RESEARCH ARTICLE

Predicting active-layer soil thickness using topographic variables at a small watershed scale

Aidi Li^{1,2}, Xing Tan¹, Wei Wu³, Hongbin Liu^{1*}, Jie Zhu¹

1 College of Resources and Environment, Southwest University, Chongqing, China, **2** Chongqing Land Resources and Housing Surveying and Planning Institute, Chongqing, China, **3** College of Computer and Information Science, Southwest University, Chongqing, China

* lhbin@swu.edu.cn

Abstract

Knowledge about the spatial distribution of active-layer (AL) soil thickness is indispensable for ecological modeling, precision agriculture, and land resource management. However, it is difficult to obtain the details on AL soil thickness by using conventional soil survey method. In this research, the objective is to investigate the possibility and accuracy of mapping the spatial distribution of AL soil thickness through random forest (RF) model by using terrain variables at a small watershed scale. A total of 1113 soil samples collected from the slope fields were randomly divided into calibration (770 soil samples) and validation (343 soil samples) sets. Seven terrain variables including elevation, aspect, relative slope position, valley depth, flow path length, slope height, and topographic wetness index were derived from a digital elevation map (30 m). The RF model was compared with multiple linear regression (MLR), geographically weighted regression (GWR) and support vector machines (SVM) approaches based on the validation set. Model performance was evaluated by precision criteria of mean error (ME), mean absolute error (MAE), root mean square error (RMSE), and coefficient of determination (R^2). Comparative results showed that RF outperformed MLR, GWR and SVM models. The RF gave better values of ME (0.39 cm), MAE (7.09 cm), and RMSE (10.85 cm) and higher R^2 (62%). The sensitivity analysis demonstrated that the DEM had less uncertainty than the AL soil thickness. The outcome of the RF model indicated that elevation, flow path length and valley depth were the most important factors affecting the AL soil thickness variability across the watershed. These results demonstrated the RF model is a promising method for predicting spatial distribution of AL soil thickness using terrain parameters.

OPEN ACCESS

Citation: Li A, Tan X, Wu W, Liu H, Zhu J (2017) Predicting active-layer soil thickness using topographic variables at a small watershed scale. PLoS ONE 12(9): e0183742. <https://doi.org/10.1371/journal.pone.0183742>

Editor: Claudionor R Silva, Universidade Federal de Uberlandia, BRAZIL

Received: November 10, 2016

Accepted: August 10, 2017

Published: September 6, 2017

Copyright: © 2017 Li et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: The authors received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

Introduction

Active-layer (AL) soil thickness, defined as the top layer of [soil](#), is one of the most important factors affecting soil quality and productivity, vegetation growth, soil moisture pattern, surface and subsurface flow, and shallow landslide [1–6]. Information about the spatial distribution of

AL soil thickness is beneficial to hydro-ecological modeling, precision agriculture, and land resource management [7]. In most cases, the information of AL soil thickness is derived from conventional soil survey. But this method cannot offer sufficient information to satisfy the special planning (e.g. land-use planning and cropping pattern), and the application is obstructed by the high costs on time, expense and labors. Furthermore, some authors used a constant value to represent AL soil thickness across the entire landscape [8, 9] that the variability of it is ignored consequently. However, AL soil thickness is not constant over space in areas with complex terrain conditions. Therefore, alternative approaches are proposed to meet the demands of many applications.

In order to overcome the limitations of conventional soil mapping, many statistical methods have been developed for modeling the spatial pattern of soil attributes within the digital soil mapping (DSM) framework. Among them, multiple linear regression (MLR) has been widely used to quantify the soil thickness variability [5, 10–12] due to its simplicity and efficiency in computation, and easy interpretation. Ziadat established the MLR model between soil thickness and terrain variables in the Al-Muwaqqar watershed in Jordan and reported that the model was a promising approach for predicting soil attributes [11, 13]. Yang et al. investigated the relationship between soil thickness and terrain attributes using linear regression in peak-cluster depression region and showed that more than one-half of the variation in the soil thickness could be explained [12]. However, the MLR model cannot reveal the non-linear relationship between soil thickness and environmental variables that usually exists in reality [14]. Therefore, a more efficient method is required. Recently, machine learning approaches, such as support vector machines [15, 16], geographically weighted regression [17–20] and random forest [19, 20], have been applied to map soil nutrients. Previous studies have demonstrated that these methods were more intelligent due to their abilities for solving non-linearity problem [15, 16, 21]. But these applications were mostly concentrated on mapping the spatial distribution of soil organic carbon or soil organic matter.

The spatial distribution of AL soil thickness has high heterogeneity as a function of parent material, topography, climate, vegetation cover, and intense human activities [8, 17, 22–25]. Among them, topography plays a crucial role in the soil attributes variability [11, 25, 26–28], especially at a small watershed area under relatively uniform parent material, vegetation cover, and climate conditions (e.g. precipitation and temperature) [5, 29]. Topography has direct or indirect effects on the spatial distribution of soil attributes by affecting surface and underground runoff, soil erosion, soil temperature and soil formation [28], especially for slope fields with sharply rolling terrain [10, 29]. There is a strong relationship between AL soil thickness and terrain variables [5, 10, 29–32], and previous studies have demonstrated that the variables derived from digital elevation models (DEMs) could improve the prediction of AL soil thickness in different landscapes [11, 26, 33, 34]. For example Moore et al. predicted the spatial distribution of soil horizon thickness using terrain variables (e.g. slope, wetness index, and maximum flow path length), they reported that about one-half of the variation of soil attributes were explained by terrain variables [35]. Conversely, Penížek and Borůvka found the improvement of prediction models using three primary terrain variables including elevation, slope and aspect as covariates was relatively small in Southern Bohemia, they attributed this to low correlation between soil thickness and the primary variables. In the end, they conceded that the use of more terrain variables (e.g. flow path length, topographic wetness index) would provide more accurate predictions of AL soil thickness [33].

Random Forest (RF), a relatively new method, is developed from Classification and regression trees (CARTs) for improving the prediction accuracy [36]. Compared with other machine learning methods, RF has several advantages, such as resistance to overfitting, insensitive to noise features, robust error estimation, simplicity in parameters defined, and quantifying

variable importance [20, 36, 37]. In this paper, the overall purpose was to predict the spatial pattern of AL soil thickness based on terrain variables at a small watershed scale using RF model. More specifically, the objectives of this study were to (1) investigate the capability of RF model for predicting the AL soil thickness by comparing it with MLR, GWR and SVM based on a validation set, (2) estimate uncertainty using Monte Carlo simulation method, (3) model the spatial distribution of AL soil thickness, and (4) explore the main factors affecting the AL soil thickness variability.

Materials and methods

Study area

The study site ($31^{\circ}6'26''$ – $31^{\circ}11'41''$ N, $109^{\circ}35'9''$ – $109^{\circ}41'11''$ E) covering 5376.374 ha is located within a small watershed area of Fengjie County of Chongqing, southwest China (Fig 1). The topography of the study area is mainly mountainous with elevation varying between 238 and 1592 m above sea level. The climate is moderate subtropical with vertical characteristic. The average annual precipitation is 1132 mm which mainly happens in summer and the mean annual temperature decreases along with the increasing of elevation (16.4°C ,

Fig 1. Maps of the watershed location and soil sampling sites.

<https://doi.org/10.1371/journal.pone.0183742.g001>

13.7–16.4°C, and < 10.8°C for areas below 600 m, within the range 1000–1400 m, and above 1400 m, respectively). The average annual sunlight length of the area is 1639 h. The predominant parent materials are feldspar-quartz sandstone, limestone and dolomite. The soil types consist of yellow earths, purplish soils, and limestone soils [38]. The dominant cropping system is wheat-maize rotation across the study site. The soil is neutral with a pH value of 7.2 and content of organic matter is 17.4 g/kg for the watershed.

Soil data

A total of 1113 soil samples were collected from slope fields over the watershed using a manual auger in October of 2007 when crops were harvested. The study did not involve private land, protected land, endangered or protected species. No specific permissions were required for these locations/activities. The position and elevation of each soil sample were georeferenced using a global positioning system (GPS), and other relevant information such as cropping system was recorded at the same time. AL soil thickness was determined by digging soil profiles and measuring the depth to the non-soil portion of the soil profile. The samples were randomly divided into calibration (770 soil samples) and validation (343 soil samples) sets. The calibration set was used to develop models for predicting AL soil thickness and the validation set was used to evaluate the performance of the models.

Terrain variables

ASTER GDEM (Global Digital Elevation Model) was used in this study (Fig 2). It produced by NASA and METI of Japan and can be obtained from Geospatial Data Cloud hub published by Chinese Academy of Sciences (2009). Pre-production estimated accuracies for this product are 20 meters at 95% confidence for vertical data and 30 meters at 95% confidence for horizontal data. The coordinate system of the data is WGS84 and the original data is in IMG format. Seven terrain variables, namely, elevation, aspect, relative slope position (RSP), valley depth (VD, m), flow path length (FPL, m), topographic wetness index (TWI), and slope height (SH, m), were derived from the DEM (S1 Table) using SAGA GIS software [39]. In order to determine an appropriate prediction model in this study, all seven terrain variables will be used for these models.

Multiple linear regression model

As one of the most widely used regression methods, multiple linear regression model (MLR) based on ordinary least square method was adopted to predict the AL soil thickness spatial distribution in this study. The MLR model is expressed as follow:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} \dots + \beta_n X_{in} \quad (1)$$

where Y_i ($i = 1, 2, 3, \dots, n$) and β_j ($j = 0, 1, 2, \dots, n$) are the dependent variable and regression coefficient, respectively. X_i ($i = 1, 2, 3, \dots, n$) is the independent variable. The model will be built based on the linear relationship existing between dependent and independent variables. Moreover, input and output data must be numeric. In the current study, the selected terrain variables were independent variables and the AL soil thickness was dependent variable.

Geographically weighted regression

Geographically weighted regression (GWR) is an exploratory technique that mainly intends to indicate where non-stationarity is taking place on the map [17], that is, where locally weighted regression coefficients move away from their global values. The technique is fully described by

Fig 2. The digital elevation model of the study area.

<https://doi.org/10.1371/journal.pone.0183742.g002>

Fotheringham et al. (2002) and involves first selecting a bandwidth for an isotropic spatial weights kernel, typically a Gaussian kernel with a fixed bandwidth chosen by leave-one-out cross-validation. A standard GWR can be defined as follow:

$$Y_i = \beta_0(u_i, v_i) + \beta_1(u_i, v_i)X_{i1} + \beta_2(u_i, v_i)X_{i2} + \dots + \beta_n(u_i, v_i)X_{in} \quad (2)$$

where Y_i ($i = 1, 2, 3, \dots, n$) is the predicted AL soil thickness at the i th location; (u_i, v_i) is the coordinate at the i th location; and β_n ($n = 0, 1, 2, \dots, n$) is the regression coefficient; X_{in} ($in = i1, i2, \dots, in$) is environmental variable for the i th location; and n is the number of environment variables.

Support vector machine

Support vector machine (SVM), a machine learning algorithm, is based on statistical theory and structural risk minimization principle. SVMs map the sample space into a high-dimensional or infinite-dimensional feature space based on Mercer's nuclear [40]. Then, the highly non-linear problem of sample space could be solved using linear learning technique in the feature space [41]. The function ($f(x)$) for non-linear regression is the extension of linear regression by introducing a kernel function.

Given a set of training data, $T = \{(x_i, y_i), i = 1, 2, \dots, n\}$, where x_i and y_i are the input and output values of the i th data, respectively, the $f(x)$ is defined as an approximately linear regression problem and expressed as,

$$f(x) = \omega \cdot \varphi(x) + b \quad (3)$$

where $\varphi(x)$ is the mapping function, ω is an adjustable weight vector, b is the scalar threshold.

In order to avoid the generation of dimension disaster problem during the computation in feature space and solve the non-linear problem, a kernel function is introduced. In the current study, a radial basis function (Eq 4) which satisfied the Mercer's condition was used as the kernel function.

$$K(x, x_i) = \exp^{-(2\sigma^2)^{-1} \|x - x_i\|^2} \quad (4)$$

where σ is the bandwidth parameter. After introducing the kernel function, the sample data can be converted into linear separable data in a high-dimensional feature space. Support vector regression is a kind of regression algorithm based on penalty learning [37]. An ε -insensitive loss function [38] is introduced in SVM and the formula is,

$$l_\varepsilon = |y - f(x)|_\varepsilon = \begin{cases} 0 & \text{if } |y - f(x)| \leq \varepsilon \\ |y - f(x)| - \varepsilon & \text{otherwise} \end{cases} \quad (5)$$

where ε is the insensitive coefficient, which is used to control the accuracy of the prediction. Based on the structural risk minimization, the regression problem can be transformed to an optimization problem by the following way.

$$R(C) = Cn^{-1} \sum_{i=1}^n l_\varepsilon + 2^{-1} \|\omega\| \quad (6)$$

where C is the penalty weight, which controls the extent of penalty on the samples (which their prediction errors are beyond the range of ε) and realizes the trade-off between the number of samples and the model complexity. In generally, the value of C should not be too large in order to avoid overlearning problem. In view of the existence of samples with prediction errors are larger than the non-negative constant ε , the non-negative slack variables (ξ_i, ξ_i^*) are introduced. Therefore, the optimization function for computing the empirical risk minimization problem is equivalent to solve the following constrained optimization problem.

$$\begin{aligned} \text{Min. } R(\omega, \xi_i, \xi_i^*) &= 2^{-1} \|\omega\|^2 + C \sum_{i=1}^n (\xi_i + \xi_i^*) \\ \text{Subjected to } &\begin{cases} y_i - f(x_i) \leq \varepsilon + \xi_i \\ f(x_i) - y_i \leq \varepsilon + \xi_i^* \end{cases} \quad i = 1, 2, 3, \dots, n \end{aligned} \quad (7)$$

In fact, the above Eq (7) is a typical quadratic programming problem. According to the dual

theory, Lagrangian function (a_i, a_i^*) and kernel function ($K(x, x_i)$) are introduced to transform the above problem into a dual optimization problem. The dual form can be expressed as,

$$Q(a_i, a_i^*) = \sum_{i=0}^n y_i (a_i + a_i^*) - 2^{-1} \sum_{i=0}^n \sum_{j=0}^n (a_i - a_i^*) (a_j - a_j^*) K(x_i, x_j) - \epsilon \sum_{i=0}^n (a_i + a_i^*)$$

$$\text{Subjected to } \begin{cases} \sum_{i=1}^n (a_i - a_i^*) = 0 \\ 0 \leq a_i, a_i^* \leq C \end{cases} \quad i = 1, 2, 3 \dots n \quad (8)$$

where a_i and a_i^* are Lagrange multipliers. Support vectors consist of the sample points which satisfy the equality of $a_i - a_i^* \neq 0$. Therefore, the optimization hyperplane linear regression function can be written as,

$$f(x) = \sum_{i=1}^n (a_i - a_i^*) K(x, x_i) + b \quad (9)$$

In this study, the parameters C and σ were set to be 0.87 and 13 reckoned by Grid search approach, respectively. The SVM model was calculated in Matlab 7.12.0.

Random forest

Random forest (RF) is a forest which combines multiple unrelated classification or regression trees (CARTs) [34]. Bootstrap sampling (with replacement) is used to generate a random sample from original dataset for each tree. Each tree is built based on a random subset which is selected from training dataset by bootstrap sampling (with replacement). The number of trees (n_{tree}) defined by users is identical to the random subsets. The trees are grown by binary recursive partitioning [42], the output variable at each tree is partitioned into multiple binary subsets (each subset with maximum homogeneity). For each node, a random subset selected from the corresponding input variables is used. The best variable is searched for each binary split by the least impurity of node. The size of each randomly selected subset of input variables (m_{try}) remains unchanged within the whole forest growing procedure and needs specification. Furthermore, the bias of each tree can be reduced as it is grown without pruning [36]. RF can be used to establish classification and regression model. The final output in classification model is a majority vote of the outputs of the trees. However, in regression model, the prediction value is an average of the outputs of all trees and the formula is expressed as,

$$\bar{h}(x) = (1/k) \sum_{i=1}^k h(X; \theta_i) \quad (10)$$

where $\bar{h}(x)$ is a prediction value, θ_i is a random vector with independent and identically distributed, X is an input matrix, $h(X; \theta_i)$ is an output of a tree, k is the number of trees. In regression model, three parameters (n_{tree} : the number of trees, nodesize : the minimum size of observations at each terminal node, m_{try} : the number of input variables at each node) are defined by users. Theoretically, the prediction accuracy increases as the number of trees increases. However, the computational load increases and the speed of performance improvement decreases with the increasing of the forest size. As a randomly selected subset of input variables is used for splitting at each node, the correlation between the trees is decreased, the calculation time is reduced and the performance of the forest is improved.

Each tree is trained on about two-third of the training data and tested on the reminder which is called out-of-bag (OOB) samples. Then, the out-of-bag error (OOB error) is estimated [36]. The OOB error is an unbiased estimator and is identical with the prediction error which is estimated by a test set with the same size of the random subset [36]. The mean square

error (MSE_{OOB}) of the forest is calculated by aggregating the OOB predictions and the expression is given as,

$$MSE_{OOB} = \frac{1}{n} \sum_{i=1}^n (z_i - \hat{z}_{i_{OOB}})^2 \quad (11)$$

where z_i is the OOB prediction of the i th point, $\hat{z}_{i_{OOB}}$ is the average of OOB predictions. Moreover, the variable importance also can be assessed based on the OOB error. The increased in MSE_{OOB} (InMSE) is used to quantify the importance of variable. The variable with a higher value of InMSE was more important than others.

For the RF model we used the “random Forest” package [43] in R [44]. After several experiments, n_{tree} , $nodesize$, and m_{try} were set to be 2000, 3, and 5, respectively, in this study.

Accuracy assessment

In order to assess the performance of models for AL soil thickness mapping, mean error (ME), mean absolute error (MAE), root mean squared error (RMSE), and determination coefficient (R^2) were calculated by the measured and predicted values from the validation set. The formulas were defined as follows.

$$ME = \frac{1}{n} \sum_{i=1}^n (A_i - B_i) \quad (12)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |A_i - B_i| \quad (13)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (A_i - B_i)^2} \quad (14)$$

$$R^2 = \frac{\sum_{i=1}^n (A_i - \bar{A})^2 (B_i - \bar{B})^2}{\sum_{i=1}^n (A_i - \bar{A})^2 \sum_{i=1}^n (B_i - \bar{B})^2} \quad (15)$$

where n is the number of validation sites, A_i and B_i represent the measured and predicted values of AL soil thickness, respectively. $\bar{A}$ and $\bar{B}$ are mean values of the measured and predicted AL soil thickness, respectively. Models with lower values of ME, MAE, RMSE and a higher value of R^2 perform better.

Statistical analysis

Descriptive statistical analyses of minimum, maximum, mean, standard deviation (SD), and coefficient of variation (CV) were applied to test the variation of AL soil thickness and terrain variables. Correlation analysis was used to find the relationship between AL soil thickness and terrain variables. All of the statistical analysis and the multiple linear regression were computed using SPSS 19.0 software.

Uncertainty analysis

Uncertainties within the best performed model, which were caused by terrain variables and collected AL soil thickness data, were analyzed through Monte Carlo method [18]. A number of input and output variables were generated based on their probability distribution functions. In this study, 1000 estimates of DEM and AL soil thickness were generated. The seven terrain indicators were derived from the created DEMs. The uncertainties in DEM and AL soil thickness were then estimated using the best performed model. This analysis was conducted by using R [44] and Feature Manipulate Engine 2015 (FME) software.

Results

Descriptive statistical analysis

The descriptive statistics of AL soil thickness and terrain variables were summarized in (Table 1). The AL soil thickness values of all sampling sites in the study area varied from 15 to 60 cm with a mean of 35.39 cm and standard deviation (SD) of 17.495 cm. The coefficient of variation (CV) was used to estimate the variability of AL soil thickness. The CV for AL soil thickness was 49.44%, which indicated that AL soil thickness had intermediate variability. For terrain variables, all had intermediate variability with CV varying between 28.82% and 84.46%. It seems that the variability of AL soil thickness might be controlled by the complexity of topography of the study area. The statistical results of AL soil thickness for both calibration and validation sets were very similar to that of all sampling data, which indicated the small samples were suitable for developing and verifying the prediction models in this work.

Correlation analysis

The relationships between AL soil thickness and terrain variables were explored by correlation analysis (S2 Table). As the table shown, there were significant correlations between AL soil thickness and terrain variables in this study site (Table 2). AL soil thickness was negatively correlated with elevation ($r = -0.510$, $p < 0.01$), RSP ($r = -0.207$, $p < 0.01$) and SH ($r = -0.098$, $p < 0.01$), and positively correlated with aspect ($r = 0.089$, $p < 0.01$), VD ($r = 0.225$, $p < 0.01$), FPL ($r = 0.272$, $p < 0.01$), TWI ($r = 0.061$, $p < 0.05$). This implied that elevation had higher impacts on the AL soil thickness variability, whereas TWI had little impacts on the AL soil thickness variability. The results suggested that topography had strong effects on AL soil thickness.

Model performance

In order to quantify the performance of models, the accuracy indicators (mean error, mean absolute error, root mean squared error, and R^2) were calculated using the validation data set

Table 1. Descriptive statistics of AL soil thickness and terrain variables.

Variable	N	Min.	Max.	Mean	SD	CV(%)
Elevation (m)	1113	249	1546	891.55	297.21	33.34
Aspect (°)	1113	2.12	360	174.2	94.37	54.17
RSP	1113	0	1	0.43	0.325	75.58
VD (m)	1113	0	465	143	107.03	74.61
FPL (m)	1113	0	3889	1207	770.26	63.80
TWI	1113	3.76	21.82	6.45	1.86	28.82
SH (m)	1113	2.76	4596	94	79.58	84.46
ST: All sampling sites (cm)	1113	15	60	35.39	17.495	49.44
ST: Calibration sites (cm)	770	15	60	35.47	17.487	49.30
ST: Validation sites (cm)	343	15	60	35.2	17.537	49.82

<https://doi.org/10.1371/journal.pone.0183742.t001>

Table 2. Relationships between AL soil thickness and terrain variables.

	Elevation	Aspect	RSP	VD	FPL	TWI	SH
ST	-0.510**	0.089**	-0.207**	0.225**	0.272**	0.061*	-0.098**

* and ** denote significance levels at $p < 0.05$ and $p < 0.01$, respectively.

<https://doi.org/10.1371/journal.pone.0183742.t002>

Table 3. Accuracy assessment indices of different methods based on validation set.

	ME(cm)	MAE (cm)	RMSE (cm)	R ²
MLR	-0.32	12.83	14.95	0.27
GWR	0.01	6.23	11.48	0.59
SVM	1.42	8.57	12.67	0.49
RF	0.39	7.09	10.85	0.62

ME, mean error; MAE, mean absolute error; RMSE, root mean square error; R², coefficient of determination.

<https://doi.org/10.1371/journal.pone.0183742.t003>

(343 soil samples) (S3 Table) and the results were listed in (Table 3). The mean error (ME) was calculated to identify the bias of predictions. The ME of MLR, SVM, GWR and RF were -0.32, 1.42, 0.01, and 0.39 cm, respectively. The negative value suggested MLR model underestimated the AL soil thickness values, while the positive value revealed overestimation of SVM, GWR and RF models. The extent of AL soil thickness estimated by RF model was much narrow than that of SVM. Furthermore, the MAE (7.09 cm) and RMSE (10.85 cm) of RF model were lower than that of MLR and SVM models, and the R² of RF (62%) was much higher than that of MLR (27%) and SVM (49%) models. In terms of RMSE and R², RF model outperformed others for predicting the spatial distributions of AL soil thickness in this study area.

(Fig 3) showed that bigger errors were mostly found in the central areas with longer flow path length. The uncertainties in input and output variables were estimated and the contributions of them to the best performed model (RF) were determined (S4 Table). The results were given in (Table 4). The changes of RMSE and R² indicated that the uncertainty in AL soil thickness was bigger than that of DEM.

Spatial distribution of AL soil thickness

The spatial distribution maps of AL soil thickness produced by MLR, SVM, GWR and RF models were shown in (Fig 4). The four models produced similar distributional trends of AL soil thickness over the study area. Higher values of AL soil thickness mainly concentrated in valley bottoms and lower areas with low elevation, whereas lower values of AL soil thickness mainly located around of valley with high elevation. These results suggested that elevation was the notable variable affecting the spatial distribution of AL soil thickness across the watershed. The spatial distribution map of AL soil thickness produced by MLR and GWR reflected more smooth spatial variation and fewer details in spatial variation of AL soil thickness in the local area. On the contrary, the distribution maps predicted by SVM and RF showed more particular spatial variation information such as in the valley bottoms where the predicted values by SVM and RF were higher than those by MLR and GWR. Moreover, the range of AL soil thickness values produced by RF model was closer to the measured values than those predicted by MLR, SVM and GWR.

Variable importance

The relative importance of terrain variables yielded by the best regression model (RF) was shown in (Fig 5). The terrain variables were ranked in order of elevation > flow path length > valley depth > aspect > relative slope position > slope height > topographic wetness index. Among these terrain variables, elevation with InMSE of 237.59% was the most critical variable affecting the variation of AL soil thickness in the current area. Flow path length and valley depth with InMSEs of more than 75% were the second and the third most important variables controlling AL soil thickness variability, respectively.

Fig 3. Spatial distribution maps of prediction deviation of AL soil thickness produced by RF.

<https://doi.org/10.1371/journal.pone.0183742.g003>

Discussion

The comparison of the accuracy indicators (ME, MAE, RMSE and R^2) suggested that RF was the best method for modeling the spatial distribution of AL soil thickness. RF model outperformed others for predicting AL soil thickness benefited from its insensitivity to over-fitting and avoid noise and its predominance in dealing with the non-linear problem [19, 36]. Lark indicated that the complex non-linear relationships between soil attributes and terrain variables are subsistence, a simple linear model cannot catch these relationships [45]. This could explain why the MLR and GWR model gave worse performance in this study area. Although the SVM algorithm is able to catch the non-linear relationships, its performance is still subject to certain restrictions, such as the difficulty in finding the suitable kernel function and

Table 4. Contributions of uncertainties in DEM and AL soil thickness to RF model.

	DEM	AL Soil thickness
RMSE	41.4%	58.5%
R^2	45.2%	54.8%

<https://doi.org/10.1371/journal.pone.0183742.t004>

Fig 4. Spatial distribution maps of AL soil thickness produced by (A) MLR, (B) GWR, (C) SVM, and (D) RF.

<https://doi.org/10.1371/journal.pone.0183742.g004>

optimal parameters. Furthermore, the range of AL soil thickness estimated by RF was closer to the measurements than that of SVM (Fig 3), this might because RF performed better on avoiding the overfitting problem which involves high-dimension explanatory variables [46]. Additionally, RF algorithm also has other advantages over SVM, GWR and MLR approaches. For instance, RF does not need to preprocess the raw data before building the prediction model, and the importance of explanatory variables can be measured using the OOB dataset [19].

The spatial distribution maps of AL soil thickness produced by MLR, GWR, SVM and RF showed a similar spatial pattern trend over the watershed, where much thicker soils mostly concentrated in valley and lower areas while lower ones mainly occurred in steep slopes and higher areas. This was also found in other landscape. For example, Mehnatkesh et al. made a study in the semiarid hilly region in western Iran [5]. They reported that higher values of AL soil thickness existed in the footslope and toeslope. Scarpone et al. also found a similar distribution trend in AL soil thickness in southern British Columbia [47]. A possible explanation for this might be the effects of complex topography, since topography indirectly influence the

Fig 5. Variable importance produced by RF model.

<https://doi.org/10.1371/journal.pone.0183742.g005>

spatial distribution of AL soil thickness by affecting surface runoff, soil erosion, soil deposition and soil temperature, especially in slope fields with sharply rolling terrain. Thus, the spatial heterogeneity of AL soil thickness is a result of the interaction of soil process (soil erosion and soil deposition) and surface runoff. In the current study, elevation which showed highly negative correlation with AL soil thickness was the most important one affecting the spatial pattern of AL soil thickness over the watershed. Soil erosion could be accelerated under the increases of elevation in hilly slopes especially around the hilltop, while soil deposition is intensified in valley bottoms and lower areas [47]. Under the interaction of soil erosion and soil deposition, therefore, the valley bottoms with relatively lower elevation had thicker soils, and around of valley with higher elevation had shallower soils. In addition, the relative importance of terrain variables also showed that flow path length and valley depth were the second and the third most important factors, respectively, which both showed highly positive correlation with AL soil thickness. The values of AL soil thickness in this watershed increased with the increases of flow path length and valley depth. Generally, surface water always accumulated in concave areas and valley with high values of valley depth [10, 30, 48], where AL soil thickness is thicker resulting from fluvial deposition processes [47]. The flow path length can reflect the spatial pattern of runoff, and deeper surface water can be observed in the area with relative longer flow path length [49]. Hence, thicker soils concentrated in the concave areas and valley with longer flow path length and higher valley depth.

The uncertainty of the best performed model could be attributed to both DEM and AL soil thickness. Many studies found that some subtle details of the terrain features become less discernible and the predictive efficiency of terrain variables is lost rapidly with a coarse resolution of DEM [50, 51]. Hence, the accuracy of terrain variables could be affected by the resolution of DEM [27, 52, 53], such as smaller mean slope and higher mean wetness index are produced as

the decrease of the resolution. This could reduce the predictive capability of the models for the spatial distribution of AL soil thickness [12]. Additionally, the consideration of objects on the earth surface of DEMs also yields uncertainty in digital terrain, which could further impact on models' performance. However, this will not affect the results of the current work which was conducted at a small watershed of a rural area. Moreover, the outcomes of the sensitivity analysis suggested that the uncertainty in AL soil thickness was bigger than that of DEM. During the work, the sample locations were identified by the local trained experts and the digging depths were mainly determined by their subjective experiences.

Conclusion

In this study, we are devoted to find an acceptable, simple, effective, and low cost soil-landscape modeling approach using limited observations to predict the spatial pattern of AL soil thickness at watershed scale. Seven quantitative terrain indexes including elevation, aspect, relative slope position, valley depth, flow path length, slope height, and topographic wetness derived from DEM were applied with MLR, SVM, and RF methods. The spatial distribution maps of AL soil thickness produced by all models illustrated that thicker soils mostly concentrated in valley bottoms and lower areas and shallower soils mainly occurred in steep slopes and higher areas. According to the performance indicators, RF model which could simulate about 62% of the AL soil thickness variability outperformed others in the study area. Because the RF model can not only solve the non-linear problems between AL soil thickness and terrain variables but also avoid overfitting phenomenon and resist noise in predictors. The importance of variables calculated by RF indicated that elevation, flow path length, and valley depth were the most important terrain variables affecting the AL soil thickness variability across the watershed. These variables indirectly affect AL soil thickness mainly by controlling soil erosional and depositional processes.

The outcomes of the current study suggested that RF technique is a useful tool for generating the digital AL soil thickness map at watershed scale. Of course, the generality and transportability of the RF model to other landscapes with similar features as that in the current area still remains to be verified. Furthermore, a high-quality digital terrain model (DTM) is needed to find an optimum resolution for improving the accuracy of predictive model.

Supporting information

S1 Table. Statistical analysis of AL soil thickness and terrain variables.
(XLSX)

S2 Table. Relationships between AL soil thickness and terrain variables. Significance levels can be judged by Signif.code.
(XLSX)

S3 Table. Accuracy assessment indices of four methods: ME, MAE, RMSE and R^2 .
(XLSX)

S4 Table. Uncertainties assessment in DEM and AL soil thickness to RF model.
(XLSX)

Acknowledgments

We thank Junbo Liu, Jialin Huang and Zhengtao Peng for their excellent technical support. We are very thankful to Qiang Wang for the determination of software development. Authors

wish to express their gratitude to Dr Tintin Xu (The University of Auckland, New Zealand) for his thorough re-reading of this article.

Author Contributions

Conceptualization: Hongbin Liu.

Data curation: Aidi Li, Xing Tan, Wei Wu, Jie Zhu.

Formal analysis: Hongbin Liu.

Investigation: Aidi Li, Xing Tan, Wei Wu, Hongbin Liu, Jie Zhu.

Methodology: Hongbin Liu.

Resources: Aidi Li.

Software: Wei Wu.

Supervision: Hongbin Liu.

Validation: Aidi Li, Xing Tan, Hongbin Liu, Jie Zhu.

Writing – original draft: Aidi Li, Xing Tan.

Writing – review & editing: Aidi Li, Wei Wu, Hongbin Liu.

References

1. Stieglitz M, Shaman J, Mcnamara J, Engel V, Shanley J, Kling GW. An approach to understanding hydrologic connectivity on the hillslope and the implications for nutrient transport. *Global Biogeochemical Cycles*. 2003; 17:1105.
2. Meerveld TV, McDonnell JJ. On the interrelations between topography, soil depth, soil moisture, transpiration rates and species distribution at the hillslope scale. *Advances in Water Resources*. 2006; 29:293–310.
3. Meyer MD, North MP, Gray AN, Zald HSJ. Influence of AL soil thickness on stand characteristics in a Sierra Nevada mixed-conifer forest. *Plant and Soil*. 2007; 294:113–123.
4. Dahlke HE, Behrens T, Seibert J, Andersson L. Test of statistical means for the extrapolation of soil depth point information using overlays of spatial environmental data and bootstrapping techniques. *Hydrological Processes*. 2009; 23:3017–3029.
5. Mehnatkesh A, Ayoubi S, Jalalian A, Sahrawat KL. Relationships between soil depth and terrain attributes in a semi arid hilly region in western Iran. *Journal of Mountain Science*. 2013; 10:163–172.
6. Segoni S, Martelloni G, Catani F. Different Methods to Produce Distributed AL soil thickness Maps and Their Impact on the Reliability of Shallow Landslide Modeling at Catchment Scale. *Landslide Science and Practice*. 2013; 127–133.
7. Tesfa TK, Tarboton DG, Chandler DG, McNamara JP. Modeling soil depth from topographic and land cover attributes. *Water Resources Research*. 2009; 45:1277–1278.
8. Bathurst JC, Moretti G, Elhames A, Begueria S, Garcia Ruiz JM. Modelling the impact of forest loss on shallow landslide sediment yield, Ijezu river catchment, Spanish Pyrenees. *Hydrology and Earth System Sciences*. 2007; 11:569–583.
9. Talebi A, Troch PA, Uijlenhoet R. A steady-state analytical slope stability model for complex hillslopes. *Hydrological Processes*. 2008; 22:546–553.
10. Florinsky IV, Eilers RG, Manning GR, Fuller LG. Prediction of soil properties by digital terrain modelling. *Environmental Modelling & Software*. 2002; 17:295–311.
11. Ziadat FM. Prediction of Soil Depth from Digital Terrain Data by Integrating Statistical and Visual Approaches. *Pedosphere*. 2010; 20:361–367.
12. Yang Q, Zhang F, Jiang Z, Li W, Zhang J, Zeng F, Li H. Relationship between soil depth and terrain attributes in karst region in Southwest China. *Journal of Soils and Sediments*. 2014; 14:1568–1576.
13. Mishra U., Drewniak B., Jastrow J.D., Matamala R.M., and Vitharana U.W.A.. Spatial representation of organic carbon and active-layer thickness of high latitude soils in CMIP5 earth system models. *Geoderma*. 2016;

14. Mcbratney AB, Santos MLM, Minasny B. On digital soil mapping. *Geoderma*. 2003; 117:3–52.
15. Ballabio C. Spatial prediction of soil properties in temperate mountain regions using support vector regression. *Geoderma*. 2009; 151: 338–350.
16. Were K, Bui DT, Dick ØB, Singh BR. A comparative assessment of support vector regression, artificial neural networks, and random forests for predicting and mapping soil organic carbon stocks across an Afromontane landscape. *Ecological Indicators*. 2015; 52:394–403.
17. Mishra U., and Riley W.J.. Active-layer thickness across Alaska: comparing observation-based estimates with CMIP5 earth system model predictions. *Soil Science Society of America Journal*. 2014; 78:894–902.
18. Mishra U., Torn M.S., Ogle S., and Masanet E.. Improving regional soil carbon inventories: combining IPCC carbon inventory method with regression kriging. *Geoderma*. 2012; 189–190:288–295.
19. Grimm R, Behrens T, Märker M, Elsenbeer H. Soil organic carbon concentrations and stocks on Barro Colorado Island—Digital soil mapping using Random Forests analysis. *Geoderma*. 2008; 146:102–113.
20. Wiesmeier M, Barthold F, Blank B, Kögel-knabner I. Digital mapping of soil organic matter stocks using Random Forest modeling in a semi-arid steppe ecosystem. *Plant Soil*. 2011; 340:7–24
21. Rodriguez-Galiano V, Sanchez-Castillo M, Chica-Olmo M, Chica-Rivas M. Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. *Ore Geology Reviews*. 2015; 71: 804–818.
22. Boer M, Barrio GD, Puigdefábres J. Mapping soil depth classes in dry Mediterranean areas using terrain attributes derived from a digital elevation model. *Geoderma*. 1996; 72:99–118.
23. Kuriakose SL, Devkota S, Rossiter DG, Jetten VG. Prediction of soil depth using environmental variables in an anthropogenic landscape, a case study in the Western Ghats of Kerala, India. *CATENA*. 2009; 79:27–38.
24. Luo D, HuiJun J, Marchenko S, Romanovsky V. Distribution and changes of active layer thickness (ALT) and soil temperature (TTOP) in the source area of the Yellow River using the GIPL model. *Science China Earth Sciences*. 2014; 57:1834–1845.
25. Saito T, Yasuda H, Suganuma H, Inosako K, Abe Y, Kojima T. Predicting Soil Infiltration and Horizon Thickness for a Large-Scale Water Balance Model in an Arid Environment. *Water*. 2016; 8:96.
26. Ohnuki Y, Kimhean C, Shinomiya Y, Toriyama J. Distribution and characteristics of AL soil thickness and effects upon water storage in forested areas of Cambodia. *Hydrological Processes*. 2008; 22:1272–1280
27. Thompson JA, Bell JC, Butler CA. Digital elevation model resolution: effects on terrain attribute calculation and quantitative soil-landscape modeling. *Geoderma*. 2001; 100:67–89.
28. Emili LA, Price JS, Fitzgerald DF (2006) Hydrogeological influences on forest community type along forest-peatland complexes in coastal British Columbia. *Canadian Journal of Forest Research*. 2006; 36:2024–2037.
29. Thompson JA, Pena-Yewtukhiw EM, Grove JH. Soil-landscape modeling across a physiographic region: Topographic patterns and model transportability. *Geoderma*. 2006; 133:57–70.
30. GESSLER PE, Deceased IDM, McKENZIE NJ, RYAN PJ. Soil-landscape modelling and spatial prediction of soil attributes. *International Journal of Geographical Information systems*. 1995; 9:421–432.
31. Saulnier GM, Beven K, Obled C. Including spatially variable effective soil depths in TOPMODEL. *Journal of Hydrology*. 1997; 29:158–172.
32. Lee KT, Ho JY. Prediction of landslide occurrence based on slope-instability analysis and hydrological model simulation. *Journal of Hydrology*. 2009; 375:489–497.
33. Penížek V, Borůvka L. Soil depth prediction supported by primary terrain attributes: a comparison of methods. *Plant Soil & Environment*. 2006; 52:424–430.
34. Ho J, Lee KT, Chang T, Wang Z, Liao Y. Influences of spatial distribution of AL soil thickness on shallow landslide prediction. *Engineering Geology*. 2012; 124:38–46.
35. Moore ID, Gessler PE, Nielsen GA, Peterson GA. Soil Attribute Prediction Using Terrain Analysis. *Soil Science Society of America Journal*. 1993; 57:443–452.
36. Breiman L. Random Forests. *Machine Learning*. 2001; 45:5–32.
37. Díazuriarte R, Andrés SAD. Gene selection and classification of microarray data using random forest. *Bmc Bioinformatics*. 2006; 7:3. <https://doi.org/10.1186/1471-2105-7-3> PMID: 16398926
38. Gong Z. Chinese soil taxonomy. Science Press, Beijing, New York. 2003.
39. Cimmery V. User Guide for SAGA (version 2.0). 2007. <http://www.saga-gis.org/en/index.html>.
40. Ben-Hur A, Weston J. A user's guide to support vector machines. Humana Press. 2010; 609: 223–39.

41. Vapnik VN. Statistical learning theory {Adaptive and learning systems for signal processing, communications, and control}. John Wiley & Sons, Inc. (US). 1998.
42. Breiman L, Friedman JH, Olsen RA, Stone CJ. Classification and regression trees. Wadsworth International Group; 1984.
43. Liaw A, Wiener M. Classification and Regression by RandomForest. R News 23. 2001.
44. R Development Core Team (2010) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. 2010. <http://www.R-project.org>
45. Lark RM. Soil–landform relationships at within-field scales: an investigation using continuous classification. *Geoderma*. 1999; 92:141–165.
46. Jia S, Hu X, Sun L. The Comparison between Random Forest and Support Vector Machine Algorithm for Predicting β -Hairpin Motifs in Proteins. *Engineering*. 2013; 5:391–395
47. Scarpone C, Schmidt MG, Bulmer CE, Knudby A. Modelling AL soil thickness in the critical zone for Southern British Columbia. *Geoderma*. 2016; 282:59–69.
48. Heimsath AM, Dietrich WE, Nishiizumi K, Finkel RC. Cosmogenic nuclides, topography, and the spatial variation of soil depth. *Geomorphology*. 1999; 27:151–172.
49. Zhang ZH, Wang XF. Calculation and control of flow path length in superelevation sections. *Intelligent Automation & Soft Computing*. 2013; 19:327–333.
50. Gessler PE, Chadwick OA, Chamran F, Althouse L, Holmes K. Modeling Soil–Landscape and Ecosystem Properties Using Terrain Attributes. *Soil Science Society of America Journal*. 2000; 64:2046–2056.
51. Mora A. Small scale digital soil mapping in Southeastern Kenya. *Catena*. 2008; 76:44–53.
52. Quinn PF, Beven KJ, Lamb R, Beven KJ. The $\ln(a/\tan\beta)$ index: How to calculate it and how to use it within the topmodel framework. *Hydrological Processes*. 1997; 9:161–182.
53. Wu S, Li J, Huang GH. A study on DEM-derived primary topographic attributes for hydrologic applications: Sensitivity to elevation data resolution. *Applied Geography*. 2008; 28:210–223.