Retrieval of soil salinity from Sentinel-2 multispectral imagery

Mohammad Mahdi Taghadosi, Mahdi Hasanlou & Kamran Eftekhar


To link to this article: https://doi.org/10.1080/22797254.2019.1571870

© 2019 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

Published online: 30 Jan 2019.

Submit your article to this journal

View Crossmark data
Retrieval of soil salinity from Sentinel-2 multispectral imagery

Mohammad Mahdi Taghadosi*†, Mahdi Hasanlou* and Kamran Eftekhari‡

**School of Surveying and Geospatial Engineering, College of Engineering, University of Tehran, Tehran, Iran; ‡Department of Soil Genesis, Classification and Cartography, Soil and Water Research Institute, Karaj, Iran

**ABSTRACT

Soil salinity is a widespread environmental hazard and the main causes of land degradation and desertification, especially in arid and semi-arid regions. The first step in finding such a solution is providing accurate information about the severity and extent of the salinity spread in affected areas; this can be done by mapping the electrical conductivity (EC) of the soil. Utilizing the potential of high-resolution satellite imagery along with remote sensing techniques is a promising method to map salinity, as it allows for large-scale monitoring and provides high accuracy and efficiency. This paper, therefore, aims at assessing soil salinity by mapping the EC of soils, using satellite imagery from the newly launched Sentinel-2 satellite as well as Landsat-8 data. A field study was carried out using those data, and various salt features were extracted that relate the EC values of field samples to satellite-derived salt features. The study used two different regression approaches MLP and SVR. Additionally, two feature selection algorithms, GA and SFS, were implemented on the data to improve model performance. The study concludes that the proposed method for modeling salinity and the mapping of soil EC can be considered an effective approach for soil salinity monitoring.

**Article history**

Received 8 March 2018
Revised 22 December 2018
Accepted 16 January 2019

**Keywords**

Soil salinity; EC mapping; Sentinel-2 satellite; regression analysis; feature selection

**Introduction**

Soil salinity is one of the most serious environmental issues, causing land degradation and desertification especially in arid and semi-arid regions (Allbed & Kumar, 2013; Asfaw, Suryabhadgavan, & Argaw, 2016; Gorji, Sertel, & Tanik, 2017). This dynamic phenomenon, which can occur through natural processes (primary salinization) or as a result of human activities (secondary salinization), is a major threat to soil productivity and agricultural land (Allbed & Kumar, 2013; Gorji et al., 2017; Scudiero, Skaggs, & Corwin, 2016). According to Koohaifkan and Stewart (2008), about 397 million hectares of lands around the world are affected by salinity, and it continues to spread at a rate of up to 2 million hectares each year. Because of the negative impact of salinization on soil fertility and agricultural production, finding ways to preserve soil quality and reclaim saline soils is essential. The first step toward preserving soil as a natural resource is providing accurate information about the spatial extent and the severity of salinity in affected areas. One way to achieve this is by mapping the electrical conductivity (EC) of the soil. Traditional methods to produce soil EC maps – by conducting field surveys and measuring the EC values of soils – are costly and time-consuming. However, modern technologies and new methods are available to help map soil EC and provide information about salt-affected areas with much greater efficiency.

Using the potential of satellite data and remote sensing techniques, it becomes possible to monitor soil salinity more efficiently and more economically (Allbed & Kumar, 2013; Garcia, Eldeiry, & Elhaddad, 2005; Morshed, Islam, & Jamil, 2016; Taghadosi, Hasanlou, & Eftekhari, 2018). Determining the spectral characteristics of saline surface soils and the spatial distribution of affected areas can be achieved by analyzing data from newly launched satellites with good spatial and spectral resolution, which helps to map salinity over large scales and with high accuracy. Among the different satellite sensors used to map salinity and produce EC maps, multispectral satellite imagery has been widely studied in previous research and has been found to be a very promising tool for this task. Elnaggar and Noller (2009), in a case study in Malheur County, Oregon, used Landsat TM imagery to map salinity. Their results revealed that a high salt content in bare soils can be identified by Landsat bands 1–4 because of the high spectral reflectance of salt in this range of the electromagnetic spectrum. Another study, conducted by Katwatin and Kotrapat (2005), used three sources of ancillary data...
In recent years, various salinity indicators have been developed to detect salt-affected areas from satellite images (Alexakis, Daliakopoulos, Panagea, & Tsanis, 2018; Allbed & Kumar, 2013; Elhag, 2016), mostly based on the spectral behaviors of saline soils in different bands of satellite imagery. Vegetation indices (VI), which are indirect salinity indicators, have also been used to assess the presence of salt in the soils by its adverse effects on crop growth and plant stress. On the other hand, salinity indices, which are direct salinity indicators, highlight the spectral reflectance of salt crusts on the soil surface, especially in the visible and near-infrared (NIR) range of the electromagnetic spectrum (Gorji et al., 2017; Madani, 2005; Vermeulen & Niekerk, 2016; Wu et al., 2014). Bannari, Guedon, El-Harti, Cherkaoui, and El-Ghmari (2008) assessed the negative impact of soil salinity on plant growth, which was done by measuring salt stress using the normalized difference vegetation index (NDVI). Lobell et al. (2010) used multi-temporal MODIS data to determine the efficiency of the enhanced vegetation index (EVI) and the NDVI for assessing soil salinity. The results indicated that the EVI is more reliable for detecting salt regions than the NDVI. Khan, Rastoskuev, Sato, and Shiozawa (2005) assessed land degradation in Pakistan, using a LISS-II sensor of the IRS-1B satellite. Three salinity indices, the brightness index (BI), normalized difference salinity index (NDSI), and salinity index (SI), were proposed. Among these indices, the NDSI was the most accurate index for identifying different salt classes.

In addition to spectral indices that were obtained from a combination of image bands, different image transformation methods can be used to extract useful features for soil salinity assessment from satellite data. Furthermore, the effectiveness of using thermal bands for salinity detection has already been proven in previous studies. Abbas and Khan (2007) investigated the use of principal component analysis (PCA) along with spectral indices to monitor soil salinity. Their results showed that using PCA transformation techniques and also salinity indices, is a promising approach for predicting salinity from satellite images. Alavipanah and Goossens (2001) assessed the potential of the thermal band Landsat TM for soil salinity monitoring. The results of this study revealed that the addition of the TM thermal band as complementary information to the TM reflective bands contained some useful information that may play an important role in soil salinity studies. However, a large number of these indicators and the wide use of them in previous salinity studies lead to a degree of uncertainty and confusion in terms of their efficiency and their advantages when compared to each other. This paper aims at evaluating a wide range of these indicators in order to determine the best indices, i.e. those which can most effectively detect salt features from satellite images.

Monitoring and mapping soil salinity can be performed by relating ground-truth measurements of soil salinity and the corresponding pixel values extracted from satellite data. In recent years, a wide range of regression methods have been developed for modeling soil salinity and estimating EC values. The performance of these methods varies with the study area and the applied regression techniques (Allbed, Kumar, & Sinha, 2014; Eldeiry Ahmed & Garcia Luis, 2010; Farifteh, Van der Meer, Atzberger, & Carranza, 2007; Shahabi, Jafarzadeh, Neyshabouri, Ghorbani, & Kamran, 2017; Vermeulen & Niekerk, 2017). Gorji et al. (2017) performed linear and exponential regression analysis to relate satellite-derived salinity indices to 28 field samples, generating salinity maps for years of study. The best results were achieved with $R^2$ values of 0.93 and 0.83 for exponential and linear regression analysis, respectively. Wang, Wang, and Liu (2007) applied autocorrelation analysis, ordinary least square, and spatial regression models to explore the spatial variation of soil salinity in the Yellow River Delta. The results of this study revealed that the spatial regression model had significant ($p\_value <0.01$) estimations and was a “good fit” for the study site. Another study conducted by Qu, Jiao, and Lin (2008) established partial least squares regression (PLSR) as a method to retrieve soil salinity using hyperspectral data. The results indicated that the calibrated PLSR model could be used as a tool to retrieve soil salinity with accurate results.

Based on the results of previous research, this paper investigates the monitoring of soil salinity and the mapping of soil EC using imagery from the newly launched Sentinel-2 satellite. The Sentinel-2 satellite provides high-resolution optical images, with a resolution of 10 m, and global coverage of Earth’s land surface every 5 days. This makes it an effective tool for environmental monitoring and soil degradation management. In addition, Sentinel-2 provides spectral bands in a wide range of the electromagnetic spectrum – including visible, NIR, short wave infrared, and four red edge bands – which are freely available data that can be very effective for salinity monitoring. The original Sentinel-2 image and its derived features (e.g. salinity and VI, or transformation features) were used as a remote sensing data source to estimate salinity. At the same time, a field study was also conducted to measure salinity at the study site, providing a source of ground-truth data. Based on the two sources of data, regression analysis was performed to relate satellite remote sensing data and ground-measured salinity. Furthermore, optimization processes were done for selecting the best salinity

(topography, geology, and underground water quality) to map soil salinity using Landsat-7 ETM+. The results of this study indicated the possibility of mapping salinity using Landsat-7 ETM+ data bands 4, 5, and 7 in combination with the same three types of ancillary data, showing an overall accuracy of 83.6%.
indicators, and to improve the accuracy of the developed models. This methodology will now be discussed in more detail.

Materials and methods

Study area

Kuh Sefid is a village in the Qomrud Rural District, in the Central District of Qom County, Qom Province, Iran. This area, which has an arid climate with high evapotranspiration and little rainfall, is located adjacent to the salt Lake Qom. Severe salinization has occurred in this region, mainly due to overexploitation of groundwater and inappropriate irrigation practices. The salinity of soils in the study site varies between relatively high and highly saline, and in some areas the salinity is extreme. The soil is mainly composed of bare soils and sparse vegetation, and wheat and barley produced in few parts. The soil texture ranged from silt loam to silty clay loam, and the color of the surface soil varies from yellowish brown to dark brown (Fallahi, Banaei, & Eskandarzadeh, 1983). Figure 1 shows Qom County and the areas around the study site. Description of different soil salinity classes based on United States Salinity Laboratory, their EC range, and the used color scale for each class are also shown in Table 1.

Soil sampling and field data

The field study was conducted on 4 March 2017, and 58 samples were randomly collected from the soil surface. Soil samples were then analyzed in the laboratory and the EC of samples, in units of decisiemens per meter (dS/m), was measured. For this purpose, the method of measuring the EC of the soil-water extract, with a fixed soil solution ratio of 1:1, proposed by Richards (1954) was used. The soil samples were air-dried and passed through a 2 mm sieve.
in the laboratory. Then, 100 ml distilled water was added to each sample bottle and shook well. After a period of 24 h, the conductivity of solution was measured by conductivity-meter (WTW Cond 330i) at 25°C. The measured EC range in the study site was between 0.25 dS/m to around 44 dS/m. The spatial location of soil samples and their corresponding measured EC (dS/m) were stored to be used as ground-truth data. The field observation images and distribution of the soil samples are shown in Figure 2.

**Remote sensing data**

The potential of multispectral satellite images (such as ASTER or Landsat MSS/TM/ETM+) for soil salinity monitoring has been widely studied in recent years (Katawatin & Kotrapat, 2005; Mehrjardi, Mahmoodi, Taze, & Sahebjalal, 2008). Multispectral remote sensing data can be effectively used for large-scale salinity assessment because of its wide area coverage of the Earth’s surface, easy access, and relatively good spatial and spectral resolution of the images. However, using images with higher spectral and spatial resolution to achieve better results has always been emphasized.

This study used data from the Sentinel-2A optical imaging satellite, in 13 spectral bands and with 10 m spatial resolution. The data were acquired on 4 March 2017, alongside field observation data (Table 2) which are provided for salinity estimation. Due to a lack of

---

**Table 1. Description of soil salinity classes and their corresponding EC range based on Richards (1954), “soil salinity,” 2018.**

<table>
<thead>
<tr>
<th>Soil salinity classes</th>
<th>EC range</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>S4: Extremely saline</td>
<td>&gt; 16 dS/m</td>
<td>Only highly salt tolerant plants survive and the community will be dominated by 2 or 3 species. Extensive bare saline areas occur with salt stains or crystals evident (on some soils a dark organic stain may be visible). Top soil may be flowery or puffy with some plants surviving on small pedestals.</td>
</tr>
<tr>
<td>S3: Highly saline</td>
<td>8–16 dS/m</td>
<td>Salt tolerant species like Sea Barley-grass may dominate large areas and only salt tolerant plants remain unaffected. Large, bare saline areas may occur showing salt stains or crystals. At the upper end of the range, halophytic plants may dominate the plant community.</td>
</tr>
<tr>
<td>S2: Moderately saline</td>
<td>4–8 dS/m</td>
<td>Salt tolerant species begin to dominate the vegetation community and all salt sensitive plants are markedly affected by soil salinity levels. Small bare areas up to 1 m² may be present and salt stain/crystals may sometimes be visible on bare soil at the upper end of the range.</td>
</tr>
<tr>
<td>S1: Slightly saline</td>
<td>2–4 dS/m</td>
<td>Salt tolerant species such as Sea Barley-grass are often abundant. Salt sensitive plants in general show a reduction in number. There are no bare saline patches and no salt stain/crystals are evident on bare ground.</td>
</tr>
<tr>
<td>S0: Non saline</td>
<td>&lt;2 dS/m</td>
<td>No vegetation appears affected by salinity and a wide range of plants present.</td>
</tr>
</tbody>
</table>
a thermal band in Sentinel-2 imagery, the Landsat-8 thermal infrared sensor (TIRS) was utilized to evaluate the efficiency of thermal features in soil salinity monitoring. The Landsat images are dated 3 March 2017, covering the area of the study site. In addition, a digital elevation model (DEM) of the study area was freely obtained using 1-arc second shuttle radar topography mission data, to be used as a height feature in the analysis. It should be noted that since optical imagery measures the spectral reflectance of object’s surface, salinity of soil surface will be assessed in this study.

**Pre-processing of the satellite images**

Because of the availability of Sentinel-2 data in the Level-1C (L1C) processing level (top-of-atmosphere (TOA) reflectance in cartographic geometry), these images were acquired for salinity assessment. These products should then be converted into a Level-2A (L2A) ortho-image bottom-of-atmosphere corrected reflectance product suitable for regression analysis. Pre-processing of L1C products was performed by scene classification and atmospheric correction, which gives L2A products (Suheit & Hoersch, 2015). This was applied by Sen2Cor, a processor for Sentinel-2 L2A product generation and formatting developed by Mueller-Wilm et al. (2016).

Pre-processing of the Landsat-8 thermal bands was also conducted by converting digital number (DN) values to at-sensor radiance, and then spectral radiance to TOA brightness temperature. Determining the land surface temperature (LST) from TOA brightness temperature is also considered, and will be discussed in the next section.

**Extraction of the satellite features**

**Sentinel-2 bands**

To be able to assess soil surface reflectance in a wide range of the electromagnetic spectrum (as is provided by the Sentinel-2 data), and to be able to detect the affected regions from the spectral behavior of saline soils, various spectral bands were selected for determining the relationship between EC values of soil samples, and corresponding pixel values in the imagery. Out of 13 spectral bands of Sentinel-2, coastal aerosol band 1 utilized for imaging shallow water and tracking fine particles like dust and smoke and is not to be used together with other multispectral bands. Band 9 (water vapor) and band 10 (Cirrus) are also used for atmospheric correction and detection of high altitude clouds, respectively, making them inappropriate for model building. Thus, these three bands were excluded from the analysis.

**Salinity and VI**

Since the study site is located largely on bare ground where vegetation is sparse or non-existent, it is more effective to use SI, which highlights the surface reflectance of salt-affected regions. Using VI, which are commonly utilized to highlight vegetated areas, can be more practical in trend analysis of soil salinity and multi-year studies, because they can help track salinity changes in vegetated lands by their adverse effects on plant growth over time (Moreira, Teixeira, & Galvão, 2015).

However, in order to evaluate the correlation between EC values of soil samples and the corresponding pixel values derived from salinity and VI, several SI and VI were selected as predictor variables in our regression analysis. These indices showed the best performance for salinity detection in previous studies (Allbed & Kumar, 2013; Vermeulen & Niekerk, 2016; Weng, Gong, & Zhu, 2010).

**Land surface temperature (LST)**

To investigate the possibility of using soil surface temperature for salinity detection, the algorithm for retrieving LST using Landsat-8 satellite imagery, proposed by Avdan and Jovanovska (2016), was applied. Among two thermal bands of Landsat-8 satellite, band 10 is more suitable in case of calculating the actual evapotranspiration. Also, USGS recommends not using TIRS band 11 due to its larger calibration uncertainty. Thus, the thermal band 10 of the Landsat-8 imagery was used to retrieve the LST for each pixel in the scene. The following equations (Equations (1–3)) were used for LST \( T_s \) conversion:

\[
L_\lambda = M_L \times Q_{cal} + A_L - O_i \tag{1}
\]

\[
BT = \frac{K_2}{\ln\left(\frac{K_1}{L_\lambda}\right) + 1} - 273.15 \tag{2}
\]

\[
T_s = \frac{BT}{\{1 + (\lambda BT / \rho \ln \epsilon_s)\}} \tag{3}
\]
where $L_1$ is the TOA spectral radiance, $M_k$ is the band-specific rescaling gain factor, $Q_{cal}$ is the quantized calibrated pixel value (DN band 10), $A_k$ is the band-specific rescaling bias factor, $O_i$ is the correction for TIR band 10, $BT$ is brightness temperature in Celsius, $K_1$ and $K_2$ are calibration constants from the metadata. $\lambda$ is the wavelength of emitted radiance, $\rho \approx 1.438 \times 10^{-2} mK$, and $e_1$ is the emissivity. For more information see Avdan and Jovanovska (2016) and Chander, Markham, and Helder (2009).

Transformation-based features
In image processing and machine learning, feature transformation begins by transforming measured datasets in order to produce derived datasets (features) which should be non-redundant, informative, and facilitating subsequent learning methods (Richards, 2013). In this study, we used the most common methods, PCA and independent component analysis (ICA). These methods were shown to have a high performance in related studies (Cao, Chua, Chong, Lee, & Gu, 2003; Takiguchi & Ariki, 2006), and were used to evaluate the potential of feature extraction methods for soil salinity map generation.

PCA transformation. PCA is one of the most efficient and effective transformation techniques that converts data linearly into a low-dimensional subspace and computes new orthogonal directions in which the data have maximal variance (Ranchordas, Pereira, Araújo, & Tavares, 2010). This method, which was first proposed by Pearson (1901), provides essential information on the data and allows for better analysis and interpretation by reducing the dimensionality of a dataset. Since multispectral images are often highly correlated, using PCA can be an effective tool to produce uncorrelated bands containing the most important information of the scene. Therefore, PC transformation of Sentinel-2 bands was performed, and the first four bands with the most variance were selected for regression analysis (PC1-PC4).

ICA transformation. ICA is a statistical technique that reduces higher order statistical dependencies by decomposing a complex dataset into independent sub-parts (Hasanlou & Samadzadegan, 2012). This method initially developed to solve the blind source separation problem, can serve as an effective feature extraction method, which helps improve the performance of various classifiers, such as support vector machine (SVM), ANN, and so on. (Lee & Batzoglou, 2003). This transformation was also performed on the Sentinel-2 bands, in order to evaluate the usefulness of extracted features in regression analysis.

Regression analysis
Regression analysis is commonly used to estimate the relationship between a dependent and one or more independent variables. In recent years, different regression techniques have been developed in a wide range of applications, which can be used for prediction and model construction (Fan, Liu, Tao, & Weng, 2015). This study proposes two main regression methods for soil salinity assessment and for relating EC values of field samples to salt features extracted from satellite data: (1) multiple linear regression (MLR) techniques; and (2) kernel-based regression techniques. In the first approach, it is assumed that the relationship between predictor and dependent variables is approximately linear. A best-fit line is drawn to predict EC values from predictor variables; this is called the regression line. However, multiple linear methods cannot always produce accurate and reliable results, since the relationship between variables may not be linear. For such cases, kernel-based techniques, which map the input data into a high-dimensional feature space via nonlinear functions, can be more efficient for regression analysis. Previous studies have also revealed the effectiveness of using kernel bases, especially SVR, to solve regression problems. Based on this, SVR with different kernel functions was used as our second approach for salinity assessment. The performance of the proposed methods were tested, and the accuracy of the constructed models was estimated using root mean square error (RMSE), the coefficient of determination ($R^2$), and normalized root mean square error (NRMSE). The constructed models were then applied to the entire image, in order to predict EC values for each pixel in the scene.

Kernel-based method
The SVM, first proposed by Vapnik, Golowich, and Smola (1996), is a learning system based on the computation of a linear regression function in a high-dimensional feature space (Basak, Pal, Ch, & Patranabis, 2007). To incorporate classification purposes using the SVM technique, most researchers use support vector classification; for regression applications of the SVM technique, it is common to use SVR. Complications when incorporating these techniques arise from the estimation and selection of suitable SVR parameters $(C, \varepsilon)$, and from the utilized kernel parameter(s). To model
our salinity map, we used the \( \varepsilon \)-SVR model, which results in the creation of a suitability map. Kernel functions make it possible to perform linear separation by transforming the data into a new, higher feature space (Basak et al., 2007; Smola & Schölkopf, 2004). Different kernel functions can be used for model building: linear (LN), polynomial (PL), and (Gaussian) radial base function (RBF) kernels. For each kernel function, some parameters should be estimated in order to achieve better model fit, with high accuracy (Table 3).

In this study, we utilized the LIBSVM package in MATLAB, which is a library for SVM and SVR implementation (Chang & Lin, 2011). Following is the structure of \( \varepsilon \)-SVR and the LN, PL, and RBF kernels as implemented in LIBSVM32. Consider a set of training points \( \{(x_1, z_1), \ldots, (x_l, z_l)\} \), where \( x_i \in \mathbb{R}^n \) is a feature vector, and \( z_i \in \mathbb{R}^1 \) is the target output. Under the given parameters \( C > 0 \) and \( \varepsilon > 0 \), the standard form of SVR is (Equation (4)) (Vapnik, 2000)

\[
\begin{align*}
\min_{\omega, b, \xi, \xi^*} & \quad \frac{1}{2} \omega^T \omega + C \sum_{i=1}^{l} \xi_i + C \sum_{i=1}^{l} \xi_i^* \\
\text{subject to} & \quad \omega^T \varphi(x_i) + b - z_i \leq \varepsilon + \xi_i, \\
& \quad z_i - \omega^T \varphi(x_i) - b \leq \varepsilon + \xi_i^*, \\
& \quad \xi_i, \xi_i^* \geq 0, \quad i = 1, \ldots, l.
\end{align*}
\]

where \( \varphi(x_i) \) maps \( x_i \) into a higher dimensional space, \( \xi_i, \xi_i^* \) are slack variables to cope with otherwise infeasible constraints of the optimization problem, and \( C > 0 \) determines the trade-off between the flatness and the amount up to which deviations larger than \( \varepsilon \) are tolerated. Also, \( K(x_i, x_j) = \varphi(x_i)^T \varphi(x_j) \) and \( K(x_i, x_j) \) are the kernel function.

**Multiple linear method**

Since our modeling has more than one predictor variable, MLR should be used for analysis. MLR tries to construct a model between two or more explanatory variables and a response variable by fitting a linear equation to the observed data (Montgomery, Peck, & Vining, 2012). MLR has the following form (Equation (5)):

\[
y_i = a_0 + a_1 x_{1,i} + a_2 x_{2,i} + a_3 x_{3,i} + \ldots + a_k x_{k,i} + e_i
\]

where \( y_i \) is a dependent variable, which is estimated by the model and corresponds to the EC values. \( x_{1,i}, x_{2,i}, \ldots, x_{k,i} \) are the K predictor variables and correspond to satellite bands or feature indices. \( a_1, a_2, \ldots, a_k \) are unknown coefficients, which are determined in the analysis, and \( e_i \) is the error term for each regression point.

In order to make comparison possible between the results obtained by the SVR technique and the MLR approach, the same train and test datasets were used for model building and validation.

**Optimization procedures**

To improve the accuracy of the constructed models and selecting the best explanatory variables, optimization processes were considered in two main steps: First, the parameter selection of the kernel-based method, which was performed based on all explanatory variables. Second, performing feature selection (FS) algorithms to select the best satellite features and reducing the dimensionality. Both steps improve the model accuracy by minimizing RMSE and maximizing \( R^2 \) values.

**Parameter selection of SVR**

Various methods have been developed for tuning the SVR hyperparameters. In this study, we used the grid search method, which is one of the most common and reliable techniques for model selection (Ma, Zhang, & Wang, 2015). The range of selective parameters between \( (2^{-15}, 2^{15}) \) were chosen in order to search for the best setting over the parameter space.

**FS algorithms**

FS methods are commonly used to improve the performance and the accuracy of constructed models by selecting relevant features and reducing the amount of data, which also makes the model more stable (Stańczyk & Jain, 2014). Different optimization algorithms can be applied when performing FS; in our case, to help select the best bands that properly characterize the salt features of the study site. In this paper, we implemented (1) genetic algorithm (GA) and (2) sequential feature selection (SFS) methods on satellite data, and salinity models were reconstructed based on the selected features.

**Sequential feature selection (SFS).** In this method, the algorithm searches for a subset of predictor variables by adding or removing features from a candidate subset while evaluating the criterion, which is done by minimizing RMSE. This algorithm can be applied in two ways. The first is SFS, in which features are added sequentially until the

<table>
<thead>
<tr>
<th>Kernel function</th>
<th>Formula</th>
<th>Selective parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear</td>
<td>( x^2 + y^2 )</td>
<td>( C, \varepsilon )</td>
</tr>
<tr>
<td>Polynomial</td>
<td>( (yx^2 + r)^d )</td>
<td>( C, y, d, r, \varepsilon )</td>
</tr>
<tr>
<td>Radial base function (Gaussian)</td>
<td>( \exp(-\gamma</td>
<td></td>
</tr>
</tbody>
</table>

**Table 3.** Kernel functions and their selective parameters of SVR.
highest accuracy is achieved. The second way is sequential backward selection (SBS), in which features are sequentially removed from a full candidate set until further removal of features leads to a loss of model accuracy (De Silva & Leong, 2015).

**Genetic algorithm (GA).** The GA is an effective optimization approach based on a natural selection process which can be considered a powerful technique for FS (Lei, Peng, & Yang, 2013). In this paper, GA was used to improve the salinity models by selecting the best features and reducing data redundancy.

While the inputs of regression techniques (MLR and SVR) have different features, a binary FS was used as the encoding strategy of GA. As a result, $n$ bits of the binary chromosome ($n$ input feature sets) were found. If the selection is filled with zeroes, it means that the associated feature must be ignored. Each individual has a size of $n$ features (genes), with each gene representing bit strings of zeroes and ones for coding. Since $n$ features form one combination, the chromosome is arranged as comprised of $n$ individual feature sequence numbers, arranged in a serial mode (Figure 3).

**Accuracy assessment**

To evaluate the efficiency of the constructed models, three criteria were chosen: RMSE, the coefficient of determination ($R^2$), and NRMSE. The precision and the accuracy of the constructed models were then compared based on these criteria (Equations (6–8)):

$$RMSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$  \hspace{1cm} (6)

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$  \hspace{1cm} (7)

$$NRMSE = \frac{RMSE}{\bar{y}}$$  \hspace{1cm} (8)

where $\hat{y}_i$ is a vector of predicted dependent variables with $n$ data points, $y_i$ is the vector of observed values of the variable being predicted, and $\bar{y}$ is the mean of the observed dependent variables. Flowchart of the proposed methodology is shown in Figure 4.

**Results and discussion**

**Satellite-derived features**

Based on Table 4, 35 features in different categories were extracted from the satellite data to be used in the analysis. Figure 5 illustrates some of these features, which are derived from Sentinel-2 bands, SI/VI indices, and transformation features.

**Modelling soil EC**

Considering the spatial location of ground-truth data at the study site, the corresponding pixel values in the satellite data were distinguished and extracted for analysis. For each soil sample, 35 features were
considered, and a matrix including all ground-truth measurements and satellite features was created to be used for model building. The data matrix was then divided into two parts: 70% for training and 30% for testing, i.e. for model development and model validation, respectively. However, due to a large number of extracted features as predictor variables, the initial model with 35 distinct features was unable to predict EC values, and some instability was observed in both regression techniques. In the MLR method, the $R^2$ value was less than 0.60 and RMSE was greater than 10. In the case of SVR, although the $R^2$ values were reliable in most of the kernel functions (greater than 0.70), the obtained RMSE were too large to be acceptable (more than 20). We also compared the estimated EC values with ground-truth measurements to evaluate model consistency, and the results were unsatisfactory. Our solution to this was to select fewer predictor variables, choosing those with the most efficiency.

In order to determine which features should be excluded from the analysis, 2-D scatterplots were drawn between predictor variables and ground-truth data to evaluate the probable relationship between them, and thus help decide which features should be removed (Figure 6).

The obtained results are as follows:

- For VI, there was not any apparent relationship between explanatory variables and measured EC; this was to be expected since most of the soil

---

**Table 4.** Soil salinity features that are extracted from Sentinel-2 satellite data.

<table>
<thead>
<tr>
<th>Sentinel band</th>
<th>Wavelength (µm)</th>
<th>Index</th>
<th>Formula</th>
<th>Image transformation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blue</td>
<td>0.490</td>
<td>NDVI</td>
<td>$\frac{\text{NIR} - \text{Red}}{\text{NIR} + \text{Red}}$</td>
<td>PC 1 Sentinel-2</td>
</tr>
<tr>
<td>Green</td>
<td>0.560</td>
<td>NDSI</td>
<td>$\frac{\text{R} - \text{NIR}}{\text{NIR} + \text{R}}$</td>
<td>PC 2 Sentinel-2</td>
</tr>
<tr>
<td>Red Edge 1</td>
<td>0.705</td>
<td>EVI</td>
<td>$\frac{\text{2.5(NIR - Red)}}{\text{(NIR + Red) - 0.165(NIR + Red) + 0.5}}$</td>
<td>PC 3 Sentinel-2</td>
</tr>
<tr>
<td>Red Edge 2</td>
<td>0.740</td>
<td>GDOI</td>
<td>$\frac{\text{NIR} - \text{R}}{\text{NIR} + \text{R}}$</td>
<td>PC 4 Sentinel-2</td>
</tr>
<tr>
<td>Red Edge 3</td>
<td>0.783</td>
<td>SI-1</td>
<td>$\frac{\text{Red + NIR}}{\text{Red + NIR} + \text{Green}}$</td>
<td>IC 1 Sentinel-2</td>
</tr>
<tr>
<td>NIR</td>
<td>0.842</td>
<td>SI-2</td>
<td>$\frac{\text{Blue - Red}}{\text{Green}}$</td>
<td>IC 2 Sentinel-2</td>
</tr>
<tr>
<td>Red Edge 4</td>
<td>0.865</td>
<td>SI-3</td>
<td>$\frac{\text{Blue - Red}}{\text{NIR}}$</td>
<td>IC 3 Sentinel-2</td>
</tr>
<tr>
<td>SWIR 1</td>
<td>1.610</td>
<td>RVI</td>
<td>$\frac{\text{NIR}}{\text{Red}}$</td>
<td>IC 4 Sentinel-2</td>
</tr>
<tr>
<td>SWIR 2</td>
<td>2.190</td>
<td>SI-7</td>
<td>$\frac{\text{Blue - Red}}{\text{NIR}}$</td>
<td>LST Landsat-8</td>
</tr>
</tbody>
</table>

---

**Figure 4.** Flowchart of the proposed methodology for mapping soil EC.
samples were from bare ground areas, and VI (as indirect indicators) were thus unable to detect salt presence by its adverse effect on plants. Moreover, the negative effects of salinity on plant growth can be seen well in multi-year studies because salinity affects vegetated lands over time.

- Assessing the maps for the study site retrieved from LST and DEM reveals that surface temperature and height values are very close to each other, which makes it difficult to relate EC values to these features. Using LST and height features may be useful for assessing salinity in areas with more variable climate, which helps to characterize salinity by changes in meteorological variables.
- For salinity indices and transformation features (ICA and PCA), although some heterogeneity can be observed at some points, there was a significant relationship between the variables, which can be useful in analysis. However, we decided to remove IC3 and IC4, and also PC3 and PC4, since they contained less information and were noisy in some parts.

Given the above problems, Sentinel-2 bands, salinity indices, and transformation features were selected to construct models based on reliable

Figure 5. Representation of some satellite-derived features in the study site. (a) Blue band, (b) red edge 3 band, (c) SWIR 1 band, (d) NDVI, (e) CRSI, (f) SI-3, (g) SI-6, (h) PC1, and (i) IC2.
Table 5 shows the results of the regression analysis, based on 22 predictor variables. Selected parameters of the kernel-based method are also shown for each kernel function, as a result of using the grid search method.

**Optimization and model reconstruction**

Parameter selection of kernel-based method was first performed to create models with the most accuracy, based on all predictor variables, as shown in Table 5.

Table 5. Results of regression analysis based on all explanatory variables.

<table>
<thead>
<tr>
<th>Regression method</th>
<th>Train/test</th>
<th>$R^2$(%)</th>
<th>RMSE</th>
<th>NRMSE</th>
<th>Selected parameters of SVR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple Linear Regression</td>
<td>Test</td>
<td>63.13</td>
<td>6.38</td>
<td>0.44</td>
<td>——</td>
</tr>
<tr>
<td></td>
<td>Train</td>
<td>90.70</td>
<td>5.99</td>
<td>0.52</td>
<td></td>
</tr>
<tr>
<td>SVR linear kernel</td>
<td>Test</td>
<td>84.55</td>
<td>6.18</td>
<td>0.43</td>
<td>$C = 32768, \varepsilon = 0.1$</td>
</tr>
<tr>
<td></td>
<td>Train</td>
<td>70.22</td>
<td>4.12</td>
<td>0.36</td>
<td></td>
</tr>
<tr>
<td>SVR polynomial kernel degree 1</td>
<td>Test</td>
<td>87.25</td>
<td>5.96</td>
<td>0.41</td>
<td>$C = 128, \gamma = 256, d = 1, r = 256, \varepsilon = 0.10$</td>
</tr>
<tr>
<td></td>
<td>Train</td>
<td>70.54</td>
<td>4.09</td>
<td>0.36</td>
<td></td>
</tr>
<tr>
<td>SVR polynomial kernel degree 2</td>
<td>Test</td>
<td>63.52</td>
<td>7.02</td>
<td>0.48</td>
<td>$C = 1, \gamma = 0.70, d = 2, r = 128, \varepsilon = 0.10$</td>
</tr>
<tr>
<td></td>
<td>Train</td>
<td>71.97</td>
<td>3.98</td>
<td>0.35</td>
<td></td>
</tr>
<tr>
<td>SVR polynomial kernel degree 3</td>
<td>Test</td>
<td>63.95</td>
<td>6.86</td>
<td>0.47</td>
<td>$C = 0.003, \gamma = 0.25, d = 3, r = 128, \varepsilon = 0.10$</td>
</tr>
<tr>
<td></td>
<td>Train</td>
<td>72.12</td>
<td>4.12</td>
<td>0.36</td>
<td></td>
</tr>
<tr>
<td>SVR RBF kernel</td>
<td>Test</td>
<td>87.42</td>
<td>5.19</td>
<td>0.36</td>
<td>$C = 512, \gamma = 0.06, \varepsilon = 0.50$</td>
</tr>
<tr>
<td></td>
<td>Train</td>
<td>99.67</td>
<td>0.49</td>
<td>0.04</td>
<td></td>
</tr>
</tbody>
</table>

Figure 6. Representation of some of the 2-D scatterplots between measured soil EC values (y-axis) and predictor variables (x-axis) including (a) blue, (b) red edge 1, (c) NIR, (d) SI-1, (e) SI-2, (f) SI-6, (g) NDVI, (h) EVI, (i) SAVI, (j) PC1, (k) IC2, and (l) LST.
Table 5. In the next step, FS algorithms were implemented to reduce the amount of data and selecting the best explanatory variables, which also helps to improve the model precision. For the GA, some parameters should be initially determined to run the algorithm, as described in Table 6. It should be noted that these parameters need to remain unchanged in all regression methods in order for the results of GA to be comparable in terms of efficiency and selected features. The search process of the GA for each regression method is shown in Figure 7.

**Verification and accuracy of the reconstructed models**

The results obtained from implementing FS algorithms to salinity models are shown in Tables 7 and 8. In SFS method, since SBS had better accuracy than SFS, results of the backward method were represented.

**Prediction of soil EC maps**

Considering the results of the regression analysis, our constructed models were used to map EC values for each pixel in the scene. For each regression method, three models were developed based on corresponding selected predictors (18 in total). Among these models, the best of each regression technique, based on \( R^2 \) and RMSE values, were considered to map soil EC for the entire image. The produced maps were then compared, which will be discussed in the next section. Figure 8 shows the predicted EC maps for each regression method.

Evaluating the performed analysis and the predicted EC maps, we can observe the following results:

- Among the different regression methods proposed in this study, the SVR technique with RBF kernel function shows the most accuracy, with the highest \( R^2 \) of 87.42% and the lowest RMSE of 5.1962 (when using all features).

**Table 6. Genetic algorithm parameters.**

<table>
<thead>
<tr>
<th>Parameters/definition</th>
<th>Assigned values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population size: specifies how many individuals there are in each generation.</td>
<td>20</td>
</tr>
<tr>
<td>Elitism ratio: specifies the number of individuals that are guaranteed to survive to the next generation.</td>
<td>1</td>
</tr>
<tr>
<td>Crossover fraction: specifies the fraction of the next generation.</td>
<td>0.70</td>
</tr>
<tr>
<td>Crossover method</td>
<td>One point</td>
</tr>
<tr>
<td>Mutation ratio: specify how the genetic algorithm makes small random changes in the individuals in the population to create mutation children.</td>
<td>0.05</td>
</tr>
<tr>
<td>Maximum generation: Specifies the maximum number of iterations for the genetic algorithm to perform.</td>
<td>100</td>
</tr>
</tbody>
</table>

- Using FS algorithms improves the accuracy of all regression models, which can be considered a useful method to provide results that are more reliable, and to enhance model performance. Between the two selected methods, GA is more suitable than SFS because it shows higher accuracy in most of the cases and a lower number of selected features. Using GA in conjunction with SVR with a polynomial of degree 1 shows the best performance, increasing \( R^2 \) value from 87.25% to 96.16% and decreasing RMSE from 5.9652 to 4.1353.

- Evaluation of the predicted EC maps, which were obtained from polynomial kernels, reveals that using high degrees of polynomials (more than 1) decreases model accuracy, and causes over-fitting problems. It also increases the range of variation of EC values in the scene, which leads to intensified salinity maps. This also happens in the MLR model, and significant changes in EC values can be observed in the predicted maps; we interpret this as a weakness of the model.

- For the MLR method, despite improving the accuracy of the constructed models through FS algorithms, the generated EC maps show major differences and are uncorrelated in all three cases, especially in the areas located in the top left of the scene. This is due to a weakness of the linear model when it comes to correctly predict EC values from satellite data, making the linear method unreliable and less effective.

- In the central part of the study site, where Kuh Sefid itself is located, the high spectral reflectance of the surface crust evident in the image causes most of the developed models to represent high EC values for those pixels (SVR with RBF kernel being an exception). However, this is not necessarily related to the accumulation of salts in the soil surface. In this region, the EC values of pixels may not be valid; more investigation is necessary to predict conductivity of soils from remotely sensed data. Using a variety of hydrological and meteorological data, including soil moisture, air humidity, the rate of precipitation, wind speed, etc., and ancillary data such as geologic maps or historical data is recommended.

- A comparison between the selected features, which were obtained from different regression methods, indicates that “Green, Red Edge 1, SI2, SWIR2, BI, and PC1” had the best performance in model development, revealing the importance of these features in monitoring soil salinity. These features could be used as optimal salinity indicators for monitoring soil salinity through satellite imagery in future studies.

- The proposed methodology reveals the potential of the Sentinel-2A optical imaging satellite to produce...
soil EC map with high accuracy; now it is possible to estimate salinity for each 10 m × 10 m area at very short intervals of about 5 days. This represents the remotely sensed data as a useful tool for land management studies and soil reclamation programs.

**Conclusion**

This paper deals with soil salinity mapping and monitoring using multispectral satellite imagery from the newly launched Sentinel-2A satellite. Soil samples were randomly collected throughout the

---

**Figure 7.** Search process of GA for different regression techniques, (a) MLR, (b) SVR (linear), (c) SVR (poly deg. 1), (d) SVR (poly deg. 2), (e) SVR (poly deg. 3), and (f) SVR (RBF).
study site, and the EC of the soil was measured. At the same time, corresponding satellite data were acquired, and various salt features were extracted from the satellite imagery to be used for salinity detection. The data used include the Sentinel-2 bands, vegetation and salinity indices, and different transformation features. Different regression methods were used to relate the EC values of field samples to satellite-derived salt features. Results revealed that among the different regression methods, the SVR technique with RBF kernel showed the most accuracy for modeling soil salinity, with $R^2 = 87.42\%$ and RMSE = 5.162. Additionally, FS algorithms were used to further improve the results of the analysis and to select the best features that properly characterize the salt-affected regions in the scene. Results show the advantage of using GA over SFS because it provides higher accuracy, with a lower number of selected features. It can be concluded that it is possible to map soil salinity at very short intervals of about 5 days for each 10 m $\times$ 10 m area, using the potential of the Sentinel-2 satellite data and the SVR technique, which confirms remote sensing as a powerful technology for salinity detection and mapping soil EC.

### Disclosure statement

No potential conflict of interest was reported by the authors.

### ORCID

Mohammad Mahdi Taghadosi [http://orcid.org/0000-0003-0773-1570](http://orcid.org/0000-0003-0773-1570)

Mahdi Hasanlou [http://orcid.org/0000-0002-7254-4475](http://orcid.org/0000-0002-7254-4475)

### References


---

#### Table 7. Accuracy assessment of reconstructed models based on GA feature selection.

<table>
<thead>
<tr>
<th>Regression method</th>
<th>Train/ test (%)</th>
<th>RMSE</th>
<th>NRMSE</th>
<th>Selected features</th>
</tr>
</thead>
</table>

#### Table 8. Accuracy assessment of reconstructed models based on sequential feature selection.

<table>
<thead>
<tr>
<th>Regression method</th>
<th>Train/ test (%)</th>
<th>RMSE</th>
<th>NRMSE</th>
<th>Selected features</th>
</tr>
</thead>
</table>

---
Figure 8. Predicted soil EC maps based on the results of regression analysis using the best developed models, (a) linear regression, (b) SVR (linear kernel), (c) SVR (polynomial deg.1 kernel), (d) SVR (polynomial deg.2 kernel), (e) SVR (polynomial deg.3 kernel), and (f) SVR (RBF kernel).


