Extension of linear 3D trend models of soil variables

by using penalized interaction models

Milutin Pejović

3D soil mapping

Soil mapping in 3D is a natural extension of commonly used 2D soil mapping.

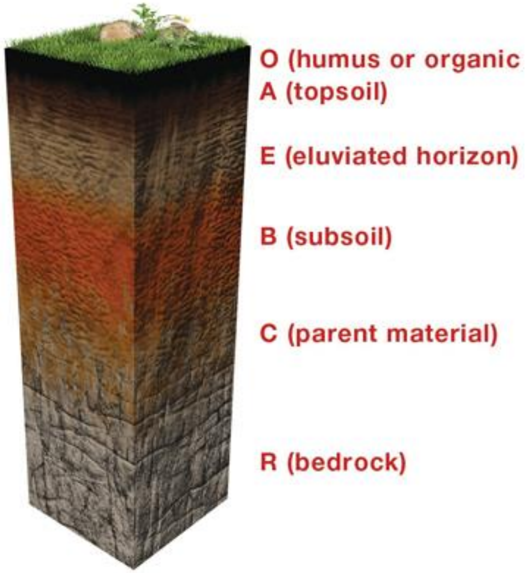
Producing maps of soil properties related to different soil depth.

Soil mapping in 3D space (2D+depth) was recognized as one of the main methodological challenge facing the developers of statistical soil models.

What is needed? 3D soil data

Soil data are often consists of samples collected from many locations at several different depths.

3D data (longitude, latitude and depth)



Environmental variables

(Soil forming factors)

3D trend model of soil variables

Most work in digital soil mapping is based on building a statistical model relating field soil observations and environmental variables.

From a geostatistical point of view, it can be considered as a trend.

This is the point where soil mapping and machine learning techniques are facing each other.

Methodological obstacles?

Lack of environmental covariates known in 3D space largely limits the development of 3D trend models of soil properties.

Environmental variables, like DEM, DEM derivates, satellite images, are relate to soil surface not to depth.

How to find the 3D model which describes both the lateral as well as vertical deterministic variation?

Benchmark model

Linear two-component 3D model

$$\mu(\mathbf{s}, d) = \beta_0 + \underbrace{\sum_{i=1}^{n} \beta_i \mathbf{x}_i(\mathbf{s})}_{\text{Lateral (2D)}} + \underbrace{\beta_{n+1} d}_{\text{Vertical (d)}}$$

Lateral and vertical components have very distinct role in model!

Extension of 3D linear model

Extending the 3D model by:

- *Polynomial expansion of depth term*
- *Inclusion of interactions between spatial covariates and depth*

$$\mu(\mathbf{s}, d) = \beta_0 + \sum_{i=1}^{n} \beta_i \mathbf{x}_i(\mathbf{s}) + \sum_{j=1}^{3} \sum_{i=1}^{n} \theta_{ji} \mathbf{x}_i(\mathbf{s}) d^j$$

Deeper understanding of relationships between environmental variables and modeled variable!

Model is flexible!

What problems occur?

Number of predictors arises by considering the interactions.

Hierarchy principle:

> *interaction effect should have nonzero parameter value only if the both (strong hierarchy) or at least one (weak hierarchy) of the main effects has a nonzero parameter value.*

What problems occur??

Which variables along with associated interactions should be included in the model?

How to ensure the hierarchical principle?

Whether and in what extent the extension of linear 3D trend models improves the two-component 3D model?

Common way to solve it?

To use stepwise regression or best subset selection

To use common t-test to select important main effects and then iteratively to select the important interaction effects

Multi-stages processes computationally very demanding!

LASSO Tibshirani (1996)

Lasso is the computationally attractive one-step approach for parameter estimation and variable selection.

$$\underbrace{\min_{\beta_0,\beta}\left\{\frac{1}{N}\sum_{i=1}^{N}(y_i-\beta_0-x_i^T\beta)^2\right\}}_{\text{Least-squares loss}} \text{ subject to } \underbrace{\sum_{j=1}^{p}|\beta_j|}_{\text{L1 constaint}} \leq t$$

Sparse solution!

LASSO for hierarchical interactions Bien et al. (2013)

Software:

Two R packages were used:

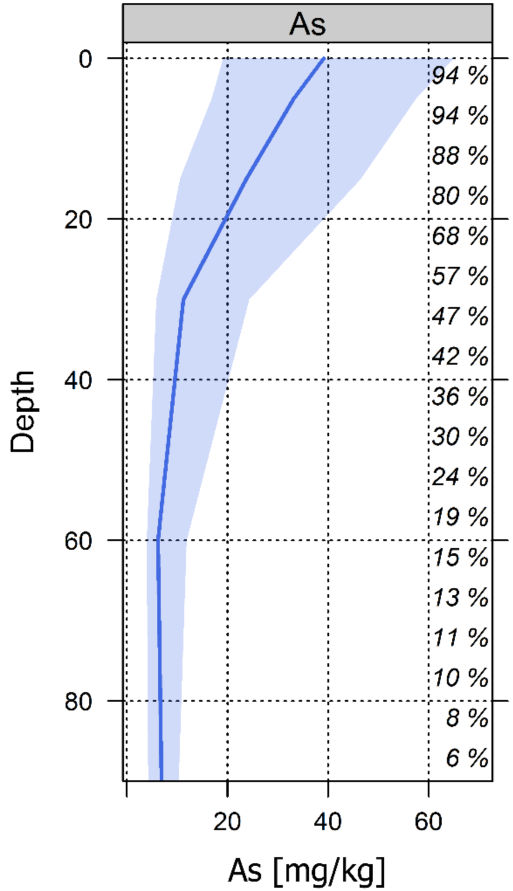
- *glmnet - extremely efficient fitting procedure*
- *hierNet – lasso for hierarchial interactions*

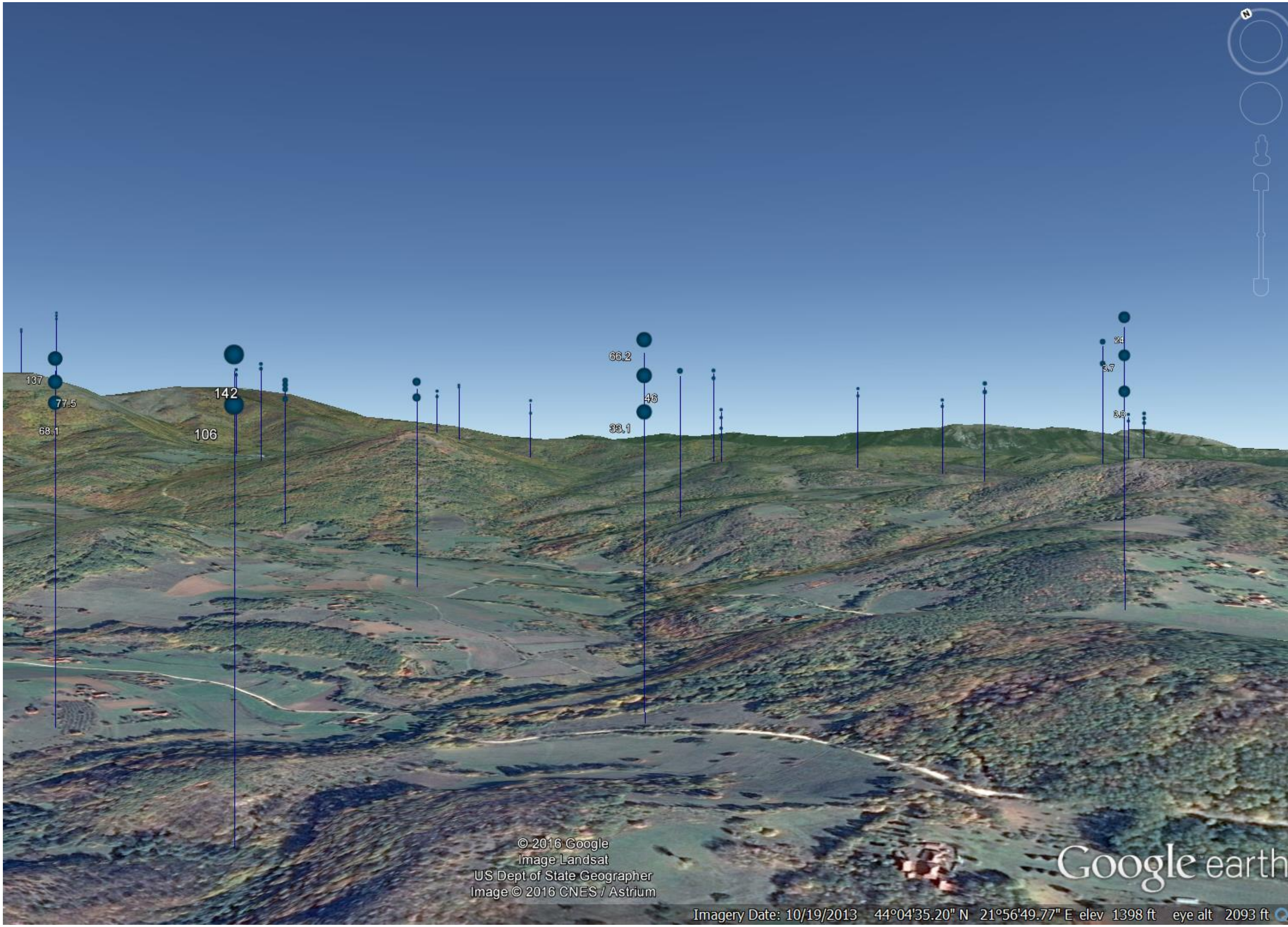
Case study

200 soil profiles with more than 450 observations of Arsenic concentration.

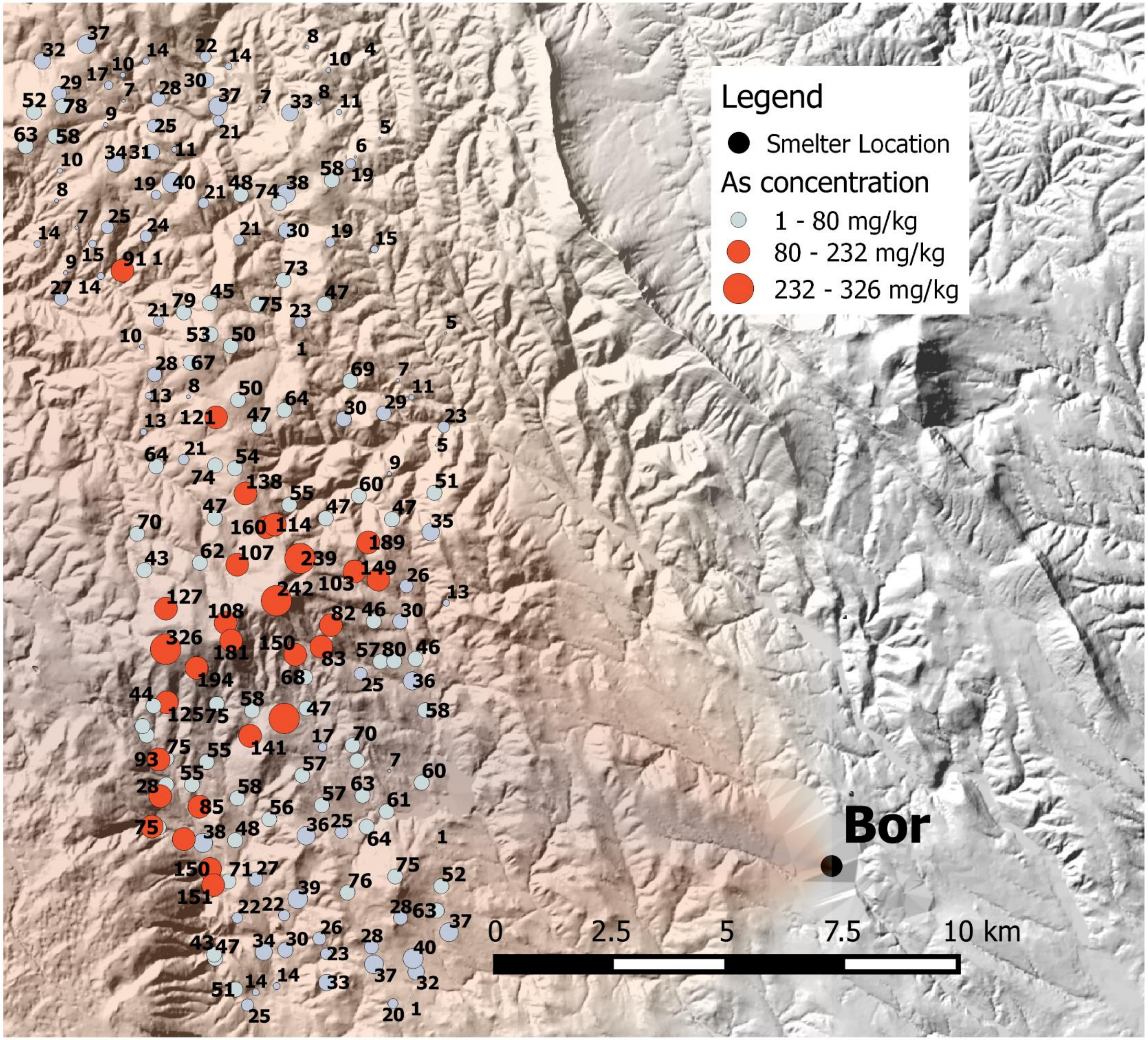
Data are collected in vicinity of copper mining complex in Bor in Serbia

Area: 10x20 km





plotKML (Hengl et.al. 2014)


Legend
Smelter Location
As concentration
1 - 80 mg/kg
80 - 232 mg/kg
232 - 326 mg/kg
Bor
0
2.5
5
7.5
10 km

	Name	Predictor Name	Source	Range	Type
	TERRAIN ATTRIBUTES				
1	Digital Elevation Model	DEM		300-1045	C
2	Aspect	Aspect		0-6.283	C
3	Slope	Slope		1-1.027	C
4	Topographic Wetness Index	TWI		2.077-21.751	C
5	Convergence Index	ConvInd		-97.5-94.4	C
6	Cross Sectional Curvature	CrSectCurv		-0.038-0.04	C
7	Longitudinal Curvature	LongCurv		-0.028-0.04	C
8	Channel Network Base Level	ChNetBLevel		301.2-974.8	C
9	Vertical Distance to Channel Network	VDistChNet		0-281.86	C
10	Negative Openness	NegOp		0.796-1.835	C
11	Positive Openness	PosOp		0.809-1.726	C
12	Wind Effect (East)	WEeast		0.756-1.323	C
13	Wind Effect (North-West)	WEnw		0.749-1.323	C
14	Down-wind Dilution	DD		0.202-0.646	C
15	Cross-wind Dilution	CD		0.389-1	C
16	CORINE LAND COVER 2006				
	Pastures	clc.231		0-1	F
	Complex cultivation patterns	clc.242		0-1	F
	Land principally occupied by agriculture	clc.243		0-1	F
	Broad-leaved forest	clc.311		0-1	F
	Transitional woodland-shrub	clc.324		0-1	F
17	SOIL TYPE				
	Dystric Leptosol	LPdy		0-1	F
	Eutric Leptosol	LPeu		0-1	F
	Mollic Leptosol	LPmo		0-1	F
	Dystric Cambisol	CMdy		0-1	F
	Eutric Cambisol	CMeu		0-1	F
	Calcaric Cambisol	CMca		0-1	F
	Dystric Regosol	RGdy		0-1	F
	Vertisol	VR		0-1	F
18	Depth	d		0-1.25	C

Model selection

Model selection was done through the process of 5-fold cross-validation

The folds were stratified according to the three criteria:

- *Geographical locations*
- *Depth-wise distribution*
- *Range of observed variable*

Model accuracy assessment

Model assessment was done through the process of 5-fold nested cross-validation.

Same sampling strategy

RMSE and R^2 were used as accuracy measures

Results

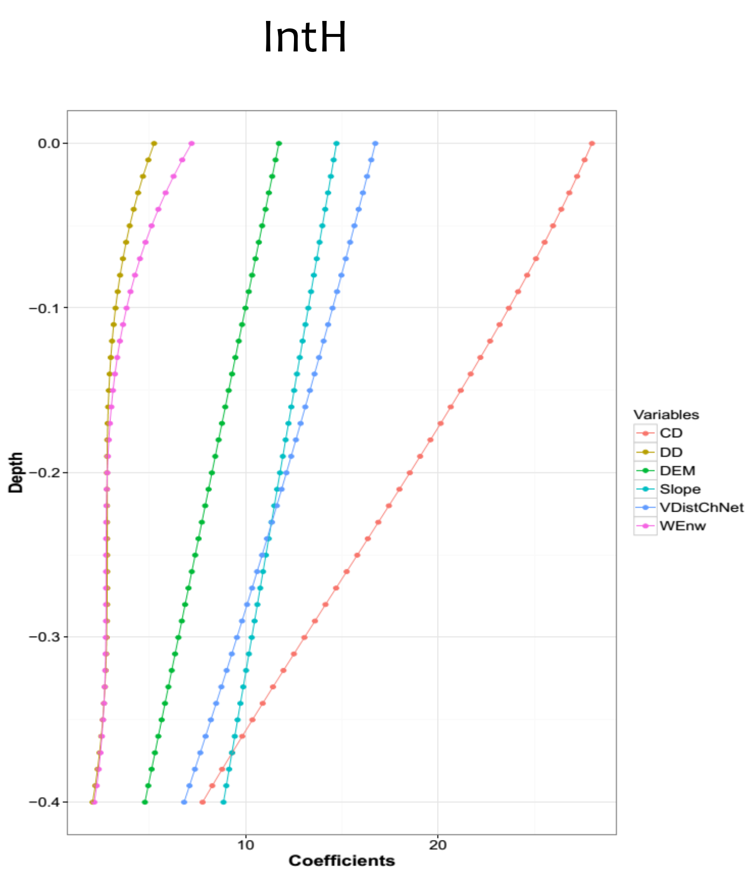
Final model (Int)

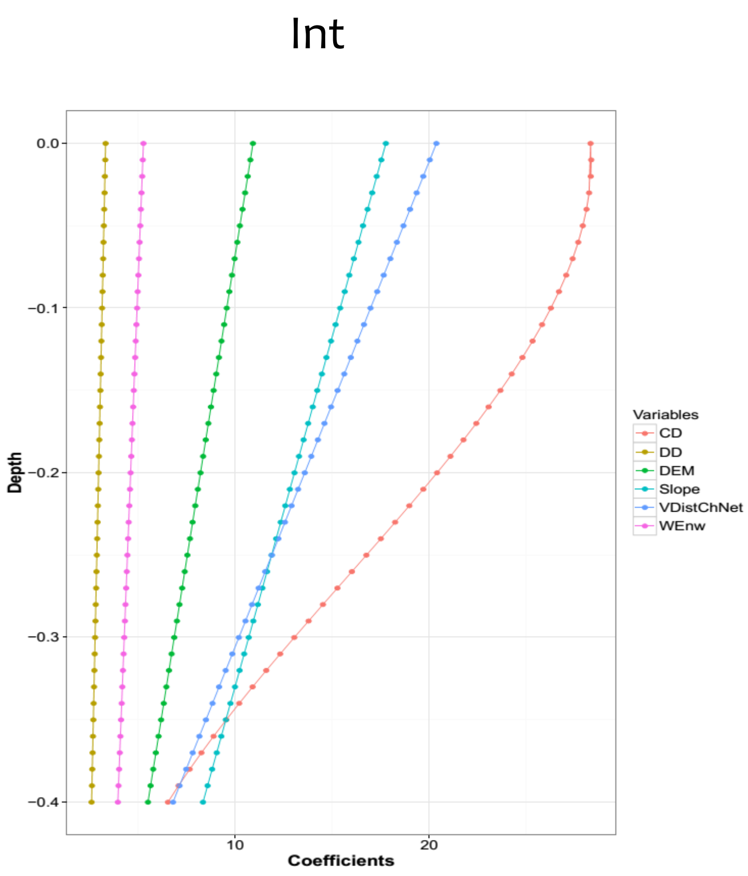
variable	As			
	me	$ie(d)$	$ie(d^2)$	$ie(d^3)$
DEM	7.350	2.996	0	0
Aspect	1.437	3.095	0	0
Slope	11.552	5.218	0	0
TWI	6.710	0.955	0	0
ConvInd	0	-2.762	0	3.625
CrSectCurv	-3.987	-1.052	0	0
LongCurv	0	0	0	0
ChNetBLevel	0	0	0	0
VDistChNet	11.419	7.516	0	0
NegOp	-0.328	-1.045	0	0
PosOp	0	0	0	1.785
WEEast	-1.967	0	0	0
WEnw	4.404	0.730	0	0
DD	2.848	0.401	0	0
CD	15.735	16.558	0	-4.231
clc.231	-22.685	0	1.793	-4.586
clc.242	0	7.121	0	0
clc.243	-6.779	0	0	0
clc.311	2.289	9.301	0	0
clc.324	23.371	3.701	0	-31.590
CMca	43.280	0	0	-30.626
CMdy	0	0	0	-1.057
LPdy	0	7.883	0	-12.964
RGdy	-35.590	0	0	0
CMeu	-1.836	0	0	-1.653
LPeu	-0.197	0	0	0
LPmo	0	3.151	0	0
VR	8.543	3.554	0	0
d	15.166	0	0	0
d^2	0	0	0	0
d^3	-4.739	0	0	0

Nested cv results

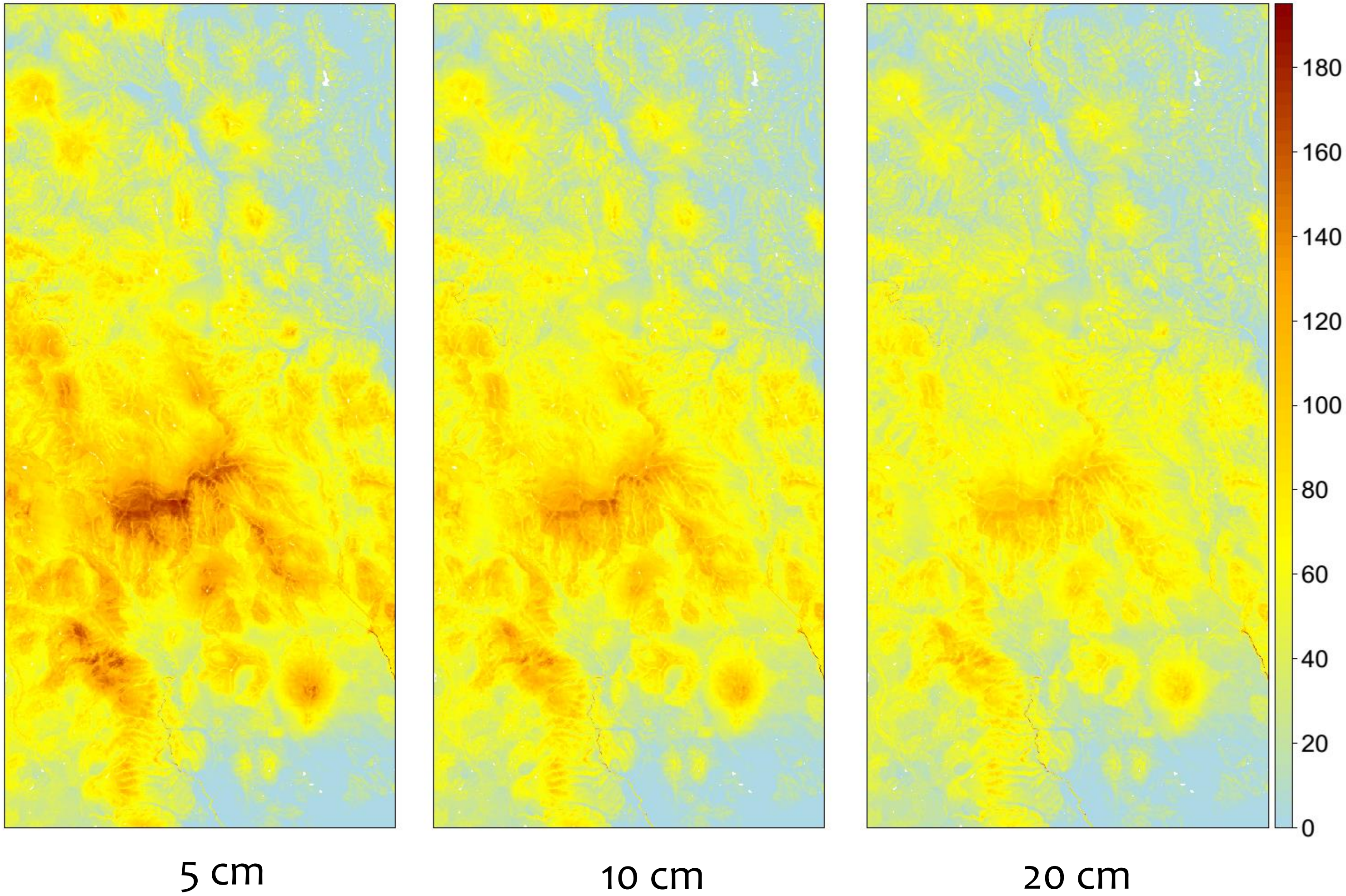
Model	As	
	RMSE	R^2
Base	42.80	0.34
Int	40.67	0.41
IntH	40.92	0.40

Cofficients paths:





Maps:



Thank you