



RESEARCH ARTICLE

WILEY

Soil salinity prediction and mapping by machine learning regression in Central Mesopotamia, Iraq

Weicheng Wu¹ | Claudio Zucca² | Ahmad S. Muhaimed³ | Waleed M. Al-Shafie⁴ | Ayad M. Fadhil Al-Quraishi⁵ | Vinay Nangia² | Minqiang Zhu¹ | Guangping Liu⁶

¹Key Laboratory of Digital Land and Resources, East China University of Technology, Nanchang, Jiangxi, PR China

²ICARDA (International Center for Agricultural Research in the Dry Areas), Rabat, Morocco

³College of Agriculture, University of Baghdad, Baghdad, Iraq

⁴GIS Division, Ministry of Agriculture, Baghdad, Iraq

⁵Earth Sciences Department, Faculty of Science, and Remote Sensing Center, University of Kufa, Kufa, Iraq

⁶Faculty of Science, East China University of Technology, Nanchang, Jiangxi, PR China

Correspondence

W. Wu, Key Laboratory of Digital Land and Resources, East China University of Technology, Nanchang, 330013 Jiangxi, PR China.
Email: wuwc030903@sina.com; wuwch@ecit.cn

Funding information

East China University of Technology, Grant/Award Number: DHTP2018001; Australian Agency for International Development, Grant/Award Number: LWR/2009/034

Abstract

Soil salinization affects crop production and food security. Mapping spatial distribution and severity of salinity is essential for agricultural management and development. This study was aimed to test the effectiveness of machine learning algorithms for soil salinity mapping taking the Mussaib area in Central Mesopotamia as an example. A combined dataset consisting of Landsat 5 Thematic Mapper (TM) and ALOS L-band radar data acquired at the same time was used for fulfilling the task. Relevant biophysical indicators were derived from the TM images, and the soil component was retrieved by removing the vegetation contribution from the L-band radar backscattering coefficients. Field-measured salinity at the three corner plots of triangles were averaged to represent the salinity of these triangular areas. These averaged plots were converted into raster by either direct rasterization or buffering-based rasterization into different cell size to create the training set (TS). One of the three triangle corners was randomly selected to constitute a validation set (VS). Using this TS, the support vector regression (SVR) and random forest regression (RFR) algorithms were then applied to the combined dataset for salinity prediction. Results revealed that RFR performed better than SVR with higher accuracy (93.4–94.2% vs. 85.2–89.4%) and less normalized root mean square error (NRMSE; 6.10–7.69% vs. 10.29–10.52%) when calibrated with both TS and VS. In comparison, prediction by multivariate linear regression (MLR) achieved in our previous study using the same datasets also showed less NRMSE than SVR. Hence, both RFR and MLR are recommended for soil salinity mapping.

KEYWORDS

combined optical–radar dataset, field sample rasterization, random forest regression, soil salinity prediction, support vector regression

1 | INTRODUCTION

Soil salinization is one of the most active land degradations and environmental hazards in irrigated lands worldwide, especially, in dry areas (Farifteh, Farshad, & George, 2006; Metternicht & Zinck, 2003),

such as Central and Western Asia (Ivushkin, Bartholomeus, Bregt, & Pulatov, 2017; Qadir et al., 2009; Qadir, Qureshi, & Cheraghi, 2008; Wu, Al-Shafie, et al., 2014; Wu, Muhaimed, et al., 2014). On average, 20% of the world's irrigated lands are affected by salinization, and this number increases to more than 30% in Iran and Egypt (Metternicht &

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2018 The Authors Land Degradation & Development Published by John Wiley & Sons Ltd

Zinck, 2003) and 50–51% in Uzbekistan (Ivushkin et al., 2017; Qadir et al., 2009). Salinity has greatly influenced crop production, which has declined, for example, by 30–60% in comparison with that in the nonaffected croplands in Mesopotamia, Iraq (Wu, Al-Shafie, et al., 2014). Therefore, it is of prime importance to investigate the severity and distribution of soil salinity in space and time to support decision makers in planning agriculture development to mitigate food security issues in the salt-affected countries.

In the past decades, a great number of remote sensing (RS)-based soil salinity mapping studies have been conducted (Allbed & Kumar, 2013; Bannari, El-Battay, Bannari, & Rhinane, 2018; Dwivedi & Rao, 1992; Farifteh et al., 2006; Fernández-Buces, Siebe, Cram, & Palacio, 2006; Gorji, Tanik, & Sertel, 2015; Ivushkin et al., 2017; Mougnot, Pouget, & Epema, 1993; Wu, Al-Shafie, et al., 2014; Wu, Mhaimed, et al., 2014). These studies have not only identified the relevant salinity indicators, for example, different vegetation indices, combined spectral response index, principal components, and land surface temperature (LST), but also proposed operational approaches such as best band combination and multiyear maxima-based multivariate regression modeling.

Several authors have explored the possibility to detect soil salinity by microwave radar data as they are independent of weather condition (Gong, Shao, Brisco, Hu, & Tian, 2013; Sreenivas, Venkataratnam, & Rao, 1995). The laboratory-based simulations conducted by these authors suggested that it is possible to use the microwave P, C, and especially L bands for detecting salinity in different settings because the signal can penetrate through the surface and reach the subsoil to a depth of up to 150 cm or more, depending on the wavelength/frequency of the emitted waves and soil moisture. However, satisfactory radar-based salinity mapping has been rarely reported probably due to the difficulty to separate the soil salinity from the moisture within the radar backscattering coefficients. Wu, Muhaimed, Al-Shafie, and Fadhil (in press) employed the leaf area index (LAI) and vegetation water content (VWC) derived from the optical data to remove the effects of vegetation cover on the backscattering coefficients of soil and found that these corrected backscattering coefficients were highly correlated with the measured soil salinity ($R^2 = 0.565\text{--}0.677$).

Recently, a strong momentum has been gained in RS-based land cover mapping including extraction of saline land by machine learning classifiers such as artificial neural network (ANN), support vector machines (SVMs), and random forests (RFs; Belgiu & Dragut, 2016; Huang, Davis, & Townshend, 2002; Kavzoglu & Colkesen, 2009; Ritter & Hepner, 1990; Rodriguez-Galiano, Ghimire, Rogan, Chica-Olmo, & Rigol-Sanchez, 2012; Wu, Zucca, Karam, & Liu, 2016). The advantage of these algorithms over the traditional parametric classifiers lies in their capacity to separate nonparametric signatures by determining the hyperplane in a high-dimensional space or by growing ensembles of decision trees and letting them vote for the most popular class (Breiman, 2001) making the nonseparable clusters in the parametric space separable (Wu et al., 2016). Comparing the most frequently applied and promising machine learning algorithms, Wilkinson (2005), Mas and Flores (2008), and Wu et al. (2016) found that ANN was often outperformed by other classifiers such as SVM and RF, and even by maximum likelihood. Pal (2005) and Wu et al. (2016) noted that SVM and RF could achieve equally well land cover mapping with a very high accuracy of 95.7–96.8% for local sites though they took much longer processing times than maximum likelihood.

Recently, Abdel-Rahman, Ahmed, and Ismail (2013), Wang, Zhou, Zhu, Dong, and Guo (2016), and so forth applied the random forest regression (RFR) to biophysical prediction such as leaf nitrogen concentration and biomass estimation. Farifteh, van der Meer, Atzberger, and Carranza (2007) used partial least squares regression and ANN, and Taghizadeh-Mehrjardi, Minasny, Sarmadian, and Malone (2014) employed regression tree to predict pixel-based soil salinity. This aroused our strong interest to explore the possibility to use the hotspotted machine learning regression algorithms, RFR, and support vector regression (SVR) for predicting and mapping soil salinity.

Actually, application of SVR and RFR for RS-based soil salinity prediction and mapping has been rarely reported. For this reason, the main objective of our study was to ascertain the applicability of these machine learning regression algorithms for such purpose. One specific objective was to compare their performance (mapping accuracy and reliability) with that of multivariate linear regression (MLR) using the same dataset (a single date of optical and radar dataset) used by Wu et al. (in press). The research was implemented in the Mussaib site in Central Mesopotamia.

2 | METHODS AND MATERIAL

2.1 | Study area

The study area is located in between the Tigris and the Euphrates Rivers in Central Mesopotamia, Iraq (Figure 1), where the main land use is croplands. This area has been a national agriculture development project site since 1950s for grain production including irrigated wheat and barley in spring and corn, vegetables, and fruits in summer. Perennial alfalfa and permanent tree crop such as date palm are also locally cultivated. Long-term fallows or abandoned croplands (uncultivated in the past 15–20 years) and unmanaged bare lands exist, and built-up areas are very local. The total area of the project site is around 250,000 ha. The dominant soil types are Aridisols and Entisols with texture class ranging from silt clay loam to silty loam with more than 20% of lime. The soils are mostly saline with electrical conductivity (ECe) ranging from 4 (low) to 30 (strong) dS m^{-1} (Wu et al., in press; Wu, Al-Shafie, et al., 2014).

Climatically, the Mussaib site is characterized by short cool winter and long hot summer. Rainfall is concentrated in winter and early spring from December to March with an annual average of about 82.5 mm during the past 60 years (recorded in the adjacent station Hillah). The mean minimum temperature is about 6.25°C in December–February, whereas the mean maximum temperature is around 43.2°C in July–August.

As a part of the Mesopotamian Plain, the landform of the study area is mostly flat with elevation varying from 25 to 31 m above sea level.

2.2 | Data

2.2.1 | Field data

Field surveys were conducted from July 2011 to July 2012 including soil sampling (July–November 2011), apparent electrical conductivity (ECa in millisiemens per meter or mS m^{-1}) measurements by EM38-MK2 (Geonics Ltd.; EM38 hereafter) in March–July 2012 and June 2013.

Soil samples were taken from 13 pedons (0- to 30-cm horizon of the profiles up to 150 cm in depth) and 17 auger holes of 0–30 cm in

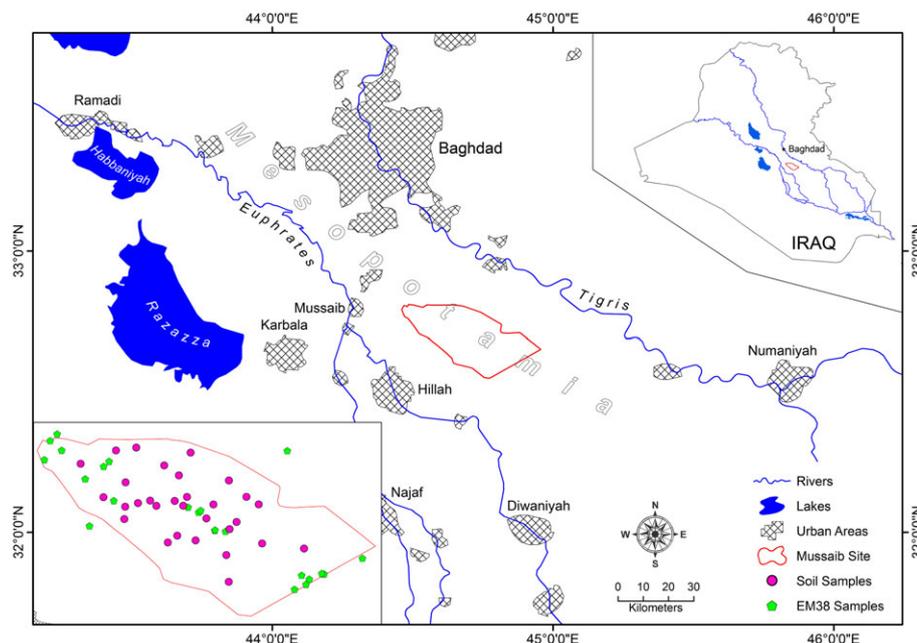


FIGURE 1 Location of the study area and distribution of the field sampling points [Colour figure can be viewed at wileyonlinelibrary.com]

depth in the study area in July–November 2011, when EM38 instruments were not available. The soil samples were analyzed in laboratory to measure soil electrical conductivity (ECe; 1:1 dilution method). Samples were taken mainly in croplands or under halophytes, which are normally problematic for soil salinity mapping by RS (Metternicht & Zinck, 2003).

After the arrival of the instruments, EM38 readings were conducted in three campaigns, in spring (March–April) 2012, with 45 (3×15) pairs of vertical (V) and horizontal (H) readings, and early summer (June–July 2012, when dry season started after harvesting wheat and barley), with 21 (3×7) pairs of V and H readings as supplementary sampling. V and H EM38 readings (EM_V and EM_H) were taken in small plots ($1 \text{ m} \times 1 \text{ m}$ in size) distributed at the three corners of triangles. The designed distance of any two corners of a triangle was about 15–20 m to ensure that the triangle could approximately represent one Thematic Mapper (TM) pixel. However, due to accessibility problem in field, EM38 readings could not be measured at the same points as soil samples, and it was also difficult to control the sampling triangles as equilateral, and their actual side lengths ranged between 25 and 52 m, so that the triangles covered an area of about 470–920 m^2 . The averaged EM_H and EM_V of the three pairs of readings were considered as the representative values of the observed triangular areas or, rather, of the corresponding TM pixels. Two additional triangles (3×2 pairs) of measurements surveyed near the site in June 2013 were also integrated in this study. Hence, totally 24 averaged pairs of EM38 readings including EM_V and EM_H were used as ground-truth training set (TS) for this study.

For validation purpose, any one pair of the three triangle corners was selected to compose a ground-truth validation set (VS), which was slightly different in both EM_V and EM_H readings and spatial locations from their averaged TS. The VS also contains 24 pairs of samples as above. As for land use/cover-related distribution, five of these samples were located in the long-term fallows or abandoned croplands, three in bare lands, and the remained ones in mixed croplands including alfalfa.

The lab-analyzed soil samples were used neither for calibrating the above EM38 readings nor for model training because of different locations from the EM38 sampling points (Figure 1) and could not represent the salinity of the TM pixels due to high spatial variability of salinity. Thus, these soil samples were only used for verification of the classified grades of salinity (ECe) converted from the predicted ECa (see Section 2.3.7 for detail).

2.2.2 | Satellite data

Level 1.5 product of PALSAR data of the Japanese ALOS satellite with a spatial resolution of 12.5 m were obtained from the European Space Agency (<https://alos-palsar-ds.eo.esa.int>). The L-band images were produced by a microwave radar sensor with a wavelength of 23 cm and frequency of 1.27 GHz in Fine Beam Double polarization mode (HH/HV). The images were acquired with an off-nadir angle of 34.3° and an incidence angle of $7.5\text{--}60^\circ$ on November 26, 2010, when summer crops, mainly maize, became mature and winter wheat and barley were to be sown. Rainy season had not yet started in the study area.

Landsat 5 TM images dated November 23, 2010, acquired almost on the same date as ALOS images, were also obtained from European Space Agency (<https://landsat-ds.eo.esa.int>).

It is noted that in the surrounding weather stations of the study area, namely, Baghdad, Karbala, Diwaniyah, and Hillah, no rainfall was recorded in the period from May to November 2010 (<https://fr.tutiempo.net/climat/iraq.html>). Thus, rainfall-induced moisture problem (Wu, Al-Shafie, et al., 2014; Wu, Mhaimeed, et al., 2014) could be avoided in our analysis.

2.3 | Approaches and processing procedures

2.3.1 | TM image processing

The Landsat 5 TM images were radiometrically calibrated, and a FLAASH (Fast Line-of-sight Atmospheric Analysis of Spectral Hypercubes) model (Perkins et al., 2012) was applied to remove the additive

atmospheric effects. The produced reflectance was rescaled to 0–1 for each band.

Biophysical indicators recognized in our previous studies as most relevant for salinity mapping (Wu, Al-Shafie, et al., 2014; Wu, Mhaimeed, et al., 2014) were produced. They were the normalized difference vegetation index (NDVI), the normalized difference infrared index (NDII; Hardisky, Klemas, & Smart, 1983) from TM bands 4 and 5, the generalized difference vegetation index (GDVI; Wu, 2014) with power number of 2 and 3 (denoted, respectively, GDVI2 and GDVI3), the LST from the thermal band, and the Tasseled Cap brightness (Crist & Cicone, 1984).

2.3.2 | L-band radar processing

The Level 1.5 radar product has been geometrically corrected, and pixels resampled to 12.5 m in size to rectify deformation by the provider. The digital number (DN) of the two HH and HV bands were, respectively, calibrated and converted into backscattering coefficients (σ_{HH}^0 and σ_{HV}^0), expressed in decibel (dB) following Shimada, Isoguchi, Tadono, and Isono (2009):

$$\sigma^0[\text{dB}] = 10 \log_{10}(\text{DN})^2 - 83.0. \quad (1)$$

An enhanced Lee filter (3×3 in size; Lee, 1980) was then applied to remove speckles or noises. σ_{HH}^0 and σ_{HV}^0 were hence derived and resampled to 30 m pixels to match the TM data.

2.3.3 | Removal of the influence of vegetation cover

As mentioned above, the difficulty to use backscattering coefficients to characterize soil salinity is related to the effects of soil moisture, especially, where vegetation cover is present. Attema and Ulaby (1978) have proposed the water cloud model for characterizing the effect of VWC on radar backscattering coefficient, which can be expressed as follows (Kumar, Prasad, & Arora, 2012; Moran, Vidal, Troufleau, Inoue, & Mitchell, 1998):

$$\sigma^0 = \sigma_{\text{veg}}^0 + L^2 \sigma_{\text{soil}}^0, \quad (2)$$

with

$$\sigma_{\text{veg}}^0 = AV_1 \cos(\theta_i) (1 - L^2), \quad (3)$$

$$L^2 = \exp(-2BV_2 \sec(\theta_i)), \quad (4)$$

$$\sigma_{\text{soil}}^0 = (\sigma^0 - \sigma_{\text{veg}}^0) / L^2, \quad (5)$$

where σ^0 is the total backscattering coefficient from both vegetation canopy and soil (either σ_{HH}^0 or σ_{HV}^0 in our case), σ_{veg}^0 is the backscattering contribution of the vegetation cover, and σ_{soil}^0 is that of soil; L^2 is the two-way vegetation attenuation; θ_i is the incidence angle of the radar beam; A and B are the vegetation parameters; and V_1 and V_2 are the vegetation descriptors. Kumar et al. (2012) applied LAI ($\text{m}^2 \text{m}^{-2}$) for V_1 and VWC (kg m^{-2}) for V_2 , respectively.

After numerous fittings, the LAI-GDVI2 model of Wu (2014), was found to perform better than other LAI-NDVI models given the same VWC (V_2), A , and B parameters. This model is shown as follows:

$$\text{LAI} = 0.091 \exp(3.7579 \text{GDVI2}) \quad (R^2 = 0.932). \quad (6)$$

Using this LAI model, vegetation-removed backscattering coefficient, σ_{soil}^0 , was better correlated to the field-measured apparent soil salinity. It was hence adopted for this study.

Similarly, we selected the VWC-NDVI model developed by Jackson et al. (2004) for maize for our analysis, that is,

$$\begin{aligned} \text{VWC} = & 192.64 \text{NDVI}^5 - 417.46 \text{NDVI}^4 + 347.96 \text{NDVI}^3 \\ & - 138.93 \text{NDVI}^2 + 30.699 \text{NDVI} \\ & - 2.822 \text{ (kg m}^{-2}\text{)} \quad (R^2 = 0.990), \end{aligned} \quad (7)$$

which outperformed other VWC-NDII and VWC-NDVI models given the same LAI, A , and B .

As for A and B , those obtained by Dabrowska-Zielinska, Inoue, Kowalik, and Gruszczynska (2007) for ALOS L-band radar data were tested in this study. We found that the second case of L band, that is, $A = 0.0045$ and $B = 0.4179$, could maximize the correlation between the vegetation-removed backscattering coefficient (σ_{soil}^0) and the field-measured salinity given the same LAI and VWC. This pair of A and B was finally selected for our study.

Inputting the selected A , B , LAI, and VWC models, and 34.3° as the mean incidence angle, the vegetation-removed backscattering coefficients ($\sigma_{\text{HH}(\text{soil})}^0$ and $\sigma_{\text{HV}(\text{soil})}^0$) were obtained. This removal procedure gained an increase of 16.6–25.6% in the correlation coefficient of $\sigma_{\text{HH}(\text{soil})}^0$ with the field-measured salinity in respect to that of σ_{HH}^0 and 11.5–21.4% in that of $\sigma_{\text{HV}(\text{soil})}^0$ in comparison with σ_{HV}^0 (Wu et al., in press).

2.3.4 | Combined dataset

The produced NDVI, GDVI2, GDVI3, NDII, LST, Tasseled Cap brightness, σ_{HH}^0 , σ_{HV}^0 , $\sigma_{\text{HH+HV}}^0$, $\sigma_{\text{HH}(\text{soil})}^0$, $\sigma_{\text{HV}(\text{soil})}^0$, and their sum $\sigma_{\text{HH+HV}(\text{soil})}^0$ were stacked together to compose an optical-radar combined 12-band dataset.

2.3.5 | Rasterization of the field measurements

To model salinity using machine learning regression, it is essential to create a TS based on the field measurements, that is, to rasterize the field plots. Two kinds of rasterization were conducted. One was a direct rasterization, that is, using Point to Raster conversion tool within ArcGIS to convert the averaged field measurement plots into raster cells of 30, 60, and 90 m in size and then resampled to 30-m pixels. The other was to first use a buffering function to convert the averaged field points into circular buffers with a radius of 30, 60, and 90 m and then apply a Feature to Raster function to convert these buffers into raster with an initial cell size of 10 m to catch the buffer forms; and at last, these cells were resampled to 30-m pixels to match the combined dataset.

The objective to rasterize sample plots into such different extents (30, 60, and 90 m) was to find the optimal spatial presentation of samples for machine learning regression modeling taking both the representativeness of samples and spatial variability of salinity into account.

2.3.6 | Application of SVR and RFR for salinity prediction

Both SVR and RFR modeling were conducted within EnMap-Box (Waske et al., 2012; van der Linden et al., 2014), an image processing and analysis package designed by IDL (Interactive Data Language).

Support vector regression

SVR (Vapnik, Golowich, & Smola, 1997) is a learning regression algorithm extended from the SVM (Vapnik & Lerner, 1963). The strength of SVR is to model the complex nonlinear relationships in the multidimensional or hyperdimensional feature space and estimate the linear dependency of the variables to be predicted on the predictive covariates by fitting an optimal approximating hyperplane to the training data. For linearly nonapproximable problems, the training data are implicitly mapped by a kernel function with regularization into a higher dimensional space, wherein the new data distribution enables a better fitting of a linear hyperplane that appears nonlinear in the original feature space (van der Linden et al., 2014).

While executing SVR modeling, the parameterization is a critical procedure that requires the user to select the parameter(s) of the kernel function (γ) as well as the regularization (C) and the loss function (ϵ). As many researchers have underlined (Huang et al., 2002; Kavzoglu & Colkesen, 2009; van der Linden et al., 2014; Wu et al., 2016), radial basis function can capture best the nonparametric features. Hence, radial basis function including linear kernel was selected. And the default values were chosen for the other parameters such as C (min 0.01 and max 1,000) with a multiplier 10, three folds of cross-validation, and automatic search for ϵ .

After training, the derived SVR models were applied back to the combined dataset to produce the apparent soil salinity (ECa) maps.

Random forest regression

RFR is formed by an ensemble of growing decision trees depending on random vectors and begins with many bootstrap samples that are drawn randomly with replacement from the original training dataset (Breiman, 2001). A key procedure in RFR is to use Bagging (Bootstrap Aggregating) in tandem with random feature selection, as Bagging can dramatically reduce the variance of unstable procedures such as tree growing, leading to an improved prediction and enhanced accuracy (Breiman, 2001). More concretely, a regression tree is fitted to each of the bootstrap samples from the TS, or rather, random vectors, that govern the growth of each tree in the ensemble to grow regression forests. In these forests, random feature selection at each node to determine the split criteria is on top of Bagging. Therefore, the generalization error can be provided by out-of-bag estimation, which can be also used to estimate the importance of each variable. RFR has no overfitting problem because it applies the strong law of large numbers as RF. The more features used, the less error produced (Breiman, 2001).

While conducting RFR modeling, we kept all 12 bands as input variables with 24 observations (samples for training, TS). Some critical parameters to be set were first the number of trees (NT) depending on the complexity of the features. The default value was 100 within EnMap-Box, but tests were also conducted by setting it to 300, 500, and 1,000 in view of the spatial variability of salinity. The second one was the number of randomly selected features (or number of

variables) at each node, which can be the square root of all features or logarithm (log) of all features or a user-defined value. In this analysis, the square root of all features was selected. The third one was the stop criteria (for node splitting), where the default values of the minimum number of samples in a node, 1, and the minimum impurity calculated based on Gini index, 0, were chosen.

After parametrization using the rasterized EM_V or EM_H as TS, the produced RFR models were applied back to the combined dataset to predict the apparent soil salinity (ECa).

2.3.7 | Conversion from ECa to ECe

Since what SVR and RFR had predicted was the apparent soil salinity ($mS\ m^{-1}$), it had to be converted into the lab-measured ECe ($dS\ m^{-1}$), which would be more meaningful for land management. We applied hence our results obtained from the regional-scale sampling and lab analysis in the whole Mesopotamia for this purpose. Regional sampling includes two transects and four pilot sites, where both soil and EM38 readings were sampled at the same plots. The ECe-EM38 readings (ECa) relationships were expressed as follows (Wu, Al-Shafie, et al., 2014; Wu, Mhaimeed, et al., 2014):

$$ECe\ (dS\ m^{-1}) = 0.0005EM_V^2 - 0.0779EM_V + 12.655\ (R^2 = 0.850), \quad (8)$$

$$ECe\ (dS\ m^{-1}) = 0.0002EM_H^2 + 0.0956EM_H + 0.0688\ (R^2 = 0.791). \quad (9)$$

2.3.8 | Verification and reliability analysis

The predicted salinity by both SVR and RFR modeling was calibrated against both the TS and VS to evaluate their performance at each test of the given conditions (e.g., rasterization type and NT), either by linear regression analysis using R^2 or by the root mean square error (RMSE) and the normalized RMSE (NRMSE), which can shed light on the goodness of fit between the prediction and measurement. Mathematically, the latter can be expressed as

$$RMSE = \left[\left(\sum_{i=1}^n (\hat{S}_i - S_i)^2 \right) / n \right]^{1/2}, \quad (10)$$

$$NRMSE = \frac{RMSE}{S_{max} - S_{min}}, \quad (11)$$

where \hat{S}_i is the i th predicted soil salinity, S_i is the i th measured salinity, and n is the sample number of the observed dataset, 24 in this case; S_{max} and S_{min} are, respectively, the maximum and minimum values of the measured salinity. NRMSE is a unitless index; the lower the value, the better the fit.

In addition, the converted salinity of the typical land use types in the study area such as alfalfa, mixed croplands, long-term fallows, and bare saline soil and built-up area were also sampled through definition of their corresponding polygons to check the reliability of prediction.

3 | RESULTS AND DISCUSSION

3.1 | Effects of rasterization procedure

3.1.1 | Effects of buffering field samples

As revealed in Tables 1 and 2, the buffering-based rasterization produced better modeling results (i.e., higher R^2) for both RFR and SVR algorithms (Table 2) than the direct rasterization (Table 1) when calibrated against the ground-truth TS and VS. This is because the direct rasterization (Figure 2a,c) resulted in irrational presentation of the training sample plots in space (small pink plots were not enclosed in the centers after rasterization), and the buffering-based rasterized pixels were able to envelop better the sampling plots, and hence more spatially representative (Figure 2b,d).

3.1.2 | Effect of rasterization cell size

Different rasterization of cell sizes led to a different performance of salinity prediction (Tables 1 and 2). As shown in Figure 2a,c, the original sample plots were distributed on the borders or close to the borders of the rasterized cells of 30, 60, and 90 m, indicating a poor representation of the samples after direct rasterization. For the buffering-based rasterization, sample plots (Figure 3b,d) were fully encompassed inside the resampled pixels, which could represent well the sample plots leading to a relevant salinity prediction, that is, generally high R^2 in Table 2. As for RFR, both circular buffers with radius of 30 and 60 m produced equally good prediction, better than that of 90 m (Tables 2 and 3). Probably in the latter case, the buffer size was too large (about 2.5 ha in area) and hence shaded the spatial variability of salinity. In case of SVR, the buffer cell with a radius of 60 m outperformed the other two cases. Overall, a 60 m of initial buffer size will be recommended for both RFR and SVR modeling.

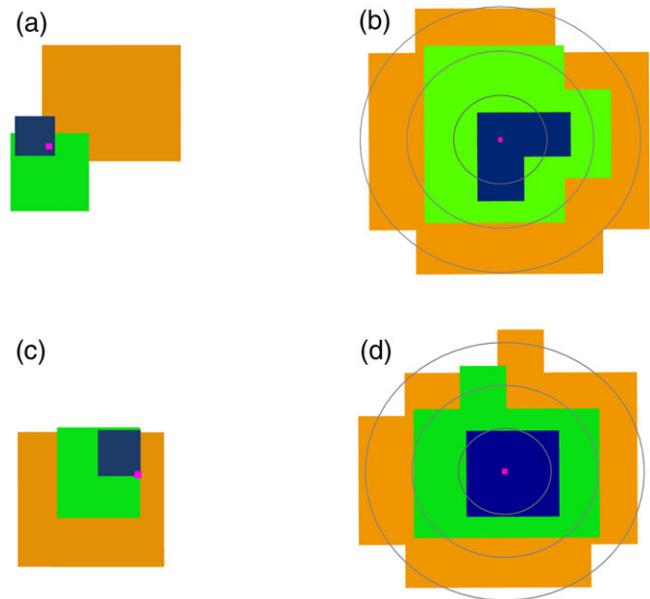


FIGURE 2 Difference between the direct rasterization (a,c) and buffering-based rasterization (b,d) of the field averaged sample plots 2 and 11 (the smallest pink squares). (a,c) With initial rasterized cells of 30 (blue), 60 (green), and 90 m (brown); (b,d) first into circular buffers of a radius of 30, 60, and 90 m to which the same colors as the former were assigned and then were finally resampled to 30-m pixels [Colour figure can be viewed at wileyonlinelibrary.com]

3.1.3 | NT with RFR

The NT affected the prediction results when applying RFR algorithm (Table 3). Despite its capacity to capture most of the features when NT was set to 100, the prediction results (R^2) were better when it was set to 300 and 500 for buffers with a radius of both 30 and 60 m, and R^2 slightly decreased when it was 1000. Hence, 300 or 500 is recommended for NT when dealing with salinity mapping in general case.

TABLE 1 Agreement (R^2) between the predicted soil salinity (EM_H) and field-measured salinity (EM_H) with direct rasterization of the field samples for training (RFR was run by setting the number of trees [NT] to 100)

Salinity prediction	RFR predicted soil salinity (EM_H)			SVR predicted soil salinity (EM_H)		
	30	60	90	30	60	90
Initial rasterized cell (m)	30	60	90	30	60	90
Resampled pixel (m)	30	30	30	30	30	30
R^2 against training set (TS; EM_H)	0.4904	0.7238	0.7922	0.6529	0.8007	0.8009
R^2 against validation set (VS; EM_H)	0.4889	0.7147	0.7812	0.6500	0.7905	0.7795

Note. RFR: random forest regression; SVR: support vector regression.

TABLE 2 Agreement (R^2) between the predicted soil salinity (EM_H) and field-measured salinity (EM_H) resulted from the buffering-based rasterization of the field samples (RFR was run by setting the number of trees [NT] to 100)

Salinity prediction	RFR predicted salinity (EM_H)			SVR predicted salinity (EM_H)		
	30	60	90	30	60	90
Buffer size (radius in m)	30	60	90	30	60	90
Initial rasterized cell (m)	10	10	10	10	10	10
Resampled pixel (m)	30	30	30	30	30	30
R^2 against TS (EM_H)	0.9206	0.9283	0.8727	0.8353	0.8606	0.7903
R^2 against VS (EM_H)	0.9285	0.9075	0.8590	0.8493	0.8888	0.8102

Note. RFR: random forest regression; SVR: support vector regression; TS: training set; VS: validation set.

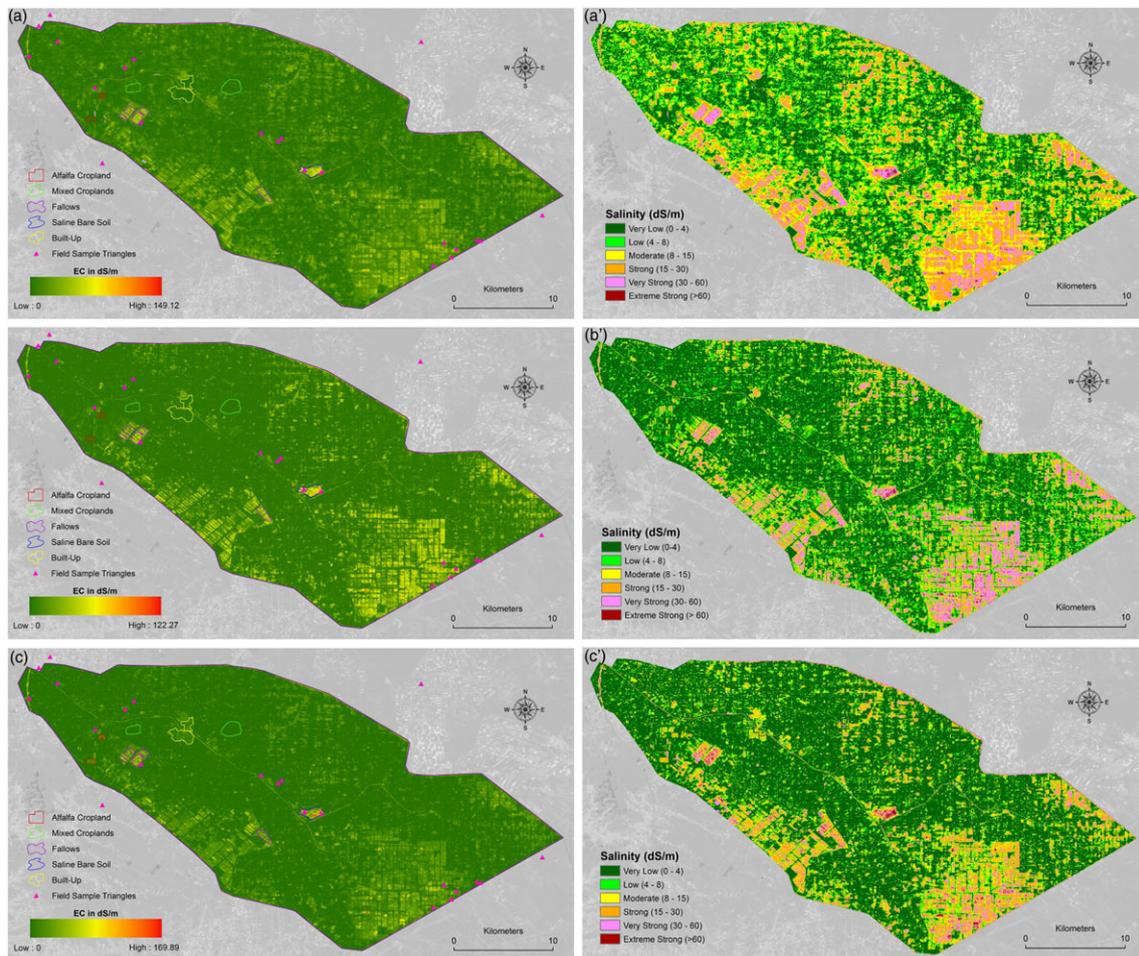


FIGURE 3 Soil salinity maps of the study area predicted by different regression algorithms and converted into EC_e ($dS\ m^{-1}$) [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 3 Agreement (R^2) between the RFR predicted salinity (EM_H) and field-measured one (EM_H) given the different numbers of trees (NT) when buffering-based rasterization was conducted

Number of trees	Rasterization and calibration	Predicted EM_H vs. measured EM_H		
		30	60	90
	Buffer size (radius in m)	30	60	90
	Initial rasterized cell (m)	10	10	10
	Resampled pixel (m)	30	30	30
100	R^2 against TS (EM_H)	0.9206	0.9283	0.8727
	R^2 against VS (EM_H)	0.9285	0.9075	0.8590
300	R^2 against TS (EM_H)	0.9325	0.9235	0.8700
	R^2 against VS (EM_H)	0.9432	0.9019	0.8546
500	R^2 against TS (EM_H)	0.9349	0.9331	0.8867
	R^2 against VS (EM_H)	0.9416	0.9141	0.8697
1000	R^2 against TS (EM_H)	0.9246	0.9189	0.8802
	R^2 against VS (EM_H)	0.9352	0.8978	0.8639

Note. RFR: random forest regression; TS: training set; VS: validation set.

3.1.4 | Prediction from EM_V

As seen in Table 4, the predictivity of soil salinity by RFR and SVR with EM_V seems slightly lower than that with EM_H (Tables 2 and 3) given the same buffering-based rasterization procedure. Table 4 also indicated that rasterization with 60-m buffering procedure delivered the best prediction for both RFR and SVR algorithms when EM_V dataset was used as TS.

3.2 | Soil salinity maps and their reliability

3.2.1 | Salinity maps

The best predicted apparent soil salinity maps by RFR on EM_H (e.g., NT = 500, buffer size = 30 m; Table 3) and by SVR on EM_V (buffer size = 60 m; Table 4), and those by MLR on EM_H were converted into EC_e ($dS\ m^{-1}$). They were presented in Figure 3 either in continuous ramp (Figure 3a–c) or classified severity grades (Figure 3a'–c'), respectively, by MLR, RFR, and SVR.

Although performing differently in different land use types, RFR estimated salinity was closer to the field-measured ones than SVR in built-up areas and alfalfa cropland in the defined polygons (Figure 3a–c and the mean values in Table 5). Theoretically, the salinity should be zero in the built-up areas and very low in the vigorously performing croplands (e.g., $<4\text{--}8\ dS\ m^{-1}$), including the salt-tolerant crops such as alfalfa. SVR seemed to have overestimated salinity in these two types of land use (Table 5). In comparison with RFR and SVR, salinity predicted by MLR is also close to the measured ones for these two land use categories.

For mixed croplands, all three algorithms predicted reasonably well salinity when compared with measured EC_e (Table 5).

Regarding the long-term fallows including the abandoned croplands, uncultivated during the past 15–20 years, the three algorithms performed equally well, $31.9\text{--}37.9\ dS\ m^{-1}$, approximate to the field-

TABLE 4 Performance of RFR and SVR in salinity prediction with different buffer size rasterization

	RFR predicted salinity (EM _V)			SVR predicted salinity (EM _V)		
	30	60	90	30	60	90
Buffer size (radius in m)	30	60	90	30	60	90
Initial rasterized cell (m)	10	10	10	10	10	10
Resampled pixel (m)	30	30	30	30	30	30
R ² against TS (EM _V)	0.8807	0.9360	0.8516	0.7848	0.8943	0.7683
R ² against VS (EM _V)	0.8860	0.8937	0.8131	0.7280	0.8525	0.7170

Note. RFR was run at NT of 100. RFR: random forest regression; SVR: support vector regression; TS: training set; VS: validation set; NT: number of trees.

TABLE 5 Predicted salinity (dS m⁻¹) by different algorithms for different land use types under the sample polygons defined in Figure 3a–c

Land use types	RFR			SVR			MLR			Mean converted ECe from EM _H readings	Mean lab-analyzed soil ECe
	Min	Mean	Max	Min	Mean	Max	Min	Mean	Max		
Alfalfa (green cropland)	1.213	4.045	11.672	0.001	16.394	32.750	0.010	2.010	8.813	3.880 (two triangles)	3.1 (one sample)
Mixed croplands (incl. newly sown)	0.341	2.946	23.089	0.001	1.142	21.875	0.045	4.958	32.162	4.216 (14 triangles)	4.0 (25 samples)
Built-up	0.703	3.186	16.909	0.000	8.377	32.280	0.002	1.491	33.085	N/A	N/A
Long-term fallows (incl. abandoned croplands)	2.619	31.996	101.578	3.984	36.405	158.900	10.030	37.967	126.351	39.515 (five triangles)	38.8 (four samples)
Saline bare soil	2.582	43.650	122.270	5.240	52.113	169.890	13.885	47.182	149.120	88.929 (three triangles)	N/A

Note. RFR: random forest regression; SVR: support vector regression; MLR: multivariate linear regression.

measured mean, 38.8–39.15 dS m⁻¹. For the saline bare soil, all algorithms predicted a salinity ranging from 43.65 to 52.11 dS m⁻¹, lower than the measured mean, 88.93 dS m⁻¹. Probably, our field sampling was not enough (only three pairs) to cover the full spectrum of the spatial variability of salinity in this land use unit.

Salinity expressed in continuous ramp (a) and severity levels (a') predicted by MLR modeling using the combined Model 2 of Wu et al. (in press), with an accuracy of 83.7% and 81.5% versus the TS and the VS, respectively; the same meaning for (b) and (b') predicted by RFR, with an accuracy of 93.5% and 94.2% versus TS and VS, respectively (Table 3); and (c) and (c') by SVR, with an accuracy of 89.4% and 85.2% versus TS and VS, respectively (Table 4). Polygons defined in Figure 3a–c were the sample areas of the main land use categories used for evaluating the reliability of the predicted salinity (Table 5).

3.2.2 | Prediction reliability

Calibration by linear regression revealed that the reliability of prediction was high as R² of the RFR and SVR prediction versus TS and VS were, respectively, 0.9349 and 0.9416 (Table 3), 0.8606 and 0.8888 (Table 2) based on EM_H, or 0.8943 and 0.8525 (Table 4) based on

EM_V. The R² of the MLR prediction was 0.8371 and 0.8135 versus TS and VS, respectively. Generally, all these regression algorithms could achieve reasonable estimation, and RFR performed best.

Table 6 presents the verification results by RMSE and NRMSE, another frequently applied indicator to evaluate the reliability. The same as revealed by the linear regression analysis, salinity prediction by RFR has the least RMSE and NRMSE, followed by MLR having less NRMSE than SVR.

3.3 | Approach assessment

To use field samples as TS for classification and regression modeling is a common procedure. Our study revealed that the buffering-based rasterization of samples, for example, with a buffer radius of 30–60 m for RFR and 60 m for SVR, is an efficient procedure to use point data as such rasterization can better preserve spatial locations and representativeness of the sample plots.

Among the tested machine learning algorithms, RFR outperformed SVR and generated maps with higher reliability. Unlike RF and SVM classification, RFR and SVR can run fast, from tens of seconds to several minutes on a normal personal computer depending on the NT for RFR and on the kernel function type for SVR. One

TABLE 6 RMSE and NRMSE of prediction by different regression algorithms

Field-measured sample sets	RFR		SVR		MLR	
	RMSE (dS m ⁻¹)	NRMSE (%)	RMSE (dS m ⁻¹)	NRMSE (%)	RMSE (dS m ⁻¹)	NRMSE (%)
TS	5.275	6.10	9.410	10.29	8.208	9.09
VS	6.793	7.69	9.651	10.52	8.280	9.19

Note. RFR: random forest regression; SVR: support vector regression; MLR: multivariate linear regression; TS: training set; VS: validation set; RMSE: root mean square error; NRMSE: normalized RMSE.

disadvantage of the machine learning algorithms is that they cannot produce intuitive models as MLR does.

Farifteh et al. (2007) and Taghizadeh-Mehrjardi et al. (2014) have already predicted soil salinity using machine learning algorithms. The tests of Farifteh et al. (2007) were carried out in very small areas (about 5–6 ha) in the Netherlands and Hungary. Whether their approaches were applicable to larger areas was not clear. We tested partial least squares regression in our research site, and the accuracy of the resulted maps was low, only 69.5–72.3% ($R^2 = 0.69–0.72$) corresponding to TS of 30 and 60 m of buffering size, much lower than our machine learning results ($R^2 = 0.85–0.94$).

The study conducted in a remote site in Iran by Taghizadeh-Mehrjardi et al. (2014) seemed comparable with ours. But they used EM38 readings to produce ECa maps by Kriging interpolation, and these maps were then input as independent variables with others for salinity prediction. Our concern lies in the uncertainty of their interpolated ECa maps because EM38 readings were limited and the ECa in most pixels was 'predicted.' In our opinion, using such uncertain ECa as inputs to predict salinity seems irrelevant. Moreover, the algorithm they used, regression tree, is only a part of the RFR and less predictively powerful than the latter (Breiman, 2001). We believe thence our approaches and results would be more robust.

4 | CONCLUSIONS

This study applied machine learning regression algorithms to soil salinity prediction and mapping using a combined optical–radar dataset and field measurements. The results showed that it was effective and practical to employ thematic biophysical indicators from both optical and radar data to achieve the objectives. The removal of vegetation impact on the radar backscattering coefficients increased substantially the predictivity of the radar data. Rasterization of the field samples with buffering radius of 60 m was the most effective procedure for creating the TSs.

Among the tested regression algorithms, RFR performed best with the highest correlation coefficients and least RMSE (5.275 and 6.793 dS m^{-1}) and NRMSE (6.10% and 7.69%) against TS and VS. The main RMSE was produced in the strongly salinized areas such as the saline bare soil, where more field samples will be needed in future to improve the prediction performance. It was also noted that MLR can predict salinity with acceptable NRMSE (<10%), and its advantage lies in the possibility to deliver intuitive models. Hence, we concluded that RFR and MLR are two good regression predictors of salinity and recommended for application elsewhere.

ACKNOWLEDGMENTS

The authors would like to thank AusAID (Australian Agency for International Development) for funding our previous research at ICARDA (International Center for Agricultural Research in the Dry Areas) in 2010–2014 (Project LWR/2009/034) and East China University of Technology for their financial support to Dr Weicheng Wu (DHTP2018001, 2018–2021). The European Space Agency (ESA) is acknowledged for their provision of the ALOS PALSAR and Landsat 5 TM data. A special gratitude will go to Dr Manzoor Qadir (UNU-

INWEH), and Dr Theib Oweis (ICARDA) for their cooperation in the early stage of the works.

ORCID

Weicheng Wu  <http://orcid.org/0000-0001-6362-9840>

Ayad M. Fadhil Al-Quraishi  <http://orcid.org/0000-0001-7732-129X>

REFERENCES

- Abdel-Rahman, E. M., Ahmed, F. B., & Ismail, R. (2013). Random forest regression and spectral band selection for estimating sugarcane leaf nitrogen concentration using EO-1 Hyperion hyperspectral data. *International Journal of Remote Sensing*, 34(2), 712–728. <https://doi.org/10.1080/01431161.2012.713142>
- Allbed, A., & Kumar, L. (2013). Soil salinity mapping and monitoring in arid and semi-arid regions using remote sensing technology: A review. *Advances in Remote Sensing*, 2, 373–385. <https://doi.org/10.4236/ars.2013.24040>
- Attema, E. P. W., & Ulaby, F. T. (1978). Vegetation modeled as a water cloud. *Radio Science*, 13, 357–364. <https://doi.org/10.1029/RS013i002p00357>
- Bannari, A., El-Battay, A., Bannari, R., & Rhinane, H. (2018). Sentinel-MSI VNIR and SWIR bands sensitivity analysis for soil salinity discrimination in an arid landscape. *Remote Sensing*, 10, 855. <https://doi.org/10.3390/rs10060855>
- Belgiu, M., & Dragut, L. (2016). Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing*, 114, 24–31. <https://doi.org/10.1016/j.isprsjprs.2016.01.011>
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32. <https://doi.org/10.1023/A:1010933404324>
- Crist, E. P., & Cicone, R. C. (1984). Application of the tasseled cap concept to simulated thematic mapper data. *Photogrammetric Engineering & Remote Sensing*, 50, 343–352.
- Dabrowska-Zielinska, K., Inoue, Y., Kowalik, W., & Gruszczynska, M. (2007). Inferring the effect of plant and soil variables on C- and L-band SAR backscatter over agricultural fields, based on model analysis. *Advances in Space Research*, 39, 139–148. <https://doi.org/10.1016/j.asr.2006.02.032>
- Dwivedi, R. S., & Rao, B. R. M. (1992). The selection of the best possible Landsat TM band combination for delineating salt-affected soils. *International Journal of Remote Sensing*, 13(11), 2051–2058. <https://doi.org/10.1080/01431169208904252>
- Farifteh, J., Farshad, A., & George, R. J. (2006). Assessing salt-affected soils using remote sensing, solute modelling, and geophysics. *Geoderma*, 130(3–4), 191–206. <https://doi.org/10.1016/j.geoderma.2005.02.003>
- Farifteh, J., van der Meer, F., Atzberger, C., & Carranza, E. (2007). Quantitative analysis of salt affected soil reflectance spectra: A comparison of two adaptive methods (PLSR and ANN). *Remote Sensing of Environment*, 110, 59–78. <https://doi.org/10.1016/j.rse.2007.02.005>
- 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 Environment*, 65, 644–667. <https://doi.org/10.1016/j.jaridenv.2005.08.005>
- Gong, H., Shao, Y., Brisco, B., Hu, Q., & Tian, W. (2013). Modeling the dielectric behavior of saline soil at microwave frequencies. *Canadian Journal of Remote Sensing*, 39(1), 1–10. <https://doi.org/10.5589/m13-004>
- Gorji, T., Tanik, A., & Sertel, E. (2015). Soil salinity prediction, monitoring and mapping using modern technologies. *Procedia Earth and Planetary Science*, 15, 507–512. <https://doi.org/10.1016/j.proeps.2015.08.062>
- Hardisky, M. A., Klemas, V., & Smart, R. M. (1983). The influences of soil salinity, growth form, and leaf moisture on the spectral reflectance of

- Spartina alterniflora* canopies. *Photogrammetric Engineering and Remote Sensing*, 49, 77–83.
- Huang, C., Davis, L. S., & Townshend, J. R. G. (2002). An assessment of support vector machines for land cover classification. *International Journal of Remote Sensing*, 23, 725–749. <https://doi.org/10.1080/01431160110040323>
- Ivushkin, K., Bartholomeus, H., Bregt, A. K., & Pulatov, A. (2017). Satellite thermography for soil salinity assessment of cropped areas in Uzbekistan. *Land Degradation & Development*, 28, 870–877. <https://doi.org/10.1002/ldr.2670>
- Jackson, T. J., Chen, D., Cosh, M., Li, F., Anderson, M., Walthall, C., ... Hunt, E. R. (2004). Vegetation water content mapping using Landsat data derived normalized difference water index for corn and soybeans. *Remote Sensing of Environment*, 92, 475–482. <https://doi.org/10.1016/j.rse.2003.10.021>
- Kavzoglu, T., & Colkesen, I. (2009). A kernel functions analysis for support vector machines for land cover classification. *International Journal of Applied Earth Observation and Geoinformation*, 11, 352–359. <https://doi.org/10.1016/j.jag.2009.06.002>
- Kumar, K., Prasad, K. S. H., & Arora, M. K. (2012). Estimation of water cloud model vegetation parameters using a genetic algorithm. *Hydrological Sciences Journal*, 57(4), 776–789. <https://doi.org/10.1080/02626667.2012.678583>
- Lee, J.-S. (1980). Digital image enhancement and noise filtering by use of local statistics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-2(2), 165–168. <https://doi.org/10.1109/TPAMI.1980.4766994>
- Mas, J. F., & Flores, J. J. (2008). The application of artificial neural networks to the analysis of remotely sensed data. *International Journal of Remote Sensing*, 29(3), 617–663. <https://doi.org/10.1080/01431160701352154>
- Metternicht, G. I., & Zinck, J. A. (2003). Remote sensing of soil salinity: Potentials and constraints. *Remote Sensing of Environment*, 85, 1–20. [https://doi.org/10.1016/S0034-4257\(02\)00188-8](https://doi.org/10.1016/S0034-4257(02)00188-8)
- Moran, M. S., Vidal, A., Troufleau, D., Inoue, Y., & Mitchell, T. A. (1998). Ku- and C-band SAR for discriminating agricultural crop and soil conditions. *IEEE Transactions on Geoscience and Remote Sensing*, 36(1), 265–272. <https://doi.org/10.1109/36.655335>
- Mougenot, B., Pouget, M., & Epema, G. (1993). Remote sensing of salt-affected soils. *Remote Sensing Reviews*, 7, 241–259. <https://doi.org/10.1080/02757259309532180>
- Pal, M. (2005). Random forest classifier for remote sensing classification. *International Journal of Remote Sensing*, 26(1), 217–222. <https://doi.org/10.1080/01431160412331269698>
- Perkins, T., Adler-Golden, S. M., Matthew, M. W., Berk, A., Bernstein, L. S., Lee, J., & Fox, M. J. (2012). Speed and accuracy improvements in FLAASH atmospheric correction of hyperspectral imagery. *SPIE Optical Engineering*, 51(11), 111707. <https://doi.org/10.1117/1.OE.51.11.111707>
- Qadir, M., Noble, A. D., Qureshi, A. S., Gupta, R. K., Yuldashev, T., & Karimov, A. (2009). Salt-induced land and water degradation in the Aral Sea basin: A challenge to sustainable agriculture in Central Asia. *Natural Resources Forum*, 33(2), 134–149. <https://doi.org/10.1111/j.1477-8947.2009.01217.x>
- Qadir, M., Qureshi, A. S., & Cheraghi, S. A. M. (2008). Extent and characterisation of salt-affected soils in Iran and strategies for their amelioration and management. *Land Degradation & Development*, 19(2), 214–227. <https://doi.org/10.1002/ldr.818>
- Ritter, N. D., & Hepner, G. F. (1990). Application of an artificial neural network to land-cover classification of thematic mapper imagery. *Computers & Geosciences*, 16(6), 873–880. [https://doi.org/10.1016/0098-3004\(90\)90009-I](https://doi.org/10.1016/0098-3004(90)90009-I)
- Rodriguez-Galiano, V. F., Ghimire, B., Rogan, J., Chica-Olmo, M., & Rigol-Sanchez, J. P. (2012). An assessment of the effectiveness of a random forest classifier for land-cover classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 67, 93–104. <https://doi.org/10.1016/j.isprsjprs.2011.11.002>
- Shimada, M., Isoguchi, O., Tadono, T., & Isono, K. (2009). PALSAR radiometric and geometric calibration. *IEEE Transactions on Geoscience and Remote Sensing*, 47(12), 3915–3931. <https://doi.org/10.1109/TGRS.2009.2023909>
- Sreenivas, K., Venkataratnam, L., & Rao, P. V. N. (1995). Dielectric properties of salt affected soils. *International Journal of Remote Sensing*, 16, 641–649. <https://doi.org/10.1080/01431169508954431>
- 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. <https://doi.org/10.1016/j.geoderma.2013.07.020>
- van der Linden, S., Rabe, A., Held, M., Wirth, F., Suess, S., Okujeni, A., & Hostert, P. (2014). *ImageSVM regression, manual for application: imageSVM version 3.0*. Berlin: Humboldt-Universität zu Berlin.
- Vapnik, V., Golowich, S., & Smola, A. (1997). Support vector method for function approximation, regression estimation, and signal processing. In M. C. Mozer, M. I. Jordan, & T. Petsche (Eds.), *Advances in neural information processing systems 9* (pp. 281–287). Cambridge: MIT Press.
- Vapnik, V., & Lerner, A. (1963). Pattern recognition using generalized portrait method. *Automation and Remote Control*, 24, 774–780.
- Wang, L., Zhou, X., Zhu, X., Dong, Z., & Guo, W. (2016). Estimation of biomass in wheat using random forest regression algorithm and remote sensing data. *The Crop Journal*, 4(3), 212–219. <https://doi.org/10.1016/j.cj.2016.01.008>
- Waske, B., van der Linden, S., Oldenburg, C., Jakimow, B., Rabe, A., & Hostert, P. (2012). imageRF—A user-oriented implementation for remote sensing image analysis with random forests. *Environmental Modeling & Software*, 35, 192–193. <https://doi.org/10.1016/j.envsoft.2012.01.014>
- Wilkinson, G. G. (2005). Results and implications of a study of fifteen years of satellite image classification experiments. *IEEE Transactions on Geoscience and Remote Sensing*, 43, 433–440. <https://doi.org/10.1109/TGRS.2004.837325>
- Wu, W. (2014). The generalized difference vegetation index (GDVI) for dryland characterization. *Remote Sensing*, 6, 1211–1233. <https://doi.org/10.3390/rs6021211>
- Wu, W., Al-Shafie, W. M., Mhaimeed, A. S., Ziadat, F., Nangia, V., & Payne, W. (2014). Soil salinity mapping by multiscale remote sensing in Mesopotamia, Iraq. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 7(11), 4442–4452. <https://doi.org/10.1109/JSTARS.2014.2360411>
- Wu, W., Mhaimeed, A. S., Al-Shafie, W. M., Ziadat, F., 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>
- Wu, W., Mhaimeed, A. S., Al-Shafie, W. M., & Fadhil, A. M. (in press). Using radar and optical data for soil salinity modeling and mapping in Central Iraq. In A. M. Fadhil, & A. Negm (Eds.), *Environmental remote sensing and GIS in Iraq*. Berlin: Springer.
- Wu, W., Zucca, C., Karam, F., & Liu, G. (2016). Enhancing the performance of regional land cover mapping. *International Journal of Earth Observation and Geoinformation*, 52, 422–432. <https://doi.org/10.1016/j.jag.2016.07.014>

How to cite this article: Wu W, Zucca C, Mhaimeed AS, et al. Soil salinity prediction and mapping by machine learning regression in Central Mesopotamia, Iraq. *Land Degrad Dev*. 2018;29:4005–4014. <https://doi.org/10.1002/ldr.3148>