* (Vol. 35). Cham, Switzerland: Springer International Publishing.


Thank
you

NPTEL ONLINE CERTIFICATION COURSES

Machine Learning for Soil and Crop Management

Prof. Somsubhra Chakraborty
Agricultural and Food Engineering Department
Indian Institute of Technology Kharagpur

Week 12: DIGITAL SOIL MAPPING WITH CATEGORICAL VARIABLES

LECTURE 60

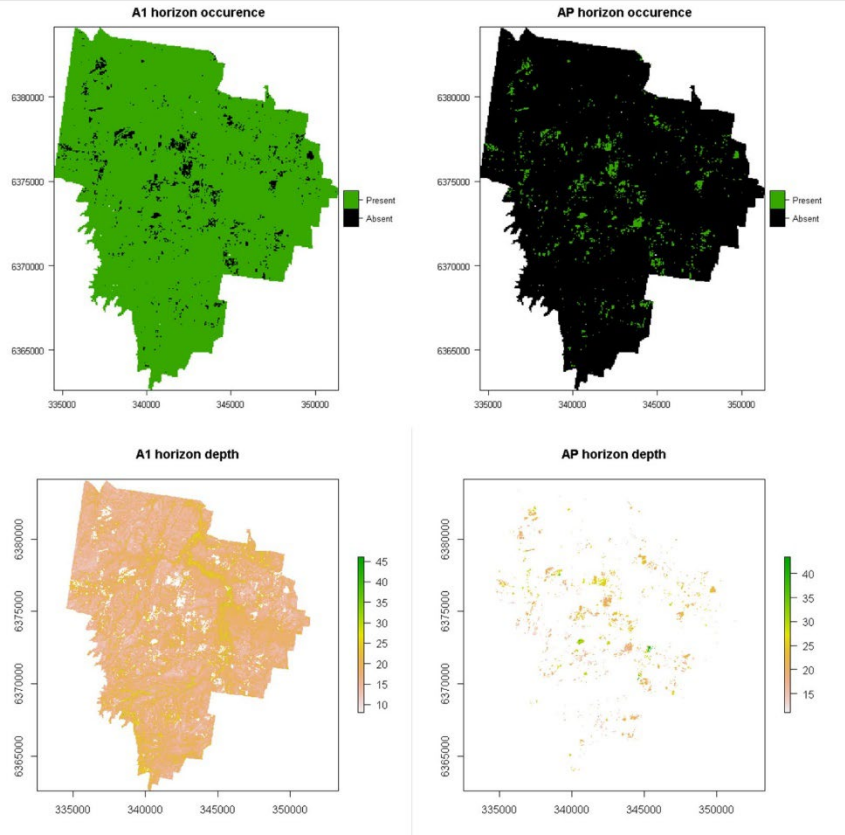
CONCEPTS COVERED

- Some important ML models
- DSM with AIML- TWO CASE STUDIES

KEYWORDS

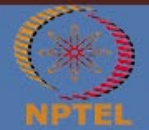
- Boosting
- Gradient boosting
- XGBoost
- DSM
- Superlearner
- Base learner

MAPPING A1 AND Ap OCCURENCE



RECALL: BOOSTING

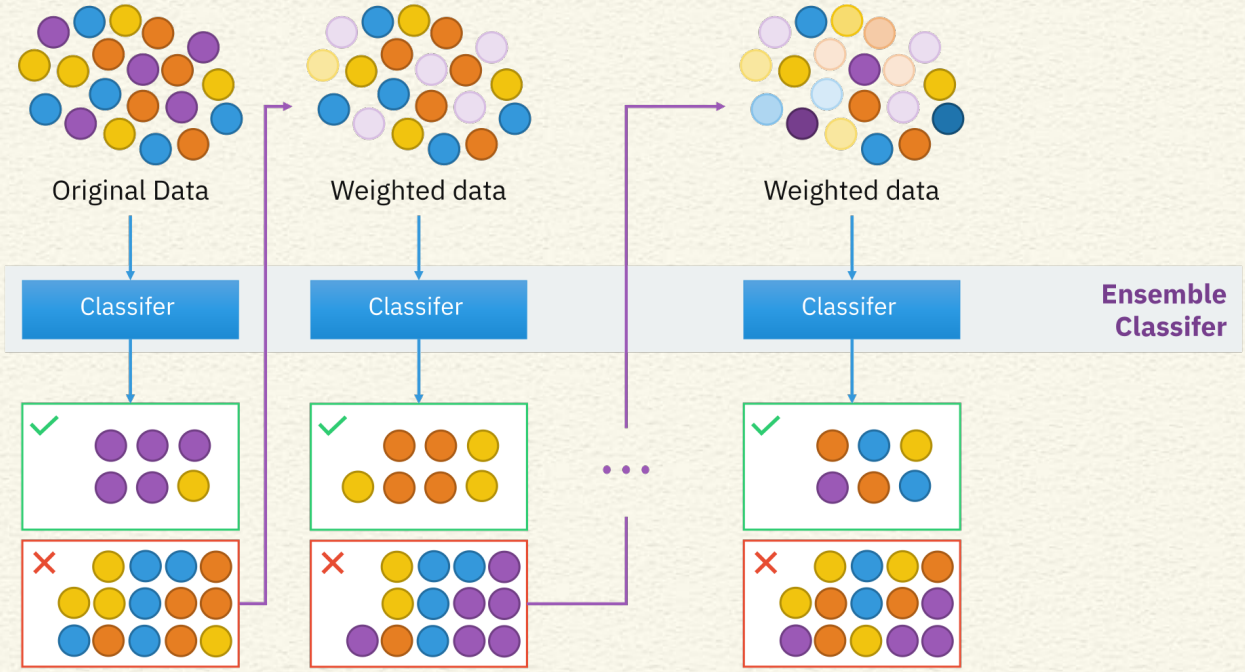
- An ensemble meta-algorithm for primarily reducing bias and variance
- Used to create a collection of predictors.
- Learners are learned sequentially with early learners fitting simple models to the data and then analysing data for errors
- Consecutive trees (random sample) are fit and at every step
- The goal is to improve the accuracy from the prior tree



RECALL: BOOSTING

- When an input is misclassified by a hypothesis, its weight is increased so that next hypothesis is more likely to classify it correctly
- This process converts weak learners into better performing model

RECALL: BOOSTING



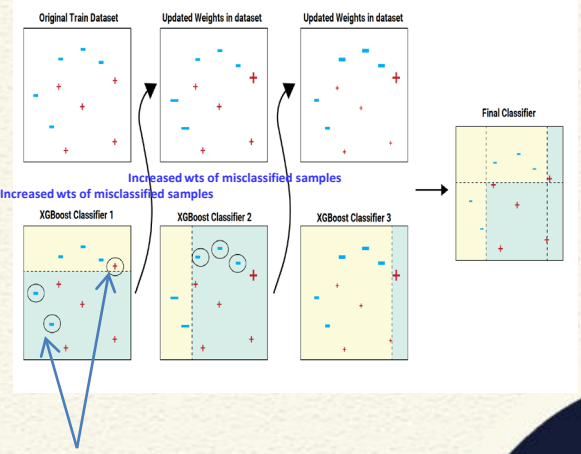
Source: Sirakorn, https://commons.wikimedia.org/wiki/File:Ensemble_Boosting.svg (CC BY-SA 4.0)

GRADIENT BOOSTING

- A popular boosting algorithm
- In gradient boosting, each predictor corrects its predecessor's error
- Each predictor is trained using the residual errors of predecessor as labels

XGBoost

- An implementation of Gradient Boosting: Python Library
- In this algorithm, decision trees are created in sequential form
- Weights play an important role in XGBoost. Weights are assigned to all the independent variables which are then fed into the decision tree which predicts results
- The weight of variables predicted wrong by the tree is increased and these variables are then fed to the second decision tree
- These individual classifiers/predictors then ensemble to give a strong and more precise model. It can work on regression, classification, ranking, and user-defined prediction problems



https://blog.quantinsti.com/xgboost-python/

DSM: APPLICATION OF AIML

Table 1
Non-exhaustive list with summary of case studies in which machine learning algorithms are used for digital soil mapping.

Spatial extent[a]	Sample size	Sampling design	Number of covariates	Machine learning model[b]	Covariate selection	Parameter tuning	Map quality indices[c]	Uncertainty quantification	Reference
Quantitative maps									
Plot	285	grid-based	19	cubist, RF	no	no	R^2, RMSE	no	Pouladi et al. (2019)
Local	47	stratified random	41	RF	yes	yes	RMSE, IQR	yes	Blanco et al. (2018)
Local	70	cLHS	19	cubist	no	no	MAE, RMSE, R^2, CCC	yes	Lacoste et al. (2014)
Local	75	grid-based	9	ANN	no	no	R^2, MSE	no	Kalambukattu et al. (2018)
Local	98	varied sources	173	RF	yes	no	RMSE, R^2	no	Shi et al. (2018)
Local	116	simple random	20	RF	no	no	R^2, RMSE, CCC	no	Dharumarajan et al. (2017)
Local	117	not specified	13	GBM	yes	yes	R^2, RMSE, MAE	yes	Hamzehpour et al. (2019)
Local	117	not specified	412	cubist	yes	no	ME, MAE, R^2, R^2_{adj}	no	Miller et al. (2015b)
Local	120	stratified random	not specified	RF	no	no	ME, RMSE, R^2, MSE	no	Wiesmeier et al. (2011)
Local	120	stratified random	22	ANN, BRT	yes	yes	R^2, RMSE, ME	no	Mosleh et al. (2016)
Local	137	systematic random	20	ANN, GEP	yes	yes	RMSE, R^2, MBE	no	Mahmoudabadi et al. (2017)
Local	138	not specified	15	RF	yes	no	RMSE, R^2 CCC	no	Zhu et al. (2019)
Local	150	grid-based	not specified	ANN	no	yes	correlation coefficient, R^2, RMSE, Willmott's index of agreement, RPIQ	no	Sergeev et al. (2019)
Local	151	not specified	not specified		no	no	R^2, NRMSD	no	Kovačević et al. (2010)
Local	153	grid-based	26	RF	yes	no	RMSE, R^2	no	Tajik et al. (2019)
Local	159/34	not specified	37	RF, cubist, QRF, NN, avNNet, ctree, evtree, GBM, *k*-NN, RT, SVM	yes	no	R^2, RMSE, MAE, MARE	yes	Rudiyanto et al. (2018)
Local	165	stratified random	18	RF	no	yes	MSE, NMSE	no	Grimm et al. (2008)
Local	173 profiles	cLHS	19	RF	no	no	ME, RMSE, R^2	no	Taghizadeh-Mehrjardi et al. (2014)
Local	188 profiles	cLHS	16	ANN, SVR, *k*-NN, RF, RT	no	yes	RMSE, CCC	no	Taghizadeh-Mehrjardi et al. (2016b)
Local	234	not specified	410	cubist	yes	no	MAE, R^2	yes	Miller et al. (2015a)
Local	330 profiles	not specified	12	BRT, ANN, least-square SVM	no	yes	R^2, R^2_{adj},RMSE, relative RMSE	no	Ottoy et al. (2017)
Local	330	simple random	10	RF, GBM	no	yes	ME, MAE, RMSE, R^2		Tziachris et al. (2019)
Local	334	cLHS	16	cubist, RF, RT	yes	no	R^2, RMSE	no	Zeraatpisheh et al. (2019)
Local	342/321	-	14	MARS, SVR, RF, Cubist, NN	-	yes	R^2	no	Behrens et al. (2018b)
Local	399	not specified	12	RF	no	no	R^2, RMSE	no	da Silva Chagas et al. (2016)
Local	440	varied sources	19	RF, SVM, ANN	no	yes	RMSE, ME	no	Were et al. (2015)
Local	460	grid-based	21	RF	no	yes	ME, MAE, RMSE	no	Pahlavan-Rad and Akbarimoghaddam (2018)
Local	568	simple random	26	QRF	no	no	R^2, RMSE, range-normalized RMSE, Moran's I	yes	Kirkwood et al. (2016)
Local	1104	expert	29	RF, SVM, SGB	no	yes	RMSE, sMAPE	no	Forkuor et al. (2017)
Local	≤ 1052/ 2050/ 2379	varied sources	300-500	BRT, RF	yes	yes	bias, RMSE, SS, R^2	no	Nussbaum et al. (2018)
Local	2388	varied sources	3	CNN, RF	no	yes	ME, RMSE, R^2, CCC	no	Wadoux et al. (2019b)
Regional	not specified	not specified	20	cubist	no	no	R^2, RMSE, bias, CCC	yes	Mulder et al. (2016)
Regional	125 profiles	purposive	12	BRT, RF	no	no	MAE, RMSE,R^2, CCC	no	Yang et al. (2016)
Regional	244	grid-based	4	ANN	no	yes	ME, MAE, RMSE, CCC	no	Dai et al. (2014)
Regional	339/961	varied sources	40	QRF	no	no	R^2, RMSE	yes	Nauman and Duniway (2019)
Regional	485 profiles	not specified	5	CNN	no	yes	R^2, RMSE	yes	Padarian et al. (2019)
Regional	500	not specified	12	RF, BRT	yes	no	R^2, RMSE	no	Beguin et al. (2017)

(*continued on next page*)

Wadoux et al. (2020)

DSM: APPLICATION OF AIML

Table 1 (*continued*)

Spatial extent[a]	Sample size	Sampling design	Number of covariates	Machine learning model[b]	Covariate selection	Parameter tuning	Map quality indices[c]	Uncertainty quantification	Reference
Regional	528	subset from a systematic grid	18	*k*-NN	yes	no	RMSE, R^2, Bias, coefficient of variance	no	Mansuy et al. (2014)
Regional	705	simple random	16	RF, BRT, SVM	yes	yes	R^2, MAE, RMSE, CCC	yes	Wang et al. (2018)
Regional	978 profiles	not specified	24	RF	no	no	R^2, ME, RMSE, CCC	no	Akpa et al. (2014)
Regional	1,014	stratified random	327	CART, BRT, BRT, RF, SVM	yes	no	R^2, RMSD, RPD, RPIQ	no	Keskin et al. (2019)
Regional	1,134	not specified	81	NN	no	no	R^2, ME, MAE, RMSE	no	Aitkenhead and Coull (2016)
Regional	1,300 profiles	not specified	6	RF	no	no	CCC, RMSE	yes	McNicol et al. (2019)
Regional	1,626	not specified	40	SVM	no	yes	R^2, MSE	no	Wu et al. (2016)
Regional	2,024 profiles	legacy data	16	QRF	no	no	ME, RMSE, R^2, accuracy plot	yes	Vaysse and Lagacherie (2017)
Regional	2,024 profiles	legacy data	16		no	yes	MSE, R^2	no	Vaysse and Lagacherie (2015)
Regional	2,943	two-stage systematic	37	CNN, RF	no	yes	ME, RMSE, R^2, CCC	yes	Wadoux (2019)
Regional	4,859	not specified	26	QRF	no	no	ME, RMSE, accuracy plot	yes	Szatmári et al. (2019)
Regional	4,859	not specified	32	QRF	no	no	ME, RMSE, accuracy plot	yes	Szatmári and Pásztor (2019)
Regional	5,386	varied sources	6	cubist, SVM	no	no	R^2, MSE, CCC		Somarathna et al. (2016)
Regional	13,000	not specified	18	RF	no	no	R^2	yes	Koch et al. (2019)
Regional	19,790	two-stage systematic	197	RF	no	no	ME	no	Wadoux et al. (2019a)
Regional	37,693	legacy soil data	74	RF, Cubist, SVM	yes	yes	R^2, RMSE, MAE	yes	Gomes et al. (2019)
Regional - Global	2,268-27,262	varied sources	34	cubist	no	yes	CCC, RMSE, SDE, ME	yes	Viscarra-Rossel et al. (2015)
Regional - Global	366,034	varied sources	> 200	RF, GBM	no	yes	R^2,ME, RMSE, MAE	yes	Ramcharan et al. (2018)
Global	11,268	legacy soil data	118	SVM, kernel weighted NN, RF	yes	no	EC, RMSE, R^2	yes	Guevara et al. (2018)
Global	150,000	legacy soil data	> 200	RF, GBM	no	yes	R^2	no	Hengl et al. (2017a)
Categorical maps									
Local	-	not specified	125	ANN	no	no	Accuracy, recall, precision	no	Behrens et al. (2005)
Local	33 profiles	not specified	16	RF, J48	no	no	not specified	no	Massawe et al. (2018)
Local	103/297/57	cLHS	130	*k*-NN, NSC, CT, BCT, RF, linear SVM, radial-basis SVM, NN, ANN	yes	yes	Kappa analysis, Brier scores, visual inspection, confusion index	no	Brungard et al. (2015)
Local	125 profiles	cLHS	17	RF	no	no	map purity, Cohen's kappa, Shannon entropy index, relative purity, relative diversity	no	Zeraatpisheh et al. (2017)
Local	151	not specified	not specified	SVM	no	no	NRMSD, micro averaged F1 measure, kappa statistics	no	Kovačević et al. (2010)
Local	175, 63 profiles	varied sources	27	*k*-NN, SVm, DT, RF	no	no	OA, PA, UA, kappa coefficient, AUROC	no	Vermeulen and Van Niekerk (2017)
Local	452 profiles	regular grid	6	DT, RF	yes	no	OA, UA, PA, Kappa coefficient of agreement	no	Sharififar et al. (2019)
Local	917	grid-based	33	RF	yes	no	Kappa index	no	Hounkpatin et al. (2018)
Local	3,121	by-polygon, equal-class, area-weighted, and area-weighted with random over sampling	20	CART, CART with bagging, RF, *k*-NN, NSC, ANN, LMT, SVM	no	yes	overall agreement, quantity disagreement, allocation disagreement, total disagreement	no	Heung et al. (2016)

(*continued on next page*)

Wadoux et al. (2020)

DSM: APPLICATION OF AIML

Table 1 (*continued*)

Spatial extent[a]	Sample size	Sampling design	Number of covariates	Machine learning model[b]	Covariate selection	Parameter tuning	Map quality indices[c]	Uncertainty quantification	Reference
Regional	89,323	random sampling	26	*k*-NN, RF	yes	no	recall, accuracy	no	Subburayalu and Slater (2013)
Regional	366,034	varied sources	> 200	RF, GBM	no	yes	OA, regional dataset	yes	Ramcharan et al. (2018)
Regional	7,664 profiles	varied sources	110	DT, RF, EGB, SVM, *k*-NN	yes	no	OA, precision, recall, F-score, K-index	no	Taghizadeh-Mehrjardi et al. (2019b)
Regional	9,924	not specified	23	RF	yes	no	error matrix	no	Häring et al. (2012)
Global	150,000	legacy data	> 200	RF, GBM	no	yes	map purity, weighted kappa metrics, AUC, true positive rate, scaled Shannon's entropy index	no	Hengl et al. (2017a)

[a] Plot: 0-1 km^2; Local: > 1 km^2–10^4 km^2; Regional: > 10^4 km^2–10^7 km^2; Global: > 10^7 km^2.

[b] RF: random forest; ANN: artificial neural networks; CNN convolutional neural networks; GBM: gradient boosting machine; BRT: boosted regression tree; GEP: gene expression programming; QRF: quantile regression forest; avNNet: neural networks using model averaging; ctree: conditional inference trees; evtree: evolutionary algorithm for classification and regression tree; NN: neural networks; GBM: generalized boosted regression; *k*-NN: *k*-nearest neighbours; RT: regression tree; SVM: support vector machine; MARS: multivariate adaptive regression splines; SGB: stochastic gradient boosting; CART: classification and regression tree; NSC: nearest shrunken centroids; CT: classification tree; BCT: bagged classification tree; DT: decision tree; LMT: logistic model tree; EGB: extreme gradient boosting.

[c] R^2 coefficient of determination; R^2adj: adjusted coefficient of determination; RMSE: root mean square error; IQR: interquartile range; MAE: mean absolute error; CCC: Lin's concordance correlation coefficient; MSE: mean square error; ME: mean error; MBE: mean bias error; RPIQ: ratio of performance to interquartile distance; NRMSD; normalized root mean squared deviation; MARE: median absolute relative error; NMSE: normalized mean square error; sMAPE: symmetric mean absolute percentage error; SS: skill score; RMSD: minimum root mean square deviation; RPD: residual prediction deviation; SDE: standard deviation of the error; EC: overall ratio; OA: overall accuracy; PA: producer accuracy; UA; user accuracy; AUROC: area under receiver operating characteristic curve; AUC: area under the curve.

Wadoux et al. (2020)

DSM WITH SUPERLEARNER- A CASE STUDY

- **12 base learners were used to create a superlearner**

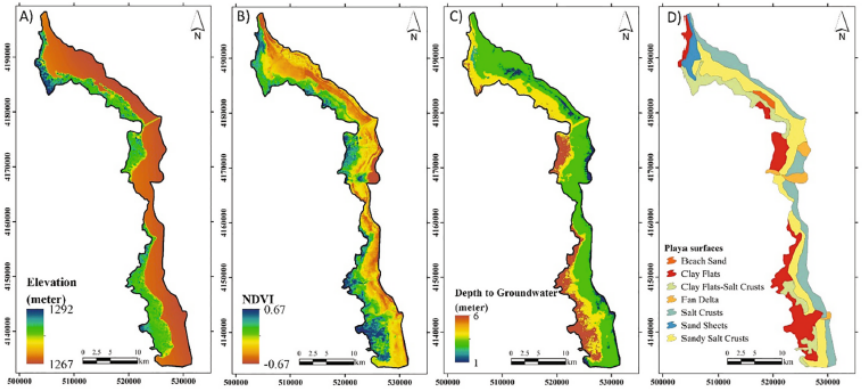


Fig. 2. Four examples of environmental covariates: (A) elevation derived from digital elevation model; (B) normalized difference vegetation index-NDVI calculated from Landsat 8 images; (C) depth to the groundwater calculated by regression-kriging model; and (D) geomorphic map based on the Iranian classification system.

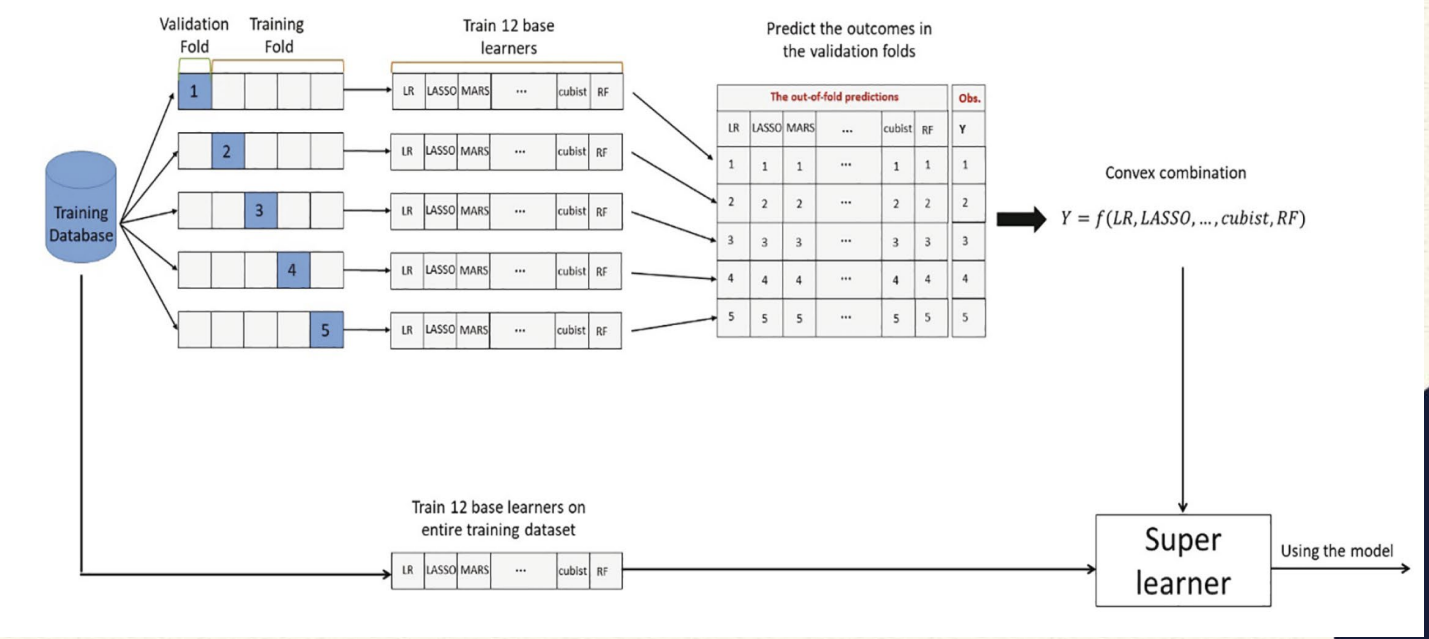
Models	Definition	Hyperparameters
LR	Multi-linear regression	*none*
LASSO	The least absolute shrinkage and selection operator	lambda
MARS	The multivariate adaptive regression splines	nprune, degree
kNN	K-nearest neighbor	k
SVR	Support vector regression	C, σ, kernel
GP	Genetic programming	operators
ANN	Artificial neural network	decay, size
ANFIS	Adaptive-network-based fuzzy inference system	number and shape of the membership function
Cubist	Cubist	committees, neighbors
RF	Random Forest	mtry, ntree
ET	Extremely randomized trees	K, M
XGBoost	Extreme gradient boosting	e.g., booster, max_depth, subsample, eta
AvEqC	Equal weight combiner	*none*
Super learning	Super learning strategy	weights

Taghizadeh-Mehrjardi et al. (2021)

DSM WITH SUPERLEARNER- A CASE STUDY

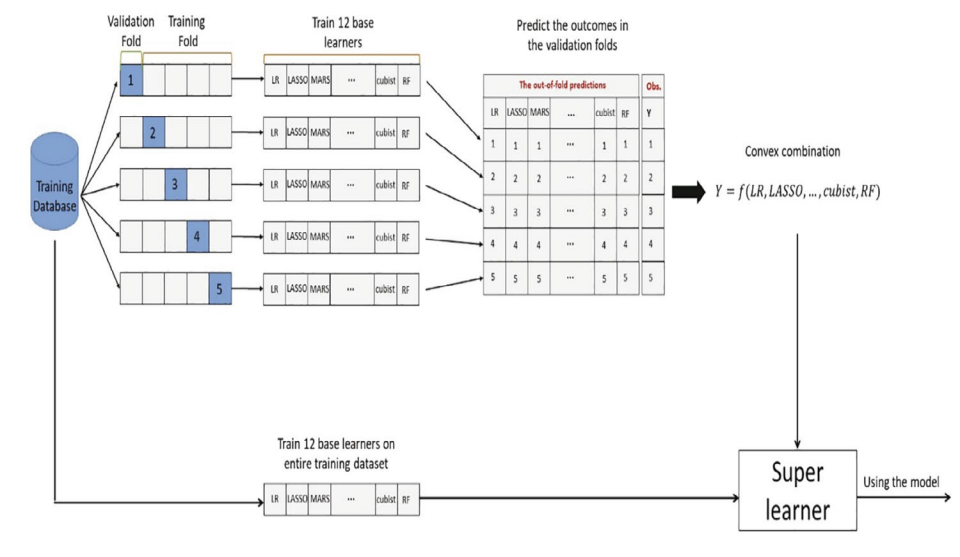
SuperLearner is an algorithm that uses cross-validation to estimate the performance of multiple machine learning models, or the same model with different settings. It then creates an optimal weighted average of those models, aka an "ensemble", using the test data performance

DSM WITH SUPERLEARNER- A CASE STUDY



Taghizadeh-Mehrjardi et al. (2021)

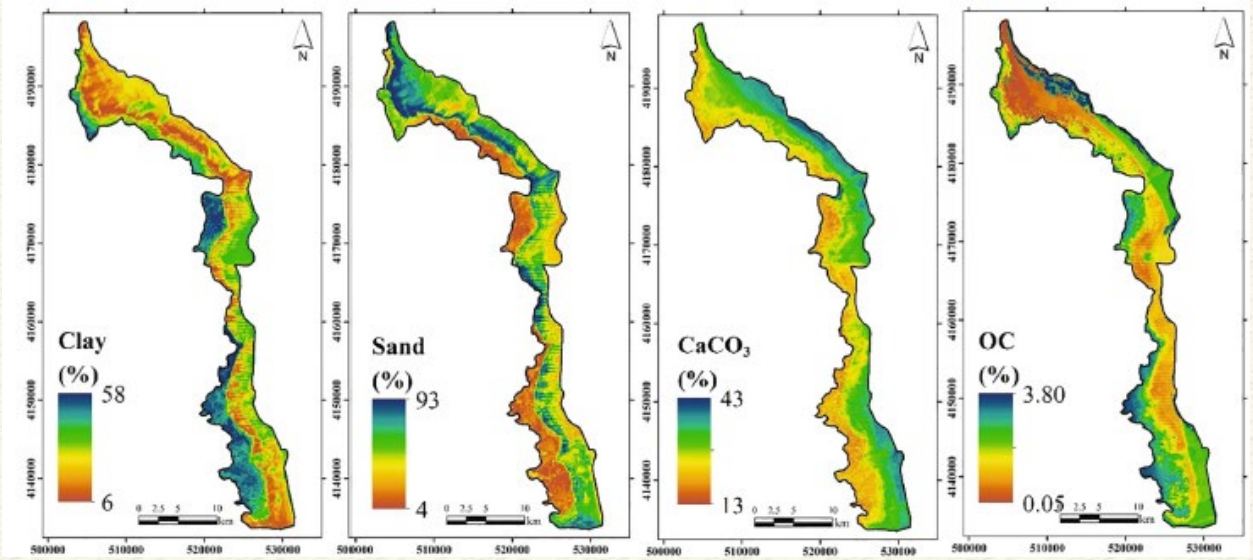
DSM WITH SUPERLEARNER- A CASE STUDY



1) Splitting the training datasets into five folds
2) Training the 12 base learners
3) Storing the out-of-fold predictions
4) Evaluating the model using the out-of-fold dataset
5) Fitting a meta-model on the out of- fold predictions to extract the weights
6) Fitting the base learners on the full training dataset and storing the predictions; and
8) Combining these predictions using the estimated weights

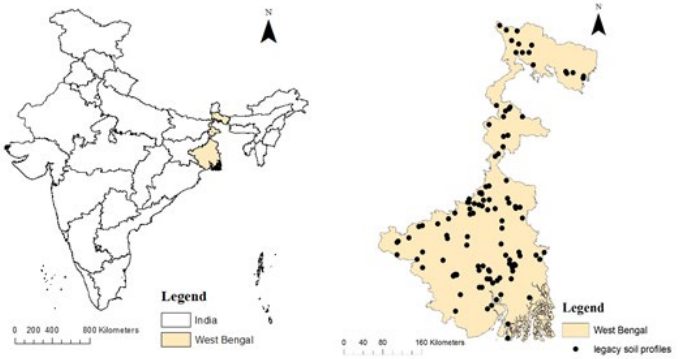
Taghizadeh-Mehrjardi et al. (2021)

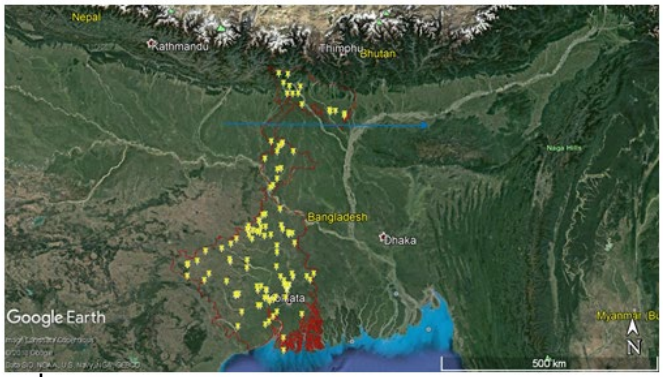
DSM WITH SUPERLEARNER- A CASE STUDY



Taghizadeh-Mehrjardi et al. (2021)

DSM WITH RK- A CASE STUDY





Terrain attributes

Slope

Altitude above channel

Hillshade

Aspect

Profile curvature

Plan curvature

Terrain Ruggedness Index

Topographic Wetness Index

Elevation

Bioclimatic variables

Annual mean temperature

Annual precipitation

Temperature seasonality

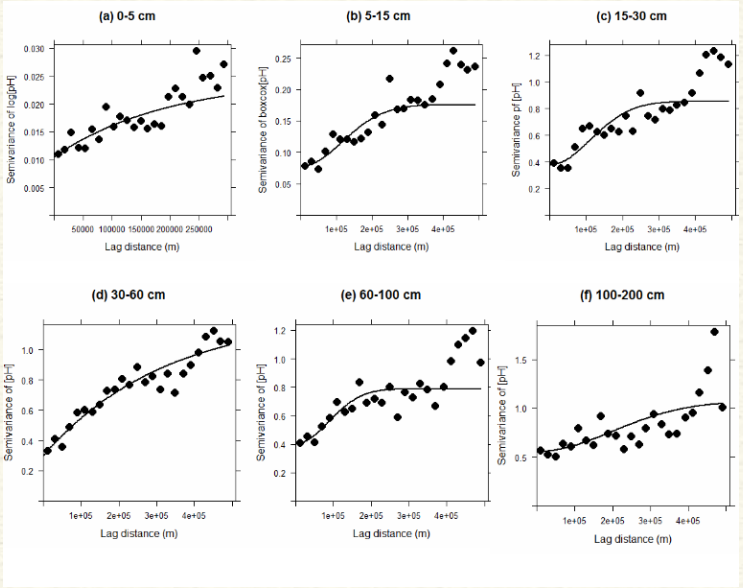
Rainfall seasonality

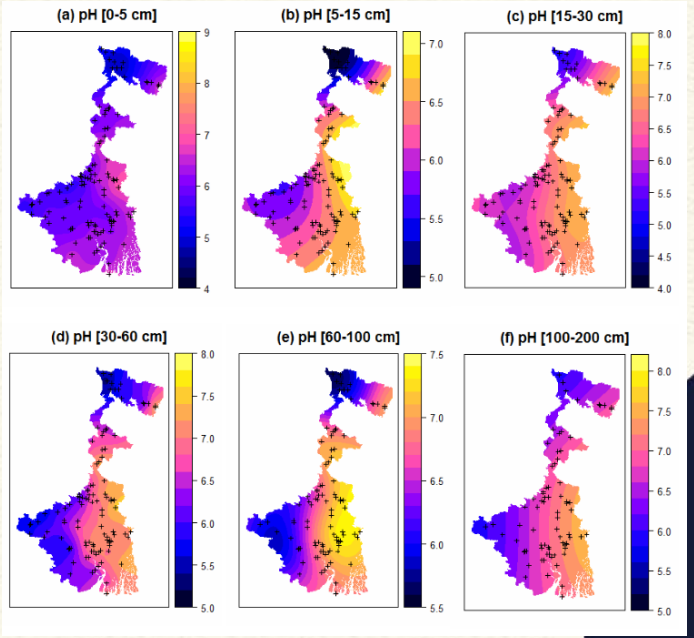
Mean diurnal range of temperature

Annual range of temperature

Rainfall of wettest quarter

DSM WITH RK- A CASE STUDY





THANKS!

Subhadip Dasgupta

Madan Jatiya

REFERENCES

Taghizadeh-Mehrjardi, R., Hamzehpour, N., Hassanzadeh, M., Heung, B., Goydaragh, M. G., Schmidt, K., & Scholten, T. (2021). Enhancing the accuracy of machine learning models using the super learner technique in digital soil mapping. Geoderma, 399, 115108.

Wadoux, A. M. C., Minasny, B., & McBratney, A. B. (2020). Machine learning for digital soil mapping: Applications, challenges and suggested solutions. Earth-Science Reviews, 210, 103359.


Thank
you