Comparisons of random forest and stochastic gradient treeboost algorithms for mapping soil electrical conductivity with multiple subsets using Landsat OLI and DEM/GIS-based data at a type oasis in Xinjiang, China

Yang Wei\textsuperscript{ab}, Jianli Ding\textsuperscript{ab}, Shengtian Yang\textsuperscript{ab}, Xiaodong Yang\textsuperscript{cd} and Fei Wang\textsuperscript{ab}

\textsuperscript{a}Xinjiang Common University Key Lab of Smart City and Environmental Stimulation, College of Resource and Environmental Sciences, Xinjiang University, Urumqi, China; \textsuperscript{b}Lab for Oasis Ecosystem, Ministry of Education, Urumqi, China; \textsuperscript{c}Ningbo University, Ningbo, China; \textsuperscript{d}Global Centre for Environmental Remediation (GCER), ATC Building, The University of Newcastle, Australia

ABSTRACT

Accurate assessment of the spatial distribution and severity of soil salinization has long plagued local governments and researchers in the arid parts of Xinjiang Uygur Autonomous Region (XJUAR). The emergence of machine learning has brought hope to this research field, such as Random Forest (RF) and Stochastic Gradient Treeboost (SGT). However, these are few applications to the quantitative assessment of soil salinization. Therefore, in order to evaluate the accuracy level of the two algorithms for predicting soil salinity, twenty-seven environmental subsets were designed. Each dataset is calculated using both RF and SGT to produce an optimal set of variables. The simulation results from 70.37% (19/27) of the subsets showed that the predicted value of soil salinity from SGT is closer to the observed value than that from RF. The statistics of all datasets showed that the average value of $R^2$ for RF and SGT were 0.38 and 0.40, the average Root Mean Squared Error (RMSE) value were 28.59 and 27.46, and the Ratio of Prediction to Deviation (RPD) averages were 1.20 and 1.24, respectively. The important dominant factor were topographic variables with coarse resolution, temperature and vegetation indices, land use and landform.

Introduction

The Xinjiang Uygur Autonomous Region (XJUAR) is the largest arid region in the country, is located in the most northwestern part of China and is one of the main areas affected by soil salinization in Asia (Wang et al., 2015). In 2014, according to a report by the Xinjiang Institute of Ecology and Geography of the Chinese Academy of Sciences, salinized farmland accounted for 37.72% of the irrigated land in Xinjiang – 6% higher than the level in 2006 (Tian et al., 2016). Moreover, the saline farmland in southern XJUAR (with the Tianshan mountains as the boundary: XJUAR is geographically divided into the Northern and Southern Territories) accounted for almost half (49.6%) of the irrigated land, seriously impacting the lives of farmers and herdsmen.

Many different remote-sensing-derived indices are used to identify the spatial distribution and variability of soil salinization in different research areas (Allbed et al., 2014; Ding & Yu, 2014; Wang et al., 2020). Local scholars from different areas have attempted to establish (or have cited existing) soil and vegetation-related indices to identify soil salinization. The results show that the above indices can play a role in the detection of soil salinization, and some indices are highly sensitive. However, due to environmental and climatic differences, it is unclear whether the indices that are suitable in other regions maintain their high sensitivity to soil salinization in the target area. No research has comprehensively compared the effectiveness of the above remote sensing derivative indices in XJUAR to detect soil salinity.

However, there are differences between this article and the above research, and the two have different research focuses. Instead, most studies investigated the responses of environmental variables to soil salinity at the oasis scale (Ding & Yu, 2014; Wang & Li, 2013) without considering whether these environment variables retain their high efficiency to quantify soil salinity under small-scale or relatively homogeneous environments (i.e., within a basin/oasis). Surface roughness, albedo and other properties that affect water and energy exchanges between land surfaces and the atmosphere (Garcia et al., 2008) will be altered by the conversion of natural ecosystems to irrigated agriculture (the most common model) (Xin et al., 2016), which leads to changes of the surface energy and net radiation that vary among different land-use types. Each environmental component or land use type (farmland, shrubbery, grassland, salt-affected land, etc.) has a different response mechanism between water/salt and the surface (Gong et al.,...
Statistical methods are an important way to acquire knowledge of the soil salinization-landscape. Multivariate linear regression is the most commonly used approach for soil salinity predictions. Because the relationship between soil properties and environmental variables can be very complex, the linear hypothesis as the basis for modeling will increase the deviation between the predicted value and the real value (Lark, 1999; Zhu et al., 2001). Machine learning is a modern statistical method based on computers that are used to discover the complex relationship between predictor variables and response variables (Hastie et al., 2009). A variety of machine-learning algorithms such as CART, Random Forest (RF), Stochastic Gradient Treeboost (SGT), K-Nearest Neighbour (KNN), Artificial Neural Networks (ANN) and the Support Vector Machine (SVM) have been tested to determine the complex relationships between environmental variables and soil properties (Chagas et al., 2016; Heung et al., 2016; Ließ et al., 2012). A comparative analysis of these studies found that RF and SGT performed better than did the other algorithms (Forkuor et al., 2017; Ingлада et al., 2015). Breiman (2001) and Friedman (2002) pointed out in their study that the above two data mining methods have some advantages in many aspects compared with most statistical modeling methods. RF has the advantage of dealing with complex nonlinear relationships, no assumptions are required refer to participating data, processing continuous and categorical variables, preventing overfitting to a certain extent, robustness to noise in the data, impartial error evaluation criteria, and the ability to determine the importance of variables, only a small number of parameters can be adjusted to implement. SGT is an improved form of Classification and Regression Trees (CART) that generally achieves accurate and robust predictions with few overfitting effects (Friedman, 2002). This powerful algorithm does not require prior assumptions about the involved variables, thus providing more flexibility than traditional generalized linear or additive models (Friedman, 2002). These characteristics may be important when examining a system defined by complex relationships. When the spatial heterogeneity of the research area is large and the data have abnormal noise, the SGT still has high prediction accuracy and excellent interpretation ability of relationship of environment variable-soil attribute (Schillaci et al., 2017). Otherwise, feature importance analysis also motivates the use of machine learning against deep learning methodologies in terms of model interpretability. The above discussion shows that it is difficult to distinguish the advantages and disadvantages of RF and SGT. Moreover, these two potentially attractive algorithms are rarely used for in soil salinity prediction in drylands.

In this paper, we assessed the potential of Landsat OLI, land use, landform and Digital elevation models (DEM)-derivatives to predict soil salinity at a depth of 0–10 cm using modern statistical techniques. The study had two goals. The first goal was to compare the relative efficiency of RF and SGT in predicting soil salinity under different subsets at the typical oasis level, which is characterized by high soil salinity variability in XJUAR. The second goal was to quantify the effects of geographic information system-based subsets on predicting accuracy level of soil salinity and comprehensive evaluate the most relevant contributors in different subsets.

Materials and methods

Study area

Oasis is the foundation of local residents and socio-economic development. In the XJUAR, the oasis occupies only 5% of the total land area, but it is the settlement area of more than 95% of the local population. The mountain-Oasis-Desert System (MODS) is a typical landscape model of the Tarim Basin in southern XJUAR (Zhang, 2001): cryosphere (glacial, frozen, alpine Meadow Forest) occupies the upper reaches of the landscape pattern (mountain), while area of the crop irrigation system in its middle reaches.oasis), the riparian ecosystem and the Oasis-Desert-belt are distributed downstream (desert).The area between the mountains (Tianshan and Kunlun) and the basin (Tarim Basin) includes several oases, which possess both independent and similar characteristics. Their difference is that each oasis is alluvial from one or two rivers originating from the Tianshan and Kunlun Mountains (Figure 1(a)). Their similarity is that the landscape structures among the multiple oases are relatively homogeneous.

The Kuqa watershed is a classic oasis located in the north part of the Tarim Basin. This study area is centred at 82.50°N and 41.38°E and consists of a low-lying alluvial fan plain, a low-altitude alluvial floodplain, a middle-altitude alluvial fan plain, a low-altitude fixed grass shrub area and a low-altitude semi-fixed grass shrub area (Figure 1(b)). The altitude in the study area ranges from 892 to 1,100 m, decreasing from northwest to southeast. The soil types are primarily Cumulic Anthrosols, Salic Fluvisols, Gypsic/Calcic Solonchaks, Cambic Arenosols, Calcic Vertisols and Calcaric Phaeozems. The vegetation cover is lower on salt-affected land and is dominated by desert species,
Figure 1. Location of the study area (a) and distribution of the field sampling plots (b).
such as Phragmites australis, Tamarix ramosissima, Alhagi sparsisfolia, Karelina caspica, and Kalidium gracile (Ding & Yu, 2014). The region is an extremely arid desert climate with an average annual rainfall of 51.6 mm, a potential evapotranspiration of 235 6 mm per year, and an annual cumulative temperature of 4,500°C. The dramatic increase in irrigation water (because of the expansion of irrigated land) has led to a significant increase in groundwater depth and groundwater mineralization in the area, which is a major contributor to drive soil salinization (Saritajane et al., 2009). Within the area of cultivated land, salinization has reached more than 50%, of where serious salinization affects 30% of the cultivated land (Wang et al., 2015).

**Material and methods**

The workflow consisted of the steps presented in following Figure 2: (1) pre-processing of Landsat and environmental variables; (2) classifying datasets according to driving and responding factors; (3) selection of the best combination of variables for each dataset; (4) hyper-parameter tuning; (5) validation using method 1 and 2; (6) accuracy assessment.

**Data collection and soil sample analysis**

To perform soil salinity analyses and model predictions, 371 samples were employed in this research (Figure 1(b)). Data collection was performed from September 09 to 30, 2016. At each location, five topsoil samples (0–10 cm) were collected using a 5-pointed star form. The sampling locations were designed based on conditioned Latin hypercube sampling (cLHS) (Minasny & Mcbratney, 2006). The main purpose of CLHS is to maximally cover the diversity of environmental variables in multidimensional feature spaces through a fixed number of samples. Details on the procedure and algorithm of cLHS are presented in Minasny and Mcbratney (2006). The sample locations are selected based on the spatial variability of the major factors affecting the formation of the soil, which represent the heterogeneity of the soil in target area, including remote sensing derivatives (soil adjusted vegetation index and salinity ratio index), elevation (Dem derivatives-Topographic wetness index, Multiresolution Index of Valley Bottom Flatness (MRVBF and MRRRTF), Valley Depth, Vertical Distance to Channel Network), climate (land surface temperature), land use and landforms. As to the accessibility of the sample points, first of all, according to historical ground traffic data to evaluate the accessibility of the pre-designed samples, and secondly, when the actual sampling point is not reachable due to the local road upgrading, according to the design of the sample point of the surrounding environment to make real-time adjustment to improve the sampling efficiency.

Soil salinity (measured as electrical conductivity in saturation paste, ECe) in this study was analyzed in a laboratory. After air drying, the samples were ground and passed through a 0.5-mm screen. The ground and sieved soil sample was mixed with water at a 1 (soil sample):5 (water) ratio at a temperature of 25°C. The leaching liquid was extracted to measure the electrical conductivity (EC1:5) and soil salinity of the soil samples as detailed in the Analysis Methods of Soil and Agricultural Chemistry (Lu, 1999). The averages of the electrical conductivity of these points are used as the base values for the sampling site. In addition, ECe :1:5 and soil salt content (g/kg) are usually highly correlated ($R^2 = 0.95$), and EC1:5 is often used as a surrogate for soil salinity in the Kuqa oasis (F. Zhang et al., 2007).

**Environmental covariates**

The Landsat OLI (Date 19 September 2016, 145/31) were acquired. Using the Fast line-of-slight Atmospheric Analysis of Spectral Hypercubes
(FLAASH) model of the Environment for Visualizing Images (ENVI) 5.1 for the study area to minimize and normalize atmospheric effects (both additive and multiplicative) and the effects of sun illumination geometry on the images. The FLAASH reflectance was rescaled to a normal range of 0 to 1.

In this study, the environmental covariates for soil salinity prediction were selected based on the SCORPAN formula (Metternicht & Zinck, 2003). These covariates included bands, climate factor (referring to land surface temperature), vegetation indices, salinity and soil-related indices, soil moist indices. The use of satellite images for determining estuarine salinity was pioneered by Khorram in a study that mapped the salinity distribution in the San Francisco Bay area (Khorram, 1982). Later, a number of researchers found divergent correlations among Landsat TM bands 2, 3, 4, 5, and 7 with in situ salinity measurements and the electrical conductivity of the saturation extract (ECE) (R² = 0.13–0.86) for the days when satellites passed over the study areas (Scudiero et al., 2015). Bands from Landsat were employed in this study based on the following rationale: the performance of the derived variables in this study were investigated to determine if they were better than these basic bands for detecting soil salinity in these three oases. TC and PCA derived from Landsat ETM+ images have been employed for soil salinity prediction in previous studies (Scudiero et al., 2014). Although TC and PCA were used to detect soil salinization, the predictive power of PCA has not been tested for a wide range of geographic areas. Temperature (TEM) has also been valuable in studies of evapotranspiration, soil moisture conditions, surface energy balance, and urban heat islands. Soil salinity at the surface soil has been estimated in numerous studies using vegetation indices (see Table 1 for details). The studies summarized several widely used soil salinity/mineral indices for soil salinity assessments (Allbed et al., 2014; Metternicht & Zinck, 2015).

Table 1. Environmental covariates derived from a 30-m spatial resolution DEM and 30-m Landsat imagery.

| Auxiliary data | Index | Abbrev | References |
|---------------|-------|--------|------------|
| Landsat OLI   | Landsant-8 OLI six bands and thermal infrared band | TC1/TC2/TC PCI |           |
|               | Tasseled Cap |                       |           |
|               | Principal Component Analysis |                     |           |
| Vegetation indices | Normalized Difference Vegetation Index | NDVI | (Rouse et al., 1973) |
|               | Soil Adjusted Vegetation Index | SAVI | (Huete, 1988) |
|               | Enhanced Vegetation Index | EVI | (Huete et al., 2002) |
|               | Generalized Difference Vegetation Index | GDVI | (Wu, 2014) |
|               | Canopy Response Salinity Index | CRSI | (Scudiero et al., 2015) |
|               | Simple Ratio vegetation index | SR | (Jordan, 1969) |
|               | Two-band enhanced vegetation index | EV2 | (Jiang et al., 2008) |
|               | Extended NDVI | ENDVI | (H. Chen et al., 2015) |
|               | Extended EVI | EEVI | (H. Chen et al., 2015) |
| Soil-related indices | Salinity index | SI | (Allbed et al., 2014) |
|               | Salinity index | S1 | (Khan et al., 2005) |
|               | Salinity index | S11 | (Douaoui et al., 2006) |
|               | Salinity index | S12 | (Douaoui et al., 2006) |
|               | Salinity index | S13 | (Douaoui et al., 2006) |
|               | Salinity index | S1A | (Allbed et al., 2014) |
|               | Salinity index | SIB | (Allbed et al., 2014) |
|               | Salinity ratio index | SAI0 | (Metternicht & Zinck, 2003) |
|               | Clay index | CLEX | (Boettinger et al., 2008) |
|               | Gypsum index | GYEX | (Nield et al., 2007) |
|               | Brightness index | BREX | (Metternicht & Zinck, 2003) |
|               | Carbonate index | CAEX | (Boettinger et al., 2008) |
|               | FSEN | FSEN | (Yu et al., 2010) |
|               | Color indices (Hue, Saturation, Value) | H/S/V | (Harsdisky et al., 1983) |
|               | Normalized Difference Infrared Index | NDII | (Ceccato et al., 2002) |
| Dem derivatives | Global Vegetation Moisture Index | GWMI |           |
|               | Valley Depth | VD | SAGA Development Team, 2011 |
|               | Vertical Distance to Channel Network LS-Factor | VDCN |           |
|               | Topographic wetness index | LSF |           |
|               | Slope Length | TWI |           |
|               | Sky View Factor | SL |           |
|               | Topographic Position Index Multiresolution Index Of Valley Bottom Flatness | SVF |           |
|               | Slope Height | TPI |           |
|               | Normalized Height | MRVBF/MRRTF |           |
|               | Standardized Height | SH |           |
|               | Mid-slope Position | NH |           |
|               | Terrain Surface Texture | STH |           |
|               | Flow Accumulation | MSP |           |
|               | Cross-Section Curvature Longitudinal Curvature | TEX |           |
|               | Relative Slope Position | FA |           |
|               | Catchment Slope | CSC |           |
|               | LC | | |
|               | RSP | | |
|               | CS | | |
2003; Scudiero et al., 2014)(see Table 1 for details). In addition, the model construction of (b5-b7)/(b5+ b7) (termed FSEN in this study) was employed in Yu et al. (2010), Zheng et al. (2014), and Kushida and Yoshino (2010) for soil-salinity mapping, crop residue cover and FAPAR estimation. Considering that this band combination has been applied to soil salinization, we wanted to investigate whether this index has potential for evaluating soil salinity in arid zones. The moisture/water content index, the Normalized Difference Infrared Index (NDII), is sensitive to not only canopy moisture but also salinity by combining band 5 (Hardisky et al., 1983). Moreover, we introduced another soil moisture-related index called the Global Vegetation Moisture Index (GVMI) (Ceccato et al., 2002), which is consistent with field measurements of water content expressed as a quantity of water per unit area.

The DEM-derived topographic variables directly or indirectly reveal the direction of movement of parent material and water, thereby changing the accumulation patterns and locations of soil salts (Pike, 2000). This may be why related research has identified a significant correlation between soil salinity and topography (Akramkhanov et al., 2011; Ellaggar & Noller, 2009). DEMs (ASTER GDEM V2, 30 m) were downloaded from the Geospatial Data Cloud (http://www.gscloud.cn/); then, preprocessing operations such as mosaicking and fill sink were applied to the regional extent of all three cases. The study uses the DEM preprocessing method established by Martz and Garbrecht (1998) to fill depressions and correct the elevation of flat areas according to topographic features to eliminate flat areas. Based on previous studies, 19 DEM derivatives, referring to channel, hydrology, and morphometry, from eight spatial resolutions (30 m, 60 m, 90 m, 120 m, 150 m, 180 m, 210 m, and 240 m) were calculated. The System for Automated Geoscientific Analyses (SAGA) software package was used to generate the DEM derivatives (see Table 1 for details). The details of computing these covariates can be found at http://www.saga-gis.org/en/index.html. The land use data were produced (based on Landsat images) by the Chinese Academy of Sciences, which can be downloaded at http://www.resdc.cn. The land-use system applies a two-level classification based on remote sensing surveys, including 6 first-level and 25 second-level classifications. The accuracy of the comprehensive evaluation of the primary and secondary categories of land use has reached 94.3% and 91.2%, which meets the accuracy of 1:100,000 scale user mapping (Liu et al., 2014). The land use data mapped in 2015 were selected for this study.

Chai et al. (2009) presented a method of geomorphic regionalization for Xinjiang, with the help of SRTM-DEM (90 m) and Landsat TM images (30 m). The mapping process mainly refers to a China 1:1, 000,000 geomorphic digital map. The geographical map of the China Geomorphic Regionalization and Xinjiang Landform and the evolutionary history of the Xinjiang geomorphic space patterns were also considered. Maintaining the integrity of the Geomorphic Genesis Unit is the basis of these data. After a comprehensive consideration, the geomorphologic regionalization hierarchy of Xinjiang resulted in three levels that included six divisions: 23 medium-landform divisions and 200 micro-landform divisions. The boundaries of the geomorphologic regionalization are accurate and reliable. These reliable data (30 m) can be downloaded from http://www.geodata.cn/.

Random forest and stochastic gradient treeboost algorithms
To construct the RF relationships between ECe and environmental covariates, the randomForest package in statistical software R was used (Liaw & Wiener, 2002). The RF classifier uses numerous decision trees, ntree, that are grown from bootstrap samples (63%) from the entire sample population, n (Breiman, 2001). RF modeling requires three user-defined parameters: the number of variables used to grow each tree (mtry), the number of trees in the forest (ntree) and the minimum number of terminal nodes (nodesize). The RF classifier uses a bootstrapped sample to create decision trees. At each binary split, the predictor that produces the best split is chosen from a random subset, mtry, of the entire predictor set, p, where the number of predictors tried at each split, mtry, is defined by the user. As a result, mtry is recognized as the main tuning parameter of RF and should therefore be optimized (Svetnik et al., 2003). The use of bootstrap sampling in RF modeling allows the remaining un-used subset (37%, i.e., the out-of-bag data (OOB)) to be used for the estimation of general errors. RF predictions are the averaged output of all aggregations. The default mtry value of randomForest relates to p/3. However, in the case of RF, this is particularly true for datasets with correlated predictor variables, for which several mtry values should be considered (Strobl et al., 2008). To optimize the primary tuning parameter, mtry values ranging from 1 to 30 were tested, and the OOB error rates from 50 replicates from each mtry value were assessed. Then, mtry was further assessed using RMSE (root mean squared error) obtained from the replication of a 5-fold cross-validation. Nodesize value was used when the RMSE value is the smallest among multiple iterations. The default value for the number of trees (ntree, 200) proved to be insufficient to yield stable results (Grimm et al., 2008). Therefore, we set ntree = 1,000 for the test that applied to the RF model.

The SGT method combines regression trees and a boosting technique to improve the predictive performance of multiple single models. 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 (Elith et al., 2008). This process introduces a stochastic gradient boosting procedure that can improve model performance and reduce the risk of overfitting (Friedman, 2002). The SGT algorithm is an iterative process in which tree-based models were fitted iteratively using recursive binary splits to identify poorly modeled observations in existing trees until a minimum model deviance was reached. The final fitted model is a linear function of the sum of all the trees multiplied by the learning rate (LR) based on all the data (Elith et al., 2008). The ‘gbm’ package for SGT was performed. SGT fits a simple parameterization function (a base learner) to pseudo residuals by least squares (after comparing multiple loss functions, least squares was finally selected) at sequential iterations to construct additive regression models. During each iteration, the residuals were calculated as the gradient of the loss function with respect to the training data being evaluated in each regression. To improve the accuracy of this process, a subsample of the training data (to avoid overfitting, a subsample fraction was set to 0.63) was selected from the entire dataset at random with replication (Angileri et al., 2016). Each replicate was randomly built by extracting 63% of the soil salinity data for calibration purposes. The remaining 37% was kept for model validation. The final optimal value of LR was set to 0.01. The number of trees was initially set to 1000, and the maximum nodes per tree needed to be tested based on the actual operation depended on multiple subsets. This operation can generate optimal nodes of 1000 trees using a 5-fold cross-validation method. The relative importance of variables was measured based on the number of times they were selected for modeling, and they were weighted by the square improvement from each split and averaged across all trees (Friedman, 2002).

RF and SGT each have different modeling strategies and advantages. The RF modeling engine is a collection of many CART trees that are not influenced by each other when constructed. The sum of the predictions made from decision trees determines the overall prediction of the forest. Random Forests’ strengths are spotting outliers and anomalies in data, displaying proximity clusters, predicting future outcomes, identifying important predictors, discovering data patterns, and replacing missing values with imputations. The SGT modeling engine adds the advantage of a degree of accuracy usually not attainable by a single model or by ensembles such as bagging or conventional boosting. As opposed to neural networks, the SGT methodology is not sensitive to data errors and needs no time-consuming data preparation, preprocessing or imputation of missing values. This type of data error can be very challenging for conventional data mining methods and will be catastrophic for conventional boosting. In contrast, the SGT model is generally immune to such errors as it dynamically rejects training data points too much at variance with the existing model. The SGT modeling engine robustness extends to data contaminated with erroneous target labels.

Designing subsets partitioned by driving and response factors
The study designed 12 subsets to compared performance of Random Forest and Stochastic Gradient Treeboost algorithms for mapping soil electrical conductivity with at type oasis. Figure 3 shows a schematic diagram of the interactions of hydrothermal and energy exchange between the oasis and desert, as well as the relative coverage of each subset (modified from (Xin et al., 2016)). We show a conceptual model. Instead of focusing on the abovementioned cycle, this paper uses this framework to help explain the effects of different parts of the cycle on soil salinity variations. The different processes of hydrothermal cycling have profound impacts on the vertical variability of soil salinity and, subsequently, its spatial distribution. The research classifies samples into 12 subsets based on the driving and responding variables: based on land uses: grassland, agricultural land and unused land; based on landforms: lower altitude alluvial fan plain and low altitude fixed grass shrub; based on vegetation cover: a threshold of NDVI = 0.22, in which the extent of vegetation coverage is divided into vegetation cover (NDVI > 0.22) and no vegetation cover (NDVI ≤ 0.22) by field measurement test, which was proven in the study by Wu et al. (2014) and (Song et al., 2017) in arid land; and soil salinity classes in terms of electrical conductivity (ECe): ECe ≤ 4 ds/m, 2 ds ECe < 16 ds/m, ECe > 4 ds/m, ECe > 16 ds/m (Richards, 1954). The settings of a subset also take into account the number of samples. These design subsets cover the oasis interior irrigated land and the oasis-desert transition zone outside the irrigation area according to Figure 2. Table 2 show the composition of land-use types for each subset and have a description of the typical situation.

Using a one-way analysis of variance (ANOVA), a statistical comparison of the mean soil salinity in the different scenarios was performed and is presented in Figure 4. There were significant differences between the scenarios in terms of soil salinity (P < 0.001).

The selection of the best combination of variables based on their contribution rankings
Variable importance assessment is a vital step when high-dimensional datasets are used. Using only the most important variables can result in higher regression/classification accuracies than using all available features (Svetnik et al., 2003). High-dimensional data-sets with a large number of predictors may result in
model overfitting (Gregorutti et al., 2017). Furthermore, removing irrelevant or redundant variables and creating a sparse subset for regression/classification helps to simplify statistical problems and shortens data processing times.

In this study, variable selection was tested in all 12 subsets to identify the minimum set of variables and the importance variables. The following variable reduction procedure was adopted from Svetnik et al. (2003) and Heung et al. (2014):

1. The machine-learning algorithms were initially applied to all 12 subsets; then, variable importance in each subset, based on the mean decrease in accuracy (Breiman, 2001), was used to rank the predictor variables.
Figure 4: Analyses of variance (ANOVA) for soil salinity at different scenarios; 1: NDVI > 0.22, 2: NDVI < 0.22, 3: farmland, 4: grassland, 5: unused land, 6: low altitude alluvial fan plain, 7: low altitude fixed grass shrub, 8: ECe ≤4 ds/m, 9: ds/m 2 < ECe < 16 ds/m, 10: ECe > 4 ds/m, 11: ECe > 16 ds/m, 12: oasis scale (values in each row with the same letter are not significantly different among scenarios).

(2) Refer to variable rankings, the least important predictors do not participate in the next calculation. Then, the remaining variables continue to iterate.
(3) Fold cross-validation (CV) is used for the optimization of parameters and the selection of the minimum set of variables. A total of 20 replicates of a 5-fold CV were performed.
(4) Steps 2 and 3 were repeated until the lowest RMSE value was achieved.
(5) Steps 2 to 4 were applied to all 12 subsets; then, the optimal variable combination for each subset was determined.

Figure 5 and Figure 6 show the optimal value of the mtry set in RF and the nodes set in SGT based on RMSE in different subsets. In the optimization of the main parameter of RF and SGT, we chose to use mtry and nodes with the lowest RMSE in red mark for model construction.

The comparison of RF and SGT with combined and reclassified subsets for the prediction of soil salinity

The following two methods were used to compare RF and SGT performance (Figure 7).

Method 1. First, according to the data classification criteria in Table 2, the samples are divided into 11 data subsets. Each subset is modeled separately based on RF and SGT and evaluated using R2 and RMSE.

Secondly, input all the samples into the RF and SGT algorithms, and also according to the classification standard of Table 2, divide the predictions into 11 data subsets, and calculate the R2 and RMSE values of each data set, as shown in left one in Figure 7.

Finally, compare and evaluate the above 11 data sets one by one, correspondingly.

Method 2. First, merge the above independent subsets, namely subset 2&3, subset 4, 5 & 6, subset 7 & 8, subset 9 & 11, as shown in in right one in Figure 7.

Secondly, according to the merging method in the first step, the prediction results of global modeling are classified into corresponding subsets.

Finally, compare the combined results with the reclassified results one by one, correspondingly.

In addition, the purpose of the above comparison is not only to compare the performance of the two algorithms, but also to illustrate the differences between global modeling and partition modeling, which may reduce uncertainty in large-scale soil mapping.

Model evaluation

A 5-fold cross-validation was used to generate optimized parameters. The advantage of this validation test is that it performs reliably and is unbiased for smaller data set, because the process requires much more computational effort than do simple train-and-test (hold out) procedures (Taghizadeh-Mehrjardi et al., 2016; K. Zhao et al., 2011). The training dataset was randomly partitioned into five subsets, with four-fifths comprising 80% of the observations used for model training and 20% for
model validation. The correlation coefficient ($R^2$), root mean squared error (RMSE) and the ratio of prediction to deviation (RPD) were employed as metrics to evaluate the performance. Predictions become increasingly optimal as the RMSE approach zero. The greater the RPD and $R^2$ value, the higher the accuracy.

**Results**

*Variations in soil salinity in different subsets*

Table 3 shows the basic statistical information regarding soil salinity in different subsets. The statistical analysis of all the samples shows that ECe has a maximum value of 184.5 ds/m, a minimum of 0.14 ds/m, and an average of 31.32 ds/m in the oasis scale according to the salinity classification used in American salinity laboratories (0–2 ds/m = no salt effects; 2–4 ds/m = slight salt effects; 4–8 ds/m = moderate salt effects; 8–16 ds/m = severe salt effects; > 16 ds/m = Extreme salt effects) (Richards, 1954). Approximately 67% of the samples (subset 11) were affected by soil salinization. The coefficient of variation (CoV) in different subsets indicated a strong variability of the soil salinity in shallow ground with a depth of 0–10 cm in the study area (CoV < 0.1 = low variability, CoV > 1.0 = high variability, 0.1 < CoV < 1.0 = moderate variability). Under different vegetation coverage, the salinity variability of subset 2 is higher than subset 3. Subset 2 may occur simultaneously in areas covered by natural vegetation and artificial vegetation. Among the three land-use types, the variability of soil salinity in cropland is much higher than that of the other two types. Although the soil salinity variability of subset 5 and subset 6 are similar, the former has medium variability, while the latter has strong variability. Moreover, the average soil salinity levels of subset 5 and subset 6 are significantly higher than that of subset 4. The soil salinity variability of subset 7 was greater than that of subset 8, and the former generally has higher soil salinity levels than the latter. In the saline subsets, the variability of all four classifications is in the middle level. Among them, the salinity variabilities in subset 9 and subset 11 are close to the strong variability level.

*Precision level of the predicted soil salinity based on RF and SGT*

The study calculated the error level of the predicted soil salinity based on RF and SGT for 10 independent subsets (Table 4) and 10 regroup subsets from all the
sample models with the oasis scale (Table 5) according to Figure 7 (Method 1). The statistical parameters $R^2$, RMSE and RPD in Table 4 show that the descriptive statistics indicate the following: (1) similar levels of prediction accuracy of the SGT and RF models for subset 2, 3 and 5; (2) 70% of the subsets (subset 2, 3, 5, 6, 7, 8, 11, and 12) had a higher prediction accuracy of SGT than that of RF. The $R^2$ value increased sequentially by 6.06%, 0.00%, 5.13%, 23.40%, −4.00%, 51.52%, 18.75% and 16.67%. The RMSE value decreased by 2.21%, 0.005%, 1.47%, 11.73%, 16.14%, 12.49%, 3.79%, and 3.20%, respectively. The RPD values increased by 2.44%, 0.00%, 1.56%, 13.67%, 18.80%, 13.82%, 4.10%, and 3.33%, respectively. Although the RGT of the SGT in subset 7 is lower than the RF, both the RMSE and RPD value of SGT are better than RF, so we believe the prediction accuracy of SGT is higher than that of RF. In contrast, RF has a higher SGT in subset 9 and 10, and the $R^2$ value increased, respectively, by 26.09% and 34.21%. The
RMSE value decreased by 4.82% and 11.18%, respectively. The RPD values increased by 5.26% and 12.50%; (3) The verification results of all the subsets reached the p < 0.01 level. Table 5 shows the following: (1) similar levels of prediction accuracy of the SGT and RF models for all the subsets except subset 2. In subset 2, compared with RF, the $R^2$ value in SGT increased by 46.67%, the RMSE value decreased by 11.61%, and the RPD value increased by 13.56%; (2) The 50% subset of prediction accuracy shows that SGT is higher than RF. On the whole, the prediction accuracy of SGT is greater than that of RF at the oasis scale. Unfortunately, the salinity value cannot be predicted for cropland by RF and SGT due to high variability.

Table 5. Validation statistics of RF and SGT to predict surface soil salinity for different subsets originating at the oasis scale according to Method 1 in Section 2.2.6 in Figure 7.

| Subsets   | $R^2$ | RMSE | RPD | P  | $R^2$ | RMSE | RPD | P  |
|-----------|-------|------|-----|----|-------|------|-----|----|
| Subset 1  | 0.46  | 29.65| 1.37| ** | 0.48  | 29.17| 1.39| ** |
| Subset 2  | 0.30  | 19.20| 1.18| ** | 0.44  | 16.97| 1.34| ** |
| Subset 3  | 0.38  | 35.30| 1.27| ** | 0.37  | 35.71| 1.26| ** |
| Subset 4  | 0.25  | 36.87| 1.24| ** | 0.24  | 36.67| 1.24| ** |
| Subset 5  | 0.51  | 27.94| 1.36| ** | 0.54  | 26.06| 1.46| ** |
| Subset 6  | 0.46  | 28.42| 1.36| ** | 0.50  | 27.31| 1.42| ** |
| Subset 7  | 0.39  | 36.78| 1.26| ** | 0.36  | 37.09| 1.25| ** |
| Subset 8  | 0.00  | 15.19| 0.06| ns | 0.00  | 18.52| 0.05| ns |
| Subset 9  | 0.20  | 21.68| 0.20| ** | 0.18  | 21.68| 0.20| ** |
| Subset 10 | 0.37  | 33.73| 1.25| ** | 0.39  | 37.17| 1.27| ** |
| Subset 11 | 0.28  | 37.88| 1.08| ** | 0.31  | 37.12| 1.10| ** |
| Average   | 0.34  | 29.31| 1.05|    | 0.36  | 28.50| 1.09|    |

ns: no significance

Table 6 shows the summary results for cross-validation according to Method 2 in Figure 7. Table 6 shows that the prediction accuracy of SGT in five subsets is higher than that of RF, except in subset 2 and subset 3. In these five subsets, the maximum increase value of $R^2$ is 0.08, and the minimum value is 0.01. The maximum reduction value of RMSE is 0.08. The minimum value is 0.01. The RPD value has a maximum value of 0.11, and the minimum value is 0.01.

Predictions overestimate soil salinity at low values and underestimate it at high values, which results in a linear trend of residuals. This happens in all subsets from both algorithms. The underestimated values are
mainly distributed on the periphery of the oasis (which receives salt transferred from the upper irrigation area), and the underestimated values are distributed in the irrigated areas inside the oasis due to human management (irrigation with low salinity water).

Comparisons of RF and SGT for soil salinity prediction with different subsets

Tables 4 and 5 show the results obtained from the first comparison method in Section 2.2.6. The study found that the following results: (1) For RF, the accuracies of six subsets were superior in the original (Table 4) compared to the regrouped subsets (Table 5; subset 2, subset 5, subset 6, subset 9, subset 10, and subset 12). Taking the RPD value as an example, the precision of the aforementioned subsets increased by 4.23%, 3.23%, 2.21%, 1900%, 620% and 11.11%, respectively. The predicted accuracy of the remaining four regrouped subsets was higher than the original (subset 3, subset 11, subset 7 and subset 8). The RPD values increased by 3.14%, 2.4%, 13.97% and 2.38%, respectively; (2) For SGT, the following six subsets had a higher predicted accuracy in the original compared to the regrouped subsets: subset 5, subset 6, subset 9, subset 10, subset 12 and subset 8. The RPD values increased by 4.84%, 8.22%, 2180%, 540%, 12.73% and 12.00%, respectively. The accuracy of the regrouped subsets was higher than that of the original in subset 2, subset 3 and subset 7. The RPD values increased by 5.97%, 2.38% and 2.11%, respectively. Moreover, the validation results in the regrouped subsets showed that the prediction accuracy based on SGT was better than that of RF for approximately 50% of the subsets (subset 2, subset 6, subset 11, subset 12 and subset 7), and the average value of the improvement from SGT was 5.75%. The improvement of RF and SGT was consistent in 20% of the subsets (subset 5 and subset 10).

Table 6 shows the difference in prediction accuracy between the combined and reclassified subsets discussed in Section 2.2.6 (Figure 7). First, the predicted values of the two subsets (subset 9 and subset 11) were combined and compared with the predicted results of the regrouped subsets. The SGT showed that the values of R² and RPD in the combined subsets increased by 14.59% and 6.47% compared to the regrouped values. The results calculated from RF showed that the values of R² and RPD in the combined subsets increased by 10.87% and 4.38% compared to the regrouped values. Second, when the predictive values of subset 2 and subset 3 were merged, the validation results showed that the predictive precision based on SGT was reduced by 2.17% (R²) and 2.90% (RPD), while the predictive accuracy based on RF was reduced by 2.08% (R²) and 1.46% (RPD). The predicted values of the two combined subsets (subset 7 and subset 8) compared with the predicted values of the regrouped subsets showed that the R² and RPD values from the former increased by 8.51% and decreased by 3.62% when obtained from the SGT. However, the simulation results of RF showed that the R² value of the former decreased by 7.00%, and the RPD value decreased by 3.03%. When the predicted values of subset 5 and subset 6 are combined, the predicted accuracy increases by 17.50% (R²) and 5.38% (RPD) when calculated from the SGT. The RF-based predicted accuracy increased by 7.50% (R²) and 2.32% (RPD).

Key variables for capturing soil salinity variations in different subsets

Figure 8 shows that the most important variables selected in the eleven environment subsets different. The relative importance of each variable was determined from SGT and RF and is shown in decreasing order and normalized to 100%. Considering that both algorithms can generate important variables, only the variables that produced higher precision are displayed in Figure 7. This result proves the hypothesis mentioned in the preface, namely, that the key variables that affect soil salinity vary in different subsets. We also calculated the contribution of each dominant factor used in the models, which is illustrated in
For all the subsets, terrain attributes with multiple resolutions had the highest influence (with an average value of 52.98%) on the model prediction, followed by Landsat-based variables (with an average value of 36.22%), landform (with an average value of 11.29%) and land use (with an average value of 10.51%) (Figure 9). In all the subsets, the Landsat-based variables with a higher frequency of occurrence
in optimized data sets are TEM, EVI2, EVI, ENDVI, and FSEN, which are 7 times, 6 times, 6 times, 4 times, and 4 times higher, respectively. Landform and land use rank in the forefront of importance in the following subsets: subset 1, subset 2, subset 3, subset 11, and subset 12.

**Spatial distribution of soil salinity**

Based on the iterative model calculated from Section 2.2.6, this study simulated the distribution of soil salinization in the study area (Figure 10). Generally, the soil salinity maps based on RF and SGT are more similar in spatial distribution. In Figure 10, a and b show the spatial distribution of soil salinity based on an all sample model, whereas c and d (subset 7 and subset 8), e and f (subset 3 and subset 2), and g and h (subset 5 and subset 6) show the soil salinity map calculated from RF and SGT, respectively, in separate regions. The range beyond the sub-regions (LAFGS & LAAFP; Grassland & Unused land) is filled with predictions from the full sample model. Overall, the major patterns shown in the above figure are consistent with (Ding & Yu, 2014) and (Wang et al., 2015), who showed corresponding changes to the surface features of land use and soil salinity transport characteristics (Wang et al., 2015). Located at the foot of the mountain, coarse soil texture is not conducive to soil salinity accumulation. Within the oasis(eregion area as maple leaf type in green), surface irrigation washes away salt from the soil since the 1970s, a large number of drainage and desalting water conservancy projects have been tested in China, and a complete drainage system has been established, which makes the agricultural irrigation area have better desalting efficiency. The Desert-Oasis transition zone in the southern part of the study area, however, there are no such outlets for the accumulated salt content within the topsoil during both dry and wet seasons, resulting in a heightened salt concentration outside the oasis(color in orange) (Ding & Yu, 2014). In the northeast of the map, the accumulation of soil salinity has been terminated due to the deep groundwater level. Comparing the above field survey results, figures a and b, e and f are more consistent with the above results than other subsets. The soil salinity values are overestimated in figures c and d at the foot of the mountain and in the northeastern region. The distribution of soil salinity in figures g and f is similar to that of figures c and d.

**Discussion**

**The causes of the spatial variability of soil salinity in the Tarim Basin oasis**

The diversity of vegetation and soil types and the high intensity of anthropogenic activity in the Tarim Basin cause the local soil salinity to exhibit strong variability. In the past 50 years, to meet the needs of population growth and economic development, a large number of land development and utilization activities both inside and outside irrigation areas have been implemented successively (Duan et al., 2010). During this period, people’s ability to regulate and control water resources has been continuously strengthened. The distribution of water resources has gradually changed from river-dominated to channel-dominated. However, due to the instability of irrigation cycles and the diversity of local topography, vegetation and soil types in the study area result in uneven water distributions, both spatially and temporally (Ding Jianli, 2017). Over the years, these natural and anthropic influences have affected the wet and dry processes of the soil at different times and spatial scales, enhancing the spatial heterogeneity of water and salt in the surface and subsurface soil (Z.Y. Zhao et al., 2008). This process may be the reason for the strong spatial variability of soil salinity in most of the above environmental conditions.

**Algorithm evaluation**

From Table 4 to Table 6, 18 out of 26 subsets (close to 70%) show that SGT predictions are closer to field measurements than to RF predictions. This result suggests that SGT is more suitable in arid regions where the variability in soil salinity is higher. However, because these two algorithms have not been compared in previous soil salinity studies, therefore, we cite other studies that illustrate the performance of SGT and RF. Naghibi et al. (2016) produced groundwater spring potential maps in Iran using three machine-learning models (Naghibi et al., 2016), including SGT, CART, and RF. The SGT model produced the best prediction results, followed by the CART and RF models. In the study of Youssef et al. (2016), four modelling techniques, including RF, SGT, classification and regression tree (CART), and the general linear (GLM) were used, and the results were compared to landslide susceptibility mapping in Saudi Arabia. The results showed that the average under the curve (AUC) for success rates were 0.783 (78.3%), 0.958 (95.8%), 0.816 (81.6%), and 0.821 (82.1%) for the RF, SGT, CART, and GLM models, respectively. Yang et al. (2016) found the spatial distribution of soil organic carbon content by comparing SGT and RF models in the northeastern Tibetan Plateau(Yang et al., 2016). The SGT and RF models yielded similar predicted SOC spatial patterns, and the spatial pattern predictions in the BRT model were slightly greater than those of the RF model in areas with denser vegetation. However, not all comparisons show that RF performs better than SGT. Ismail and Mutanga (2010) compared the performance of bagging, boosting and random forest to predict Sirex noctilio-induced water stress in Pinus patula trees using nine spectral parameters derived from hyperspectral data. Using the random forest
Figure 10. Spatial distribution of soil salinity (ds/m) predicted by SGT and RF for three combined modes, and the full sample model (Table 6) in the study area.
ensemble, they found a 15% increase in predictive accuracy when compared to single regression trees, a 4% increase in accuracy when compared to bagging and a 5% increase in accuracy when compared to SGT. Therefore, the above comparison results showed that the performance of RF and SGT in different subsets does not always have absolute advantages, and multiple studies are needed before the subsets can be used.

The phenomenon that the attribute values of soil samples are overestimated or underestimated has also occurred in other studies (Figure 11 and 12). A similar regionalization of soil properties with RF (Hengl et al., 2017; Muller & Niekerk, 2016) and SGT (Schillaci et al., 2017) has been attributed to small datasets and underrepresented values of target variables (Muller & Niekerk, 2016) or to the prediction algorithm of RF, which computes the unweighted average of the collection of trees (Liang et al., 2016). This process creates results biased towards the sample mean, and consequently, under/overestimation of large/small values of the target variable (Blanco et al., 2018). Despite the abovementioned objective facts, the impact of the sample design cannot be ignored. The same level of salinity may correspond to a variety of landscape features, while the same landscape feature corresponds to multiple salinity values. The corresponding landscapes in each sample area have high and low values, which results in overestimation or underestimation of the values in the entire oasis (final predicted values for both algorithms were the average of all the trees). The existence of this phenomenon cannot be avoided due to the quality of the data and high spatial heterogeneity of soil salinity. The complexity of the soil environment requires the development of more sensors to observe the land surface from different perspectives.

The location of the soil sample is designed based on global variables. Specific to the soil heterogeneity of each sub-data set or its spatial coverage, the location of the sample points may require further planning. It is necessary to consider both the variability of the field-scale and the actual conditions of the sub-regions. This requires to incorporate scale information and spatial heterogeneity to prevent blind spots that do not take into account into sample locations configuration. But it needs the cooperation of high spatial resolution remote sensing data. The algorithm itself is just passively mining the correlation between the data you feed it. If the texture information of the underlying surface cannot be maximized through the samples, the result may not be enough to represent the true spatial variability of soil salinity.

Comparisons to other studies regarding the accuracy of salinity prediction

The simulation and validation of soil salinity have been conducted using Landsat data in some other case studies reported in the literature (H. Chen et al., 2015; Yahiaoui et al., 2015). Reviewing the above research, the results are diverse because of differences in sampling depth (0–30 cm, 0–20 cm and unknown sampling depth), the number of samples used for modeling, variable selections (GDVI, CRSI, EEVI, FSEN, SI and so on), accuracy of fitting ($R^2 = 0.874$, $R^2 = 0.564$, $R^2 = 0.483$, $R^2 = 0.78$, $R^2 = 0.45$, $R^2 = 0.71$, $R^2 = 0.93$ and so on), prediction techniques (linear regression, multiple linear regression, backpropagation neural network, support vector machine, exponential regression and so on), verification methods (leave-one-field-out and cross-validation model/validate with a certain sample proportion), and the geographical environments of the study areas.

The prediction of soil salinity based on machine learning is rare in Xinjiang, and the existing studies were mostly based on a multivariate linear fitting equation when studying the sensitive variables of soil salinization. During a literature review, one study of the prediction of soil salinity based on remote sensing and machine learning in Xinjiang was found. S. Chen et al. (2014) attempted to predict soil salinity based on a modified soil adjusted index, humidity index, salt index, and a brightness and normalized difference salt index derived from Landsat OLI using a BP neural network algorithm (using 52 samples to establish the training data) (S. Chen et al., 2014). The results indicated that $R^2 = 0.92$ ($n = 20$, 0–30 cm). However, the contribution of each index to the final salinity prediction and the applicability of the variables in other regions was not reported. Thus, the learning process was not described, and the output is difficult to interpret, which affects the credibility and acceptability of the results. In the Kuqa Oasis, Ding and Yu (2014) (Ding & Yu, 2014) suggested that a multivariate model with a salt index and NDVI as the independent variable produced the best fit in autumn (sampled on September 22nd): $ECe = 183.83^{*}SI-84.22^{*}NDVI + 4.135$ ($R^2 = 0.44$, $n = 68$, $p < 0.01$). However, the accuracy of this model has not been verified.

Based on incomplete statistics, the depth ranges of soil salinity predictions are primarily 0–10 cm, 0–20 cm and 0–30 cm. Because of differences in soil depth and environmental variables, the comparability among the studies is low. Nevertheless, we believe that results of this study are an important supplement to the results of the abovementioned studies.

Relative importance of environmental variables

Land use is the most important variable that affects salinity changes in the Kuqa oasis. In recent decades, agricultural reclamation has gradually changed the approaches to land use in Xinjiang, and has subsequently affected the variations of water and salt in the soil. (Y.G. Wang et al., 2009) showed that the accumulation of soil salinity in different land-use types was
significantly affected by both natural processes and human activities in the Sangong River catchment of Xinjiang. A study by (Wang & Li, 2013) showed astonishing levels of land resource overexploitation in the Fubei region (oasis in Xinjiang) after 1960, with a > 40% increase in irrigated farmland area. Expanding soil salinity areas and higher degrees of saline land were significant in all land-use types in the study area, showing a 16.4% increase in the area with a soil salt content that exceeded 20 g/kg from 1982 to 2009. To investigate the changes in soil quality during the process of oasisization, Gong et al. (2015b) established five experimental fields in the oasis-desert ecotone of the Keriya River basin, including farmlands (FL) and four typical natural lands: natural forest (NF), saline and alkaline land (SAL), desert (D) and sand land (SL) of the main local land cover. The four natural lands have distinct soil characteristics, including distinct soil salinity. The studies from Xinjiang scholars have proven the effect of land use on salt migration, but they have not quantified its contribution. This study quantitatively describes the role of land use in salinity prediction. In addition, land use was an important factor in the optimized dataset of multiple environmental subsets (7/11), which also indicates its importance.

Scholars from different regions have demonstrated the indispensable role of the vegetation index in salinity prediction. However, the sensitivity of different vegetation indices to soil salinity varies. For details, please refer to the literature as cited in Section 2.2.4. By comparing the contribution of the vegetation index to the prediction of soil salinity in each subset, we found the following features: The vegetation index appeared within the optimized dataset in most subsets. In the present study, the vegetation indices that highly correlated with soil salinity, such as GDVI and SAVI, were not included in the final optimized datasets. The above two indices in (Allbed et al., 2014) and (Wu et al., 2014) were considered only to establish a linear regression but not to carry out the screening process. Highly dependent variables that were not included in the optimization dataset may be related to multicollinearity and non-linearity considerations. Notably, EVI2 and ENVI are in the early stages of application for soil salinization research in arid areas. The existing studies show that EVI2 has the ability to resolve LAI differences with different background soil reflectance (Mondal, 2011). Findings suggest that EVI2 is strongly associated with discrete land cover classes; thus, it has a potential for describing landscape characteristics without introducing subjective errors. The ENVI introduced shortwave infrared (2.11–2.29 um) into the original NDVI index to remove atmospheric effects and reduce the redundancy between bands, thus highlighting the differences in vegetation coverage and growth. In addition, T.-T. Zhang et al. (2015) (T.-T. Zhang et al., 2015) reported that the seasonal integral of EVI (EVI-SI) extracted from the smoothed EVI time series profile was the best indicator of the degree of soil salinity. Notably, one study used two recently improved vegetation indices, ENVI and EEVI (H. Chen et al., 2015) ( acronym show in Table 2), by adding a shortwave infrared band into NDVI and EVI for salinity analysis. The results of H. Chen et al. (2015) (applied to an estuary area of the Yellow River in China) indicated that the correlation between ENVI and soil salinity was heightened, and the multicollinearity between vegetation indices was greatly reduced by extending the traditional vegetation index. These two vegetation indices were first introduced for arid lands to determine their sensitivity to soil salinity.

Furthermore, model construction of (b5-b7)/(b5 + b7) (termed FSEN in this study) was employed in (Yu et al., 2010), Zheng et al. (2014) and (Kushida & Yoshino, 2010) for soil-salinity mapping, crop residue cover and FAPAR estimation. This finding led to the hypothesis that this index may be related to vegetation activity. When plant root salt content is too high, the vegetation cannot absorb water from the soil, which affects growth. LandSat’s 5th and 7th bands, the former for vegetation moisture measurement and the latter for hydrothermal mapping, are both hydrothermally related. Hydrothermal changes directly affect the movement of water and salt (Metternicht & Zinck, 2003).

(Sepaskhah & Boersma, 1979) found that the apparent thermal conductivity is independent of water content at very low water-content levels. Consequently, in the driest conditions (lowest moisture), thermal conductivity is associated with salinity (Wu et al., 2014).

Geomorphology maps are a useful source of information for assessing soil parent material and other factors of soil genesis, particularly in arid areas (Jafari et al., 2014). (Li et al., 2007) analysed the long-term changes in soil salt accumulation in different landforms, including the alluvial-proluvial fans, alluvial plain and deltas. The salinity accumulation process differs among these three landform types. (Taghizadeh-Mehrjardi et al., 2016) introduced the geomorphological features in their research area to infer soil salinization. The use of geomorphometry-terrain analysis using digital elevation data (Pike, 2000) to model areas that are susceptible to salt accumulation has produced good results (Vermeulen & Van Niekerk, 2017). The above studies also indicate that geomorphological features are important indicators of soil salt variability.

Topography fluctuations affect the runoff paths of surface and groundwater and redistribute the soil moisture and salinity. Several scholars have demonstrated the effects of topography on soil salinization
(Akramkhanov et al., 2011; Vermeulen & Van Niekerk, 2017). Considering that the terrain of the research area is relatively gentle, they suggested using multiple-resolution DEM derivatives to fully extract the potential of terrain for prediction. Notably, the highest contributors were the topographic derivatives with low spatial resolution. This finding may be related to the relatively gentle topography of the research area, where the variation distance was comparatively large. The terrain variables appeared in each subset, and their proportion in the optimized datasets was higher in several subsets (NDVI > 0.22 and LAFGS). This result shows the importance and necessity for selecting multiple-resolution DEM to compute the topographic index for modeling soil properties.

Figure 13 shows the correlations between soil salinity and environmental variables from all the samples tested. Among them, the variables in bold font were significant at P < 0.01, while the characters in red were significant at P > 0.05. A comparison of Figure 11 with Figure 6 shows that the relationship between environmental variables and soil salinity varied in different subsets. Variables such as EVI and EVI2 were not highly correlated with soil salinity at the oasis scale but had relatively high contributions under specific subsets, such as subset 3 and subset 5. Although there were significant correlations between the band variables and soil salinity, only a few occurred in the final optimized variable data set. Highly correlated variables such as GDVI did not have a high degree of contribution in specific subsets. The above results suggest that the correlation between variables and soil salinity is heterogeneous as the external environment changes, including sample size, sensitivity, and environmental characteristics. This finding also explains the necessity and importance of the analysis of specific issues related to the prediction of soil property.

Uncertain analysis

After a comprehensive analysis of the results of this paper, future research must consider the following aspects: 1. The design of the sample point under the oasis scale should consider not only the global area but also the sub-regions and take into account road accessibility and sampling costs. Although this study tested the simulation in each sub-region, and the predicted error was decreased, if more representative samples were gathered in the partition, we believe the forecast accuracy would be notably improved. 2. Existing remote sensing or digital archival data cannot be achieved with the exact matching of geographical phenomena due to the existence of mixed pixels. This process requires the use of higher spatial resolution data. 3. There is a multiplicity of species diversity and soil complexity in a certain range of soil salinity values, which exacerbates the probability of being misjudged. This phenomenon can be seen in the results of the correlation analysis referring to vegetation indices (Figure 12). 4. More sensitive variables and indices need to be constructed, which will require more advanced sensors and multiple tests. 5. Although the instrument for testing salinity was corrected many times, avoiding errors is difficult due to the accuracy of the instrument. 6. When the salt content of the soil is below 15%, a corresponding change cannot be

![Figure 11](image-url) Scatterplot of the residuals predicted by RF versus the measured soil salinity in each subset.
observed from the outside, which increases the difficulty of probing in two cases: one in the agricultural crop area and the other at the deep groundwater level, which has stopped accumulating salt. Remote sensing data, even with atmospheric corrections and topographic corrections, have certain errors that accumulate in subsequent processes.

It should be acknowledged that stratifying the study area into relatively homogeneous landform-vegetation units was important to the success of the soil salinity prediction over the study area using the proposed approach. In addition to the control of soil conditions, land surface dynamic feedbacks are influenced by landform and vegetation within different oasis-desert
ecosystem. However, within any given landform-vegetation strata, such as crop land or nature land, the differences in dynamic feedbacks among locations can be primarily attributed to the differences in soil conditions within that strata. This stratification enables the machine learning (RF and SGT) techniques to produce classes of dynamic feedback patterns that are primarily controlled by soil conditions. Stratification is not always obligatory, for application of the dynamic feedback approach but it is likely to be necessary for application to any areas that display significant differences in landform and vegetation conditions.

Conclusions
Based on Landsat, DEM-derived indices, land use and landforms, this study compared two machine learning algorithms (RF and SGT) to predict soil salinity under different subsets of a typical oasis in arid zones. After analysing the results, we can draw the following preliminary conclusions:

(1) Of the 27 subsets simulated in this study (12 original and 15 derivatives), 70.37% of the subsets proved that SGT had a higher prediction accuracy than did RF. Therefore, SGT is more suitable than RF for predicting the spatial variability in soil salinity in complex environments in arid areas. In addition, the key parameters set in RF and SGT strongly influence the uncertainty of the model and the choice of important environment variables. Appropriate parameters should be set for different environmental patterns.

(2) Using comparisons of multiple subset-optimized datasets, the study found that the following variables are relatively more important for predicting soil salinity in sub-regions and at the oasis scale in the study area: land use, landform, TEM, EVI, EVI2, FSEN, ENDVI and topographic variables with multiple resolutions.

Future research will introduce multi-temporal high spatial resolution microwave data, which contains parameters related to soil texture and soil moisture, which will help improve the prediction accuracy. At the same time, other depths of soil should also be collected to test its response to environmental variables. In addition, the research scale will be extended to multiple oases, and further explore the contribution of subregions in soil attribute prediction modeling.

Disclosure statement
No potential conflict of interest was reported by the authors.

Funding
This study was supported by the National Natural Science Foundation of China (U1603241, 41661046, 41771470, 41261090, U1303381), the Key Laboratory Special Fund of Xinjiang Uyghur Autonomous Region (2016D30001, 2014KL005), the China Postdoctoral Science Foundation (2016M602909) and the Scientific Research Foundation for Doctors of Xinjiang University (BS150246).

References
Akrakhchanov, A., Martius, C., Park, S. J., & Hendrickx, J. M. H. (2011). Environmental factors of spatial distribution of soil salinity on flat irrigated terrain. Geoderma, 163(1), 55–62. https://doi.org/10.1016/j.geoderma.2011.04.001
Albed, 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. https://doi.org/10.1016/j.geoderma.2014.03.025
Angileri, S. E., Conoscenti, C., Hochschild, V., Märker, M., Rotigliano, E., & Agnesi, V. (2016). Water erosion susceptibility mapping by applying stochastic gradient tree-boost to the imera meridionale river basin (Sicily, Italy). Geomorphology, 262, 61–76. https://doi.org/10.1016/j.geomorph.2016.03.018
Blanco, C. M. G., Gomez, V. M. B., Crespo, P., & Ließ, M. (2018). Spatial prediction of soil water retention in a Páramo landscape: Methodological insight into machine learning using random forest. Geoderma, 316, 100–114. https://doi.org/10.1016/j.geoderma.2017.12.002
Boettinger, J. L., Ramsey, R. D., Bodily, J. M., Cole, N. J., Kienast-Brown, S., Nield, S. J., Saunders, A. M., & Stum, A. K. (2008). Landsat spectral data for digital soil mapping. Digital Soil Mapping with Limited Data. Breiman, L. (2001). Random Forests. Mlear, 45(1), 5–32. https://doi.org/10.2307/A1010933404324
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: Theoretical approach. Remote Sensing of Environment, 82(2), 188–197. https://doi.org/10.1016/S0034-4257(02)00366-6
Chagas, C. D. S., de Carvalho Junior, W., Bhering, S. B., & Calderano Filho, B. (2016). Spatial prediction of soil surface texture in a semiarid region using random forest and multiple linear regressions. Catena, 139, 232–240. https://doi.org/10.1016/j.catena.2016.01.001
Chai, H., Zhou, C., Chen, X., & Cheng, W. (2009). Digital regionalization of geomorphology in Xinjiang. Journal of Geographical Sciences, 19(5), 600. https://doi.org/10.1007/s11442-009-0600-4
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. https://doi.org/10.3969/j.issn.1002-6819.2015.05.016
Chen, S., Gao, C., Bin, X. U., Jin, Y., Jinya, L. L., Hailong, M. A., Zhao, F., Guo, J., & Yang, X. (2014). Quantitative inversion of soil salinity and analysis of its spatial pattern in agricultural area in Shihzei of Xinjiang. Geographical
Ding, J., & Yu, D. (2014). Monitoring and evaluating spatial variability of soil salinity in dry and wet seasons in the Weiran–Kuqa Oasis, China, using remote sensing and electromagnetic induction instruments. Geoderma, 235–236, 316–322. https://doi.org/10.1016/j.geoderma.2014.07.028

Ding Jianli, W. F. (2017). Environmental modeling of large-scale soil salinity information in an arid region: A case study of the low and middle altitude alluvial plain north and south of the tianshan mountains, Xinjiang. Journal of Geographical Sciences, 72(1), 64–78. https://doi.org/10.11821/dxbj201701006

Douou, 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. https://doi.org/10.1016/j.geoderma.2005.10.009

Duan, J. J., Cao, X. L., Shen, Y. P., Gao, Q. Z., & Wang, S. D. (2010). Surface water resources and its trends in weigan river basin on the south slope of Tianshan,China during 1956–2007. Journal of Glaciology & Geocryology, 32(6), 1211–1219. https://doi.org/10.1007/s00472710010772

Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted trees. Journal of Animal Ecology, 77(4), 802–813. https://doi.org/10.1111/j.1365-2656.2008.01390.x

Elmagarr, A. A., & Noller, J. S. (2009). Application of remote-sensing data and decision-tree analysis to mapping salt-affected soils over large areas. Remote Sensing, 2(1), 151–165. https://doi.org/10.3390/rs2010151

Forkner, G., Houskpatin, O. K. L., Welp, G., & Thiel, M. (2017). High resolution mapping of soil properties using remote sensing variables in South-Western Burkina faso: A comparison of machine learning and multiple linear regression models. PLoS ONE, 12(1), e0170478. https://doi.org/10.1371/journal.pone.0170478

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

Garcia, M., Oyonarte, C., Villagarcia, L., Contreras, S., Domingo, F., & Puigdefabregas, J. (2008). Monitoring land degradation risk using ASTER data: The non-evaporative fraction as an indicator of ecosystem function. Remote Sensing of Environment, 112(9), 3720–3736. https://doi.org/10.1016/j.rse.2008.05.011

Gong, L., Ran, Q., He, G., & Tiyip, T. (2015a). A soil quality assessment under different land use types in keriya river basin, Southern Xinjiang. China. Soil and Tillage Research, 146, 223–229.

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

Gregorutti, B., Michel, B., & Saint-Pierre, P. (2017). Correlation and variable importance in random forests. Statistics and Computing, 27(3), 639–678. https://doi.org/10.1007/s11222-016-9646-1

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(1–2), 102–113. https://doi.org/10.1016/j.geoderma.2008.05.008

Hardsky, M., Klemas, S., & Smart, V. (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.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning data mining, inference, and prediction (Second ed.). Springer.

Hengl, T., Mendes, D. J. J., Heuvelink, G. B., Ruiperez, G. M., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M. N., Geng, X., & Baumerschallinger, B. (2017). SoilGrids250m: Global gridded soil information based on machine learning. PLoS ONE, 12(2), e0169748. https://doi.org/10.1371/journal.pone.0169748

Heung, B., Bulmer, C. E., & Schmidt, M. G. (2014). Predictive soil parent material mapping at a regional-scale: A random forest approach. Geoderma, 214–215, 141–154. https://doi.org/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. https://doi.org/10.1016/j.geoderma.2015.11.014

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. https://doi.org/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. https://doi.org/10.1016/0034-4257(88)90106-X

Inglada, J., Arias, M., Tardy, B., Hagolle, O., Valero, S., Morin, D., Dedieu, G., Sepulcre-Canto, G., Bontemps, S., Defourny, P., & Koetz, B. (2015). Assessment of an operational system for crop type map production using high temporal and spatial resolution satellite optical imagery. Remote Sensing, 7(9), 12356–12379. https://doi.org/10.3390/rs70912356

Ismail, R., & Mutanga, O. (2010). A comparison of regression tree ensembles: Predicting Sirex noctilio induced water stress in pinus patula forests of KwaZulu-Natal, South Africa. International Journal of Applied Earth Observations & Geoinformation, 12(1), S45–S51. https://doi.org/10.1016/j.jag.2009.09.004

Jafari, A., Khademi, H., Finke, P. A., Wauw, J. V. D., & Ayoubi, S. (2014). Spatial prediction of soil great groups by boosted regression trees using a limited point dataset in an arid region. Southeastern Iran. Geoderma, 232–234 (12), 148–163. https://doi.org/10.1016/j.geoderma.2014.04.029

Jiang, Z., Huete, A. R., Didan, K., & Miura, T. (2008). Development of a two-band enhanced vegetation index without a blue band. Remote Sensing of Environment, 112(10), 3833–3845. https://doi.org/10.1016/j.rse.2008.06.006

Jordan, C. F. (1969). Derivation of leaf-area index from quality of light on the forest floor. Ecology, 50(4), 663–666. https://doi.org/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. https://doi.org/10.1016/j.agwat.2004.09.038

Khorram, S. (1982). Remote sensing of salinity in the San Francisco Bay Delta. Remote Sensing of Environment, 12(1), 15–22. https://doi.org/10.1016/0034-4257(82)90004-9

Kushida, K., & Yoshino, K. (2010). Estimation of LAI and FAPAR by constraining the leaf and soil spectral characteristics in a radiative transfer model. International Journal of...
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. https://doi.org/10.1007/s11769-014-0718-x

Wang, F., Shi, Z., Biswas, A., Yang, S., & Ding, J. (2020). Multi-algorithm comparison for predicting soil salinity. *Geoderma, 365*, 114211. https://doi.org/10.1016/j.geoderma.2020.114211

Wang, Y., & Li, Y. (2013). Land exploitation resulting in soil salinization in a desert–oasis ecotone. *Catena, 100*, 50–56. https://doi.org/10.1016/j.catena.2012.08.005

Wang, Y. G., Yan, L. I., & Xiao, D. N. (2009). Effects of land use type on soil salinization at northern slope of tianshan mountain. *Journal of Soil & Water Conservation, 23*(5), 179–183. https://doi.org/10.3321/j.issn:1009-2242.2009.05.038

Wu, W. (2014). The generalized difference vegetation index (GDVI) for dryland characterization. *Remote Sensing, 6*(2), 1211. https://doi.org/10.3390/rs6021211

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

Xin, L., Yang, K., & Zhou, Y. (2016). Progress in the study of oasis-desert interactions. *Agricultural & Forest Meteorology, 230–231*, 1–7. https://doi.org/10.1016/j.agrformet.2016.08.022

Yahiaoui, I., Douaoui, A., Zhang, Q., & Ziane, A. (2015). Soil salinity prediction in the lower chelif plain (Algeria) based on remote sensing and topographic feature analysis. *Journal of Arid Land, 7*(6), 794–805. https://doi.org/10.1007/s40333-015-0053-9

Yang, R.-M., Zhang, G.-L., Liu, F., Lu, Y.-Y., Yang, F., Yang, F., Yang, M., Zhao, Y.-G., & 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. https://doi.org/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. https://doi.org/10.1007/s10346-015-0614-1

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

Zhang, F., Ding, J. L., & Tashpolat, T. G. (2007). Analysis on characteristics of soil salinization in the arid regions: A case study in the delta oasis of weigang and kuqa rivers. *Acta Pratacciutae Sinica, 16*(4), 1–12. https://doi.org/10.3321/j.issn:1009-2242.2007.04.006

Zhang, T.-T., Qi, J.-G., Gao, Y., Ouyang, Z.-T., Zeng, S.-L., & Zhao, B. (2015). Detecting soil salinity with MODIS time series VI data. *Ecological Indicators, 52*, 480–489. https://doi.org/10.1016/j.ecolind.2015.01.004

Zhang, X. S. (2001). Ecological restoration and sustainable agricultural paradigm of mountain-oasis-ecotone-desert system in the north of the tianshan mountains. *Acta Botanica Sinica, 43*(12), 1294–1299. https://doi.org/10.3321/j.issn:1672-9072.2001.12.016

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

Zhao, Z. Y., Zhang, H. Z., Huang, J. F., Feng-Ying, L. I., & Qian-Long, Y. U. (2008). Dynamic changes of water and salt and their coupling relationships in weigang river irrigation district. *Journal of Arid Land Resources & Environment, 22*(10), 142–146. http://doi.org/10.5007/1003-7578(2008)010-142-05

Zheng, B., Campbell, J. B., Serbin, G., & Galbraith, J. M. (2014). Remote sensing of crop residue and tillage practices: Present capabilities and future prospects. *Soil & Tillage Research, 138*, 26–34. https://doi.org/10.1016/j.still.2013.12.009

Zhu, A. X., Hudson, B., Burt, J., Lubich, K., & Simonon, D. (2001). Soil mapping using GIS, expert knowledge, and fuzzy logic. *Soil Science Society of America Journal, 65*(5), 1463–1472. https://doi.org/10.2136/sssaj2001.6551463x