Computers and Geosciences 118 (2018) 1–13

Contents lists available at ScienceDirect

Computers and Geosciences

journal homepage: www.elsevier.com/locate/cageo

Sparse regression interaction models for spatial prediction of soil properties in 3D

Milutin Pejović[a], Mladen Nikolić[b], Gerard B.M. Heuvelink[c], Tomislav Hengl[d], Milan Kilibarda[a], Branislav Bajat[a,*]

[a] Department of Geodesy and Geoinformatics, Faculty of Civil Engineering, University of Belgrade, Bulevar kralja Aleksandra 73, Belgrade, Serbia
[b] Faculty of Mathematics, University of Belgrade, Studentski Trg 16, 11000, Belgrade, Serbia
[c] ISRIC — World Soil Information, Wageningen, the Netherlands
[d] Envirometrix Ltd, Wageningen, the Netherlands

ARTICLE INFO

Keywords:
Spatial prediction
Lasso
Interactions
Nested cross-validation
Soil organic carbon
3D

ABSTRACT

An approach for using lasso (Least Absolute Shrinkage and Selection Operator) regression in creating sparse 3D models of soil properties for spatial prediction at multiple depths is presented. Modeling soil properties in 3D benefits from interactions of spatial predictors with soil depth and its polynomial expansion, which yields a large number of model variables (and corresponding model parameters). Lasso is able to perform variable selection, hence reducing the number of model parameters and making the model more easily interpretable. This also prevents overfitting, which makes the model more accurate. The presented approach was tested using four variable selection approaches – none, stepwise, lasso and hierarchical lasso, on four kinds of models – standard linear model, linear model with polynomial expansion of depth, linear model with interactions of covariates with depth and linear model with interactions of covariates with depth and its polynomial expansion. This framework was used to predict Soil Organic Carbon (SOC) in three contrasting study areas: Bor (Serbia), Edgeroi (Australia) and the Netherlands. Results show that lasso yields substantial improvements in accuracy over standard and stepwise regression — up to 50 % of total variance. It yields models which contain up to five times less nonzero parameters than the full models and that are usually more sparse than models obtained by stepwise regression, up to three times. Extension of the standard linear model by including interactions typically improves the accuracy of models produced by lasso, but is detrimental to standard and stepwise regression. Regarding computation time, it was demonstrated that lasso is several orders of magnitude more efficient than stepwise regression for models with tens or hundreds of variables (including interactions). Proper model evaluation is emphasized. Considering the fact that lasso requires meta-parameter tuning, standard cross-validation does not suffice for adequate model evaluation, hence a nested cross-validation was employed. The presented approach is implemented as publicly available sparsereg3D R package.

1. Introduction

The immense progress in the field of Earth observation and computer power has yielded a huge number and great variety of freely available environmental variables at high spatial resolution. Their inclusion in spatial statistical models can contribute to a deeper understanding of the complex relationship between target variables and the environment and hence to producing more accurate predictions. Linear models are still regarded useful for this purpose, mainly due to their low computational complexity and high interpretability. However, models with a large number of explanatory variables are harder to interpret. In addition, they have many parameters and are therefore more prone to overfitting, which may lead to lower prediction accuracy (Guyon and Elisseeff, 2003; Burnham and Anderson, 2003).

The problem of too many variables is even more pronounced if different model extensions, like interactions or polynomial expansion, are considered. In linear models, interactions are typically represented as element-wise products of explanatory variables. Therefore, their inclusion dramatically increases the size of the design matrix. Nevertheless, interactions are very useful when there is a need for

* Corresponding author. University of Belgrade, Faculty of Civil Engineering, Bulevar kralja Aleksandra 73, 11000, Belgrade, Serbia. Tel.: +381 11 3218579; fax: +381 11 3370 223.
E-mail addresses: mpejovic@grf.bg.ac.rs (M. Pejović), nikolic@matf.bg.ac.rs (M. Nikolić), gerard.heuvelink@wur.nl (G.B.M. Heuvelink), tom.hengl@envirometrix.net (T. Hengl), kili@grf.bg.ac.rs (M. Kilibarda), bajat@grf.bg.ac.rs (B. Bajat).

https://doi.org/10.1016/j.cageo.2018.05.008
Received 8 August 2017; Received in revised form 29 March 2018; Accepted 8 May 2018
Available online 14 May 2018
0098-3004/

modeling the joint effect of explanatory variables, while the polynomial expansion of explanatory variables is a common approach to modeling nonlinear relations.

To improve interpretability and resist overfitting, variable selection is often performed in regression-based environemental modeling studies (Hengl et al., 2004; Hoeting et al., 2006; Huang and Chen, 2007). Stepwise linear regression is, by far, the most extensively used technique for this purpose (Mueller and Pierce, 2003; Florinsky et al., 2002; Goetz et al., 2015). However, it employs greedy strategy that usually yields suboptimal solutions. Also, interaction models are often subject to condition that interaction effect can be included in the model only if both (*strong hierarchy condition*), or at least one (*weak hierarchy condition*) individual (main) effect is also included in the model (Cox, 1984). One way to avoid these pitfalls would be to use *lasso* (Least Absolute Shrinkage and Selection Operator) regularization for linear regression (Tibshirani, 1996). Lasso yields *sparse models*, with only a subset of nonzero coefficients by solving convex minimization problem, for which efficient techniques are available.

In a recent study by Fitzpatrick et al. (2016), lasso was successfully used for the modeling 2D spatial distribution of soil organic carbon in a case study where the number of predictors exceeds the number of available observations. Nussbaum et al. (2017) have also recently compared one lasso variant, called group lasso (which tends to set all parameters within each predefined group to zero, instead of setting them to zero independently), with other popular techniques for digital soil mapping. The obtained results showed that lasso competes well with other methods that are more complex. In addition, unlike the other methods, group lasso also showed the ability to avoid overfitting. Liddicoat et al. (2015) used lasso for modeling the deterministic component within regression kriging framework in a study of mapping topsoil organic carbon stock for a large study area in South Australia.

This study presents an approach of using lasso for 3D spatial prediction of soil properties. 3D spatial prediction involves fitting a regression model that relates soil observations from various spatial locations and soil depths to environmental (lateral) variables and depth as covariates. Considering that standard linear models include only additive terms for lateral and depth covariates, they cannot address the fact that the influence of lateral covariates varies by depth or that the influence of depth varies laterally. For that reason, the interactions between lateral covariates and depth terms should be included in the model (Orton et al., 2016; Brus et al., 2016). A large number of potentially useful predictors, among others caused by including interactions, makes the model selection problem more challenging. The presented approach involves hierarchical and non-hierarchical lasso procedures for model selection. Fitting a linear model using lasso necessarily implies tuning meta-parameters using cross-validation. In order to avoid overoptimistic accuracy assessment results as obtained with standard cross-validation, this study employs nested cross-validation — an error estimation procedure suited for cross-validation based model selection (Krstajic et al., 2014). The presented approach is implemented as R package sparsereg3D, which is available via GitHub.[1]

Observations of Soil Organic Carbon (SOC, [‰]) taken from three case studies, Bor (Serbia), Edgeroi (Australia) and the Netherlands were used to test the approach and to perform comparison with commonly used linear regression methods i.e., ordinary least squares (OLS) and stepwise linear regression. The methods were compared and evaluated by fitting four models that differ in whether they take the interactions between lateral covariates and depth into account or whether the polynomial expansion of depth is included, or both. In this way, several specific issues were explored: (1) how lasso influences model accuracy and how it compares to OLS and stepwise regression, (2) how lasso influences model sparsity and how it compares to OLS and stepwise regression, (3) how extending the model, through inclusion of interactions and polynomial expansion of depth, influences both the accuracy and the sparsity of the model, and (4) how demanding which of the considered methods is with respect to training time.

Compared to Fitzpatrick et al. (2016) and Nussbaum et al. (2017), we deal with 3D models instead of 2D models and take care of proper evaluation in presence of meta-parameters using nested cross-validation. In comparison to Fitzpatrick et al. (2016), we use three different case studies of much larger size and we consider hierarchical lasso. Compared specifically to Nussbaum et al. (2017), we consider interactions with or without hierarchy by using classical and hierarchical lasso instead of group lasso. In contrast to Liddicoat et al. (2015), our study deals specifically with lasso (and not with regression kriging), but in 3D setting instead of 2D. Also, our study is comparative in nature and analyzes the effects of interactions, polynomial depth expansion, and hierarchical constraints on model accuracy and model sparsity.

2. Materials and methods

2.1. Linear regression by ordinary least squares

Linear regression with ordinary least squares (OLS) is the most popular approach for predicting a continuous target variable y on the basis of a group of predictor variables (predictors) x_j, for $j = 1, ...,p$. The linear regression model is defined as:

$$y = \mu + \varepsilon = \beta_0 + \sum_{j=1}^{p} \beta_j x_j + \varepsilon \tag{1}$$

where the β_j, for $j = 0, ...,p$ are unknown parameters or coefficients, and where $\varepsilon \sim N(0, \sigma^2)$.

In ordinary least squares, which is the basic method for model estimation, the coefficients are estimated by minimizing the so called *residual sum of squares* (RSS):

$$RSS(\beta) = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \tag{2}$$

or in matrix formulation:

$$RSS(\beta) = (\mathbf{y} - \mathbf{X}\beta)^{\mathbf{T}}(\mathbf{y} - \mathbf{X}\beta) \tag{3}$$

where n is the number of observations, $\mathbf{y}$ is the vector of observations of the dependent variable, β is a vector of model coefficients, and $\mathbf{X}$ is an $n \times (p + 1)$ design matrix. Each row of $\mathbf{X}$ is an observation vector x_j extended by value 1, corresponding to β_0. The unique minimum of RSS (Equation (3)) is given by $\hat{\beta} = (\mathbf{X^T X})^{-1}\mathbf{X^T y}$ if matrix $\mathbf{X^T X}$ is invertible.

2.2. Stepwise linear regression

Stepwise linear regression is a procedure that performs a greedy search in the space of OLS models over different subsets of the set of predictors, by sequential adding (or alternative variant removing) best single predictor at a time to the model while its accuracy is improving. Clear explanation of stepwise regression is provided in the textbook by James et al. (2013).

2.3. Linear regression by lasso

Lasso is a regularization regression method that yields a sparse linear model (Tibshirani, 1996). A model is sparse if it includes only a subset of the explanatory variables. The smaller the subset, the more sparse the model is. In order to achieve this, lasso combines a well-known least squares loss function with ℓ_1 regularization via $\sum_{j=1}^{p} |\beta_j|$ norm of coefficients. Therefore, the coefficients of lasso regression are

[1] https://github.com/pejovic/sparsereg3D.

the solution of the following optimization problem (Hastie et al., 2015):

$$\min_{\beta}\{\|\mathbf{y}-\mathbf{X}\beta\|_2^2+\lambda\|\beta\|_1\} \quad (4)$$

where y_i, $i=1,\ldots,n$ are observations of the dependent variable, x_{ij}, $i=1,\ldots,n$, $j=1,\ldots,n$ is a vector of explanatory values and β_i, $i=1,\ldots,p$ is a vector of model coefficients. Note that the regularization is applied to the absolute values of the regression coefficients, which indicates that for the method to produce sensible results it is required that the explanatory variables are standardized (i.e. the mean is subtracted from each observation and the result is divided by the standard deviation) prior to lasso regression.

Meta-parameter λ (*regularization* parameter) controls the strength of the ℓ_1 penalty. For large λ, more coefficients are forced to be close to or equal to zero, while for sufficiently small λ, lasso coefficients are close to their least squares estimates. The effect of lasso regularization is two-fold. First, by forcing some coefficients to be equal to zero, lasso retains only a subset of variables and hence improves the model interpretability. Second, regularization may significantly improve prediction accuracy. The rationale behind this lies in bias-variance trade off (Hastie et al., 2015).

In the case of models obtained by lasso, standard estimates of p-values and confidence intervals are biased (Taylor and Tibshirani, 2015). Actual p-values should be more conservative and confidence intervals should be wider. Recently, methods which address this issue have been developed by Tibshirani et al. (2016, 2017).

2.3.1. Lasso for hierarchical interactions

In this study, only strong hierarchy was considered. This means that the interaction effects can have a nonzero parameter value only if both main (individual) effects have nonzero parameter values. This was achieved by using the approach proposed by Bien et al. (2013) and implemented in the hierNet R package (Bien and Tibshirani, 2014). This approach involves extending the linear model by interactions:

$$f(\beta,\theta)=\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}-\frac{1}{2}\sum_{j\neq k}\theta_{jk}x_{ij}x_{ik}\right)^2$$

where θ is a $p\times p$ matrix of interaction coefficients θ_{jk}, such that $\theta_{jj}=0$, $j=1,\ldots,p$. Also, ℓ_1 regularization is added and a set of convex constraints to ensure that strong hierarchy holds (Bien et al., 2013) (which relies on vectors β^+ and β^- of auxiliary coefficients):

$$\begin{aligned}
&\min_{\beta^+,\beta^-,\theta} \quad f(\beta^+-\beta^-,\theta)+\lambda\textstyle\sum_{i=1}^{p}(\beta_i^++\beta_i^-)+\frac{\lambda}{2}\sum_{i=1}^{p}\sum_{j=1}^{p}|\theta_{ij}|\\
&\text{subject to} \quad \theta=\theta^T\\
&\theta_{jj}=0 \quad \text{for} \quad j=1,\ldots,p\\
&\textstyle\sum_{i=1}^{p}|\theta_{ij}|\le\beta_j^++\beta_j^- \quad \text{for} \quad j=1,\ldots,p\\
&\beta_j^+\ge 0 \quad \text{for} \quad j=1,\ldots,p\\
&\beta_j^-\ge 0 \quad \text{for} \quad j=1,\ldots,p
\end{aligned}$$

2.4. Accuracy measures

Two common measures of model accuracy, or error, are the root mean square error (RMSE) and coefficient of determination (R^2). RMSE measures the prediction error in absolute terms and in units in which the predicted value is measured. It is defined by:

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

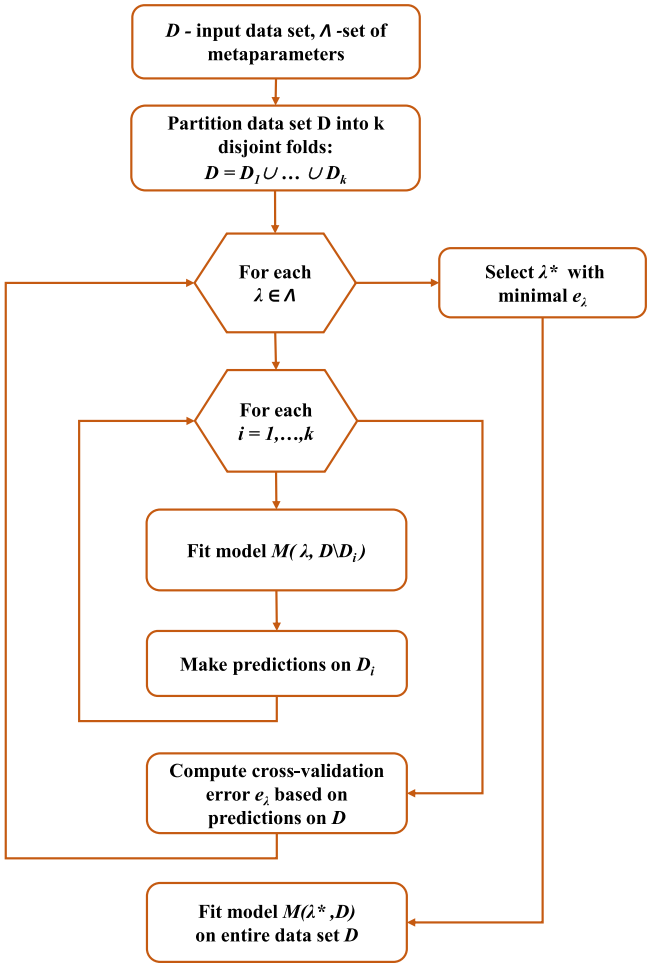

Fig. 1. Flowchart for cross-validation-based model selection procedure.

where $\hat{y}_i$ are the predictions given by the model. R^2 is the fraction of variance of the target variable explained by the model and is defined by:

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

where $\bar{y}$ is the mean of the y_i, $i=1,\ldots,n$.

2.5. Model selection

Considering that different values of λ can produce very different models, the main issue when fitting lasso is to select the optimal value of meta-parameter λ. Ideally, the optimal λ should produce a model with the lowest prediction error. We used *k*-fold cross-validation for selection of the optimal λ, and hence for selection of the optimal model (Arlot et al., 2010). Before cross-validation is applied, it is necessary to define a sequence of values of meta-parameter λ. Then, for each λ value, the error (e.g. RMSE) estimate e_λ is computed. The optimal λ value is the one which gives the lowest e_λ. The optimal value of λ is then used to fit the final model on the entire data set. Fig. 1 depicts the model selection procedure via *k*-fold cross-validation.

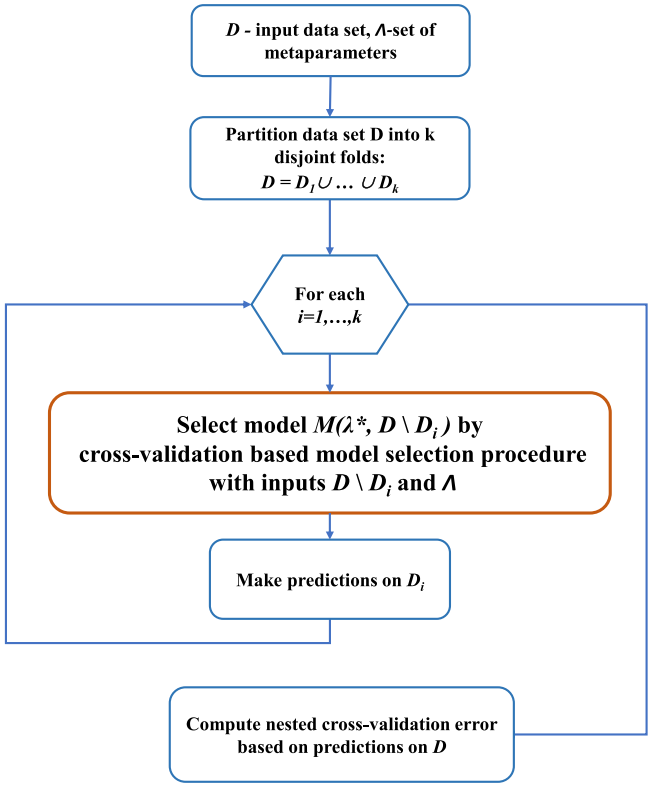

Fig. 2. Flowchart for model evaluation via nested *k*-fold cross-validation.

2.6. Model assessment

Standard *k*-fold cross-validation is often used for assessing the prediction error of a model; see e.g. Goovaerts (2001); Chung and Fabbri (2008); Robinson and Metternicht (2006); Arlot et al. (2010). However, for methods that require the optimization of meta-parameter values like lasso, an error estimate obtained during the meta-parameter selection process via standard cross-validation will produce an overoptimistic estimate of the true prediction error (Krstajic et al., 2014). This is because the choice of meta-parameter values determines the model and since the whole data set was used to select the meta-parameter values, it follows that the whole data set is used to train a model. In contrast to that, objective predictive model assessment requires that training and test data sets are separated in order to provide a fair estimate of the error.

In order to avoid such an overoptimistic estimate of model performance, we used the nested cross-validation technique as described in Krstajic et al. (2014). In short, nested cross-validation consists of two nested cross-validation loops. The outer loop serves to assess the performance of the model, which was selected in the inner cross-validation loop. For each outer fold, the model is selected on each outer training set using a standard cross-validation procedure and then applied to the outer test set. The process yields predictions for each outer fold, which are obtained from a model that was not trained on that fold. The prediction error estimate is obtained based on all obtained predictions (i.e. from the collation of outer fold cross-validation results). The nested cross-validation procedure is presented by Algorithm 1 and graphically in Fig. 2.

Algorithm 1. Nested *k*-fold cross-validation.

Input: Data set D, meta-parameter values Λ, number k

Output: Error estimate

Partition D into stratified sets D_i, $i = 1,\ldots,k$ of approximately equal size
for $i = 1$ to k **do**
  Let D' be $D \setminus D_i$
  for each $\lambda \in \Lambda$ **do**
    Partition D' into stratified sets D'_j, $j = 1,\ldots,k$ of approximately equal size
    for $j = 1$ to k **do**
      Fit the model $M(\lambda, D' \setminus D'_j)$
      Make predictions by $M(\lambda, D' \setminus D'_j)$ on D'_j
    end for
    Compute error e_λ based on predicted and real target values on D'
  end for
  Let λ^* be $\arg\min_{\lambda \in \Lambda} e_\lambda$
  Fit the model $M(\lambda^*, D')$
  Make predictions by $M(\lambda^*, D')$ on D_i
end for
Report error computed on predicted and real target values on D

Note that the standard cross-validation is a model selection procedure that, on its own, fails to provide adequate model assessment (Stone, 1974; Geisser, 1975; Varma and Simon, 2006; Krstajic et al., 2014). On the other hand, nested cross-validation is a model assessment procedure which estimates the prediction error of the model selected by cross-validation. Nested cross-validation, does not provide a model (it evaluates the performance of k approximate models in order to yield its estimate). As result, both procedures need to be used. In addition, numerous models trained in nested cross-validation could also be used to produce prediction uncertainty maps, as was done by Liddicoat et al. (2015).

3. Application to 3D spatial prediction

3.1. 3D models of soil properties

Soil samples are typically collected from multiple layers or horizons within soil profiles. Accordingly, measurements of soil properties typically represent averages over soil horizons and come with meta-data that describe their position in 3D space: 2D coordinates of the sampling location as well as the upper and lower bounds of the sampling soil horizon. If the depth interval is replaced by the horizon mid-point, soil observations have a 3D point coordinates and can be regressed against the depth and a set of environmental variables that are known for the entire sampling area. It should be noted that the actual soil observations are not point values but averages over the given depth interval. For a statistical approach that takes the specific support size of the observations into account we refer to Orton et al. (2016). In this study, we treat the soil observations as point support observations.

One common formulation that relates the soil property to depth and environmental covariates is a multiple linear regression model (Hengl et al., 2014):

$$\text{MLR: } \mu(\mathbf{s}, d) = \beta_0 + \sum_{i=1}^{p} \beta_i \mathbf{x}_i(\mathbf{s}) + \beta_{p+1} d \tag{5}$$

where $\mathbf{x}(\mathbf{s})$ represents a vector of covariates at sampling location $\mathbf{s}$, d represents depth, and the β_i, $i = 0...p + 1$ are model coefficients. In order to allow more flexibility in the depth term of MLR model (Equation (5)), it can be generalized to more complex functions of depth, such as a polynomial or a piece-wise polynomial (e.g., spline function). In this study we used a third-degree polynomial expansion of depth (MPR, Equation (6)):

$$\text{MPR: } \mu(\mathbf{s}, d) = \beta_0 + \sum_{i=1}^{p} \beta_i \mathbf{x}_i(\mathbf{s}) + \sum_{i=1}^{3} \beta_{p+i} d^i \tag{6}$$

Further extensions include interactions between lateral covariates and depth:

$$\text{MLRI: } \mu(\mathbf{s}, d) = \beta_0 + \sum_{i=1}^{p} \beta_i \mathbf{x}_i(\mathbf{s}) + \beta_{p+1} d + \sum_{i=1}^{p} \theta_i \mathbf{x}_i(\mathbf{s}) d \tag{7}$$

and

$$\text{MPRI: } \mu(\mathbf{s}, d) = \beta_0 + \sum_{i=1}^{p} \beta_i \mathbf{x}_i(\mathbf{s}) + \sum_{j=1}^{3} \sum_{i=1}^{p} \theta_{ji} \mathbf{x}_i(\mathbf{s}) d^j \tag{8}$$

where the θ_i and θ_{ji} are interaction coefficients. The MLRI model includes interactions between lateral covariates and depth, while the MPRI model includes additional interactions with polynomial (quadratic and cubic) depth terms. In this way, we obtained four models of increasing complexity (and flexibility), which is reflected in an increased number of parameters: $p + 2$ for MLR model, $p + 4$ for the MPR, $2p + 2$ for MLRI and $4p + 1$ for the MPRI model.

3.2. Case studies

To illustrate the presented methodology, we use three case studies on modeling soil organic carbon (SOC): Bor (Serbia), Edgeroi (Australia) and the Netherlands (NL). All three case studies have been previously described in the literature (listed below).

The data sets comprise soil profile observations accompanied with a list of explanatory variable layers. Table 1 reports descriptive statistics along with the numbers of observations, profiles and NA (not available/missing) observations. All observations containing NAs were excluded from further analysis. An expressed skewed distribution of SOC, reflected in the skewness coefficient, can be noticed for all case studies. To stabilize variance and improve the modeling outcome, we used log-transformed SOC as the target variable in this study. Fig. 3 shows the distribution of log (SOC) measurements by depth. As is noticeable, a typical decreasing trend with depth of log-transformed SOC can be observed for all case studies. Note also that Edgeroi has observations up to 8 m of depth, the Netherlands up to 4 m and Bor up to 1.3 m. Fig. 4 depicts the spatial extent of each case study along with the locations of soil profiles.

3.2.1. Bor

The Bor study area is the smallest but has relatively the most diverse topography. It comprises 206 soil profiles that were randomly sampled over the 10×20 km area in central Serbia (a few kilometers north-east from the town of Bor). For this case study, we use thirteen continuous and two categorical variables as environmental covariates. These include:

- 20 m resolution DEM derived from a 1:25,000 scale topographic map.
- DEM derivatives including: Aspect, Topographic Wetness Index, Slope, Curvature-planar and cross-sectional, Channel Network Base Level, Convergence Index, Vertical Distance to Channel Network

Table 1
Descriptive statistics of SOC and log (SOC) for all case studies.

Case study	Variable	Min	1st quart.	Mean	Median	3rd quart.	Max	St.Dev.	Skew.	Profiles	Obs.	NAs
Bor	SOC	0.64	12.02	28.41	21.00	35.28	234.74	26.77	2.8	206	458	13
	log (SOC)	0.49	2.57	3.09	3.08	3.59	5.46	0.78	− 0.1	206	458	13
Edgeroi	SOC	0.10	1.70	6.75	5.10	9.30	82.70	6.94	2.56	359	2,089	249
	log (SOC)	0.09	0.99	1.81	1.71	2.33	4.43	0.83	0.01	359	2,089	249
NL	SOC	0	4.64	66.54	13.92	38.86	579.47	131.45	2.52	643	2,617	0
	log (SOC)	0	1.73	2.70	2.86	3.68	6.36	1.60	0.46	643	2,617	0

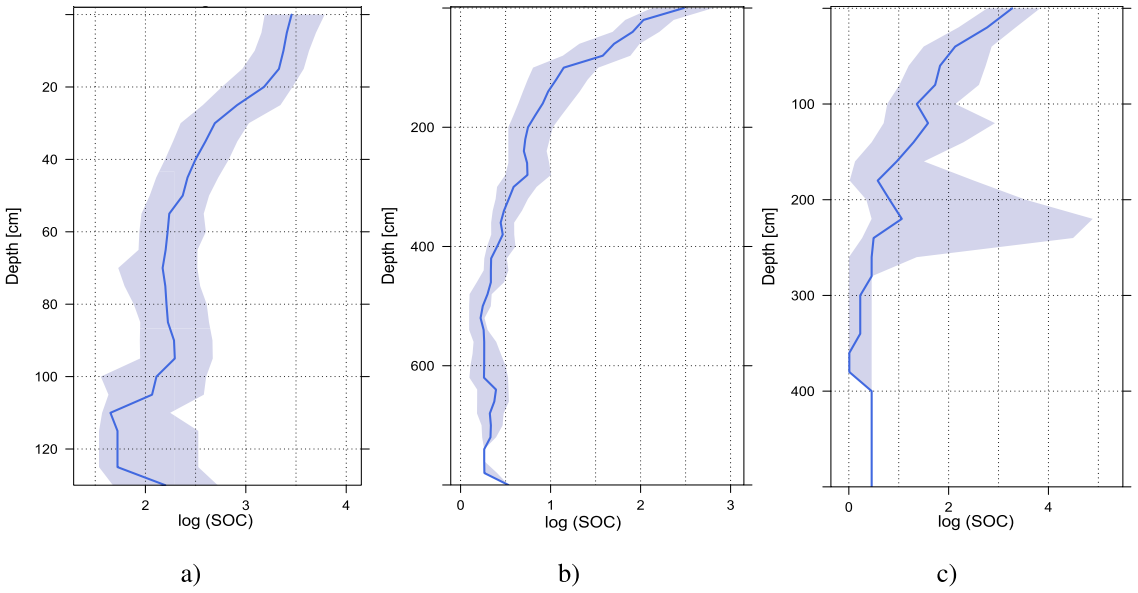

Fig. 3. Aggregated distribution of log (SOC) for the three case studies: a) Bor; b) Edgeroi; c) The Netherlands. Median value (bold line) bounded by 25th and 75th percentiles (blue area). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

(Hengl and Reuter, 2008).
- A soil type map of the area with eight classes according to the World Reference Base classification (Michéli et al., 2006).
- Corine Land Cover map (CLC) with five classes (Nestorov et al., 2007).

The Bor case study is described in Pejović et al. (2017).

3.2.2. Edgeroi

The Edgeroi case study has been previously described in McGarry et al. (1989); Minasny et al. (2006); Malone et al. (2009). The Edgeroi data set (as used in this study) is available via the GSIF package for R (http://gsif.r-forge.r-project.org/edgeroi.html). The data set comprises 359 soil profiles and 15 covariates (13 continual and 2 categorical).

3.2.3. The Netherlands

The Netherlands data set is described in detail in Kempen (2011). Maps are available for viewing and download from http://maps.bodemdata.nl (the Dutch Soil Information System). The covariates include:

- 250 m resolution DEM (based on SRTMGL3).
- Stack of MODIS EVI images for six periods (seasons) (Didan, 2015).
- Soil map of the Netherlands (based on the 1:50,000 map) with a simplified legend with 20 classes (Steur and Heijink, 1991).
- Groundwater level class map with 16 groundwater table classes derived from the 1:50,000 soil map of the Netherlands (Finke, 2000).
- Geomorphology class map with 20 landform classes derived from the 1:50,000 geomorphological map of the Netherlands (Koomen and Maas, 2004).
- Land cover maps for 1970, 1990, and 2004 with simplified legends (Knol et al., 2004; Hazeu, 2005).
- Relative elevation map derived from the 25 m resolution DEM of the Netherlands.

The relative elevation map captures local relief. The map was computed by subtracting the average elevation in the local neighborhood of each pixel (circle with 500 m radius centered on the pixel) from the actual elevation.

3.3. Modeling procedure

Methodologically, this research roughly follows the generic approach proposed by Kanevski (2013) and, in our case, involves following steps:

1. **Data preparation.** Data preparation refers to arrangement of candidate explanatory variables as sequences of regular grids (rasters), spatial overlay according to profile locations and, finally, to the creation of the *design matrix* of n observations with p explanatory variables. In the case of models with polynomial depth function (MPR and MPRI), the design matrix is extended by columns with quadratic and cubic depth terms. In the case of models with interactions (MLRI and MPRI), the design matrix is extended by columns obtained by element-wise multiplication of columns of environmental variables with columns related to depth.
2. **Data preprocessing.** In this step, categorical variables were converted into binary variables using dummy coding. That is, each categorical variable which takes M values is represented by $M - 1$ binary variables, so that one categorical value is represented by all zeros, whereas for all the others one binary variable always equals one, while all others are zero. After that, all variables (including continuous variables, depth, binary variables and interactions) are standardized to have zero mean and standard deviation equal to one.
3. **Data partitioning** Considering that cross-validation is an integral part of both the model selection and model evaluation procedures, a fundamental issue is the choice of an adequate data-partitioning strategy. With the aim of avoiding the possibility that one sample (fold) comprises the observations only from deeper (or higher) soil

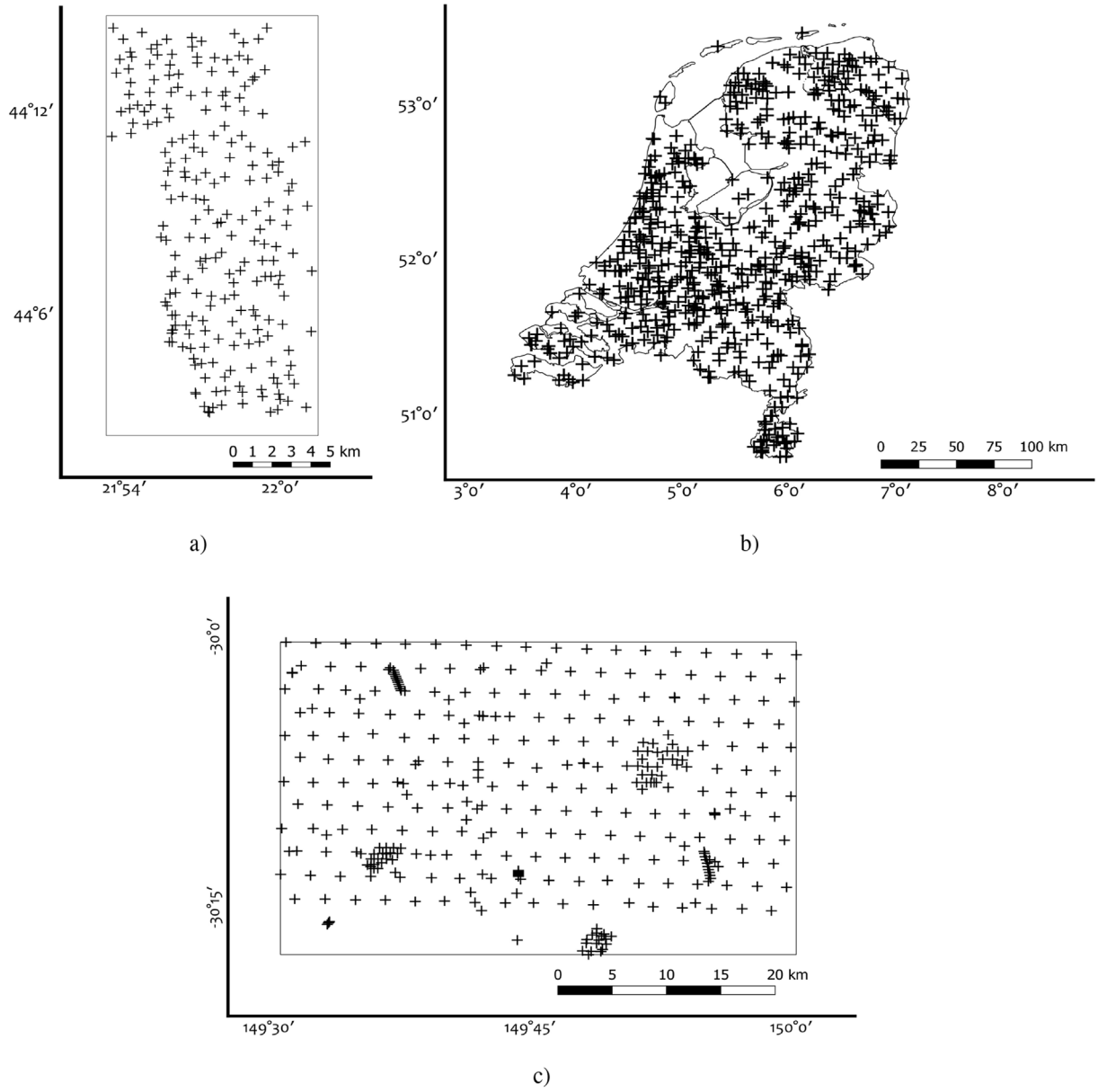

Fig. 4. Sampling locations and mapping area (black boundary) for: a) Bor; b) The Netherlands and c) Edgeroi. Coordinates are longitude and latitude expressed in WGS84.

layers or only from one part of the area, we defined a specific partitioning strategy via stratified sampling. The stratification criteria are chosen so that each fold has sufficient variation in: (1) spatial locations of soil profiles, (2) depth of the mid-points of sampled soil layers, and (3) the observed target variable (log (SOC)). The partitioning procedure is illustrated in Fig. 5. In the first step, *k*-means clustering of profiles based on spatial location was performed. Next, for each profile (set of observations sharing the same PID), the weighted mean of the target variable was computed ($\bar{y}$), using horizon thickness (d) as weight. Within each cluster profiles were sorted on $\bar{y}$ in increasing order. Next, successive profiles were cyclically assigned to different folds (each color representing one fold). It is important to note that the soil profiles, as collections of observations from different soil layers, were kept together during the partitioning, and not separated in individual observations. Criterion (1) was approximately fulfilled by forming folds from different spatial clusters. Criterion (2) was approximately fulfilled by assigning entire profiles to folds, since profiles contain measurements from different depths. Criterion (3) was approximately fulfilled by cyclical assignment of profiles sorted on the weighted mean of the target variable. Thus, each fold was assigned observations with different target variable values.

4. **Model selection.** The task of model selection is to select the optimal regularization parameter λ^*. This was performed via stratified *k*-fold cross-validation as explained in Section 2.5 (we used $k = 5$). First, the data set was split into 5 stratified folds according to the sampling

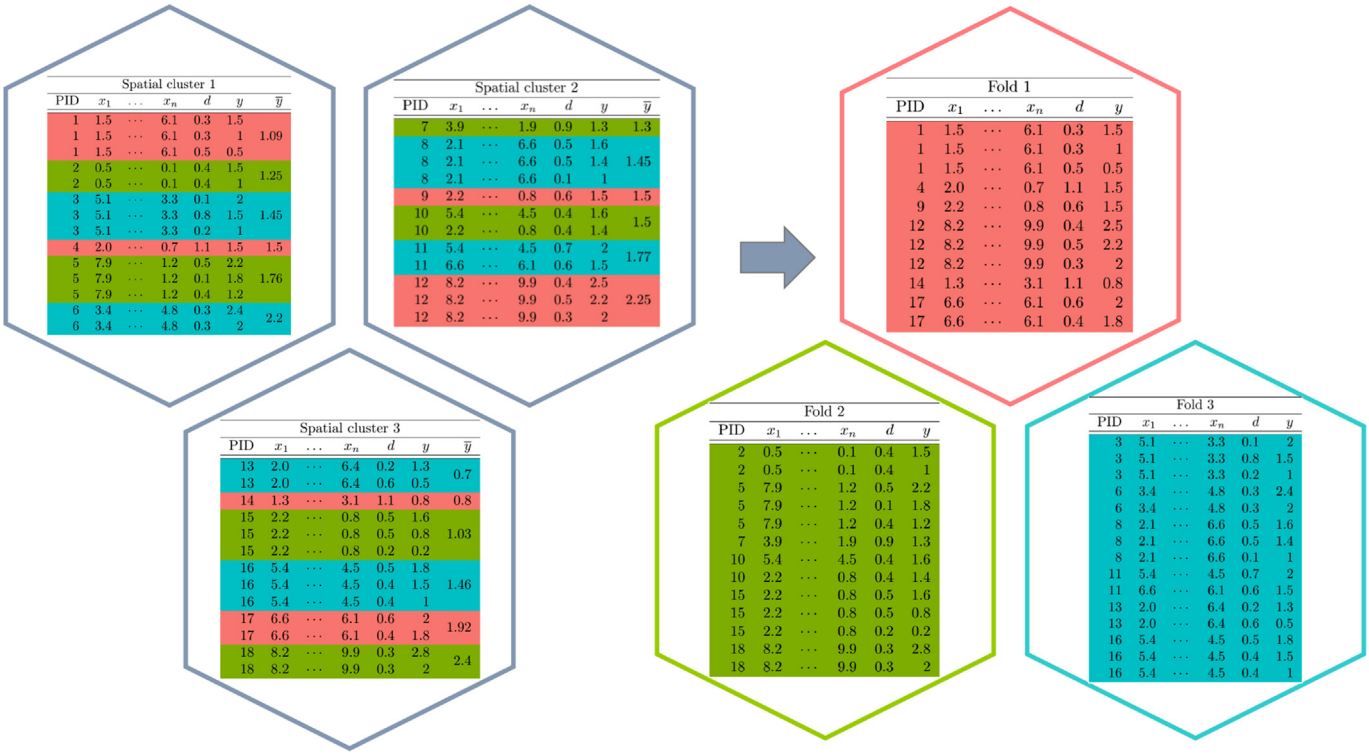

Fig. 5. Illustration of data partitioning. Left side shows three spatial clusters of soil profiles – observations sharing the same profile id (PID). x_1, ...,x_n are the values of surface variables at profile locations, d represents the soil horizon thickness within each profile, and y is the observed target variable. Right side shows three folds containing corresponding profiles from the left side.

strategy described in Step 3. Next, a sequence of λ values between 0 and 0.2 with incremental steps of 0.001 was used. This sequence of λ values is exclusively used for the models fitted by the 'glmnet' procedure (Friedman et al., 2009), whereas the 'hierNet' algorithm (Bien and Tibshirani, 2014) defines the sequence of λ automatically for hierarchical models (see Section 3.4). Mean squared error (MSE) was used as a reference metric when comparing models to obtain the optimal value of λ (i.e. corresponding to the model with the minimum MSE). The final models were selected by fitting regression models on the entire dataset using the selected λ.

5. **Model assessment.** This step involved the computation of accuracy measures – root mean square error (RMSE) and the coefficient of determination (R^2) based on 5-fold nested cross-validation for each model. The same partitioning strategy was applied to both nested cross-validation loops (outer and inner).
6. **Final prediction.** Once the final 3D model was selected, the prediction was performed on the grid defined over the entire area at specific depths.

3.4. Implementation

The presented approach was implemented in the R package sparsereg3D. The package sparsereg3D relies on functionalities of several other packages, including: glmnet (Friedman et al., 2009) for fitting lasso models, hierNet (Bien and Tibshirani, 2014) for fitting hierarchical interaction models with lasso, MASS (Venables and Ripley, 2002) for fitting stepwise regression models, aqp (Beaudette et al., 2013) for handling soil profile data, sp (Bivand et al., 2008) for handling spatial data and the tidyverse collection of R packages (Wickham, 2016) for data manipulation. The package contains procedures that perform (1) data preparation; (2) data pre-processing and data partitioning; (3) model selection; (4) model evaluation based on nested cross-validation; and (5) prediction over regular grids at different depths.

4. Results

In this section we compare OLS, stepwise regression, lasso, and hierarchical lasso methods for MLR, MPR, MLRI, and MPRI models for the Bor, Edgeroi and Netherlands case studies based on accuracy and sparsity of obtained models and on training time. One should keep in mind that very different sets of covariates were used in all three case studies and therefore it is not appropriate to make direct comparison of accuracy and sparsity between different case studies. The same data partitioning was used for all methods.

Fig. 6 summarizes changes in accuracy measures of RMSE and R^2 as model complexity increases. Since OLS and stepwise regression models do not depend on meta-parameters, the computation of accuracy measures was based on standard cross-validation. On the other hand, nested cross-validation was used for the lasso and hierarchical lasso methods. It is important to note that hierarchical lasso was only used when models included interactions.

In the same manner, the six graphs in Fig. 7 summarize the results in terms of model sparsity. Left graphs show the total number of non-zero parameters, while the right graphs show the number of distinct variables included in the final model (either through main effects or through interactions).

Table 2 summarizes computing time in seconds spent on model

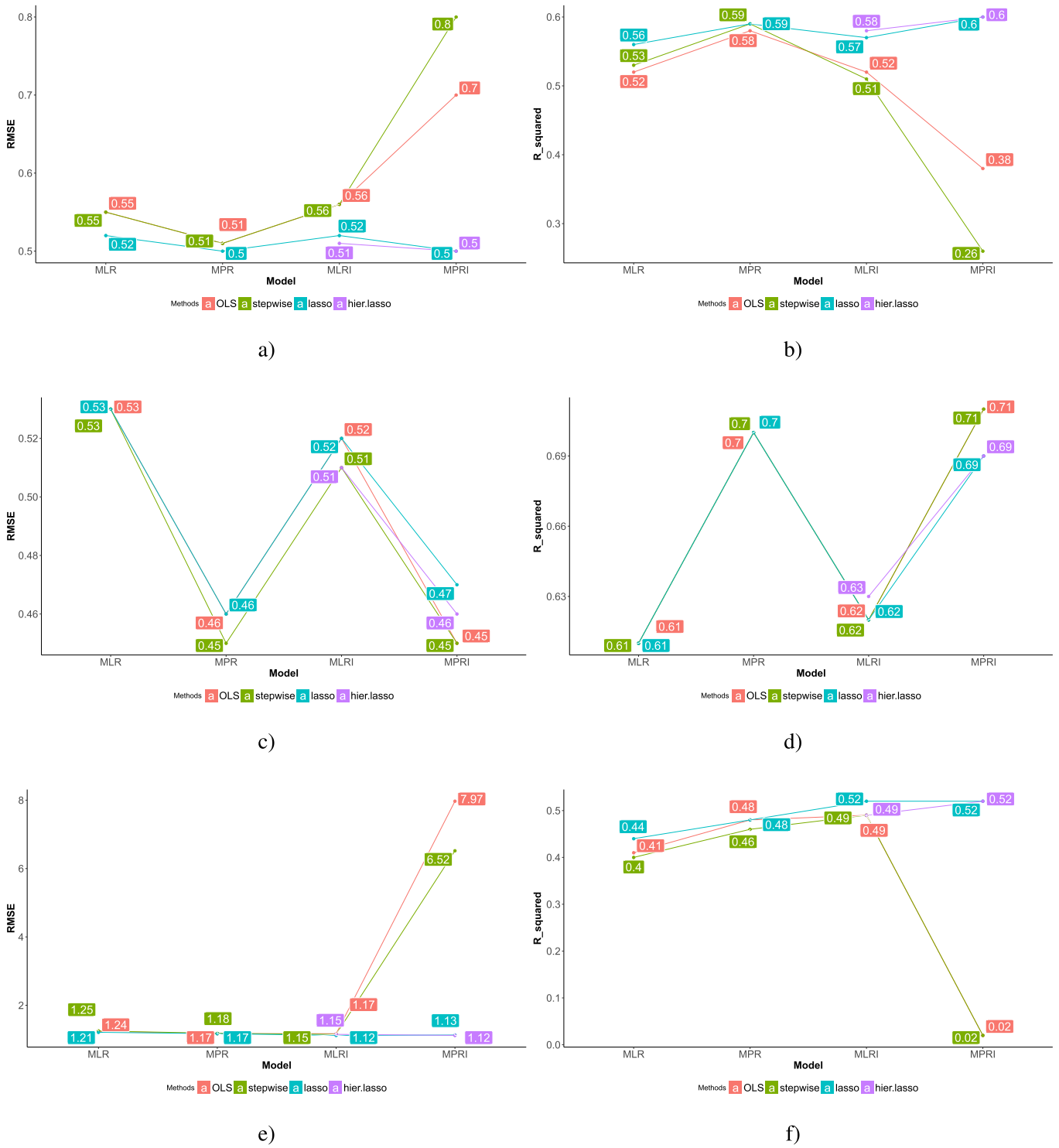

Fig. 6. Nested cross-validation accuracy indices – RMSE (left column) and R^2 (right column) of log (SOC) prediction for models of different complexity in case of Bor (first row), Edgeroi (second row), the Netherlands (third row).

training when using a computer with a 4th generation i7 processor and 16 Gb RAM.

Fig. 8 show the prediction maps of log (SOC) at 5 cm and 30 cm depth. Predictions were obtained by applying the MPRI model obtained by lasso to the gridded covariates with a spatial resolution of 20 m (Bor), 200 m (Edgeroi) and 250 m (the Netherlands).

5. Discussion

5.1. Lasso improves model accuracy

The graphs of RMSE and R^2 for Bor and the Netherlands (Fig. 6a), b), 6e) and 6f)) reveal that lasso variants (non-hierarchical and

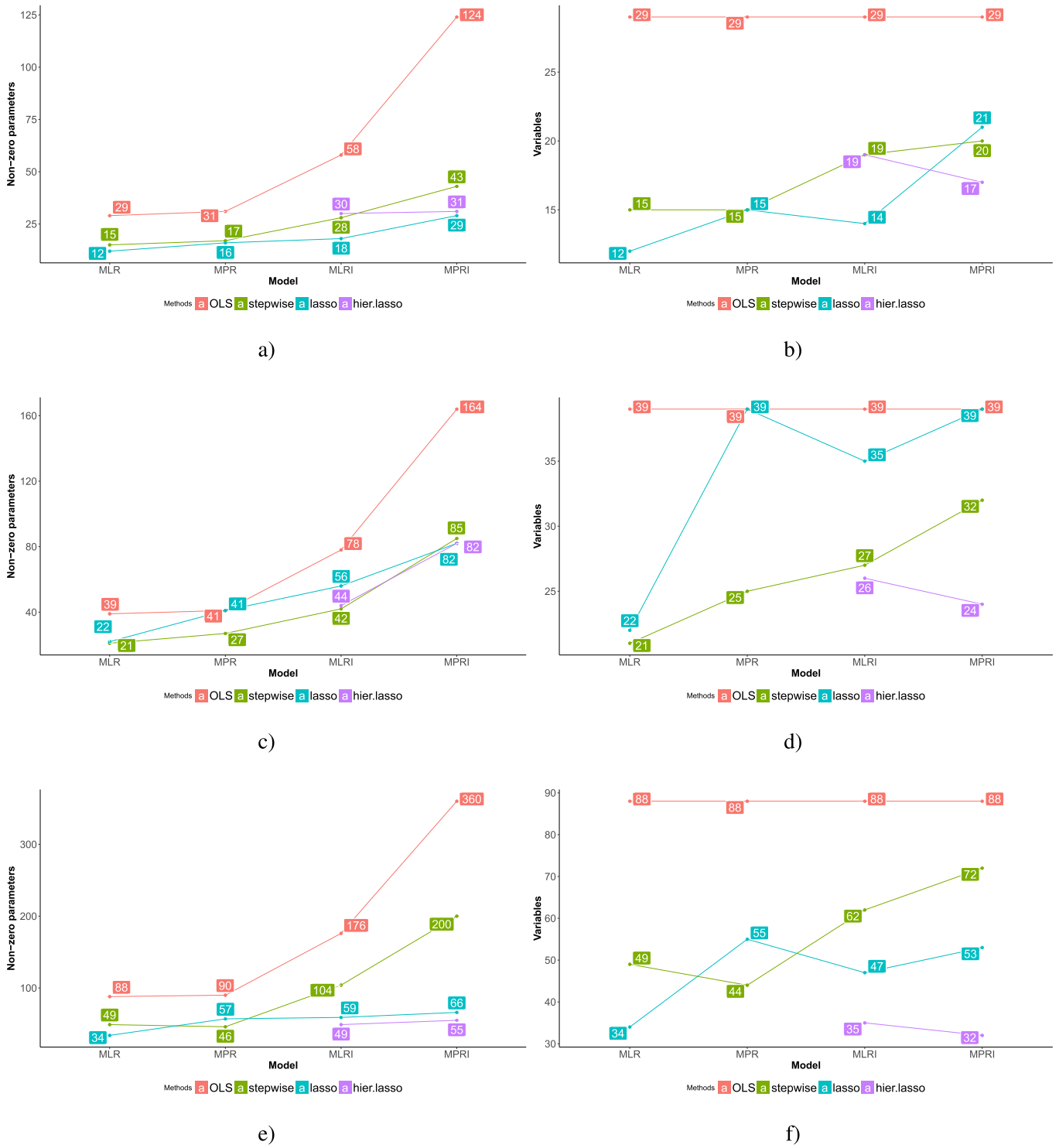

Fig. 7. Model structure indices – number of non-zero parameters (left column) and number of distinct environmental variables (right column) for models of different complexity in case of Bor (first row), Edgeroi (second row), the Netherlands (third row).

hierarchical) improve model accuracy for all models compared to OLS and stepwise regression. This conclusion is not supported by the results obtained for Edgeroi, which is less clear in all aspects. In case of Edgeroi (Fig. 6c) and d)), it appears that all methods perform almost equally since differences in the R^2 values for all models are never greater than 0.02. On the other hand, in the case of Bor, those differences (in favor of lasso methods) range from 0.01 to 0.34. Also, in the case of the Netherlands, the differences range from 0.00 to 0.50. Therefore, by rather consistent improvement in two case studies and virtually no difference in performance for the third, we conclude that, on balance, using the lasso improves model accuracy, compared to other linear modeling approaches. This result is not surprising considering that lasso is a regularization technique that is generally designed to mitigate the effects of overfitting.

Table 2
Time in seconds necessary for model training.

Model	Bor				Edgeroi				The Netherlands			
	OLS	stepwise	lasso	hier.lasso	OLS	stepwise	lasso	hier.lasso	OLS	stepwise	lasso	hier.lasso
MLR	0.01	0.6	0.1	–	0.04	2.7	1.2	–	0.05	47.4	0.2	–
MPR	0.01	0.6	0.1	–	0.04	2.8	1.1	–	0.05	48.7	0.2	–
MLRI	0.05	1.6	0.2	94	0.10	14.2	1.6	831	0.11	311	0.5	12229
MPRI	0.1	23	0.6	90	0.24	234	5.4	2410	0.35	5475	6.6	30720

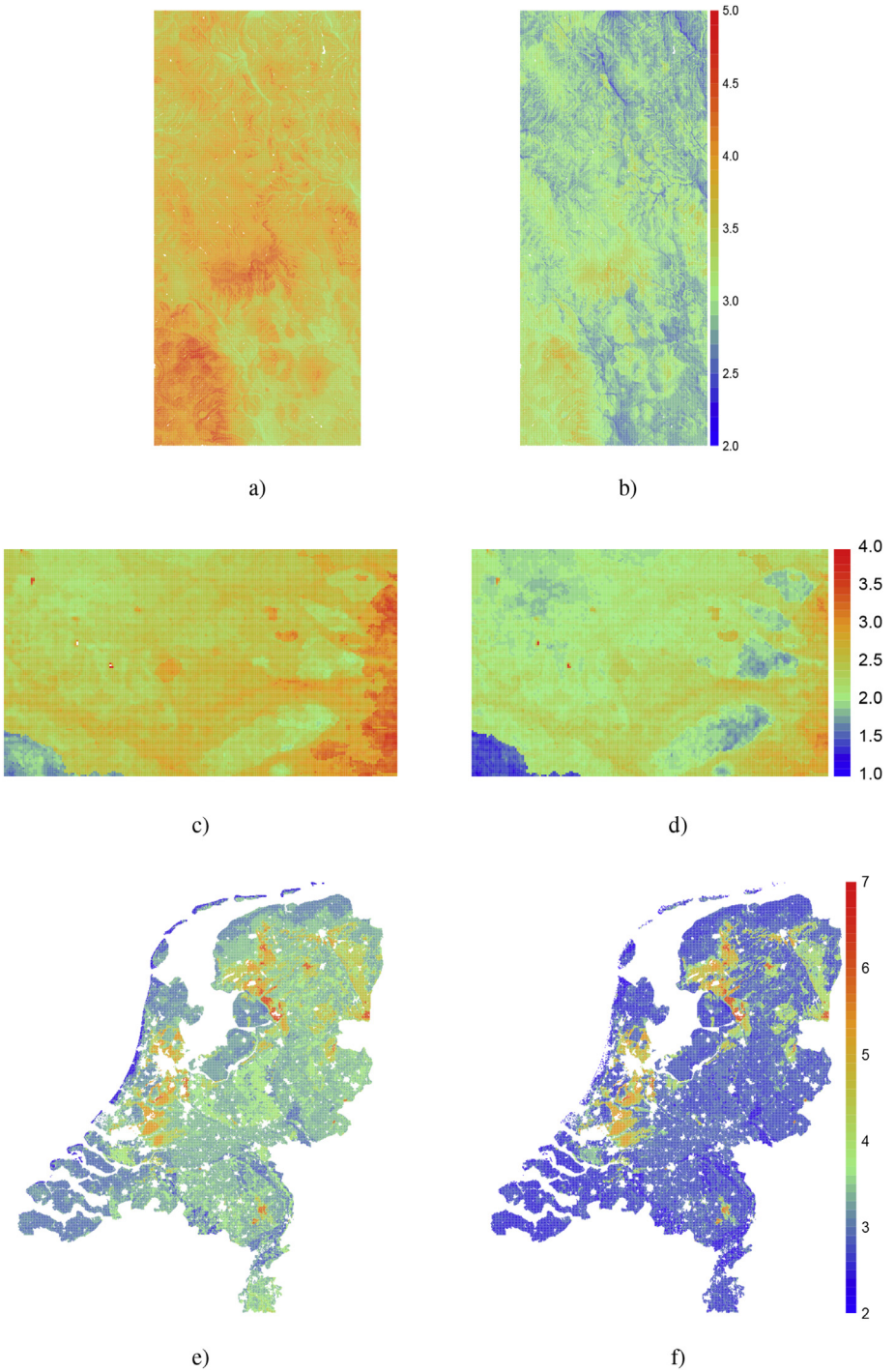

Fig. 8. Prediction maps of log (SOC) for each case study at two different depths: 5 cm (left) and 30 cm (right).

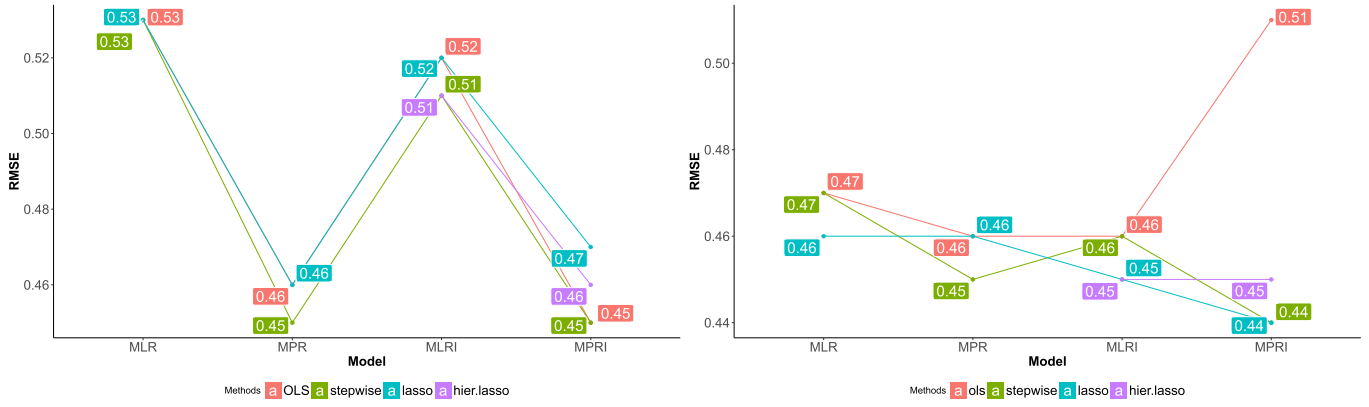

Fig. 9. RMSE of models on the original Edgeroi data set (left) and on Edgeroi data when all observations of depth larger than 2 m are discarded (right).

5.2. Lasso produces sparse models

It is evident from Fig. 7 that all methods aimed at performing variable selection (stepwise regression and lasso variants) provide models that are more or less sparse. Specifically, lasso models are typically among the sparsest, except in the case of MPR for Edgeroi (see Fig. 7c)). Compared to stepwise regression, more often than not, both lasso variants produce sparser models (especially in the case of Bor and the Netherlands). Hierarchical lasso seems to perform the best with respect to sparsity.

As an extreme, for the MPRI model, stepwise regression included 200 non-zero parameters (MPR for the Netherlands data, see Fig. 7e)), over three times more than the corresponding lasso models (which is probably one of the reasons why stepwise regression exhibited such a poor predictive performance for the MPRI model, see Fig. 6e)–f)). It is interesting to draw attention to the MPR model for Bor and the MPR and MLRI models for the Netherlands, for which all methods show similar accuracy (see Fig. 6a)–b) and Fig. 6e)–f)). In all these cases, lasso uses 2–3 times fewer parameters than OLS (see Fig. 7a)–e)). Also, the MLRI lasso variant model for the Netherlands uses almost twice as few parameters than stepwise regression.

Fig. 7b)–d) and 7f) show that a smaller number of non-zero parameters does not necessarily mean that a smaller number of distinct environmental variables are included in the model. This is particularly relevant for practical applications, especially if obtaining predictors is expensive.

5.3. Model extensions improve model accuracy if regularized

The graphs shown in Fig. 6a)–b), 6e) and 6f) show that model extensions can significantly improve prediction accuracy. Again, for Edgeroi, the results are inconclusive. However, the trends are clearer for Bor and the Netherlands. Lasso variants show a steady increase in accuracy as model complexity increases. Higher model complexity enables them to learn more effectively from the data without overfitting thanks to regularization. On the other hand, adding polynomial expansion of depth to the MLR model (which results in the MPR model) benefits OLS and stepwise regression. However, addition of interactions (resulting in MLRI and MPRI models) leads to overfitting as these are not able to control model complexity during training.

5.4. Unexpected results for edgeroi might be related to depth

As can be seen from Fig. 6c)–d), the improvement in accuracy for Edgeroi case study depends exclusively on the presence of polynomial depth in the model, regardless of the method used. This may be related to the fact that, as Fig. 3b) shows, the nonlinearity in vertical trend of log (SOC) is stronger for Edgeroi than for other case studies and that the observations from Edgeroi reach much deeper soil layers than the observations from the other case studies (up to 8 m). We conducted further experiments to check this. For all three case studies we performed OLS regression of log (SOC) based on a linear depth term and based on polynomial expansion of depth, with no other predictors included. In terms of RMSE, for Edgeroi, the polynomial model is 10.7 % better than the linear one. For Bor, it is 4.8 % better. For the Netherlands data it is even 3.8 % worse. If all observations of depth larger than 2 m are removed from the Edgeroi data set, meaning that we keep only the observations in which dependence of log (SOC) to depth is closer to linear, the polynomial model is only 3.1% better. This demonstrates that the importance of including a polynomial trend is due to these deepest observations and strong nonlinearity of the vertical trend of log (SOC).

Unlike the Bor and the Netherlands case studies, for Edgeroi OLS does not overfit, even if interactions with polynomial terms of depth are included. In Fig. 9 we compare the RMSE of the models on the original Edgeroi data and on the Edgeroi data when all observations at depth larger than 2 m (21 % of the data) are discarded. As a result, the unexpected behavior disappears. As demonstrated in the previous paragraph, the inclusion of polynomial terms of depth provides much more information to OLS in the case of the complete Edgeroi data set than in the case of Bor, the Nederlands or the reduced Edgeroi data set. Therefore, we suspect that the good behavior of OLS for the complete Edgeroi data has to do with the trade-off between additional information provided by the polynomial expansion and the increased potential for overfitting due to additional parameters, which is most favorable for the complete Edgeroi data set. The phenomenon calls for further investigation, but the insight we offer could be at least part of the explanation.

5.5. Lasso training is very efficient with or without extensions

Table 2 clearly shows that OLS and lasso dramatically outperform stepwise regression and hierarchical lasso with respect to computation time. As expected, OLS is faster than lasso. However, the extra computation time is quite acceptable considering other benefits of using lasso.

6. Conclusions

This paper presented an approach for the selection and evaluation of 3D prediction models of SOC with sparse structure, based on the lasso regression method. We demonstrated that lasso variants are generally capable of achieving better balance of accuracy, simplicity and

enhanced interpretability compared to OLS and stepwise regression. It yields sparse models, often more sparse than stepwise regression. It is observed that extension of models by interactions improves accuracy of lasso variants, but that it is detrimental for OLS and stepwise regression. The training time of non-hierarchical lasso is dramatically smaller than the training time of hierarchical lasso and stepwise regression, but is greater than that of OLS. In summary, it can be concluded that lasso is an efficient tool for obtaining sparse interaction models for spatial prediction in 3D.

In addition, we emphasize the importance of proper techniques for model evaluation. Generally, nested cross-validation is a computationally demanding technique. Nevertheless, in case of efficient regression techniques like non-hierarchical lasso, the use of nested cross-validation does not pose a computational challenge.

The perspectives of future work include the consideration of lasso-based trend models in the context of two-step geostatistical prediction methods, such as regression kriging, but also the consideration of different kinds of trend models, which might lie outside the scope of linear modeling.

Acknowledgements

This study was supported by the Serbian Ministry of Education and Science, under grants No. III 47014, TR 36035, TR 36009 and OI 174021.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.cageo.2018.05.008.

References

Arlot, S., Celisse, A., et al., 2010. A survey of cross-validation procedures for model selection. Stat. Surv. 4, 40–79.
Beaudette, D.E., Roudier, P., O'Geen, A.T., 2013. Algorithms for quantitative pedology: a toolkit for soil scientists. Comput. Geosciences 52, 258–268. https://doi.org/10.1016/j.cageo.2012.10.020.
Bien, J., Taylor, J., Tibshirani, R., 2013. A lasso for hierarchical interactions. Ann. statistics 41 (3), 1111.
Bien, J., Tibshirani, R., 2014. hierNet: a Lasso for Hierarchical Interactions. R package version 1.6. https://CRAN.R-project.org/package=hierNet.
Bivand, R.S., Pebesma, E.J., Gomez-Rubio, V., Pebesma, E.J., 2008. Applied Spatial Data Analysis with R Springer 747248717.
Brus, D.J., Yang, R.M., Zhang, G.L., 2016. Three-dimensional geostatistical modeling of soil organic carbon: a case study in the Qilian Mountains, China. Catena 141, 46–55. https://doi.org/10.1016/j.catena.2016.02.016.
Burnham, K.P., Anderson, D., 2003. Model Selection and Multi-model Inference. A Practical Informatio-theoric Approch. Sringer 1229.
Chung, C.-J., Fabbri, A.G., 2008. Predicting landslides for risk analysis-spatial models tested by a cross-validation technique. Geomorphology 94 (3), 438–452.
Cox, D.R., apr 1984. Interaction. Int. Stat. Review/Rev. Int. Stat. 52 (1), 1. http://www.jstor.org/stable/1403235?origin=crossref.
Didan, K., 2015. Mod13a3: Modis/terra vegetation indices monthly l3 global 1km sin grid v006. NASA EOSDIS Land Processes DAAC.
Finke, P., 2000. Updating the (1: 50,000) Dutch groundwater table class map by statistical methods: an analysis of quality versus cost. Geoderma 97 (3–4), 329–350.
Fitzpatrick, B.R., Lamb, D.W., Mengersen, K., 2016. Ultrahigh dimensional variable selection for interpolation of point referenced spatial data: a digital soil mapping case study. PloS one 11 (9).
Florinsky, I.V., Eilers, R.G., Manning, G., Fuller, L., 2002. Prediction of soil properties by digital terrain modelling. Environ. Model. Softw. 17 (3), 295–311.
Friedman, J., Hastie, T., Tibshirani, R., 2009. Glmnet: Lasso and Elastic-net Regularized Generalized Linear Models. R Package Version 1.
Geisser, S., 1975. The predictive sample reuse method with applications. J. Am. Stat. Assoc. 70 (350), 320–328.
Goetz, J., Brenning, A., Petschko, H., Leopold, P., 2015. Evaluating machine learning and statistical prediction techniques for landslide susceptibility modeling. Comput. Geosciences 81, 1–11.
Goovaerts, P., 2001. Geostatistical modelling of uncertainty in soil science. Geoderma 103 (1–2), 3–26.
Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. J. Mach. Learn. Res. 3 (Mar), 1157–1182.
Hastie, T., Tibshirani, R., Wainwright, M.S., 2015. Statistical Learning with Sparsity: the Lasso and Generalizations.
Hazeu, G., 2005. Landelijk grondgebruiksbestand nederland (lgn5); vervaardiging, nauwkeurigheid en gebruik. Technical Report. Alterra. pp. 28–29. http://webdocs.alterra.wur.nl/internet/geoinformatie/lgn/gin.pdf.
Hengl, T., de Jesus, J.M., MacMillan, R.A., Batjes, N.H., Heuvelink, G.B., Ribeiro, E., Samuel-Rosa, A., Kempen, B., Leenaars, J.G., Walsh, M.G., et al., 2014. Soilgrids1km - global soil information based on automated mapping. PLoS One 9 (8), e105992.
Hengl, T., Heuvelink, G.B., Stein, A., 2004. A generic framework for spatial prediction of soil variables based on regression-kriging. Geoderma 120 (1), 75–93.
Hengl, T., Reuter, H. (Eds.), 2008. Geomorphometry: Concepts, Software, Applications, vol. 33 Elsevier, Amsterdam.
Hoeting, J.A., Davis, R.A., Merton, A.A., Thompson, S.E., 2006. Model selection for geostatistical models. Ecol. Appl. 16 (1), 87–98.
Huang, H.-C., Chen, C.-S., 2007. Optimal geostatistical model selection. J. Am. Stat. Assoc. 102 (479), 1009–1024.
James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An Introduction to Statistical Learning with Applications in R.
Kanevski, M., 2013. A Methodology for Automatic Analysis and Modeling of Spatial Environmental Data (C). pp. 105–107.
Kempen, B., 2011. Updating Soil Information with Digital Soil Mapping. Phd thesis. Wageningen University, The Netherlands.
Knol, W., Kramer, H., Gijsbertse, H., 2004. Historisch grondgebruik nederland: een landelijke reconstructie van het grondgebruik rond 1900. Technical Report. Alterra. http://edepot.wur.nl/38370.
Koomen, A., Maas, G., 2004. Geomorfologische kaart nederland (gkn); achtergronddocument bij het landsdekkende digitale bestand. Technical Report. Alterra. http://edepot.wur.nl/40241.
Krstajic, D., Buturovic, L.J., Leahy, D.E., Thomas, S., 2014. Cross-validation pitfalls when selecting and assessing regression and classification models. J. Cheminformatics 6 (1), 1–15.
Liddicoat, C., Maschmedt, D., Clifford, D., Searle, R., Herrmann, T., Macdonald, L.M., Baldock, J., 2015. Predictive mapping of soil organic carbon stocks in south Australia's agricultural zone. Soil Res. 53 (8), 956–973. https://doi.org/10.1071/SR15100.
Malone, B.P., McBratney, A., Minasny, B., Laslett, G., 2009. Mapping continuous depth functions of soil carbon storage and available water capacity. Geoderma 154 (1), 138–152.
McGarry, D., Ward, W., McBratney, A.B., 1989. Soil studies in the lower namoi valley: methods and data. 1, the edgeroi data set. CSIRO Div. Soils. https://trove.nla.gov.au/work/17748968.
Michéli, E., Schad, P., Spaargaren, O., Dent, D., Nachtergaele, F., 2006. World reference base for soil resources: 2006: a framework for international classification, correlation and communication. FAO. http://www.fao.org/3/a-a0510e.pdf.
Minasny, B., McBratney, A.B., Mendonça-Santos, M., Odeh, I., Guyon, B., 2006. Prediction and digital mapping of soil carbon storage in the lower namoi valley. Soil Res. 44 (3), 233–244.
Mueller, T., Pierce, F., 2003. Soil carbon maps. Soil Sci. Soc. Am. J. 67 (1), 258–267.
Nestorov, I., Protic, D., Nikolic, G., 2007. Land cover mapping in Serbia. Wetlands 21176, 0–27.
Nussbaum, M., Spiess, K., Baltensweiler, A., Grob, U., Keller, A., Greiner, L., Schaepman, M.E., Papritz, A., 2017. Evaluation of digital soil mapping approaches with large sets of environmental covariates. SOIL Discuss. 2017, 1–32. https://www.soil-discuss.net/soil-2017-14/.
Orton, T., Pringle, M., Bishop, T., 2016. A one-step approach for modelling and mapping soil properties based on profile data sampled over varying depth intervals. Geoderma 262, 174–186.
Pejović, M., Bajat, B., Gospavić, Z., Saljnikov, E., Kilibarda, M., Čakmak, D., 2017. Layer-specific spatial prediction of as concentration in copper smelter vicinity considering the terrain exposure. J. Geochem. Explor. 179, 25–35. http://www.sciencedirect.com/science/article/pii/S0375674217303412.
Robinson, T., Metternicht, G., 2006. Testing the performance of spatial interpolation techniques for mapping soil properties. Comput. Electron. Agric. 50 (2), 97–108.
Steur, G., Heijink, W., 1991. Bodemkaart van nederland, schaal 1: 50.000. Algemene begrippen en indeligen"[Soil map of the Netherlands scale 1: 50 000. Definitions and classification], DLO Staring Centrum Wageningen.
Stone, M., 1974. Cross-validatory choice and assessment of statistical predictions. Journal of the Royal Statistical Society. Ser. B Methodol. 36 (2), 111–147.
Taylor, J., Tibshirani, R.J., 2015. Statistical learning and selective inference. Proc. Natl. Acad. Sci. 112 (25), 7629–7634.
Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B Methodol. 267–288.
Tibshirani, R., Tibshirani, R., Taylor, J., Loftus, J., Reid, S., 2017. selectiveInference: tools for Post-Selection Inference. R package version 1.2.4. https://CRAN.R-project.org/package=selectiveInference.
Tibshirani, R.J., Taylor, J., Lockhart, R., Tibshirani, R., 2016. Exact post-selection inference for sequential regression procedures. J. Am. Stat. Assoc. 111 (514), 600–620.
Varma, S., Simon, R., 2006. Bias in error estimation when using cross-validation for model selection. BMC Bioinforma. 7 (1).
Venables, W.N., Ripley, B.D., 2002. Modern Applied Statistics with S, fourth ed. Springer, New York iSBN 0-387-95457-0. http://www.stats.ox.ac.uk/pub/MASS4.
Wickham, H., 2016. Tidyverse: easily install and load 'tidyverse' packages. R package version 1.0.0. https://CRAN.R-project.org/package=tidyverse.