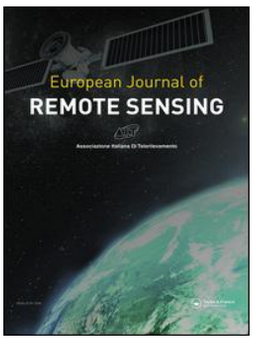

European Journal of Remote Sensing

ISSN: (Print) 2279-7254 (Online) Journal homepage: https://www.tandfonline.com/loi/tejr20

Comparison of machine learning algorithms for soil salinity predictions in three dryland oases located in Xinjiang Uyghur Autonomous Region (XJUAR) of China

Fei Wang, Shengtian Yang, Wei Yang, Xiaodong Yang & Ding Jianli

To cite this article: Fei Wang, Shengtian Yang, Wei Yang, Xiaodong Yang & Ding Jianli (2019) Comparison of machine learning algorithms for soil salinity predictions in three dryland oases located in Xinjiang Uyghur Autonomous Region (XJUAR) of China, European Journal of Remote Sensing, 52:1, 256-276, DOI: 10.1080/22797254.2019.1596756

To link to this article: https://doi.org/10.1080/22797254.2019.1596756

Published online: 02 Apr 2019.

EUROPEAN JOURNAL OF REMOTE SENSING
2019, VOL. 52, NO. 1, 256–276
https://doi.org/10.1080/22797254.2019.1596756

Comparison of machine learning algorithms for soil salinity predictions in three dryland oases located in Xinjiang Uyghur Autonomous Region (XJUAR) of China

Fei Wang[a,b], Shengtian Yang[a,b], Wei Yang[a,b], Xiaodong Yang[a,b] and Ding Jianli[a,b]

[a]Xinjiang Common University Key Lab of Smart City and Environmental Stimulation, College of Resource and Environmental Sciences, Xinjiang University, Urumqi, China; [b]Lab for Oasis Ecosystem, Ministry of Education, Urumqi, China

ABSTRACT

Many different machine learning approaches have been applied for various purposes. However, there has been limited guidance regarding which, if any, machine learning models and covariate sets might be optimal for predicting soil salinity across different oases in the Xinjiang Uyghur Autonomous Region (XJUAR) of China. This study aimed to compare five machine learning algorithms, Least Absolute Shrinkage and Selection Operator (LASSO), Multiple Adaptive Regression Splines (MARS), Classification and Regression Trees (CART), Random Forest tree ensembles (RF), and Stochastic Gradient Treeboost (SGT), to predict soil salinity in three geographically distinct areas (the Qitai, Kuqa, and Yutian oases). A total of 21 data sets from three oases were used to evaluate the performance of the algorithm and to screen the optimal variables. The results show the following indices are considered to be important indicators for quantitative assessment of soil salinity: EEVI, CSRI, EVI2, GDVI, SAIO, and SIT. Comparison results show that SGT is the most suitable algorithm for predicting soil salinity in arid areas. This study provides a comprehensive comparison of machine learning techniques for soil salinity prediction and may assist in the modeling and variable selection of digital soil mapping in the XJUAR of China.

ARTICLE HISTORY
Received 22 April 2018
Revised 12 March 2019
Accepted 14 March 2019

KEYWORDS
Soil salinity; machine learning; oasis; Landsat OLI; digital elevation model; Xinjiang Uyghur autonomous region

Introduction

Globally, soil salinization has affected approximately 831 million hectares of land (Butcher, Wick, DeSutter, Chatterjee, & Harmon, 2016; Martinez-Beltran & Manzur, 2005), 23.32% of which (193.8×10^6 hm^2) is located in Asia (FAO, 2015), and soil salinization is predicted to impact 50% of all arable land by 2050 (Wang, Vinocur, & Altman, 2003). The Xinjiang Uyghur Autonomous Region (XJUAR) in northwestern China, the largest arid region in the country, is also one of the main distribution areas of soil salinization in Asia (Wang, Chen, Luo, & Han, 2015) (Figure 1). In 2014, according to a report by the Xinjiang Institute of Ecology and Geography at the Chinese Academy of Sciences, salinized farmland accounted for 37.72% of the irrigated land in the XJUAR, representing an increase of 6% since 2006 (Tian, Mai, & Zhao, 2016). Moreover, salinized farmland in the southern XJUAR accounted for almost half (49.6%) of the irrigated land, seriously restricting the lives of farmers and herdsmen.

To evaluate the distribution and severity of soil salinization, many authors have used soil samples and environmental variables to explore the relationship between soil salinization and environmental variables by establishing models to predict soil salinization in unsampled areas. In recent studies, various indices derived from remote sensing data have been used as proxies to analyze soil salinity (Chen, Zhao, Chen, Wang, & Gao, 2015; Douaoui, Nicolas, & Walter, 2006; Fernández-Buces, Siebe, Cram, & Palacio, 2006; Khan, Rastoskuev, Sato, & Shiozawa, 2005; Metternicht & Zinck, 2003; Scudiero, Skaggs, & Corwin, 2015; Zhang et al., 2011b). Some scholars have also tried to integrate multiple variables (including temperature, vegetation, parent material, topography, soil, colour, humidity, and soil albedo) to reduce the uncertainty associated with predicting soil salinity based on a single variable (Taghizadeh-Mehrjardi, Minasny, Sarmadian, & Malone, 2014; Wu et al., 2014a). Such studies have been performed in the United States in South Dakota (Lobell et al., 2010) and the San Joaquin Valley in California (Scudiero, Skaggs, & Corwin, 2014), as well as in Dujaila, Iraq (Wu et al., 2014a), Saudi Arabia (Allbed, Kumar, & Aldakheel, 2014), the Yellow River Delta of China (Chen et al., 2015), the Chéliff Basin in Algeria (Douaoui et al., 2006), Kuqa, the XJUAR in China (Ding & Yu, 2014), and the Tuz (salt) and lake region in Turkey (Gorji, Sertel, & Tanik, 2017). However, most indices have not been comprehensively compared in the XJUAR of China, which has been termed the "World Saline Soil Museum".

CONTACT Wei Yang weiyang19870909@126.com Xinjiang Common University Key Lab of Smart City and Environmental Stimulation, College of Resource and Environmental Sciences, Xinjiang University, Urumqi 830046, China

Data mining can be defined as an automated or semi-automated process designed to uncover patterns from large digital datasets using trained models, where the resulting patterns may then be applied to new data for the purpose of prediction (Witten, Frank, & Hall, 2011). In soil science, numerous machine learning algorithms are available in the subfield of pedometrics for the development of predictive or digital soil maps, for instance, random forests (Grimm, Behrens, Märker, & Elsenbeer, 2008), multivariate adaptive regression splines (MARS) (Nawar, Buddenbaum, Hill, & Kozak, 2014); stochastic gradient treeboost(SGT) (Angileri et al., 2016), support vector machine (SVM) (Heung et al., 2016), artificial neural networks(ANN)(Heung et al., 2016), partial least squares regression (PLSR)(Nawar et al., 2014), classification and regression tree (CART)(Youssef, Pourghasemi, Pourtaghi, & Al-Katheeri, 2016),or other learner with less commonly used include least absolute shrinkage and selection operator (LASSO)(Zandler, Brenning, & Samimi, 2015). Review these literatures we found some of them have several advantages include less parameters which need user define, enables the estimation of the importance of the independent variables, own higher computational efficiency which is very important for big data operation, able to handle numerical, ordinal, or discrete predictors, such as LASSO, MARS, CART, RF and SGT. Furthermore, each of five machine learning algorithm give a different strategy for mining helpful information. LASSO is relatively recent approaches that use mathematically similar shrinkage penalties. These penalties push less important coefficients closer to zero in the case of ridge regression, or effectively set them to zero when the lasso technique is used (Tibshirani, 1996). Thus, while the lasso performs variable subset selection and therefore produces sparse models that can be applied more easily in a predictive context. MARS is a relatively new technique which combines the classical linear regression, the mathematical construction of splines, the binary recursive partitioning and brute search intelligent algorithms to produce a model capable of predicting the value of a target variable (categorical or continuous) from a set of independent variables (Friedman, 1991). The MARS algorithm works by partitioning the ranges of the explanatory variables into regions and by generating, for each of these regions, a linear regression equation. CART, is perhaps the most commonly used learners in the digital soil maps literature, which consist of nodes and leaves where each node is a partition of the training dataset that aims to maximize the within-node homogeneity and the between node heterogeneity based on node splitting rules that are generated from a set of predictor variables-a type of if-then statement. The RF learner is conceptually similar to tree-based learners(CART) and shares the same advantages; however, multiple decision trees are trained and the results are based on the predictions from an ensemble of the individual trees (Breiman, 2001). For the RF learner, each tree is trained from a randomized bootstrap sample of the entire training set and a subset of predictors used for the node-splitting rules is also randomly selected. The SGT method combines regression trees and a boosting technique to improve the predictive performance of multiple single models (Friedman, 2002). Boosting is a forward and stage-wise procedure in which a subset of the data is randomly selected to iteratively fit new tree models to minimize the loss function . This process introduces a stochastic gradient boosting procedure that can improve model performance and reduce the risk of overfitting.

Despite these five machine learning algorithms have been developed and applied in various purposes, research on soil salinization still few. Taghizadeh-Mehrjardi et al. (2014) selected a regression tree analysis to infer soil salinity attributes from nonparametric data (i.e., no assumptions regarding variable distribution), which is not sensitive to missing data. In a study by Muller and Van Niekerk (2016), relationships between image features and electrical conductivity measurements of 30 soil samples were studied using a regression analysis and classification and regression tree (CART) modelling. Vermeulen and Van Niekerk (2017) evaluated the extent to which DEM derivatives (only terrain variables were used as input) and machine learning algorithms (k-nearest neighbour, support vector machines, decision trees (DT) and random forests) can be used to predict the location and extent of salt-affected areas (where there are only two classes: salt-affected and unaffected). Among these machine learning algorithms, DT held the greatest potential for monitoring salt accumulation in irrigated areas, particularly for simulating subsurface conditions. However, few studies have compared machine learning algorithms in terms of the prediction of continuous soil salinity for more than one study area in dryland regions (such as in the XJUAR). To address this knowledge gap, our research compared the soil salinity predictions of multiple machine learning algorithms for multiple study areas using soil observations in the XJUAR. Specifically, we compared five algorithms (LASSO, MARS, CART, RF, and SGT) to infer soil salinity values in three geographical areas distributed to the south and north of the Tianshan Mountains in the XJUAR of China (specifically, the Qitai Oasis, Kuqa Oasis, and Yutian Oasis). Each study area represented a certain type of arid landscape with different characteristic salinity-environmental relationships in the soil. Among the algorithms, LASSO, MARS and SGT are not commonly used in soil salinity prediction, and LASSO and SGT, as far as we know, have never been used to predict soil salinity.

This study has two main purposes: (i) to identify sensitive variables suitable for predicting soil salinity

in the XJUAR; (ii) to evaluate and compare the efficiency of the five algorithms in predicting soil salinity in these three oases.

Materials and methods

Study area

The Qitai Oasis is located in the northern piedmont of the Tianshan Mountains, just south of the Junggar Basin in the XJUAR of China (Figure 1). It is centred at 89.60°N longitude and 44.05°E latitude. The soil types primarily include Haplic Gypsisols, Cumulic Anthrosols, Calcaric Fluvisols, Gypsic Solonchaks, and Gleyic Phaeozems. The vegetation types include temperate semi-shrub and semi-dwarf shrub, temperate salinized dwarf semi-shrub, temperate rosette dwarf grass/semi-shrub steppe, annual crops, and drought-resistant economic crops. The natural vegetation consists of *Achnatherum splendens (Trin.) Nevski, Alhagi sparsifolia Shap., Kalidium foliatum (Pall.) Moq, Halocnemum strobilaceum (Pall.) Bieb*, and *Salsola brachiata Pall.* The elevation ranges from 568 to 978 m. The mean annual precipitation is 184.8 mm, and the majority of precipitation occurs between June and August; the average annual evaporation is 2141 mm, and the mean annual temperature ranges from approximately 5.1 to 6.1°C. The salt-affected land in the Qitai Oasis accounts for 31% of the total agricultural area (Zhang, Tashpolat, Ding, Tian, & Mamat, 2009). The salt type in this area is mainly sulfate, followed by chloride-sulfate, with a small proportion of chloride.

The Kuqa Oasis is located in the northwestern part of the Tarim Basin (Figure 1). This study area is centred at 82.50°N and 41.38°E and consists of a low-lying alluvial fan plain, a low-elevation alluvial floodplain, and a mid-elevation alluvial fan plain, with low-elevation fixed grass shrub and low-elevation semi-fixed grass shrub. The elevation in this area ranges from 892 to 1100 m, decreasing from northwest to southeast. The soil types primarily include Cumulic Anthrosols, Salic Fluvisols, Gypsic/Calcic Solonchaks, Cambic Arenosols, Calcic Vertisols and Calcaric Phaeozems. The vegetation cover is lower over the salt-affected land and is dominated by desert species, such as *Phragmites australis, Tamarix ramosissima, Alhagi sparsifolia, Karelina caspica* and *Kalidium gracile* (Jiang, Ding, Tashpolat, Zhao, & Zhang, 2008). The area has an extremely arid desert climate, with a mean annual precipitation of 51.6 mm, a mean potential annual evapotranspiration of 2356 mm and a mean annual temperature of 11.3°C. In this region, more than 50% of the cultivated land exhibits salinization, with 30% exhibiting serious salinization (Wang et al., 2015). The salt types in this area are mainly chloride and sulfate, accompanied by a small proportion of chloride-sulfate.

The Yutian Oasis belongs to the Keriya River Basin (81°09–82°51′E and 35°14′-39°29′N) (Figure 1). The catchment is located between the southern margin of the Taklimakan Desert and the northern slope the central Kunlun Mountains, with rugged terrain and a gradual elevation gradient from low elevations in the south to higher elevations in the

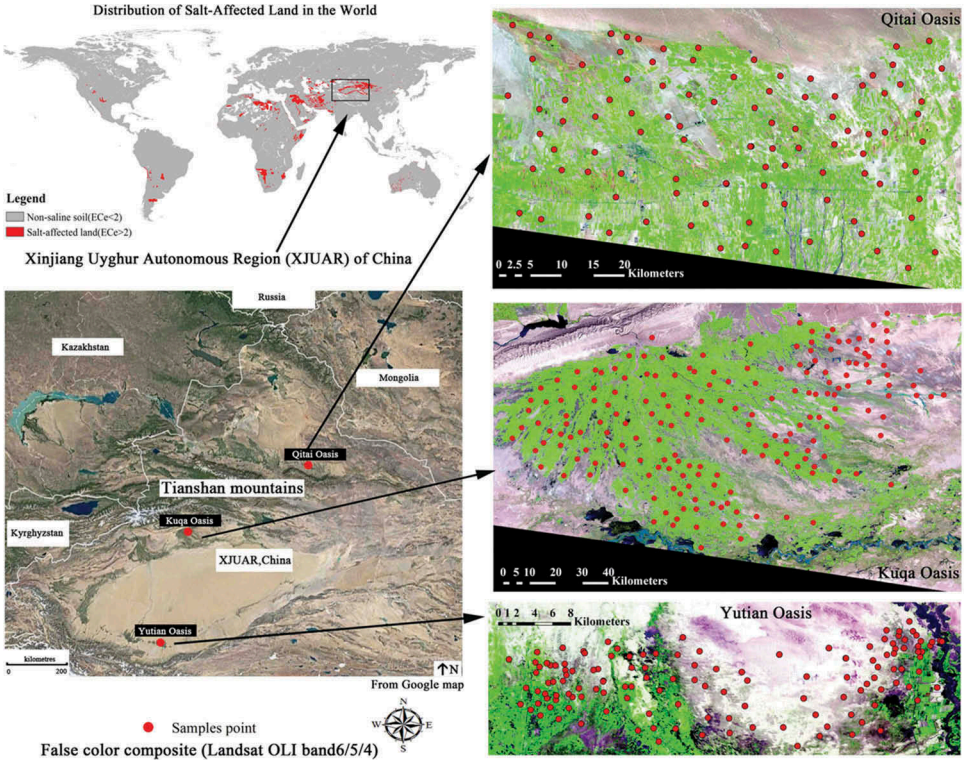

Figure 1. Location of the study area and the distribution of sampling sites.

north, ranging from 1310 m to 1491 m. The soil types primarily include Cambic Arenosols, Gleyic Phaeozems and Mollic Gleysols. Cotton, maize and wheat are the main crops. The native vegetation consists primarily of *Populus euphratica, Tamarix chinensis* and *Phragmites australis*. This area is a typical oasis-desert ecosystem, with a dry climate, scarce precipitation, intense evaporation, water shortages, and extreme vulnerability. Meteorological data statistics (1998–2015) show that the average annual precipitation, mean annual temperature and mean potential annual evapotranspiration are 47.1 mm, 12.4°C and 2498 mm, respectively. The degree of mineralization of underground water is high (Gong, Ran, He, & Tashpolat, 2015b). Hence, the intense evaporation has deposited dissolved salts at the surface, resulting in high salt concentrations.

The aforementioned three oases as test targets were mainly selected for the following reasons: First, Xinjiang covers three climatic zones: the middle temperate zone, the warm temperate zone, and the plateau climate zone (Shi et al., 2014). Among them, the annual accumulated temperature in the middle temperate zone ranges from 1600 to 3400℃, and the growth period is 100 to 171 days and for the accumulated temperature in the warm temperate zone ranges from 3400 to 4500℃, and the growth period is 171 to 218 days. The Qitai oasis is in the middle temperate zone, while the Kuqa oasis and Yutian Oasis belong to the warm temperate zone. Second, the differences in crop types and farming systems lead to different water resource allocation patterns in the three oases. The Qitai Oasis is mainly planted with grain crops, such as wheat and corn. The Kuqa and Yutian oases are mainly planted with cash crops, such as cotton. Third, the saline soil types are different among the three oases. According to Zhang, Xiong, Tian, & Luan (2011a), the proportion of sulfate in the surface soil of the Qitai Oasis is more than 65%, followed by chloride-sulfate (32%) and chloride (less than 2%). Chloride and sulfate are the main saline soils in the Kuqa Oasis, accompanied by a small proportion of sulfate-chloride soils (Zhang, Tashpolat, & Ding, 2007). Analysis results of saline soil types in the Yutian Oasis showed that the chloride-sulfate and sulfate-chloride types were the main types, followed by the chloride and sulfate-chloride type(Gong, Liu, & Tashpolat, 2015a). Therefore, the study shows that these three target areas basically represent the soil environment in the arid regions of Xinjiang.

Landsat OLI and preparation

The satellite imagery used in this study to establish the model was Landsat OLI images in Qitai Oasis, the Kuqa Oasis and the Yutian Oasis. See Table 1 for detail. All of the satellite images were obtained using the FLAASH (Fast Line-of-sight Atmospheric Analysis of Spectral Hypercubes) model from the Environment for Visualizing Images (ENVI 5.1) for the three study areas to minimize and normalize the additive and multiplicative effects of the atmosphere and the sun illumination geometry on the imagery. The FLAASH reflectance was rescaled to a normal range of 0 to 1.

Soil sampling and analysis

The sampling locations were randomly selected (relatively evenly distributed) based on soil scientists' knowledge of the Qitai Oasis, including soil type, landscape, geomorphology and accessibility (these factors were also considered for the Kuqa Oasis and the Yutian Oasis). See Table 1 for the specific sampling period of each oasis. Variations in field conditions, such as soil salinity levels, vegetation types and land use types, were fully examined by comparisons with laboratory results and existing geographic maps. The sampling point selection in the field assumed that the soil properties and vegetation species were relatively consistent, the environmental factors were similar, and the heterogeneity was relatively minimal. Finally, 101 soil sampling (0–10 cm) locations were available. The distribution of the sample plots is shown in Figure 1.

A total of 189 soil samples (0–10 cm) were collected at the Kuqa Oasis. The sites spanned the full range of local geographic landforms, soil conditions, land use types and vegetation types. The distribution of the sample plots is shown in Figure 1.

Surface soil samples (0–10 cm) were collected from local five land landscape types in the Yutian Oasis. These samples were located in farmland that had been cultivated for 12 years with irrigation water (mainly groundwater), desert-oasis ecotone covered with salinized dwarf grass near the farmland, flood plain near the end of the river characterized by high levels of vegetation coverage and a high groundwater table, site near the Keriya River bank that was adjacent to the farmland, barren land site with very low vegetation cover, and a desert site with low and sparse plant cover. A total of 100 samples points

Table 1. Date of image acquisition (Landsat OLI) and sampling period in this study.

Location	Path-row number	Date of Image acquisition	Sampling period
Qitai Oasis	141–29	19 April 2017;8 July 2017; 9 August 2017	August 1 to 13, 2017
Kuqa Oaisis	145–31	15 June 2016; 2 August 2016; 19 September 2016	September 15 to 28, 2016
Yutian Oasis	145–34	15 April 2017; 2 June 2017; 22 September 2017	September 05 to 15, 2017

(Figure 1), covering the main local landforms, were obtained in this area.

The soil salt content (g/kg) was analysed in a laboratory. The composite soil samples were air dried, ground, and sieved using a 2-mm sieve mesh. The soil salt content was determined by the residue drying method. A 5:1 soil: water mixture was extracted by the residue drying method and was used to determine the major salt ions in the soil. Na^+, K^+, Ca^{2+}, and Mg^{2+} cations were determined by atomic absorption spectrometry (AA-6800, Daojin, Japan) and Cl–, CO_3^{2-}, HCO_3^-, and SO_4^{2-} ions were determined by ion chromatography (IC-2000, Diane, America). These measurements were successively performed at the State Key Laboratory of Desert and Oasis Ecology in the Xinjiang Institute of Ecology and Geography at the Chinese Academy of Sciences.

Environmental covariates

In this study, the environmental covariates for soil salinity prediction were selected based on the SCORPAN formula (Mulder, De. Bruin, Schaepman, & Mayrc, 2016). These covariates included bands, climate factor (referring to land surface temperature), vegetation indices, salinity and soil-related indices, soil moist indices. See Table 2 for details.

Covariate selection using an inferior eliminated mechanism (IEM)

Variable reduction has been previously shown to result in slight error reductions (Svetnik et al., 2003) through the removal of potentially irrelevant predictor variables. This process allows the algorithm to progressively increase the accuracy of the prediction by reducing the chance of obtaining outliers since weak learners also produce weak outliers. In this study, all covariates were divided into 3 groups: Landsat OLI derived covariate sets, DEM-derived covariate sets and full covariate sets. Variable reduction was tested in the first three groups to examine whether or not a smaller set of predictors (optimal dataset) would lead to improvements in the five algorithms based on the following procedure, which was mainly adopted (with only small differences) from Svetnik et al. (2003) and Heung, Bulmer, and Schmidt (2014):

(1) The machine learning algorithms were initially applied to the first three variable groups; the variable importance, based on the mean decrease in accuracy, was used to rank the predictor variables.
(2) Using the variable rankings, the least important predictors were removed. In the studies of Svetnik et al. (2003) and Heung et al. (2014), the three least important predictors were removed each time. This study considered that the initial stage of variable selection could be manipulated in this way. However, when the number of variables was reduced to a threshold according to changes in the root mean square error (RMSE) and R^2, even important variables could be deleted; thus, only one variable at a time was deleted in this study.
(3) The training data were then partitioned into five cross-validation (CV) segments, and the error rates for each of the 5 CV partitions were aggregated into a mean error rate. A total of 10 replicas of the 5-fold CV were performed.
(4) Steps 2 and 3 were repeated until a balance was achieved between the number of predictors and the mean error.
(5) Steps 2 to 4 were applied to all five machine learning algorithms for three variable groups at each of the three study sites.
(6) Then, the OCG was calculated from all covariate groups for each oasis optimized by Steps 1 through 4.

Figure 2 shows iterative processes and precision trajectories ranging from the last 40 variables to the last two variables from full covariate sets using the SGT at three oases. According to the trajectory changes of R^2 and RMSE value, the study concluded that when the number of variables in the optimal data set is seven, seven, two, the prediction accuracy of SGT is relatively highest in a specific period of three oases cross all 40 tests.

Machine learning approaches and parameter initialization

Four machine learning approaches were performed using the following R packages: "glmnet" for LASSO (Friedman, Hastie, & Tibshirani, 2010), "caret" for CART (Kuhn, Leeuw, & Zeileis, 2008), "randomFForest" for RF (Liaw & Wiener, 2002), and "gbm" for SGT (Ridgeway, 2015). MARS was run in Matlab 2014a.

Tibshirani (1996) developed LASSO, a penalized likelihood approach, for linear regression. LASSO is a combination of ridge regression and subset selection developed to improve the ordinary least squares (OLS) technique by shrinking the coefficient values and setting several values equal to zero. As a result, LASSO simultaneously achieves variable selection and regression modelling (Yan & Yao, 2015). A great advantage of LASSO is that it produces simpler models than ridge regression. In LASSO, the three parameters refer to the type of loss function in the regression (least squares), points (200) and steps (5000) that need be set; the default values were adopted.

MARS is a relatively new approach and is typically known as a nonparametric method that estimates complex nonlinear relationships among independent

Table 2. Environmental covariates derived from a 30-m spatial resolution DEM and 30-m Landsat OLI imagery.

Auxiliary data	Index	Abbrev	Formulations	References
	All bands			
Landsat	Tasseled Cap	TC1,2,3		
OLI	Principal Component Analysis	PC1		
	Normalized Difference Vegetation Index	NDVI	(B5 − B4)/(B5 + B4)	Rouse, Haas, Schell, & Deering, 1973 Huete, 1988
	Soil Adjusted Vegetation Index	SAVI	[(B5 − B4) × (1 + L)]/ (B5 + B4 + L)	
	Enhanced Vegetation Index	EVI	g × (B5−B4)/(B5 + C1× B4−C2× B2 + L)	Huete et al., 2002
	Generalized Difference Vegetation Index	GDVI	$(B5^2 - B4^2)/(B5^2 + B4^2)$	Wu, 2014b
VD	Canopy Response Salinity Index	CRSI	$([(B5\times B4)-(B3\times B2)]/[(B5\times B4) + (B3\times B2)])^{0.5}$	Scudiero et al. (2014)
	Simple Ratio vegetation index	SR	B5/B4	Jordan, 1969
	Two-band enhanced vegetation index	EVI2	2.5*(B5-b4)/(B5 + 2.4*B4 + 1)	Jiang et al., 2008
	Extended NDVI	ENDVI	(B5+ B7-B4)/(B5+ B7+ B4)	Chen et al., 2015
	Extended EVI	EEVI	/(B5 + 2.5×(B6 + 6× B5+ B4-7.5× B6-B4) B2 + 1)	Chen et al., 2015
	Salinity index	SIT	(B4/B5)*100	Allbed et al., 2014
	Salinity index	SI	(B4−B5)/(B5 + B4)	Khan et al., 2005
	Salinity index	SI1	$(B4\times B3)^{0.5}$	Douaoui et al., 2006
	Salinity index	SI2	$[(B5)2 + (B4)2 \times(B3)2]^{0.5}$	Douaoui et al., 2006
SI	Salinity index	SI3	$[(B4)^2 + (B3)^2]^{0.5}$	Douaoui et al., 2006
	Salinity index	SIA	B2/B4	Allbed et al., 2014
	Salinity index	SIB	(B2−B4)/(B2+ B4)	Allbed et al., 2014
	Salinity ratio index	SAIO	(B4−B5)/(B3+ B5)	Metternicht & Zinck, 2003
	Clay index	CLEX	B6/B7	Boettinger et al., 2008
	Gypsum index	GYEX	(B6−B5)/(B6+ B5)	Nield, Boettnger, and Ramsey (2007)
	Brightness index	BREX	$(B3^2+ B4^2)^{0.5}$	Metternicht & Zinck, 2003
	Carbonate index	CAEX	B4/B3	Boettinger et al., 2008
	FSEN	FSEN	(B6-B7)/(B6+ B7)	
	Colour indices (Hue, Saturation, Value)	H/S/V		Yu et al., 2010
	Normalized Difference Infrared Index	NDII	(B5-B6)/(B5+ B6)	
	Global Vegetation Moisture Index	GVMI	((B5 + 0.1)-(B6 + 0.02))/((B5 + 0.1) +(B6 + 0.02))	Hardisky, Klemas, & Smart, 1983 Ceccato, Gobron, Flasse, Pinty, & Tarantola, 2002
	Valley Depth	VD		
	Vertical Distance to Channel Network	VDCN		
	LS-Factor	LSF		
	Topographic wetness index	TWI		SAGA
	Slope Length	SL		Development
	Sky View Factor	SVF		Team, 2011
	Topographic Position Index	TPI		
	Multiresolution Index Of Valley Bottom	MRVB		
	Flatness	F/MRRTF		
DD	Slope Height	SH		
	Normalized Height	NH		
	Standardized Height	STH		
	Mid-slope Position	MSP		
	Terrain Surface Texture	TEX		
	Flow Accumulation	FA		
	Cross-Section Curvature	CSC		
	Longitudinal Curvature	LC		
	Relative Slope Position	RSP		
	Catchment Slope	CS		

VD: Vegetation indices; SI: Soil-related indices; DD: DEM derivatives

and dependent variables (Friedman, 1991). This algorithm has only been applied to the field of visible and near-infrared reflectance spectroscopy, which was employed to predict soil salinity (Nawar et al., 2014). The MARS approach was executed using MATLAB software. During the process, input variables were divided into intervals (subsets), and basic functions were fitted to each interval. The basis function represented information about the independent variables, which was defined over a specific range; its initial and last points were called knots. A knot represents a point where the function behaviour changes. Therefore, parameters that have knot and basic functions have an important role to play in obtaining optimum results in MARS. More detailed information about MARS can be found in Cheng and

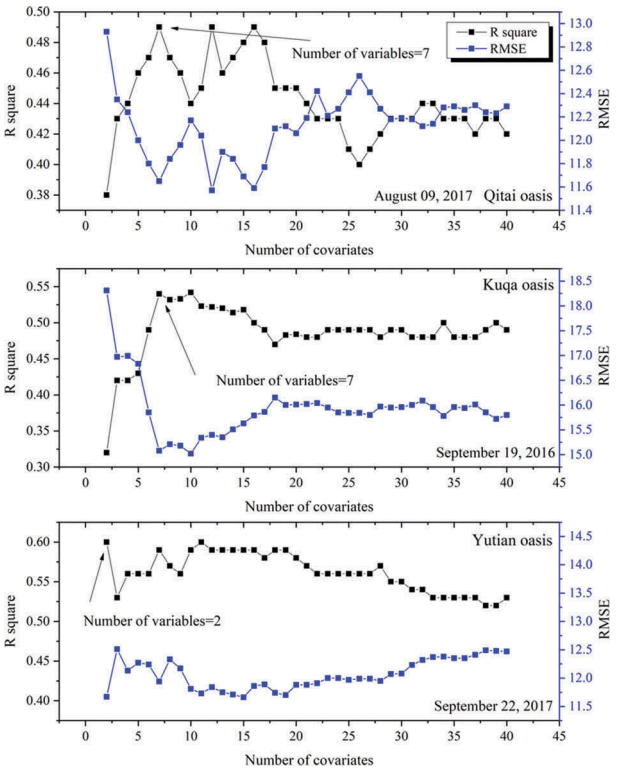

Figure 2. Iterative processes and precision trajectories ranging from the last 40 variables to the last two variables from full covariate sets using the SGT at three oases.

Cao (2014). Ultimately, the initial value of the knot was set to 3, and the maximum basis functions were set to 15.

CART models are very attractive due to the interpretability of node splits (i.e., rules), the avoidance of parametric assumptions (i.e., distribution and independent residuals) and their ability to handle noisy data. The common statistics of least squares are used in regression trees for training datasets. Moreover, the maximum depth was set to 5.

The RF learner is conceptually similar to tree-based learners and shares the same advantages; however, multiple decision trees are trained, and the results are based on predictions from an ensemble of individual trees (Breiman, 2001). More detailed information can be found in Heung et al. (2016). To construct the relationships for RF, only two parameters are defined by the user before running the RF algorithm: the minimum size of the terminal nodes and the number of variables randomly selected for each node, which are defined by m (commonly defined as the square root (sqrt) of the number of independent variables). We set this to 2*sqrt after repeated calculations and comparisons. Furthermore, the number of proximal cases and bootstrap sample sizes were changed to AUTO in the options. The default value of ntree (200) was demonstrated to be insufficient for yielding stable results (Grimm et al., 2008) in the three oases after testing. Therefore, we set the ntree value to 1000 for all tests with the RF model.

Overall, SGT (a tree-based method) was an improvement over CART and generally resulted in accurate and robust predictions (with low overfitting effects) of spatial heterogeneity and outliers (Brenning, 2005; Friedman, 2002). This powerful algorithm allows for calculations without a priori assumptions regarding the variables, which might influence the predicted values; this provides more flexibility than traditional generalized linear or additive models (Friedman, 2002). So far, the SGT has been implemented mainly for quantifying the spatial distribution of plants and animals (Mullet, Gage, Morton, & Huettmann, 2016), water erosion susceptibility (Angileri et al., 2016), topsoil carbon stocks (Schillaci et al., 2017) and oil pollution (Fox et al., 2016). However, little information is available concerning soil salinity modelling using the SGT in dryland areas. The SGT fits a simple parameterization function (i.e., base learner) for pseudo-residuals by the least squares method (after comparing multiple loss functions, the least squares was finally chosen) using sequential iterations to construct additive regression models. To avoid overfitting, a subsample fraction was set to 0.75 (Angileri et al., 2016). The number of trees was set to 1000, and the maximum number of nodes per tree was set to 6 (Schillaci et al., 2017).

Model validation

For each algorithm, a 5-fold cross validation (CV) was used to generate optimized parameters (Heung

et al., 2014). The advantage of this method is that it exhibits reliable performance and is unbiased for smaller data sets because the process requires much more computational effort than simple trained-and-tested (i.e., hold out) procedures (Taghizadeh-Mehrjardi, Nabiollahi, & Kerry, 2016; Zhao, Popescu, & Meng, 2011). The training dataset was randomly partitioned into five subsets, where four of the five subsets comprised 80% of the observations and were used for model training, and the fifth subset with 20% of the observations was used for model validation. Based on a study of machine learning presented by Heung et al. (2016), this process was repeated 10 times, using each round for validation once for all five algorithms (LASSO, MARS, CART, RF and SGT) at each of the three study areas. Three validation measurements were used to quantify the model performance of the simulations: coefficient of determination (R^2), root mean square error (RMSE) and relative root mean square error (RRMSE). The predictions were considered increasingly optimal as the RMSE and RRMSE values decreased and as the R^2 value increased.

Results

Descriptive statistics of soil salinity

Summary statistics of soil salinity for the three oases are presented in Table 3. Field samples included all salinity levels (i.e., non-saline soil (<7 g/kg), low-salinity soil (7–9 g/kg), moderate-salinity soil (9–13 g/kg), high-salinity soil (13–16 g/kg) and saline soil (>16 g/kg)) (Soil Survey Staff of Xinjiang, 1996) in all three oases. A coefficient of variation (CoV) equal to 0.66 indicated moderate variability in the soil salt content in the surface soil of the Qitai Oasis (a CoV lower than 0.1 indicated low variability, whereas a CoV higher than 1.0 indicated great variability). In the Kuqa Oasis, approximately 50% of the samples were from non-salinized land due to the relatively large agricultural area in this oasis, and 37.57% of the samples were from extremely salt-affected land. This sampling scheme was in accordance with the local conditions of the land use/cover. The CoV equalled 1.23, indicating great variability in the soil salt content in the Kuqa Oasis. This result is the same as that in Gao, Ding, Ha, and Zhang (2010) for the Kuqa Oasis. In the Yutian Oasis, 52 of the 100 samples belonged to the category of extreme salinization. This result was similar to Nurmemet et al. (2015), who found a total area of 79,763.8 ha of salinized soil within the study area (41.43%), indicating that soil salinity had already become one of the major threats to local agriculture and local communities.

Comparison of prediction accuracy of machine learning

Tables 4–7 show the predicted results of the five algorithms for the 21 datasets of the three oases. Because there are too many datasets to explain one by one, the study uses average values of R^2, RMSE and RRMSE to express the comprehensive performance of the algorithm. The R^2 average values of CART, LASSO, MARS, RF and SGT calculated from 21 datasets were 0.29 (0.18–0.42), 0.27 (0.07–0.46), 0.30 (0.16–0.43), 0.36 (0.11–0.53) and 0.39 (0.2–0.55), respectively. The average values of RMSE are 15.67 (10.41–18.42), 15.86 (10.94–21.28), 15.73 (9.74–20.91), 14.96 (9.84–22.01) and 14.54 (8.80–19.01). The average values of RRMSE are 0.66 (0.42–1.02), 0.67 (0.44–1.18), 0.66 (0.39–1.16), 0.63 (0.39–1.22) and 0.61 (0.35–1.05). The aforementioned results show that the highest prediction accuracy is that of SGT, followed by RF, and the difference between the two is small. CART and MARS have similar prediction accuracy. The prediction accuracy of LASSO is the worst among all of the algorithms.

SGT and RF show stronger information mining capabilities (compared to those of MARS, CART, and LASSO) when the environmental datasets used for modelling are more complex. As can be seen from

Table 3. Descriptive statistics of the soil salinity in the three oases.

Measurement (g/kg)	Qitai oasis (N = 101)	Kuqa oasis (N = 189)	Yutian oasis (N = 100)
Mean	25.08	18.06	23.76
Maximum	67.10	88.30	84.20
Minimum	0.75	0.1	1.90
CV	0.66	1.23	0.81
Percentile			
25%	12.00	0.7	8.62
50%	22.50	7.5	17.90
75%	35.55	30.70	36.51
Salinity classification (g/kg)			
Non salt-affected (<7)	11.88%	48.68%	20.00%
Slight salt-affected (7–9)	4.95%	4.23%	8.00%
Moderate salt-affected (9–13)	9.90%	6.35%	9.00%
Severe salt-affected (13–16)	7.90%	3.17%	11.00%
Extreme salt-affected (>16)	65.37%	37.57%	52.00%

Table 4. Comparison of machine learning algorithms for soil salinity predictions with DEM derivatives in the Qitai Oasis, Kuqa Oasis and Yutian Oasis.

		Qitai			Kuqa			Yutian		
Covariate sets	Algorithm	R^2	RMSE	RRMSE	R^2	RMSE	RRMSE	R^2	RMSE	RRMSE
DEM derivatives	CART	0.18	15.05	0.60	0.39	17.38	0.96	0.16	17.43	0.73
	Lasso	0.07	16.01	0.64	0.15	21.28	1.18	0.35	15.38	0.65
	MARS	0.16	15.23	0.61	0.18	20.91	1.16	0.19	17.13	0.72
	RF	0.30	13.88	0.55	0.11	22.01	1.22	0.30	15.95	0.67
	SGT	0.20	14.85	0.59	0.27	19.01	1.05	0.38	14.65	0.62

Table 5. Comparison of machine learning algorithms for soil salinity predictions using R^2 and RMSE accuracy metrics in Qitai oasis.

		Landsat- based index			Full Environmental covariates		
Date	Algorithm	R^2	RMSE	RRMSE	R^2	RMSE	RRMSE
April 19	CART	0.26	12.16	0.48	0.26	12.16	0.48
	LASSO	0.40	10.94	0.44	0.40	10.94	0.44
	MARS	0.38	11.12	0.44	0.34	11.43	0.46
	RF	0.29	10.35	0.41	0.35	9.84	0.39
	SGT	0.24	12.30	0.49	0.45	10.50	0.42
July 08	CART	0.28	10.41	0.42	0.28	10.41	0.42
	LASSO	0.19	11.04	0.44	0.19	11.04	0.44
	MARS	0.37	9.74	0.39	0.37	9.74	0.39
	RF	0.29	10.35	0.41	0.35	9.84	0.39
	SGT	0.25	10.61	0.42	0.46	8.80	0.35
August 09	CART	0.41	12.73	0.51	0.42	12.61	0.50
	LASSO	0.31	13.75	0.55	0.33	13.53	0.54
	MARS	0.41	12.72	0.51	0.41	12.64	0.50
	RF	0.43	12.41	0.49	0.48	12.02	0.48
	SGT	0.40	12.78	0.51	0.49	11.65	0.46

Table 6. Comparison of machine learning algorithms for soil salinity predictions using R^2 and RMSE accuracy metrics in Kuqa oasis.

		Landsat- based index			Full Environmental covariates		
Date	Algorithm	R^2	RMSE	RRMSE	R^2	RMSE	RRMSE
June 15	CART	0.26	16.34	0.90	0.26	16.34	0.90
	LASSO	0.32	15.76	0.87	0.21	17.05	0.94
	MARS	0.19	17.11	0.95	0.19	17.11	0.95
	RF	0.32	15.81	0.88	0.43	14.47	0.80
	SGT	0.28	16.05	0.89	0.41	14.72	0.82
August 02	CART	0.19	18.42	1.02	0.19	18.42	1.02
	LASSO	0.21	18.09	1.00	0.21	18.09	1.00
	MARS	0.27	17.42	0.96	0.27	17.42	0.96
	RF	0.25	17.71	0.98	0.52	14.10	0.78
	SGT	0.31	16.97	0.94	0.55	13.66	0.76
September 19	CART	0.36	17.66	0.98	0.33	17.13	0.95
	LASSO	0.37	16.21	0.90	0.46	16.39	0.91
	MARS	0.43	16.86	0.93	0.42	16.98	0.94
	RF	0.44	16.85	0.93	0.46	16.79	0.93
	SGT	0.44	16.38	0.91	0.54	15.08	0.83

Table 7. Comparison of machine learning algorithms for soil salinity predictions using R2 and RMSE accuracy metrics in Yutian oasis.

		Landsat-based index			Full Environmental covariates		
Date	Algorithm	R^2	RMSE	RRMSE	R^2	RMSE	RRMSE
April 15	CART	0.28	13.22	0.56	0.284	13.22	0.56
	LASSO	0.22	13.77	0.58	0.42	11.85	0.50
	MARS	0.24	13.60	0.57	0.24	13.60	0.57
	RF	0.38	12.34	0.52	0.56	10.42	0.44
	SGT	0.44	11.64	0.49	0.46	11.50	0.48
June 02	CART	0.48	11.24	0.47	0.48	11.24	0.47
	LASSO	0.47	11.32	0.48	0.48	11.21	0.47
	MARS	0.52	10.75	0.45	0.52	10.75	0.45
	RF	0.43	11.80	0.50	0.59	9.96	0.42
	SGT	0.53	10.68	0.45	0.59	9.90	0.42
September 22	CART	0.57	12.47	0.52	0.61	11.96	0.50
	LASSO	0.49	13.54	0.57	0.44	14.25	0.60
	MARS	0.52	13.19	0.56	0.52	13.49	0.57
	RF	0.57	12.45	0.52	0.61	11.95	0.50
	SGT	0.58	12.01	0.51	0.63	11.55	0.49

Tables 4 to 7, the Landsat-based indices (dataset 1) carry more information than the digital elevation model (DEM) derivatives (dataset 2). The former has an average R^2 of 0.36 (0.19–0.58), and the latter has an average R^2 of 0.23 (0.07–0.39). After the aforementioned two datasets are simultaneously input (dataset 3 with whole environmental covariates) into the algorithm, the average prediction accuracy of the soil salinity increases (Tables 5–7). The maximum value of R^2 is 0.63, the minimum value is 0.19, and the average value is 0.41. In addition, the dates of the Landsat images and soil sampling are not consistent, which might also reduce the accuracy.

The study also found that no algorithm in the 21 datasets could be 100% better than the others. According to statistics (R^2), the percentages of the prediction accuracy of CART, LASSO, MARS, RF, and SGT that are higher than one algorithm in the 21 datasets are 23.81%, 33.33%, 42.80%, 0% and 9.52%, respectively. The percentages that are higher than two algorithms are 28.57%, 14.28%, 14.28%, 28.57% and 0%, respectively. The percentages that are higher than three algorithms are 0%, 14.28%, 19.04%, 42.85% and 19.04%, respectively. The percentages that are higher than four algorithms are 0%, 0%, 4.76%, 19.04% and 60.90%, respectively.

Stability of machine learning

The accuracy ranks of the five algorithms were used to investigate the stability of their performance in three oases (Figure 3). The research evaluated the stability of the five algorithms in the three oases from the following aspects: R^2, RMSE and RRMSE. Tables 4–7 show that each oasis has seven datasets (three Landsat-based index datasets, three whole environmental covariates datasets, and one DEM-derived dataset). The R^2, RMSE and RRMSE values of the aforementioned seven datasets were standardized by Z-score one-by-one in each oasis.

$$Z_{RRMSE} = \frac{O - \bar{O}}{\sigma} \tag{1}$$

where O is validation measurements in this paper, $\bar{O}$ is mean value, σ is standard deviation. The number of standardized data involved in each dataset was five. Then, each algorithm has a Z_R^2 value calculated based on five R^2 values, a Z_{RMSE} value calculated based on five RMSE values and a Z_{RRMSE} value calculated based on five RRMSE values in each dataset. Take the Qitai Oasis as an example; the Z_R^2 and Z_{RMSE} formulas of CART are as follows:

$$Z_{R^2} = \sum_{i=1}^{n} Z_{S_n(R^2)} \tag{2}$$

$$Z_{RMSE} = \sum_{i=1}^{n} Z_{S_n(RMSE)} \tag{3}$$

$$Z_{RRMSE} = \sum_{i=1}^{n} Z_{S_n(RRMSE)} \tag{4}$$

where S represents the dataset, n represents the nth dataset, and maximum of n value = 7. The remaining algorithms use the same method to calculate their Z_R^2, Z_{RMSE} and Z_{RRMSE} values. The aforementioned processes were cyclized in the Kuqa Oasis and Yutian Oasis, respectively, and the Z_R^2 values, Z_{RMSE} values and Z_{RRMSE} of the five algorithms for three oases were obtained. The larger the Z_R^2 value and the smaller the Z_{RMSE} and Z_{RRMSE} value, the better the performance of the algorithm. As can be seen from the Figure 3, the Z_R^2, Z_{RMSE} and Z_{RRMSE} values of SGT rank first for the Kuqa and Yutian oases, and second for the Qitai Oasis. This represents the best performance of SGT in predicting accuracy and stability. RF's performance as observed in the three oases was only inferior to that of SGT, followed by MARS and CART. The comprehensive ranking of LASSO was the worst among all of the algorithms.

Mapping of soil salinity in three oases

The prediction results of five algorithms in the three oases are shown in Figures 4–6. The southern part of the Qitai Oasis is mainly farmland, and the soil salinity is low. With the decrease in altitude from north to south, the groundwater level gradually rises, and soil salinization is common (Zhang et al., 2011a). Saline soil in this area mainly occurs in the oasis-desert ecotone and is dot-shaped in the irrigated farmland. In the northern semi-fixed dune area, the groundwater level declines and the soil salinization level is less compared to that of the oasis-desert ecotone, but still higher than that of the farmland. Comparing the aforementioned findings, SGT's prediction results are more consistent with the actual situation, followed by those of RF. The spatial distribution pattern of the soil salinity is not clearly shown by CART's prediction results. Although the results of LASSO can reflect the distribution pattern of soil salinity, the following two aspects are quite different from the actual survey: the range of soil salinity and the salinity content of the saline-alkali soil in the oasis is at the same level as that of the semi-fixed dunes. The results of MARS show that the soil salinity tends to the same level in a desert area and in farmland, which is not consistent with the reality. The soil salinization of the Kuqa Oasis is mainly distributed in the oasis-desert ecotone outside the

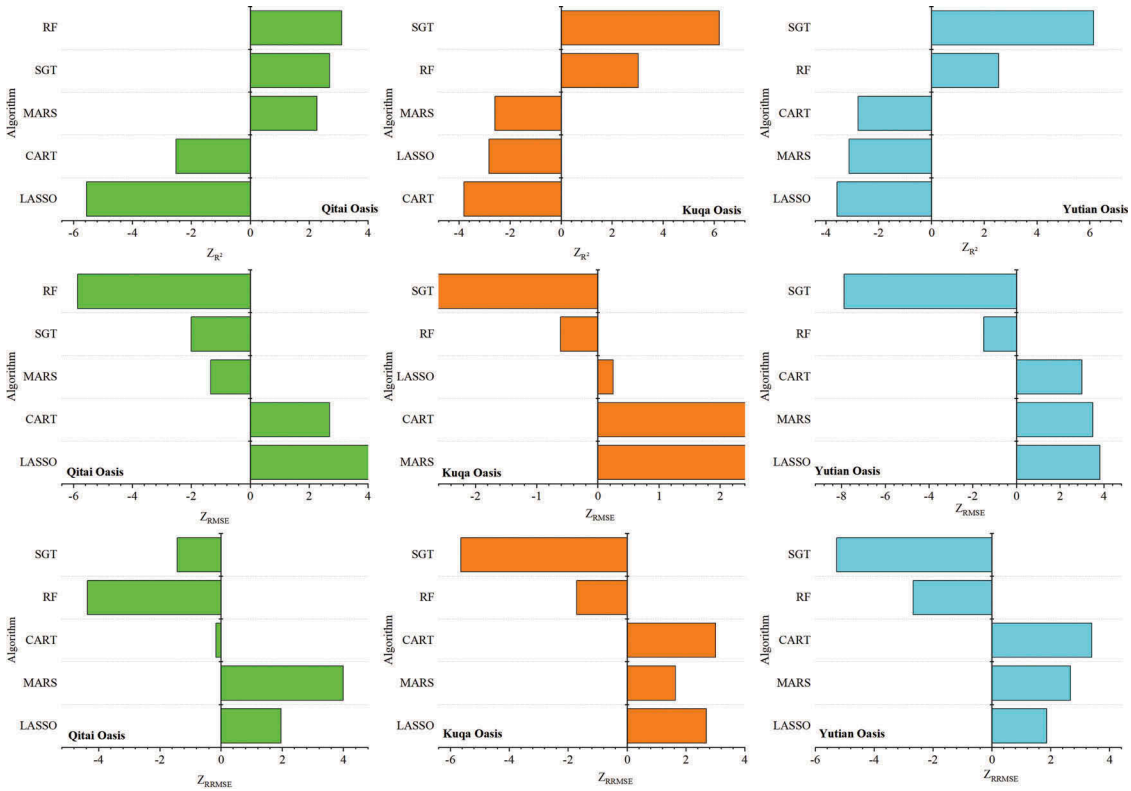

Figure 3. Accuracy ranks of the five algorithms used to investigate the stability of their performance in three oases. The R^2 and RMSE values of the seven datasets are standardized by Z-score one-by-one for each oasis. The number of standardized data involved in each dataset was five. Then, each algorithm has a $Z_R{}^2$ value calculated based on five R^2 values and a ZRMSE value calculated based on five RMSE values in each dataset. Finally, seven Z_{R2} values or seven ZRMSE values were added to represent the performance of each algorithm in each oasis.

irrigation area (Ding & Yu, 2014). The results of SGT and RF are more consistent with our understanding of the distribution of saline soil in this area. The prediction results of CART obviously show binarization. At the same time, there is no textural information in the area with a similar range of soil salt content. The severity of the soil salinity in the Kuqa Oasis is not accurate. Extreme outliers are found in the prediction results of LASSO and MARS, but the distribution pattern of soil salinity can be distinguished. In the Yutian Oasis, the saline soil mainly occurs in bare land around the irrigation area (the middle of the study area) and the buffer zone on both sides of the river (the east side of the study area) (Hu, Tashpolat, Yu, & Zhang, 2017). The prediction results of the five algorithms are consistent with the survey only in terms of pattern distribution. From the range and textural information, the prediction results of SGT and RF are closer to the actual situation. However, negativity and outliers were found in the prediction results of CART, LASSO and MARS. In addition, with the passage of time, the range of prediction results of these three algorithms greatly fluctuates. In summary, combined with the results of R^2, RMSE and RRMSE, it is considered that SGT is the preferred algorithm for soil salinity prediction in arid areas, followed by RF.

Important variables in the prediction of soil salinity

Figure 7 shows the environmental variables with a frequency of occurrence greater than 1, counting all 21 optimal datasets from the three oases. After five different algorithms iteration, a small number of variables will appear in five optimal data sets at the same time, but the importance of them is different. In the same area, the variables of the optimal dataset will change during different seasons. Tables 4–7 show a large data collection and it is not easy to see the law of change. Therefore, the frequency of occurrence of the variables and their importance (%) in various datasets were used to determine which variables contributed more to the prediction accuracy of the soil salinity in the arid areas. The specific approach was to add the frequency of occurrence to characterize its importance in the prediction of soil salinity in an arid oasis. In addition, the range of relative importance (%) of each variable (frequency > 1) was calculated to better understand its contribution to the modelling process. Although these contributions

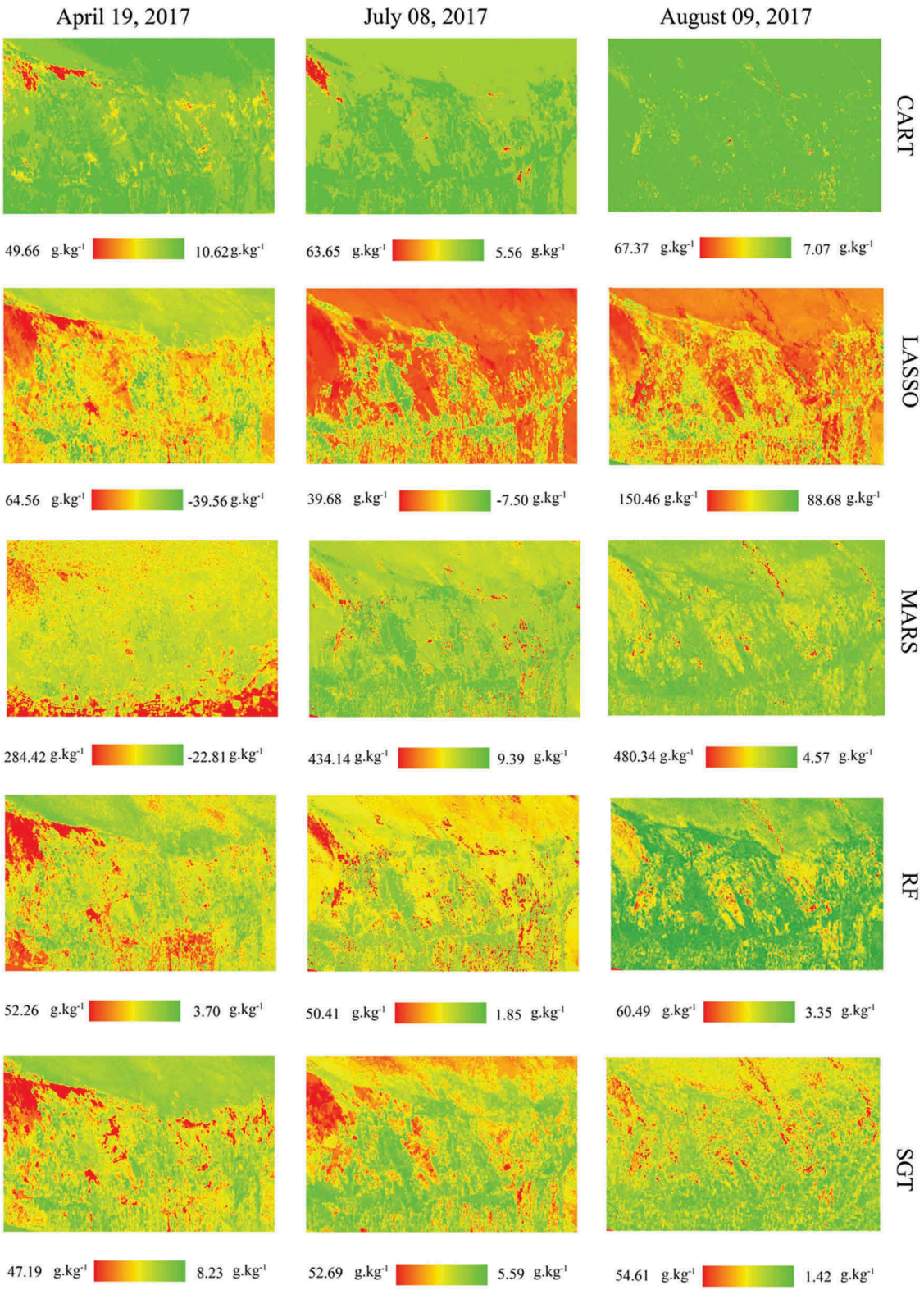

Figure 4. Distribution of soil salinity (g/kg) predictions from LASSO, MARS, CART, RF and SGT in Qitai.

originated from different algorithms and datasets, this study assumed that the variable has the potential to identify saline soils as long as it appears in the optimal datasets. Figure 7 shows that the top five environmental variables in the Qitai Oasis were ENDVI (nine times), EEVI (five times), CSRI (four times), B1 (four times) and EVI2 (two times). The importance value (%) of ENDVI and CSRI is higher than that of EEVI and B1. In the Kuqa Oasis, EVI2 (six times), GDVI (six times), NDII (six times), SAIO (six times) and EEVI (four times) were the top five environmental variables. The overall contribution of EVI and GDVI is relatively high and the contribution is relatively stable among the various datasets. The results of the Yutian Oasis show that EEVI, CSRI, EVI2,

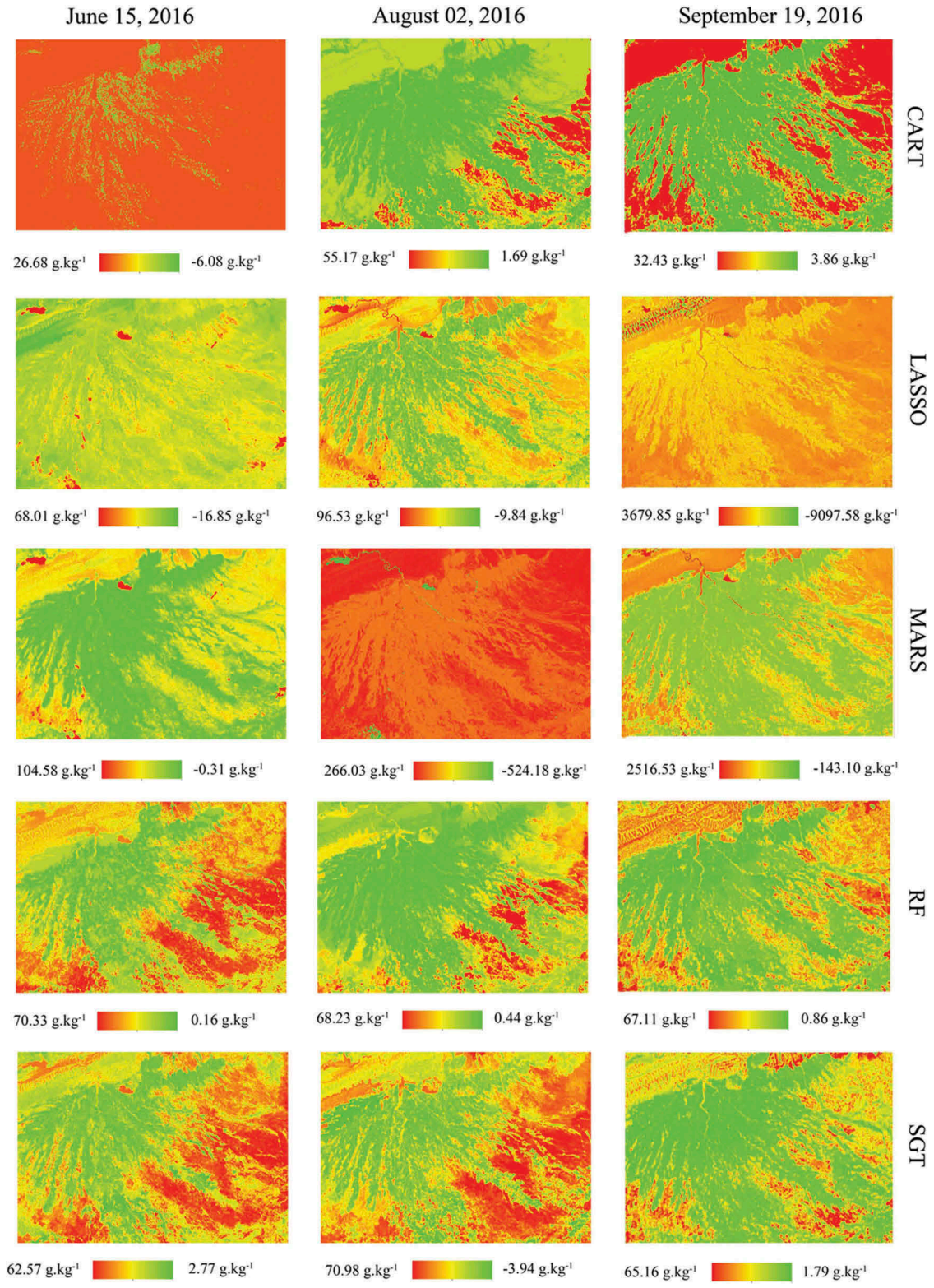

Figure 5. Distribution of soil salinity (g/kg) predictions from LASSO, MARS, CART, RF and SGT in Kuqa.

ENDVI and SAIO play an important role in the establishment of soil salinity prediction models. The relative contribution of EEVI and CSRI in multiple optimal datasets is 100%. When all the variables (frequency of occurrence of variable> 1) from the three regions were added together, the order of frequency from high to low was (frequency of occurrence of the variable > 10 times) as follows: EEVI, EVI2, ENDVI, CSRI, SAIO and GDVI.

Discussion

Performance of machine learning

From the aforementioned results (Table 4–7), we can see that the performances of SGT and RF are

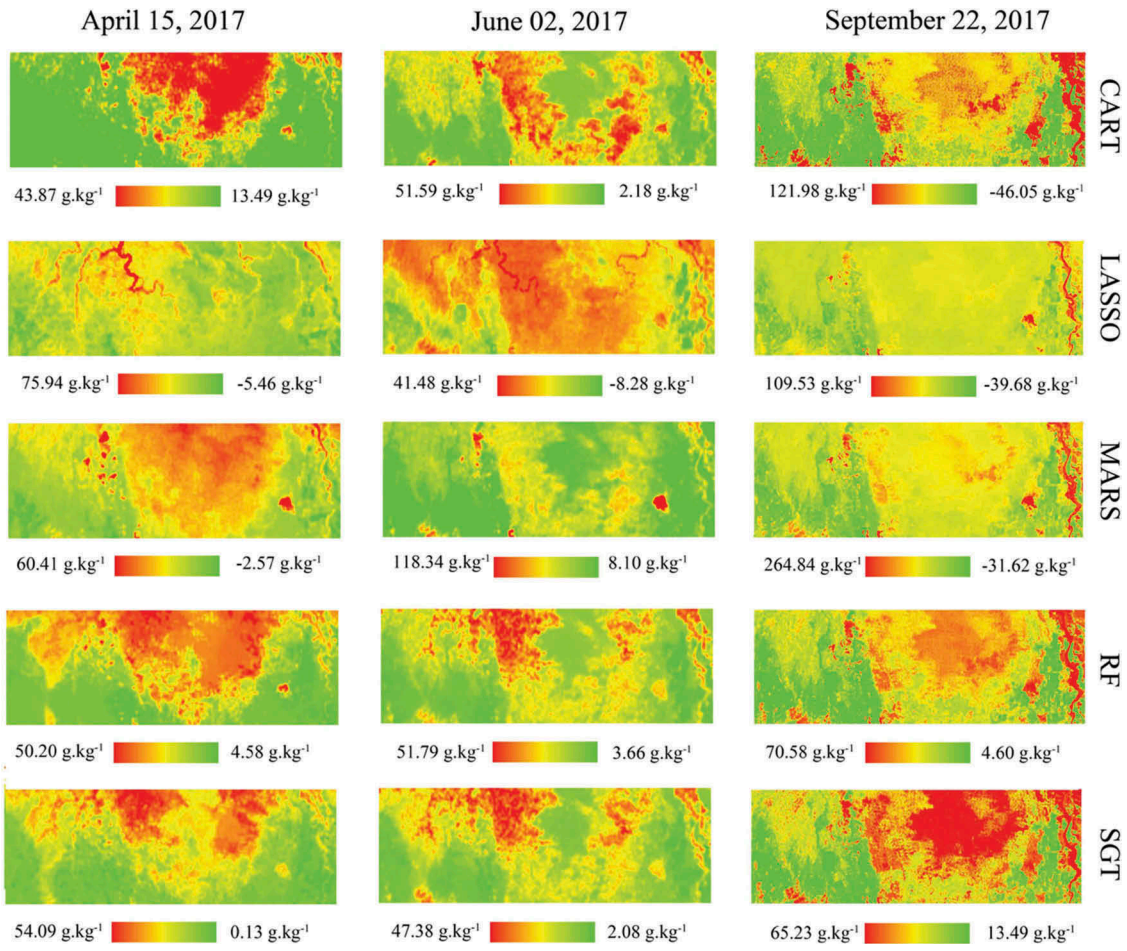

Figure 6. Distribution of soil salinity (g/kg) predictions from LASSO, MARS, CART, RF and SGT in Yutian.

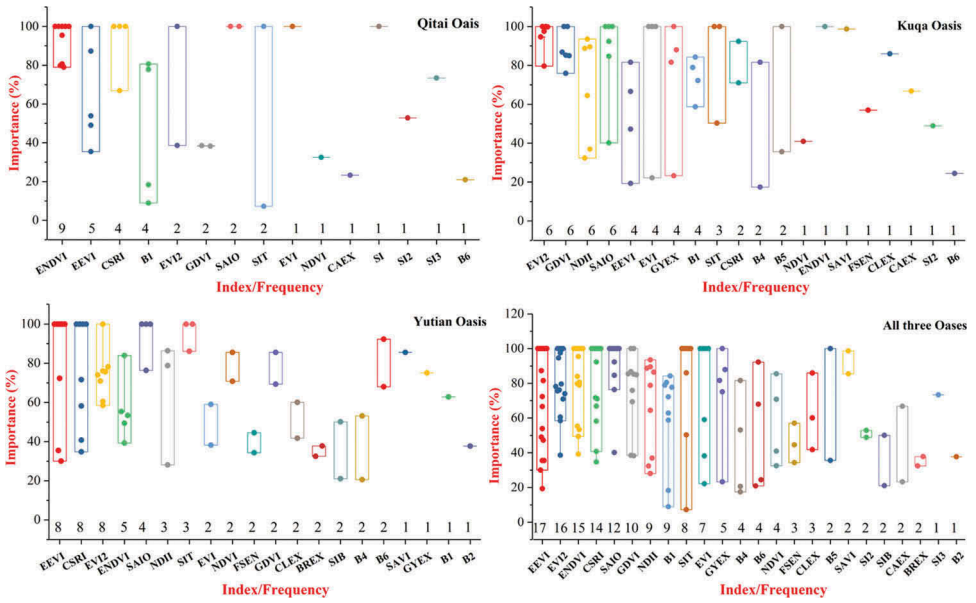

Figure 7. Excellent indices of soil salinity with frequency of occurrence greater than 1 from 21 optimal variables set in three oases.

obviously higher than those of CART, MARS and LASSO. Referring to the average R^2, RMSE and RRMSE values, the performances of SGT and RF in the 21 data sets are significantly better than those of CART, MARS, and LASSO. Comparing the prediction results of datasets 3 and 1, we found that the prediction accuracy of RF and SGT (R^2) increased by 32.94% and 39.03%, on average. The modelling accuracy of CART, LASSO and MARS increased by −0.19%, 9.08% and −1.38%, respectively. The reason

is that dataset 3 contains relatively higher information dimensions, including vegetation type and vitality, surface reflectance, surface texture (different scales), terrain variables that indirectly represent hydrological changes, parent materials and so on. Furthermore, the results also show that CART, LASSO and MARS are obviously inadequate in mining useful information from complex variable datasets in this study, that is the data utilization rate is lower than that of both RF and SGT. In all 21 data sets, the probability that the predictive accuracy of SGT and RF is better than three or four algorithms is significantly higher than that for the remaining three algorithms. The reasons for the aforementioned performance differences may be as follows. The relationships between soil property variations and the underlying environmental variables can be very complex and an assumption of linearity is often difficult to meet. RF and SGT have more power to model highly nonlinear dimensional relationships compared to that of CART, LASSO and MARS. For RF and SGT, as Breiman (2001) stated, weaker standalone models tend to be more effective when combined. By aggregating multiple models, the instability of a single-tree model is minimized, which leads to an improvement in consistency (Breiman, 1996). For example, although DEM-derived variables have relatively weak explanatory power for the spatial variability of soil salinity (average R^2 value = 0.23), the dataset still has a certain amount of information (Table 4). When ensemble-learning methods were introduced to the model for RF and SGT, consistency drastically increased. Heung et al. (2016) stated that when the relationship between environment variables and soil properties is more complex, the introduction of a random variable selection technique is also an effective means to improve the consistency between the predicted results and the measured data. In contrast, CART only uses a single tree to learn the complex relationship between the spatial variability of soil salinity and a large number of environmental variables. A lower R^2 value and higher RMSE and RRMSE value imply that CART is incapable of addressing such complex relationships. Strobl and Augustin. (2009) stated that CART is known to be very unstable; small changes in the learning sample can produce completely different trees. MARS essentially builds flexible models by fitting piecewise linear regressions. That is, the nonlinearity of the model can be approximated through the use of separate regression slopes over distinct intervals in the predictor-variable space. However, this study found that in most cases, the number of variables in the 21 optimal datasets produced by the MARS iteration is only one. We believe that even using a nonlinear approach to building models, it is difficult to explain the spatial variability of soil salinity by using a lower information dimension or sparse variables. LASSO is ultimately a regression procedure that builds a linear model. It cannot, on its own, discover nonlinearities or interactions. This explains why this algorithm, on the whole, is far inferior to the SGT and RF.

Compared to RF, the modelling method of SGT is more suitable for spatial prediction of soil salinity in arid regions. Of the 21 datasets from three regions (Table 4–7), 19.04% showed RF with a higher accuracy than that of CART, LASSO, MARS, and SGT. However, 61.90% of the datasets showed that the prediction accuracy of SGT was higher than that of the remaining four algorithms, and the performance improved by 225.10%. In addition, although the Z_R^2 values of RF and SGT for the Qitai Oasis were similar, the Z_R^2 values of SGT for the Kuqa and Yutian oases were much higher than those of RF (Figure 3). As for the Z_{RMSE} and Z_{RRMSE} ranking, the same situation occurs in the Kuqa and Yutian oases. The aforementioned results prove that SGT is more suitable for spatial prediction of soil salinity in arid areas from the perspective of the accuracy and stability of the model. Until now, only Vermeulen and Van Niekerk (2017) have used different combinations of geomorphometric covariates for predicting soil salinity with the aid of RF. RF achieved a kappa of 0.28 for Vaalharts and a kappa of 0.5 for the Breede River. In our study, RF only used topographically derived variables to predict soil salinity in the three oases with R^2 values of 0.30, 0.11, and 0.30, respectively. The difference between these two studies was that the former was used for classification and the latter for quantitative studies. Other algorithms are not covered in the field of soil salinity prediction research. Therefore, our research results are not easy to compare to the results obtained by other authors. There are two reasons: first, to retrieve the existing research results on soil salinization modelling, we noted that the results were diverse because of differences in sampling depth, selection of variables, number of observations, prediction techniques, prediction accuracy, verification methods (a linear fit with no validation, s training set/verification set with certain proportions, and a spatial leave-one-field-out cross-validation have been used), and geographical environments (plains, arid lands, coastlands, inland locations and river valleys). Second, comparative studies of RF and SGT in soil salinity prediction were not involved. Therefore, we quote previous research results in other fields to illustrate the reliability of this study. Naghibi and Pourghasemi (2015) used SGT, CART, and RF to study the potential distribution of groundwater fountains in Afghanistan. The results showed that SGT had the highest prediction accuracy, followed by CART and RF. Youssef et al. (2016) assessed landslide hazard in Saudi Arabia

based on generalized linear models (GLM), CART, SGT, and RF. The AUC (area under the curve) showed that SGT had the highest value of 0.958, followed by GLM at 0.821, CART at 0.816, and finally RF at 0.783. The greater the value, the higher the precision. Yang et al. (2016) compared the spatial distribution of soil organic matter predicted using SGT and RF in the high vegetation coverage area of the northeastern Qinghai-Tibet Plateau. The results showed that the prediction accuracy of SGT was slightly higher than that of RF.

On the whole, MARS shows better predictive accuracy than that of CART and LASSO. Referring to the R^2, RMSE and RRMSE values of the 21 datasets (Table 4–7 and Figure 3), the CART and MARS predictions are similar in terms of accuracy. Comparing the distribution characteristics of the R^2 values of the 21 datasets, the probability that the prediction accuracy of CART and MARS is higher than that of three and four algorithms is 0% and 14.28% and 0% and 4.76%, respectively. The results show that MARS performs better than CART. In terms of ranking of the $Z_R{}^2$ values, the total scores of MARS, CART, and LASSO in the three regions are −3.44, −9.10, and −11.96, respectively. The ranking of the Z_{RMSE} values is 5.34, 8.44, and 10.57, respectively. The ranking of the Z_{RRMSE} values is 6.24, 6.52, and 8.31, respectively. These results imply that the comprehensive ability of MARS is relatively excellent (it has higher prediction accuracy and better stability), followed by CART, and the worst is LASSO. The literature search found that there was no relevant field at the same time to carry out comparative studies between MARS, CART, and LASSO. Here, we tried to quote the results of previous studies in different fields to illustrate the credibility of the aforementioned analysis. Gretchen and Tracey (2002) compared five modelling techniques for predicting forest characteristics in the United States, and obtained better results using MARS than when using GAM, ANN, simple linear models (LM) and CART. Álvaro, Schnabel, and Contador (2009) showed better performance for MARS in predicting gullying with areas under the ROC curve of 0.98 and 0.97 for the validation datasets, while CART presented values of 0.96 and 0.66. The results of Gregory, Jamieson, Bezanson, and Hansena (2013) indicated that the MARS models outperformed LASSO for predicting E.coli particle attachment and virulence marker occurrence.

Excellent indices of soil salinity modelling in arid area

Several important soil salinity-sensitive variables were found by comparison of the frequency of occurrence in all oases (Figure 7). The sensitivity of these variables to soil salinity in each oasis shows the following characteristics. At the same time, the indexes with the recognition function of soil salinization and a frequency greater than 1 were as follows: CSRI, ENDVI, EEVI, B1, EVI2, SAIO, SIT, EVI, NDVI and B6. Among these variables, the frequency of occurrence of CSRI, EEVI, EVI2, GDVI, SAIO and SIT in each oasis is greater than 2. Only the frequency of EEVI appears more than four times in each oasis: five times/four times/eight times. Most of the aforementioned 10 variables have a certain degree of soil salinity environmental bias; i.e. the frequency is higher in one or two oases at the same time. For example, the frequency of occurrence of ENDVI, GDVI, EVI2, CSRI and B1 in the Qitai, Kuqa and Yutian oases were nine times/one time/five times, two times/six times/two times, two times/six times/eight times, four times/two times/eight times and four times/four times/one time, respectively. We also found that the variables with the highest frequency rank also maintained a relatively high contribution in each preferred dataset, but this situation was not absolute. From the results of the frequency ranking of the three oases, the distribution of the contribution values of each variable shows a certain fluctuation. This also indicates that the relationship between soil salinity and the aforementioned variables changes when the geographical environment changes. Therefore, although the vegetation and salinity spectral indices showed satisfactory results in monitoring salinity throughout the world, notably there is no universal spectral index that can show a satisfactory result under different environmental conditions (Allbed et al., 2014). In summary, it is suggested that variables should be considered in soil salinity mapping and unknown field sampling design in arid areas as follows: EEVI, EVI2, ENDVI, CSRI, GDVI, SAIO and SIT.

The aforementioned indices work well in the identification of saline soils in arid areas, and it is speculated that this may be related to the complexity of its formula, the adjustment factor, and the number of bands involved in constructing an environmental index model. For example, CSRI, EEVI, ENDVI and SAIO. Formula constructs are more complex, such as that of EVI2. This can involve more information than a simple environmental index model, such as the normalized difference vegetation index (NDVI). Adrianv and Gaiusr (2009) indicated that EVI2 has several advantages over NDVI, including the ability to resolve leaf area index differences for vegetation with different background soil reflectance. The spectral reflectance of a surface results from a mixture of green vegetation and soil “background” reflectance. Vegetation cover in arid areas is relatively sparse. The larger the bare soil area, the greater the influence of soil background on vegetation information

extraction. Although EVI2 uses the same information as that of NDVI, the additional weight on the red reflectance in the denominator of 2.5*(B5-b4)/(B5 + 2.4*B4 + 1) allows EVI2 to be less sensitive to soil darkening (Adrianv & Gaiusr, 2009). In areas covered by vegetation, several studies have shown that the vitality of vegetation (which can be indirectly reflected by the vegetation index) can mitigate the extent by which soil is affected by salinization. In areas with high proportions of bare soil, Peng et al. (2018) proved the reflectance of soil in an oasis of Xinjiang increased with an increase of electrical conductivity in the costal to SWIR1 band. This was the basis for constructing a soil salinity index on bare land. Among all soil salinity indices, the performances of SAIO and SIT were outstanding in the study area.

Conclusion

This study compared five machine learning techniques for mapping soil salinity with environmental covariates representing topography, climate, soil and vegetation (derived from Landsat OLI and (Advanced Spaceborne Thermal Emission and Reflection Radiometer Global Digital Elevation Model V2 data) in three oases. The key findings are summarized as follows:

(1) After a test of 21 datasets from three oases (based on the analysis of the frequency of variables), the following indices are considered to be important indicators for the identification of saline soils in arid areas and for quantitative assessment of soil salinity: CSRI, EEVI, EVI2, GDVI, SAIO and SIT.

(2) We evaluated the performance of five algorithms from the following two aspects: 1) the R^2, RMSE and RRMSE values between the predicted and measured values and 2) the consistency between the pattern of the soil salinity prediction map and the actual survey results. The results show that SGT is the most suitable algorithm for predicting soil salinity in arid areas, followed by RF with less of a performance gap compared to SGT. Negativity and outliers, binarization, an unreasonable range of values, and instability during multi-period predictions (large fluctuations in the range of soil salinity values) appear to varying degrees in the MARS, CART, and LASSO prediction maps. However, SGT and RF effectively avoid the aforementioned phenomenon, particularly the former.

However, there are several limitations to this manuscript that should be point out when building the salinity assessment model(s). Across the multitude of fields that comprise large regions, variations in management, pests, disease, climate and other soil properties can have a far greater influence on soil salinity, thus limiting the utility of remote sensing for salinity assessment. In addition, crop rotation/fallow practice which may also lead to significant change in spectral reflectance and vegetation indices whereas salinity may not subsequently change. For this reason, we suggest a fusion processing approach, that is, the multiyear maxima, integral or mean-based modelling approach rather than a single date image for salinity mapping to minimize the aforementioned challenges or problematic issues based on the achievements of other authors. Meanwhile, soil type, landform and vegetation type were may also need to be considered as covariates for more accurate soil salinity prediction.

In the future, based on the conclusions of this study, we will use SGT to predict soil salinity over multiple periods, and then summarize the spatial variation in soil salinity during the past decades. Finally, this research is expected to provide practical assistance for the rational use of land resources and ecological environment management.

Acknowledgments

We thank Professor Lu Gong and Fang Zhang from Xinjiang University for their valuable comments and data support. This study was supported by the National Natural Science Foundation of China (U1603241, 41661046, 41771470, 41261090, U1303381), the Scientific Research Foundation for Doctors of Xinjiang University (BS150246) and Tianchi Doctor Program of Xinjiang Uygur Autonomous Region (2016).

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the Tianchi Doctor Program of Xinjiang Uygur Autonomous Region [2016];Scientific Research Foundation for Doctors of Xinjiang University [BS150246];National Natural Science Foundation of China [41661046,41771470,41261090,U1303381,U1603241];

References

Adrianv, R., & Gaiusr, S. (2009). Advantages of a two band EVI calculated from solar and photosynthetically active radiation fluxes. *Agricultural & Forest Meteorology*, *149* (9), 1560–1563. doi:10.1016/j.agrformet.2009.03.016

Allbed, A., Kumar, L., & Aldakheel, Y.Y. (2014). Assessing soil salinity using soil salinity and vegetation indices derived from IKONOS high-spatial resolution imageries: Applications in a date palm dominated region. *Geoderma*, *230–231*, 1–8. doi:10.1016/j.geoderma.2014.03.025

Álvaro, G.G., Schnabel, S., & Contador, J.F.L. (2009). Using and comparing two nonparametric methods (CART and MARS) to model the potential distribution of gullies. *Ecological Modelling*, *220*(24), 3630–3637. doi:10.1016/j.ecolmodel.2009.06.020

Angileri, S.E., Conoscenti, C., Hochschild, V., Märker, M., Rotigliano, E., & Agnesi, V. (2016). Water erosion susceptibility mapping by applying stochastic gradient treeboost to the imera meridionale River Basin (Sicily, Italy). *Geomorphology*, *262*, 61–76. doi:10.1016/j.geomorph.2016.03.018

Boettinger, J.L., Ramsey, R.D., Bodily, J.M., Cole, N.J., Kienast-Brown, S., Nield, S.J., … Stum, A.K. (2008). Landsat spectral data for digital soil mapping. In A.E. Hartemink, A.B. McBratney, & M.L. Mendonca-Santos (Eds.), *Digital soil mapping with limited data* (pp. 193–203). Australia: Springer Science.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*, 123–140. doi:10.1007/BF00058655

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. doi:10.1023/A:1010933404324

Brenning, A. (2005). Spatial prediction models for landslide hazards: Review, comparison and evaluation. *Natural Hazards And Earth System Sciences*, *5*(6), 853–862. doi:10.5194/nhess-5-853-2005

Butcher, K., Wick, A.F., DeSutter, T., Chatterjee, A., & Harmon, J. (2016). Soil salinity: A threat to global food security. *Agronomy Journal*, *108*(6), 2189–2200. doi:10.2134/agronj2016.06.0368

Ceccato, P., Gobron, N., Flasse, S., Pinty, B., & Tarantola, S. (2002). Designing a spectral index to estimate vegetation water content from remote sensing data: Part 1. *Remote Sensing of Environment*, *82*(2), 188–197. doi:10.1016/S0034-4257(02)00037-8

Chen, H., Zhao, G., Chen, J., Wang, R., & Gao, M. (2015). Remote sensing inversion of saline soil salinity based on modified vegetation index in estuary area of Yellow River. *Transactions of the Chinese Society of Agricultural Engineering*, *31*(5), 107–114. (in chinese).

Cheng, M.Y., & Cao, M.T. (2014). Accurately predicting building energy performance using evolutionary multivariate adaptive regression splines. *Applied Soft Computing*, *22*, 178–188. doi:10.1016/j.asoc.2014.05.015

Ding, J., & Yu, D. (2014). Monitoring and evaluating spatial variability of soil salinity in dry and wet seasons in the Werigan–Kuqa Oasis, China, using remote sensing and electromagnetic induction instruments. *Geoderma*, *235–236*, 316–322. doi:10.1016/j.geoderma.2014.07.028

Douaoui, A.E.K., Nicolas, H., & Walter, C. (2006). Detecting salinity hazards within a semiarid context by means of combining soil and remote-sensing data. *Geoderma*, *134*(1), 217–230. doi:10.1016/j.geoderma.2005.10.009

Fernández-Buces, N., Siebe, C., Cram, S., & Palacio, J.L. (2006). Mapping soil salinity using a combined spectral response index for bare soil and vegetation: A case study in the former lake Texcoco, Mexico. *Journal Of Arid Environments*, *65*(4), 644–667. doi:10.1016/j.jaridenv.2005.08.005

Food and Agricultural Organization. (2015). Status of the world's soil resources. FAO, Rome

Fox, C.H., O'Hara, P.D., Bertazzon, S., Morgan, K., Underwood, F.E., & Paquet, P.C. (2016). A preliminary spatial assessment of risk: Marine birds and chronic oil pollution on Canada's Pacific coast. *Science of the Total Environment*, *573*, 799–809. doi:10.1016/j.scitotenv.2016.08.145

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1–22.

Friedman, J.H. (1991). Multivariate adaptive regression splines. *Annals of Statistics*, *19*(1), 1–67. doi:10.1214/aos/1176347963

Friedman, J.H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, *38*(4), 367–378. doi:10.1016/S0167-9473(01)00065-2

Gao, T.T., Ding, J.L., Ha, X.P., & Zhang, F. (2010). The spatial variability of salt content based on river basin scale: Acase study of the delta oasis in Weigan-Kuqa Watershed. *Acta Ecologica Sinica*, *30*(10), 2695–2705. (in chinese).

Gong, L., Liu, Z.Y., & Tashpolat, T. (2015a). Soil salinity characteristic and its determinant factors at different soil types in oasis of extreme arid region. *Arid Zone Research*, *32*(4), 657–662. (in chinese).

Gong, L., Ran, Q., He, G., & Tashpolat, T. (2015b). A soil quality assessment under different land use types in Keriya river basin, Southern Xinjiang, China. *Soil & Tillage Research*, *146*, 223–229. doi:10.1016/j.still.2014.11.001

Gorji, T., Sertel, E., & Tanik, A. (2017). Monitoring soil salinity via remote sensing technology under data scarce conditions: A case study from Turkey. *Ecological Indicators*, *74*, 384–391. doi:10.1016/j.ecolind.2016.11.043

Gregory, P., Jamieson, R., Bezanson, G., & Hansena, L. T., & Chris, Y., (2013). Evaluation of statistical models for predicting Escherichia coli, particle attachment in fluvial systems. *Water Research*, *47*(17), 6701–6711. doi:10.1016/j.watres.2013.09.003

Gretchen, G.M., & Tracey, S.F. (2002). Comparing five modelling techniques for predicting forest characteristics. *Ecological Modelling*, *157*(2), 209–225. doi:10.1016/S0304-3800(02)00197-7

Grimm, R., Behrens, T., Märker, M., & Elsenbeer, H. (2008). Soil organic carbon concentrations and stocks on barro colorado island-digital soil mapping using random forests analysis. *Geoderma*, *146*(1), 102–113. doi:10.1016/j.geoderma.2008.05.008

Hardisky, M.S., Klemas, V., & Smart, M.R. (1983). The influence of soil salinity, growth form, and leaf moisture on the spectral radiance of Spartina Alterniflora canopies. *Photogrammetric Engineering And Remote Sensing*, *48*(1), 77–84.

Heung, B., Bulmer, C.E., & Schmidt, M.G. (2014). Predictive soil parent material mapping at a regional-scale: A random forest approach. *Geoderma*, *214*, 141–154. doi:10.1016/j.geoderma.2013.09.016

Heung, B., Ho, H.C., Zhang, J., Knudby, A., Bulmer, C.E., & Schmidt, M.G. (2016). An overview and comparison of machine-learning techniques for classification purposes in digital soil mapping. *Geoderma*, *265*, 62–77. doi:10.1016/j.geoderma.2015.11.014

Hu, J., Tashpolat, T., Yu, S., & Zhang, F. (2017). Spatial heterogeneity of key parameters of salinized soil in Yutian oasis. *Chinese Journal of Soil Science, 2015*, *49*(1), 162–170. (in chinese).

Huete, A., Didan, K., Miura, T., Rodriguez, E.P., Gao, X., & Ferreira, L.G. (2002). Overview of the radiometric and biophysical performance of the MODIS vegetation indices. *Remote Sensing of Environment*, *83*(1), 195–213. doi:10.1016/S0034-4257(02)00096-2

Huete, A.R. (1988). A soil-adjusted vegetation index (SAVI). *Remote Sensing of Environment*, *25*(3), 295–309. doi:10.1016/0034-4257(88)90106-X

Jiang, H.N., Ding, J.L., Tashpolat, T., Zhao, R., & Zhang, F. (2008). Extracting salinized soil infor- mation in arid areas using ETM + data. *Acta Pedologica Sinica*, *45*(2), 222–228. (in chinese).

Jordan, C.F. (1969). Derivation of leaf-area index from quality of light on the forest floor. *Ecology*, *50*(4), 663–666. doi:10.2307/1936256

Khan, N.M., Rastoskuev, V.V., Sato, Y., & Shiozawa, S. (2005). Assessment of hydrosaline land degradation by using a simple approach of remote sensing indicators. *Agricultural Water Management,77*, (1), 96–109. doi:10.1016/j.agwat.2004.09.038

Kuhn, M., Leeuw, J.D., & Zeileis, A. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software*, *28*(5), 1–26. doi:10.18637/jss.v028.i07

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News*, *2*, 18–22.

Lobell, D.B., Lesch, S.M., Corwin, D.L., Ulmer, M.G., Anderson, K.A., Potts, D.J., … Baltes, M.J. (2010). Regional-scale assessment of soil salinity in the red river valley using multi-year MODIS EVI and NDVI. *Journal Of Environmental Quality*, *39*(1), 35.

Martinez-Beltran, J., & Manzur, C.L., (2005). Overview of salinity problems in the world and FAO strategies to address the problem, in: Proceedingsof the international salinity forum, Riverside, USA,311–313, 2005

Metternicht, G.I., & Zinck, J.A. (2003). Remote sensing of soil salinity: Potentials and constraints. *Remote Sensing of Environment*, *85*(1), 1–20. doi:10.1016/S0034-4257(02)00188-8

Mulder, V.L., De. Bruin, S., Schaepman, M.E., & Mayrc, T. R. (2016). The use of remote sensing in soil and terrain mapping — A review. *Geoderma*, *162*(1), 1–19. doi:10.1016/j.geoderma.2010.12.018

Muller, S.J., & Van Niekerk, A. (2016). Identification of WorldView-2 spectral and spatial factors in detecting salt accumulation in cultivated fields. *Geoderma*, *273*, 1–11. doi:10.1016/j.geoderma.2016.02.028

Mullet, T.C., Gage, S.H., Morton, J.M., & Huettmann, F. (2016). Temporal and spatial variation of a winter soundscape in south-central Alaska. *Landscape Ecology*, *31*(5), 1117–1137. doi:10.1007/s10980-015-0323-0

Naghibi, S.A., & Pourghasemi, H.R. (2015). A comparative assessment between three machine learning models and their performance comparison by bivariate and multivariate statistical methods in groundwater potential mapping. *Water Resources Management*, *29*(14), 5217–5236. doi:10.1007/s11269-015-1114-8

Nawar, S., Buddenbaum, H., Hill, J., & Kozak, J. (2014). Modeling and mapping of soil salinity with reflectance spectroscopy and landsat data using two quantitative methods (PLSR and MARS). *Remote Sensing*, *6*(11), 10813. doi:10.3390/rs61110813

Nield, S.J., Boettnger, J.L., & Ramsey, R.D. (2007). Digital mapping gypsic and nitric soil areas using Landsat ETM data. *Soil Science Society of America Journal*, *71*, 245–252. doi:10.2136/sssaj2006-0049

Nurmemet, I., Ghulam, A., Tiyip, T., Elkadiri, R., Ding, J. L., Maimaitiyiming, M., … Sun, Q. (2015). Monitoring soil salinization in keriya river basin, northwestern china using passive reflective and active microwave remote sensing data. *Remote Sensing*, *7*(7), 8803. doi:10.3390/rs70708803

Peng, J., Biswas, A., Jiang, Q.S., Zhao, R.Y., Hu, J., Hu, B.F., & Zhou, S. (2018). Estimating soil salinity from remote sensing and terrain data in southern Xinjiang Province. *China. In Press, Corrected* Proof. doi:10.1016/j.geoderma.2018.08.006.

Ridgeway, G., (2015). gbm: Generalized boosted regression models R package version 2.1.1. https://CRAN.R-project.org/package=gbm.

Rouse, J.W., Haas, R.H., Schell, J.A., & Deering, D.W., (1973). Monitoring vegetation systems in the great plains with ERTS. *Proceedings of the Third ERTS-1 Symposium,* Washington, D.C. NASA SP-3511, pp. 309–317

Schillaci, C., Lombardo, L., Saia, S., Fantappiè, M., Märker, M., & Acutis, M. (2017). Modelling the topsoil carbon stock of agricultural lands with the stochastic gradient treeboost in a semi-arid mediterranean region. *Geoderma*, *286*, 35–45. doi:10.1016/j.geoderma.2016.10.019

Scudiero, E., Skaggs, T.H., & Corwin, D.L. (2014). Regional scale soil salinity evaluation using Landsat 7, western San Joaquin Valley, California, USA. *Geoderma Regional*, *2–3*, 82–90. doi:10.1016/j.geodrs.2014.10.004

Scudiero, E., Skaggs, T.H., & Corwin, D.L. (2015). Regional-scale soil salinity assessment using Landsat ETM + canopy reflectance. *Remote Sensing of Environment*, *169*, 335–343. doi:10.1016/j.rse.2015.08.026

Shi, Q.D., Wang, Z., He, L.M., Shi, Q.S., Anayeti, A., Liu, M., & Chang, S.L. (2014). Landscape classification system based on climate, landform, ecosystem: A case study of Xinjiang area. *Acta Ecologica Sinica*, *34*(12), 3359–3367. (in chinses).

Soil Survey Staff of Xinjiang (1996). Soil of Xinjiang. Beijing: Science Press.

Strobl, C., & Augustin., T. (2009). Adaptive selection of extra cutpoints-towards reconciling robustness and interpretability in classification trees. *Journal of Statistical Theory & Practice*, *3*(1), 119–135. doi:10.1080/15598608.2009.10411915

Svetnik, V., Liaw, A., Tong, C., Culberson, J.C., Sheridan, R.P., & Feuston, B.P. (2003). Random forest: A classification and regression tool for compound classification and QSAR modeling. *Journal Of Chemical Information And Computer Sciences*, *43*(6), 1947–1958. doi:10.1021/ci034160g

Taghizadeh-Mehrjardi, R., Minasny, B., Sarmadian, F., & Malone, B.P. (2014). Digital mapping of soil salinity in Ardakan region, central Iran. *Geoderma*, *213*, 15–28. doi:10.1016/j.geoderma.2013.07.020

Taghizadeh-Mehrjardi, R., Nabiollahi, K., & Kerry, R. (2016). Digital mapping of soil organic carbon at multiple depths using different data mining techniques in Baneh region, Iran. *Geoderma*, *266*, 98–110. doi:10.1016/j.geoderma.2015.12.003

Tian, C., Mai, W., & Zhao, Z. (2016). Study on key technologies of ecological management of saline alkali land in arid area of Xinjiang. *Acta Ecologica Sinica*, *36*, 22. (in chinese).

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. journal of the royal statistical society. *Series B (Methodological)*, *58*(1), 267–288. doi:10.1111/j.2517-6161.1996.tb02080.x

Vermeulen, D., & Van Niekerk, A. (2017). Machine learning performance for predicting soil salinity using different combinations of geomorphometric covariates. *Geoderma*, *299*, 1–12. doi:10.1016/j.geoderma.2017.03.013

Wang, F., Chen, X., Luo, G., & Han, Q. (2015). Mapping of regional soil salinities in Xinjiang and strategies for amelioration and management. *Chinese Geographical Science*, *25*(3), 321–336. doi:10.1007/s11769-014-0718-x

Wang, W., Vinocur, B., & Altman, A. (2003). Plant responses to drought, salinity and extreme temperatures: Towards genetic engineering for stress tolerance. *Planta*, *218*(1), 1–14. doi:10.1007/s00425-003-1105-5

Witten, I.H., Frank, E., & Hall, M.A. (2011). Data mining: Practical machine learning tools and techniques, second edition (Morgan Kaufmann series in data management systems). *Acm Sigmod Record*, *31*(1), 76–77. doi:10.1145/507338.507355

Wu, W. (2014b). The Generalized Difference Vegetation Index (GDVI) for Dryland Characterization. *Remote Sensing*, *6*(2), 1211. doi:10.3390/rs6021211

Wu, W., Mhaimeed, A.S., Al-Shafie, W.M., Ziadat, F., Dhehibi, B., Nangia, V., & De Pauw, E. (2014a). Mapping soil salinity changes using remote sensing in Central Iraq. *Geoderma Regional*, *2–3*, 21–31. doi:10.1016/j.geodrs.2014.09.002

Yan, Z., & Yao, Y. (2015). Variable selection method for fault isolation using least absolute shrinkage and selection operator (LASSO). *Chemometrics And Intelligent Laboratory Systems*, *146*, 136–146. doi:10.1016/j.chemolab.2015.05.019

Yang, R.M., Zhang, G.L., Liu, F., Lu, Y.Y., Yang, F., Yang, F., … Li, D.C. (2016). Comparison of boosted regression tree and random forest models for mapping topsoil organic carbon concentration in an alpine ecosystem. *Ecological Indicators*, *60*, 870–878. doi:10.1016/j.ecolind.2015.08.036

Youssef, A.M., Pourghasemi, H.R., Pourtaghi, Z.S., & Al-Katheeri, M.M. (2016). Landslide susceptibility mapping using random forest, boosted regression tree, classification and regression tree, and general linear models and comparison of their performance at Wadi Tayyah Basin, Asir Region, Saudi Arabia. *Landslides*, *13*(5), 839–856. doi:10.1007/s10346-015-0614-1

Yu, R., Liu, T., Xu, Y., Zhu, C., Zhang, Q., Qu, Z., … Li, C. (2010). Analysis of salinization dynamics by remote sensing in Hetao irrigation district of North China. *Agricultural Water Management*, *97*(12), 1952–1960. doi:10.1016/j.agwat.2010.03.009

Zandler, H., Brenning, A., & Samimi, C. (2015). Quantifying dwarf shrub biomass in an arid environment: Comparing empirical methods in a high dimensional setting. *Remote Sensing of Environment*, *158*(1), 140–155. doi:10.1016/j.rse.2014.11.007

Zhang, F., Tashpolat, T., & Ding, J.L. (2007). Analysis on characteristics of soil salinization in the delta oasis of Weigan and Kuqa Rivers. *Agricultural Research in the Arid Areas*, *2007*, 1–12. (in chinese).

Zhang, F., Tashpolat, T., Ding, J.L., Tian, Y., & Mamat, S. (2009). Relationships between soil salinization and spectra in the delta oasis of Weigan and Kuqa Rivers. *Research of Environmental Sciences*, *22*(2), 227–235. (in chinese).

Zhang, F., Xiong, H.G., Tian, Y., & Luan, F.M., (2011a). Impacts of regional topographic factors on spatial distribution of soil salinization in qitai oasis. *Research Of Environmental Sciences*, *24*(7): 731–739.(in chinese)

Zhang, T.T., Zeng, S.L., Gao, Y., Ouyang, Z.T., Li, B., Fang, C.M., & Zhao, B. (2011b). Using hyperspectral vegetation indices as a proxy to monitor soil salinity. *Ecological Indicators*, *11*(6), 1552–1562. doi:10.1016/j.ecolind.2011.03.025

Zhao, K., Popescu, S., & Meng, X. (2011). Characterizing forest canopy structure with lidar composite metrics and machine learning. *Remote Sensing of Environment*, *115*(8), 1978–1996. doi:10.1016/j.rse.2011.04.001