



## Prediction of soil surface salinity in arid region of central Iran using auxiliary variables and genetic programming

Ruhollah Taghizadeh-Mehrjardi, Shamsollah Ayoubi, Zeinab Namazi, B. P. Malone, Ali A. Zolfaghari & F. Roustaei Sadrabadi

To cite this article: Ruhollah Taghizadeh-Mehrjardi, Shamsollah Ayoubi, Zeinab Namazi, B. P. Malone, Ali A. Zolfaghari & F. Roustaei Sadrabadi (2016) Prediction of soil surface salinity in arid region of central Iran using auxiliary variables and genetic programming, Arid Land Research and Management, 30:1, 49-64, DOI: [10.1080/15324982.2015.1046092](https://doi.org/10.1080/15324982.2015.1046092)

To link to this article: <http://dx.doi.org/10.1080/15324982.2015.1046092>



Published online: 08 Feb 2016.



Submit your article to this journal [↗](#)



Article views: 65



View related articles [↗](#)



View Crossmark data [↗](#)

## Prediction of soil surface salinity in arid region of central Iran using auxiliary variables and genetic programming

Ruhollah Taghizadeh-Mehrjardi<sup>a</sup>, Shamsollah Ayoubi<sup>b</sup>, Zeinab Namazi<sup>b</sup>, B. P. Malone<sup>c</sup>, Ali A. Zolfaghari<sup>d</sup>, and F. Roustaei Sadrabadi<sup>a</sup>

<sup>a</sup>Faculty of Agriculture and Natural Resources, Ardakan University, Ardakan, Iran; <sup>b</sup>Department of Soil Science, College of Agriculture, Isfahan University of Technology, Isfahan, Iran; <sup>c</sup>Department of Environmental Sciences, Faculty of Agriculture and Environment, The University of Sydney, New South Wales, Australia; <sup>d</sup>Faculty of Desert Studies, Semnan University, Semnan, Iran

### ABSTRACT

Spatial information on soil salinity is increasingly needed for decision making and management practices in arid environments. In this article, we attempted to investigate soil salinity variation via a digital soil mapping approach and genetic programming in an arid region, Chah-Afzal, located in central Iran. A grid sampling strategy with 2-km distance was used. In total, 180 soil surface samples were collected and then analyzed. A symbolic regression was then adopted to correlate electrical conductivity ( $EC_e$ ) with a suite of auxiliary data including predicted maps of apparent electrical conductivity (vertical:  $EC_{av}$  and horizontal:  $EC_{ah}$ ), Landsat spectral data and terrain attributes derived from a digital elevation model. The accuracy of the genetic programming model was evaluated using root mean square error (RMSE), mean error (ME), and coefficient of determination ( $R^2$ ) based on an independent validation data set (20% of database or thirty soil samples). In general, results showed that  $EC_{ah}$  had the strongest influence on the prediction of soil salinity followed by salinity index wetness index, Landsat Band 3, multi-resolution valley bottom flatness index, elevation, and normalized difference vegetation index. Furthermore, results indicated that the genetic programming model predicted  $EC_e$  over the study area accurately ( $R^2 = 0.87$ ,  $ME = -1.04$  and  $RMSE = 16.36 \text{ dSm}^{-1}$ ). Overall, it is suggested that similar applications of this technique could be used for mapping soil salinity in other arid regions of Iran.

### ARTICLE HISTORY



Received 18 October 2014  
Accepted 25 April 2015

### KEYWORDS

Auxiliary data; digital soil mapping; EM38; Iran

## Introduction

High levels of soil salinity negatively affect crop growth and productivity leading, ultimately, to land degradation in central Iran. Therefore, to manage this problem, accurate and reliable salinity mapping is needed. However, soil mapping at a high-resolution by conventional methods is expensive and time consuming. In recent times Digital Soil Mapping (DSM) has been able to overcome these limitations in the form of numerical and model-based analysis of soil-landscape relationships (McBratney, Mendonça-Santos,

**CONTACT** Ruhollah Taghizadeh-Mehrjardi  [rtaghizadeh@ardakan.ac.ir](mailto:rtaghizadeh@ardakan.ac.ir)  BLVD. Khatami, Faculty of Agriculture and Natural Resources, Ardakan University, P.O. Box 89516-56767, Ardakan, Iran.

Color versions of one or more of the figures in the article can be found online at [www.tandfonline.com/uasr](http://www.tandfonline.com/uasr).

and Minasny 2003). In practice DSM may be distilled down to the use computer assisted methods to harmonize soil data with more readily measured auxiliary variables.

The first and probably the most common auxiliary variable used in digital mapping of soil salinity is remote sensing data (Mariappan 2010; Bilgili et al. 2011; Allbed, Kumar, and Sinha 2014). For example, Mariappan (2010) applied Landsat Thematic Mapper (TM) Bands 1 through 7 for identifying salt minerals and found that Landsat Band 3 was particularly important for salinity detection. Metternicht and Zinck (2008) also confirmed Landsat spectral data could serve as useful covariate information to predict soil salinity distribution. Allbed, Kumar, and Sinha (2014) indicated that salinity index (SI) and red band (B3) had a good relationship with soil electrical conductivity. The second auxiliary variables that have been widely used in digital mapping are terrain attributes. For example, Sheng et al. (2010) used exclusively terrain attributes to monitor soil salinity variations in China. In addition, terrain attributes could also be combined with Landsat spectral data for spatial modelling of soil salinity (e.g., Akramkhanov and Vlek 2012; Hamzhepour et al. 2013; Taghizadeh-Mehrjardi et al. 2014). Another auxiliary variable (for salinity mapping) used by some researchers is electromagnetic induction data. Good relationships between soil salinity and  $EC_a$  values have been reported by Lesch, Herrero, and Rhoades (1998), Urdanoz and Aragüés (2011), Li et al. (2013), Taghizadeh-Mehrjardi et al. (2014), and Ding and Yu (2014).

Various DSM techniques including artificial neural networks (Behrens et al. 2005; Aitkenhead et al. 2012), decision trees (Bui and Moran 2001; Jafari et al. 2014), K-nearest neighbors (Nemes et al. 1999; Nemes, Rawls, and Pachepsky 2006; Coopersmith et al. 2014), support vector machines (Kovacevic, Bajat, and Gajic 2010; Li, Im, and Beier 2013), and random forests (Stum et al. 2010; Heung, Bulmer, and Schmidt 2014) have been developed and introduced to link soil properties and auxiliary variables in order to predict various soil attributes and soil classes.

A relatively new, yet potentially powerful modeling algorithm, which has had little traction to date in the DSM literature is genetic programming (GP). GP is defined as an evolutionary computation technique inspired from the fundamentals and rules of biological evolution (Poli, Langdon, and McPhee 2008). Briefly, GP is a systematic, domain-independent method for computers to untangle moot point automatically beginning from a high-level statement of what needs to be done (Koza 2010). GP starts with an initial population of randomly generated programs. It has been utilized in photogrammetric issues, multispectral analysis and remote sensing problems (Puente et al. 2011). It has been successfully employed in the domain of image analysis (Yang 2007; Fonlupt and Robilliard 2000; Brumby et al. 2001; Puente et al. 2011). In soil science, GP has been used successfully to develop pedo-transfer functions (Johari, Habibagahi, and Ghahramani 2006; Parasuraman, Elshorbagy, and Si 2007; Padarian, Minasny, and McBratney 2012).

As previously mentioned, despite of the rapid progression of DSM in many places, few studies have been conducted to digitally map soil salinity in arid regions in Iran (Taghizadeh-Mehrjardi et al. 2014). The only available soil map in Iran is a recently prepared one, at a scale of 1:1,000,000, which is not suitable for farm management planning. Thus, the lack of high-resolution soil maps for agricultural areas in Iran is particularly felt. Therefore, the objective of the present study is to predict the spatial distribution of soil salinity using DSM methods; in particular, focusing on the implementation of GP to fulfill that purpose.

## Materials and methods

### *Description of the study area*

The study area is located in Yazd Province, central Iran (Figure 1). It is situated between 53.8°E and 54.1°E, and between 32.35°N and 32.65°N. The area covers approximately 80,000 ha and is located 12-km north from Ardakan City. The elevation ranges from 953 to about 1900 m a.s.l. Average annual rainfall and annual temperature are approximately 80 mm and 18.3°C, respectively. Soil moisture and temperature regimes are aridic and thermic, respectively. Land use in the area includes agriculture, natural rangeland, and bare lands. The major geological units are comprised of red gypsiferous marls and brown to grey limestone. The major landforms of the region are mountain, alluvial fans, playa, and coalescing alluvial fans.

### *Data collection and soil sample analyses*

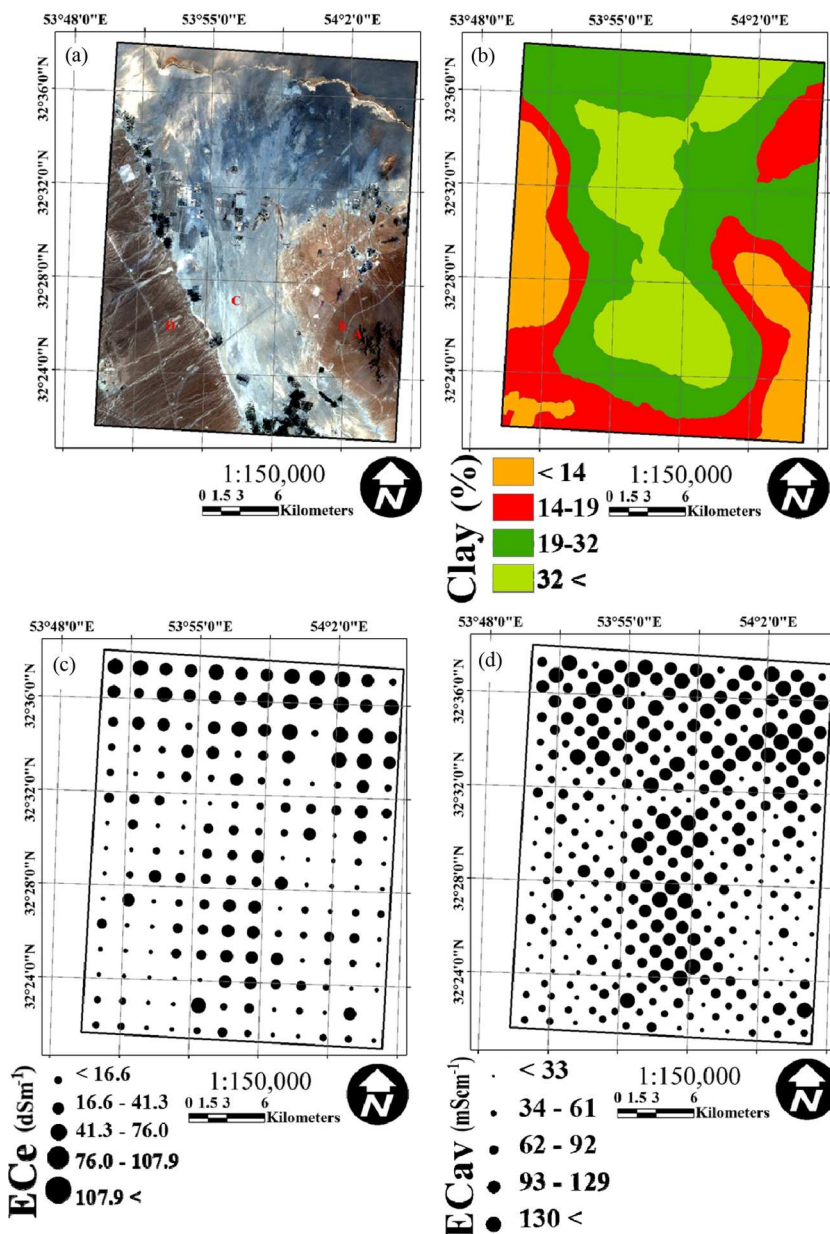
In order to adequately cover the study area, a 2-km grid sampling strategy was used in this study; totaling 180 composite soil surface samples from across the target area (Figure 2a). The samples were collected in summer 2012. Each composite soil sample was comprised of four core sub-samples that were collected at a distance of 10-m north, south, east and west of the center sampling point. The sub-samples were collected from the surface horizon (0–20 cm) with a hand auger (10-cm diameter) and were crushed and mixed together to form one sample. Soil samples were dried, and ground to pass through a 2-mm sieve. Soil organic matter (SOM) content was determined by the Walkley-Black method (Nelson and Sommers 1986). Particle size distribution in the soil samples (clay, silt, and sand) was measured using the procedure described by Gee and Bauder (1986) and calcium carbonate equivalent (CCE) content was determined by the back-titration method (Nelson 1982). Electrical conductivity (ECe) and pH were measured in the soil saturation extracts according to standard methods (Richards 1954; Sparks et al. 1996).

### *Auxiliary variables*

The environmental variables used in the present study included three kinds of auxiliary variables; remote sensing data, proximally sensed apparent electrical conductivity, and topographic variables derived from a digital elevation model.

### *Remote sensing*

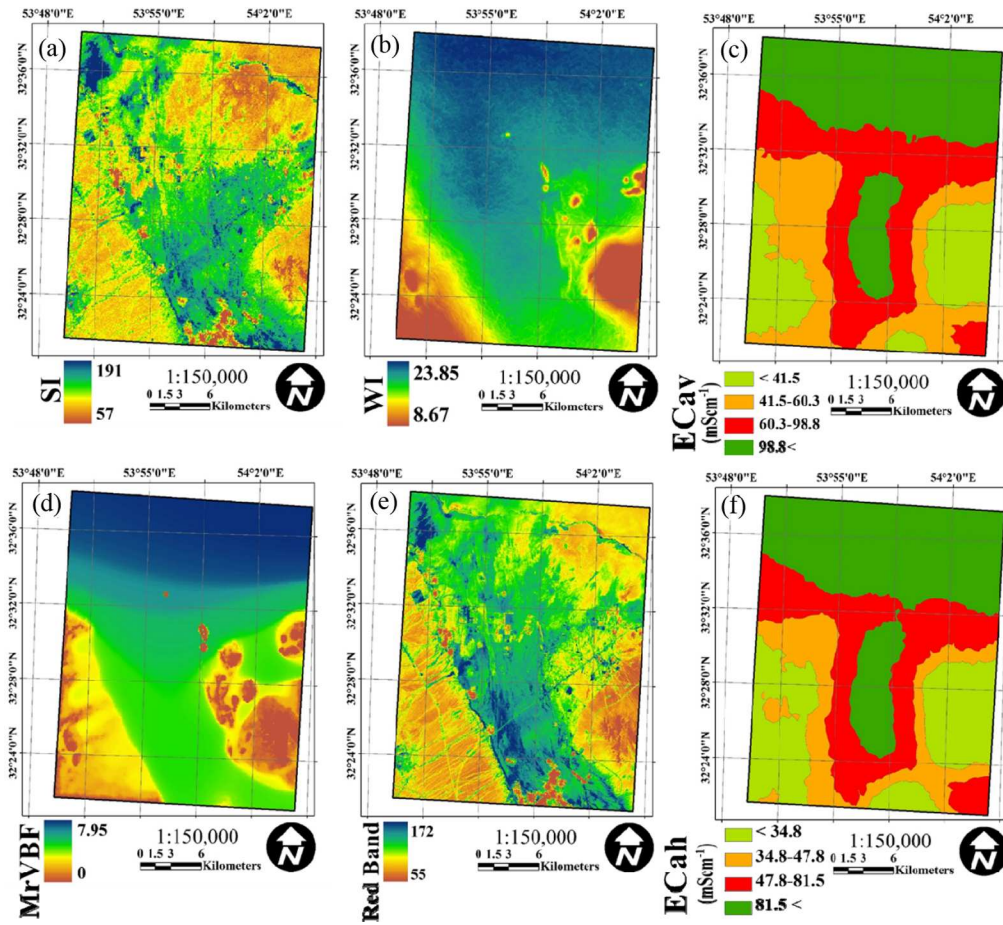
Cloud-free Landsat 7 ETM<sup>+</sup> images were used in this study and were acquired near the actual soil sampling date (August 2012). The imagery consisted of six spectral bands with pixel resolution of 30 m: B1, B2, B3, B4, B5, and B7. From these data, a number of additional variables were derived including: Normalized difference vegetation index (Rouse et al. 1973); ratio vegetation index (Pearson and Miller 1972); soil-adjusted vegetation index (Huete 1988); clay index (Boettinger et al. 2008); gypsum index (Nield, Boettinger, and Ramsey 2007); and salinity index and brightness index (Metternicht and Zinck 2008). The images were geo-rectified to a Universal Transverse Mercator (UTM) coordinate system using World Geodetic System (WGS) 1984 datum assigned



**Figure 1.** (a) Location of the study area located in Chah-Afzal region in central Iran (A: mountain; B: alluvial fans; C: playa D: coalescing alluvial fans). The false color composite was prepared using combination of three bands: B1, B2, and B3. (b) The spatial distribution of clay percentage over the study area; (c) the spatial distribution of  $EC_e$ ; and (d) the spatial distribution of apparent electrical conductivity in the vertical mode.

to north UTM zone 40. Atmospheric correction was performed using the Dark-Object Subtraction (DOS) technique (Chavez 1996). All the remote sensing data processing was performed using the Environment for Visualizing Images (ENVI) version 4.8 software.





**Figure 2.** Spatial distribution of some auxiliary data including (a) salinity index (SI); (b) wetness index (WI); (c) apparent electrical conductivity in the vertical mode; (d) multi-resolution valley bottom flatness index (MrVBF); and (e) red band of  $ETM^+$ ; (f) apparent electrical conductivity in the horizontal mode.

### Apparent electrical conductivity

The apparent electrical conductivity ( $EC_a$ ) of the bulk soil was measured using an electromagnetic conductivity meter. In order to generate  $EC_a$  maps efficiently for the study area, field  $EC_a$  data were collected in different campaigns. The first of which were 180  $EC_a$  readings taken at the sites used for the soil sampling. These data were also used to investigate the calibration of  $EC_a$  data with  $ECE$  measurements at these sites. This calibration was calculated using multi-linear regression equations using MATLAB software (Mathworks 2010). An additional 24  $EC_a$  readings were taken from six transects. These transect were selected randomly and in each, four  $EC_a$  readings were taken with the mean separation distance of 30-m. A further 156 readings were gathered based on a grid sample with a mean separation distance of 2000-m between sites in order to adequately cover the study area (Figure 1). In total, 360  $EC_a$  data in both vertical and horizontal modes were taken for this study. The device measures in two modes: vertical ( $EC_{av}$ ) and horizontal ( $EC_{ah}$ ) modes working to depth of 1.5 and 0.75 m, respectively. Then electromagnetic induction data in both vertical and horizontal angles were interpolated spatially across the study area to

derive continuous maps of  $EC_a$  at the aforementioned integrated depths with  $30 \times 30$  m pixel size. The interpolation was performed using locally fitted variograms and kriging, and implemented through the VESPER geostatistical software (Minasny, McBratney, and Whelan 1999).

### ***Terrain attributes***

In this study eleven terrain parameters were obtained from a digital elevation model (pixel size of  $10 \times 10$  m; National Cartographic Center 2010), which included: elevation, altitude above channel network, modified catchment area, mid-slop position, multi-resolution ridge top flatness index (MrRTF), and multi-resolution index of valley bottom flatness (MrVBF) (Gallant and Dowling 2003), topographic wetness index, valley depth and catchment aspect, slope, and height. The digital terrain analysis was performed using the open source SAGA GIS software (Olaya 2004). All auxiliary variables were co-registered to the same raster grid size of 30 m.

The next step was to select the relevant auxiliary variables for modelling. Removing irrelevant and redundant auxiliary variables not only reduces the dimensionality of the data but also may allow learning algorithms to operate faster and more effectively. Moreover, irrelevant and redundant information may decrease the prediction accuracy in common machine learning algorithms (Hall 1999; Hall et al. 2009; Mollazade, Omid, and Arefi 2012). Different techniques can be used to rank the relevance of auxiliary variables, including correlation-based feature selection (CFS), principal component analysis (Omid, Mahmoudi, and Omid 2010), factor analysis, and sensitivity analysis. Herein, we applied correlation-based feature selection (CFS) technique with a best first algorithm using the *CfsSubsetEval* algorithm in WEKA software to select the best subset of auxiliary variables (Hall et al. 2009). Correlation-based feature selection is a fully automatic algorithm, not requiring any predefined thresholds or the number of features. Correlation-based feature selection ranks auxiliary variables according to a correlation based heuristic evaluation function. This algorithm keeps relevant auxiliary variables that are highly correlated with the target variable (in our case soil salinity) and screens out irrelevant auxiliary variables that have low correlation with the  $EC_e$  and exhibit a high degree of co-linearity with other variables. From analysis for this study, the CFS algorithm reduced the number of auxiliary variables from 26 potentials to 7 accepted, of which included: elevation, NDVI,  $EC_{ah}$ , topographic wetness index, MrVBF, red band, and salinity index (Figure 2).

### ***Spatial modelling***

The relationship between soil salinity and auxiliary variables was implemented by applying empirical models (McBratney, Mendonça-Santos, and Minasny 2003). Here in this study, genetic programming (GP) was evaluated. The implementation of GP was performed using the MATLAB software (Mathworks 2010). To overcome the limitation of Genetic Algorithm (GA) in evolving complex models, Koza (1992) introduced Genetic Programming (GP) technique as an alternative technique. Like other evolutionary algorithms, GP initializes a population consisting of a random population of individuals. This population of programs is progressively evolved, according to the principle of Darwin's natural selection theory in evolution, over a series of generations. The fitness of each chromosome is evaluated with respect to a target value. First, GP selects a proportion of the existing

population to breed a new generation. Then, a second population generated using common variation operators—crossover and mutation. This process of selection, reproduction, and variation iterates until a user-defined “stopping criterion” is satisfied. In this work we used a specific method called symbolic regression, which uses GP to fit a function to a specific dataset, going from simple functions to increasingly more complex functions. Further description of this technique are provided in Koza et al. (1999) and Koza (1992). GP selects the most relevant variable subsets for modelling. It calculates the relative impact of auxiliary variables on the target variable ( $EC_e$ ). Given a model equation of the form  $z = f(x, y \dots)$ , the influence metrics of  $x$  on  $z$  are defined as follows:

$$\left[ \frac{\partial z}{\partial x} \right] \cdot \frac{\partial(x)}{\partial(z)} \quad (1)$$

$\left[ \frac{\partial z}{\partial x} \right]$  is the partial derivative of  $z$  with respect to  $x$ ;  $\partial(x)$  is the standard deviation of  $x$  in the input data;  $\partial(z)$  is the standard deviation of  $z$ ;  $|x|$  denotes the absolute value of  $x$  (Koza 1992).

### Model evaluation

In order to evaluate the accuracy of prediction, the data was divided randomly into two data sets. The larger data set (80%) was used for training and the smaller data set (20%) was set aside for external validation. Validation criteria which are popularly used in digital soil mapping: root mean square error (RMSE), mean error (ME), and coefficient of determination ( $R^2$ ) were applied in this study.

## Results and discussion

### Data summary

A data summary of soil salinity and EM38 readings in both modes ( $EC_{av}$  and  $EC_{ah}$ ) are presented in Table 1. According to Table 1, the average soil salinity levels are more than  $4 \text{ dSm}^{-1}$ , indicating soils in the study area are severely affected by salt (Chhabra 2006). The coefficient of variation (CVs) for soil salinity level is high (144.44), which indicates a wide range of values across the study area. The  $EC_e$  data across the study area ranged from 1.00 to  $229 \text{ dSm}^{-1}$ . The mean  $EC_e$  data was  $50.29 \text{ dSm}^{-1}$ . Taghizadeh-Mehrjardi et al. (2014) reported high values of  $EC_e$  in Yazd province. They reported  $EC_e$  as varying between 1 and  $245 \text{ dSm}^{-1}$  from top to the bottom of soil profiles. The coefficients of

**Table 1.** Descriptive statistics of soil salinity and electromagnetic induction readings ( $n = 180$  and  $360$  samples for  $EC_e$  and  $EC_{av}$ , respectively).

Layer (cm)	No.	Min	Max	Average	SD	CV%	Skew.	Kurt.
$EC_e$ (0–20) $\text{dSm}^{-1}$	180	1.00	229.00	50.29	72.64	144.44	0.54	0.11
$EC_{ah}$ (0–75) $\text{dSm}^{-1}$	360	0.01	1.66	0.57	0.41	66.99	0.81	−0.45
$EC_{av}$ (0–150) $\text{dSm}^{-1}$	360	0.03	2.27	0.69	0.46	71.97	0.61	−0.28

$EC_e$ : electrical conductivity;  $EC_{ah}$ : apparent electrical conductivity (horizontal mode);  $EC_{av}$ : apparent electrical conductivity (vertical mode); No: Number of observation; Min: minimum; Max: maximum; SD: standard deviation; CV: coefficient of variation; Skew: skewness; Kurt: kurtosis.



variation for  $EC_{ah}$  and  $EC_{av}$  were 66.99 and 71.97, respectively. These values were to some extent smaller than that for  $EC_e$  (144.44). This is likely to be due to that the instrument is measuring apparent electrical conductivity in larger volume of soil, while electrical conductivity was measured in smaller soil samples taken from the 0 to 20 cm. According to the spatial distribution of soil samples in the study area (Figure 1), soil samples found in upper land have coarse texture and low salinity, while soils characterized with more saline and fine texture are located in lower part of the area. The samples with highest salinity level were located in the middle and northern part of study area.

### Calibration of $EC_a$ measurements

In our study,  $EC_e$  and  $EC_a$  were considered as dependent and independent variables, respectively. The linear regression model was fitted to the data from all 180 sampling sites. Consequently, three predictive equations were achieved (Table 2). Similar approaches already were used by Lesch, Herrero, and Rhoades (1998), Urdanoz and Aragüés (2011), Li et al. (2013), Taghizadeh-Mehrjardi et al. (2014), and Ding and Yu (2014). According to this table, our results indicated  $EC_{ah}$  and  $EC_{av}$  had significant linear relationship with soil electrical conductivity. With regard to direct relationships among  $EC_a$  and soil salinity, we can imply that using both  $EC_{ah}$  and  $EC_{av}$  are useful predictor variables given an observed  $R^2$  of 0.75. Furthermore, comparisons showed that correlation coefficients between  $EC_{ah}$  and the  $EC_e$  ( $R^2 = 0.74$ ) were higher than those between  $EC_{av}$  and the  $EC_e$  ( $R^2 = 0.70$ ). This might be attributed to the fact that salts accumulate more intensively in the upper soil layers, and hence, the stronger response of  $EC_{ah}$  to superficial layers compared to  $EC_{av}$ . With regard to these results, it was inferred that accuracy of model just a little bit enhance as compared to simple linear regression; however, we used a simpler and parsimonious model (only  $EC_{ah}$ ) to predict the salinity.

However, the best regression model (i.e.,  $EC_{ah}$  was the only input variable) for soil surfaces recorded only an  $R^2$  value of 0.74, which is much lower when compared to other researchers' findings (Slavich 1990) who reported  $R^2$  values of around 0.90. Therefore, we concluded that we cannot use the equations in Table 2 directly for the reconstruction the soil salinity profile across the study area. This might be attributed to the fact that  $EC_a$  is influenced not only by soil salinity but also by many other factors such as soil texture, temperature and moisture content (Slavich 1990; Lesch, Herrero, and Rhoades 1998). Clay content in soils of some parts of the study area particularly, in middle part and the north, exceeds 50% and, hence this might be one reason for lowering the correlation between  $EC_a$  and  $EC_e$ . Low water content in the studied arid region, where the water content was less

**Table 2.** Regression relationships between apparent electrical conductivity and measured soil salinity ( $n = 180$ ).

Layer (cm)	$EC_e = a + b.EC_{ah}$			$EC_e = a + b.EC_{av}$			$EC_e = a + b.EC_{av} + c.EC_{ah}$			
	a	b	$R^2_{adj}$	a	b	$R^2_{adj}$	a	b	c	$R^2_{adj}$
$EC_e$ (0–20) $dSm^{-1}$	-1.85	0.88	0.74 <sup>b</sup>	-3.16	0.75	0.70	-3.11	0.15	0.71	0.75 <sup>b</sup>

$EC_e$ : electrical conductivity;  $EC_{ah}$ : apparent electrical conductivity (horizontal mode);  $EC_{av}$ : apparent electrical conductivity (vertical mode); a is intercept; c & b are the coefficients of regression;  $R^2$ : coefficient of determination.

<sup>a</sup>The relation is significant at the 0.05 level.

<sup>b</sup>The relation is significant at the 0.01 level.

than 5%, lead to a lowering correlation between  $EC_e$  and  $EC_a$  (Lesch, Herrero, and Rhoades 1998). However, temperature seemed to have the minimum effect on our  $EC_a$  values due to the fact that the survey was conducted during summer, when average soil temperature within the upper 1 m profile was around 25°C (Slavich 1990).

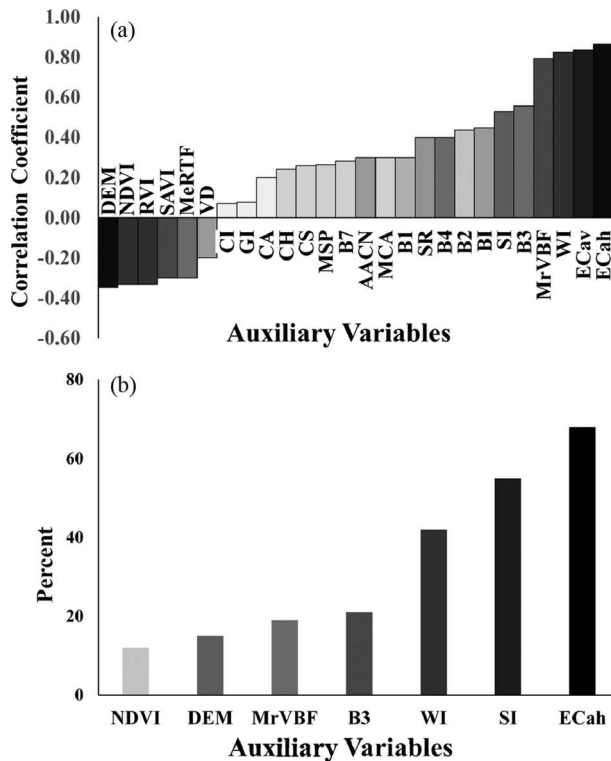
We proceeded to predict spatial distribution of  $EC_a$  at the study area using ordinary kriging. Evaluation of kriging models were integrated based on a cross-validation method. Results showed that the accuracy of models based on RMSE and  $R^2$  criteria for  $EC_{ah}$  were 25.15 and 0.57, respectively. RMSE and  $R^2$  for  $EC_{av}$  were 27.32 and 0.56, respectively. These results are consistent with previous findings from Taghizadeh-Mehrjardi et al. (2014) who reported RMSE and  $R^2$  as 37.74 and 0.49, respectively. As can be seen in Figure 2, the maps clearly illustrate that there are two distinct areas of high  $EC_a$  values located in center of region (which mostly has soils of fine texture), while low  $EC_a$  values located in east and west parts of the study area (which are generally soils with coarse texture).

### **Importance of auxiliary variables**

The correlation coefficient between  $EC_e$  and all auxiliary variables is shown in Figure 3a. As can be seen from this figure, the coefficients range between  $-0.34$  and  $0.86$ . The best correlation is between the  $EC_e$  and EM conductivity information; the correlation between  $EC_e$  data and clay index, gypsum index, and mid-slope position is considerably lower. The relative impact of auxiliary data on the electrical conductivity was also assessed by GP. Analysis of GP (based on Equation 1) showed that  $EC_{ah}$  (68%) had the strongest influence on the prediction of soil salinity followed by salinity index (55%), wetness index (42%), red band (21%), MrVBF (19%), elevation (15%), and NDVI (12%) (Figure 3b).

$EC_{ah}$  was identified as the most powerful predictor for surface electrical conductivity. Apparent electrical conductivity has been successfully used to improve the spatial distribution of soil salinity (Ding and Yu 2014; Taghizadeh-Mehrjardi et al. 2014). The second most important predictor was the salinity index (55%). Red band (Landsat Band 3) and NDVI were also encountered in the model, though at a lower rate of 21% and 12%, respectively. As most of the area was bare (Figure 1), the presence of salts at the surface can be directly detected by Landsat spectral data. However, in the vegetated area direct approach becomes complicated and may yield unreliable results. But, the present scattered vegetation or halophytes on the soil surfaces can serve as a sign of the salinity problem, making it possible to indirectly detect and map areas that are affected by soil salinity using the reflectance from vegetation. Normally, unhealthy vegetation has a lower photosynthetic activity, causing increased visible reflectance and the reduced near-infrared reflectance from the vegetation (Allbed, Kumar, and Sinha 2014). Therefore, based on our findings, NDVI has been used as indirect indicators to assess and map soil salinity. Similar results were obtained by Allbed, Kumar, and Sinha (2014) who indicated salinity index (SI) and red band (B3) had a good relationship with electrical conductivity. Mariappan (2010) also found that among the Landsat ETM<sup>+</sup> bands 1–5 and 7, the visible red band (B3) performed the best at characterizing the pattern and features of soil salinity. Shamsi, Sanaz, and Abtahi (2013) confirmed that salinity index could contribute to improving the predictive models of soil salinity.

Terrain attributes derived from the DEM such as MrVBF and wetness index had also influence in mapping of soil salinity (19 and 42%, respectively). Topographic

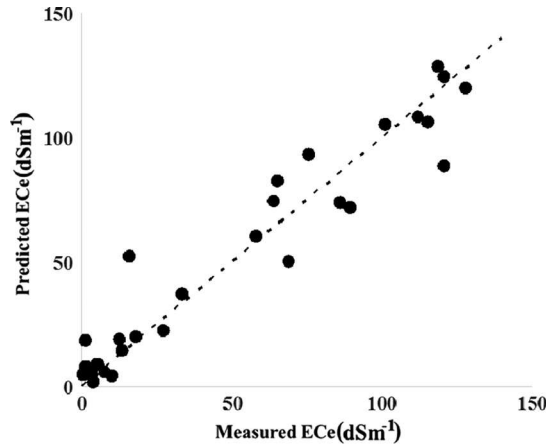


**Figure 3.** Correlation coefficients between  $EC_e$  and auxiliary variables (a) and the percentage contribution of most important auxiliary variables used in GP model (b). ( $EC_{ah}$ : horizontal;  $EC_{av}$ : vertical readings; WI: wetness index; MrVBF: multi-resolution valley bottom flatness index; B1–B7: spectral data of  $ETM^+$ ; SI: salinity index; BI: brightness index; SR: salinity ratio; MCA: modified catchment area; AACN: altitude above channel network; NDSI: normalized difference salinity index; MSP: mid-slope position; GI: gypsum index; CI: clay index; VD: valley depth; MrRTF: multi-resolution ridge top flatness index; SAVI: soil-adjusted vegetation index; RVI: ratio vegetation index; NDVI: normalized difference vegetation index; DEM: elevation; CS: catchment slope; CA: catchment aspect; CH: catchment height).

Wetness index, which can depict stationary water content in soils, indicated the potential areas where salic horizons may be presented. Our results indicated a correlation coefficient of 0.82 between  $EC_e$  and wetness index (Figure 3a–b). Similarly, Moore, Grayson, and Ladson (1991) reported a strong relationship between soil salinity and wetness index. MrVBF was also an effective index in the flat areas, especially for identifying flat valley bottoms and, consequently, indicated potential zones of transport for sediment. Taghizadeh-Mehrjardi et al. (2014) and Jafari et al. (2012) also confirmed the potential of these covariates for the identification of saline soils.

### Genetic programming

A symbolic regression was used to model the relationship between  $EC_e$  and auxiliary data. Four basic arithmetic operators (+, −, \*, and /) and more complex operators ( $\sqrt{\quad}$ ,  $x^2$ , power, Sin and Cos) were utilized. The functional set and operational parameters used in GP



**Figure 4.** Scatter gram of predicted versus measured ECe using proposed model (GP) based on validation data set.

modeling during this study are summarized in Table 3. A symbolic regression was attempted, generating a model expressed as:

$$\begin{aligned}
 EC_e = & 3.59 \times \cos(B_3) + 0.08 \times (WI) \times EC_{ah} \\
 & + \frac{0.0023}{NDVI - 0.138} + 11.9EC_{ah} \\
 & \times \sin(\cos(B_3) - 0.0049 \times DEM \times (WI)) \\
 & - EC_{ah} - \sin(SI) - M\gamma VBF \times \sin(SI) - 173 \\
 & \times NDVI \times EC_{ah} \times \sin(\cos(B_3) \\
 & - 0.0049 \times DEM \times (WI))
 \end{aligned} \quad (2)$$

where  $EC_e$  is electrical conductivity;  $EC_{ah}$  denotes apparent electrical conductivity in horizontal mode;  $MrVBF$  is multi-resolution valley bottom flatness index;  $WI$  is wetness index,  $SI$  denotes salinity index.  $DEM$  is digital elevation model; and  $NDVI$  is Normalized difference vegetation index.

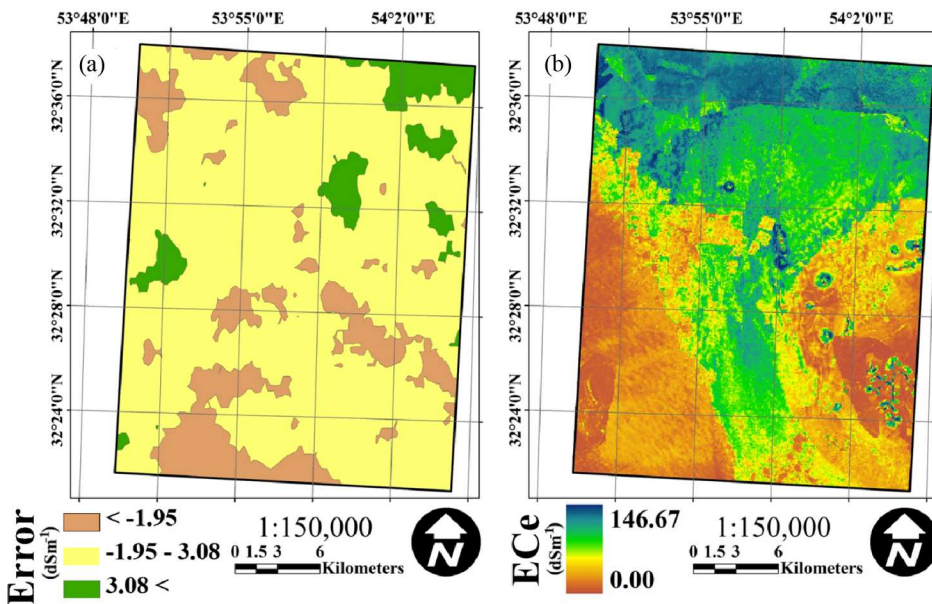
The scatter plots of the measured against predicted  $EC_e$  for the validation data set is shown in Figure 4. According to this figure, the best fitted line is close to the a 1:1 line, indicating high accuracy of estimation by the proposed model (Padarian, Minasny,

**Table 3.** Parameters of the optimized GP model.

Parameter	Description of parameter	Setting of parameter
P1	Function set	+, -, *, /, $\sqrt{\quad}$ , $\times 2$ , power, Sin, Cos
P2	Population size	400
P3	Mutation frequency %	95
P4	Crossover frequency %	20
P5	Number of replication	10
P6	Block mutation rate %	25
P7	Instruction mutation rate %	25
P8	Instruction data mutation rate %	50
P9	Homologous crossover %	85
P9	Program size	Initial 85, maximum 420

and McBratney 2012). This outcome implies that the GP model can provide a good relationship between auxiliary data and soil salinity. The statistical results showed the GP model predicted  $EC_e$  in the Chah-Afzal with  $R^2 = 0.87$ ,  $ME = -1.04$  and  $RMSE = 16.36$ . These results are acceptable when compared with similar previous researcher, for example, Allbed, Kumar, and Sinha (2014) and Taghizadeh-Mehrjardi et al. (2014) who reported  $R^2$  values around 0.65 and 0.78, respectively. Nevertheless, for digital soil mapping, these results indicate a good accuracy where  $R^2$  values over 70% are not very common and values of 50% or less are more common (Malone et al. 2009). Furthermore, our results indicate that incorporating  $E_{Ca}$  with Landsat data and terrain information to predict soil salinity can enhance the predictive power of fitted spatial models (from 0.78 to 0.87). This result is consistent with findings of Taghizadeh-Mehrjardi et al. (2014).

Map of the residuals (the difference between measured and predicted values) was also calculated to assess the uncertainty of the model (Figure 5a). Using the fitted GP model, a soil salinity map over the study area was generated (Figure 5b). As can be seen in the figure, we can easily imply that soils with the highest salinity are located in the north, whereas the soil with the lowest salinity can be found in the east and west of the region. As a matter of fact, most of the saline soils across the study area are located in the lower part of the region, which is a playa landform. This is likely due to that the playa receives more soluble salts from upper areas in the landscape. In addition, the concavely shaped plain could help to move ground water toward the north of area in which the soils with the highest electrical conductivity generally occur. In the north of area, soils are heavy or fine textured and this might facilitate capillary movement of groundwater to the soil surface and consequently lead to accumulation of saline materials in soil surfaces (Metternicht and Zinck 2008). Furthermore, as shown in Figure 1, the poor vegetation cover in the lower part of the region could be further justification for highly affected saline soils there.



**Figure 5.** Spatial distribution of residuals (a) and spatial distribution soil salinity in Chah-Afzal, central Iran, using the genetic programming across the study area (b).



## Conclusions

In the present article, we described the prediction process for mapping the spatial distribution of soil salinity in Chah-Afzal, central Iran, using a symbolic regression technique. We believe to the best of our knowledge, that this study is the first attempt to adopt a genetic programming for digital mapping of soil salinity in an arid region of Iran. Its use for other DSM studies could be invaluable given the relatively high accuracy of the modeling based on an external validation. During the application of GP, we used a variety of auxiliary variables including the predicted maps of  $EC_{av}$  and  $EC_{ah}$ ,  $ETM^+$  images and terrain variables.  $EC_{ah}$  had the highest influence on the model prediction followed by salinity index, wetness index, red band, and MrVBF. Overall, fine-resolution soil salinity maps are useful for many soil and environmental scientists and land managers in Iran. Therefore, we recommend the use of the approach applied for the study area (i.e., Genetic programming) to map the soil salinity in other parts of Iran.

## References

- Aitkenhead, M. J., M. C. Coull, W. Towers, G. Hudson, and H. I. J. Black. 2012. Predicting soil chemical composition and other soil parameters from field observations using a neural network. *Computers and Electronics in Agriculture* 82: 108–16. doi:10.1016/j.compag.2011.12.013
- Akrakhanov, A., and P. L. G. Vlek. 2012. The assessment of spatial distribution of soil salinity risk using neural network. *Environmental Monitoring and Assessment* 184: 2475–85. doi:10.1007/s10661-011-2132-5
- Allbed, A., L. Kumar, and P. Sinha. 2014. Mapping and modelling spatial variation in soil salinity in the Al Hassa oasis based on remote sensing indicators and regression techniques. *Remote Sensing* 6: 1137–57. doi:10.3390/rs6021137
- Behrens, T., H. Forester, T. Scholten, U. Steinrucken, and E. D. Spies. 2005. Digital soil mapping using artificial neural networks. *Journal of Plant Nutrient and Soil Science* 168: 21–33. doi:10.1002/jpln.200421414
- Bilgili, A. V., M. A. Cullu, H. van Es, A. Aydemir, and S. Aydemir. 2011. The use of hyperspectral visible and near infrared reflectance spectroscopy for the characterization of salt-affected soils in the Harran Plain, Turkey. *Arid Land Research and Management* 25: 19–37. doi:10.1080/15324982.2010.528153
- Boettinger, J. L., R. D. Ramsey, J. M. Bodily, N. J. Cole, S. Kienast-Brown, S. Nield, J. Saunders, and A. K. Stum. 2008. Landsat spectral data for digital soil mapping. In *Digital soil mapping with limited data*, ed. A. E. Hartemink, A. B. McBratney, and M. L. Mendonca-Santos, 193–203. Rio de Janeiro, Brazil: Springer.
- Brumby, S. P., J. Theiler, S. Perkins, N. R. Harvey, and J. J. Szymanski. 2001. *Genetic programming approach to extracting features from remotely sensed imagery*. Montreal, QC, Canada: International Conference Image Fusion (FUSION).
- Bui, E. N., and C. J. Moran. 2001. Disaggregation of polygons of surficial geology and soil maps using spatial modelling and legacy data. *Geoderma* 103: 79–94. doi:10.1016/s0016-7061(01)00070-2
- Chavez, P. S. 1996. Image-based atmospheric corrections-revisited and improved. *Photogrammetry and Remote Sensing* 62: 1025–35.
- Chhabra, R. 2006. Classification of salt-affected soils. *Arid Land Research and Management* 19: 61–79. doi:10.1080/15324980590887344
- Coopersmith, E. V., B. S. Minsker, C. E. Wenzel, and B. J. Gilmore. 2014. Machine learning assessments of soil drying for agricultural planning. *Computers and Electronics in Agriculture* 104: 93–104. doi:10.1016/j.compag.2014.04.004
- Ding, J., and D. Yu. 2014. Monitoring and evaluating spatial variability of soil salinity in dry and wet seasons in the Werigan-Kuqa Oasis, China, using remote sensing and electromagnetic induction instruments. *Geoderma* 235–36: 316–22. doi:10.1016/j.geoderma.2014.07.028

- Fonlupt, C., and D. Robilliard. 2000. Genetic programming with dynamic fitness for a remote sensing application. *Computer Science* 1917: 191–200. doi:10.1007/3-540-45356-3\_19
- Gallant, J. C., and T. I. Dowling. 2003. A multiresolution index of valley bottom flatness for mapping depositional areas. *Water Resource Research* 39: 1347–60. doi:10.1029/2002wr001426
- Gee, G. W., and J. W. Bauder. 1986. Particle size analysis. In *Methods of soil analysis: Part 1: Agronomy handbook no 9*, ed. A. Klute 383–411. Madison, WI: American Society of Agronomy and Soil Science Society of America.
- Hall, M. 1999. *Correlation-based feature selection for machine learning*, 198. Hamilton, New Zealand: The University of Waikato.
- Hall, M., E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. 2009. The WEKA data mining software: An update. *ACM SIGKDD Explorations Newsletter* 11: 10–18. doi:10.1145/1656274.1656278
- Hamzehpour, N., M. K. Eghbal, P. Bogaert, N. Toomanian, and R. S. Sokouti. 2013. Spatial prediction of soil salinity using kriging with measurement errors and probabilistic soft data. *Arid Land Research and Management* 27: 128–39. doi:10.1080/15324982.2012.724144
- Heung, B., C. E. Bulmer, and M. G. Schmidt. 2014. Predictive soil parent material mapping at a regional-scale: A random forest approach. *Geoderma* 214–15: 141–54. doi:10.1016/j.geoderma.2013.09.016
- Huete, A. R. 1988. A soil-adjusted vegetation index (SAVI). *Remote Sensing Environment* 25: 295–309. doi:10.1016/0034-4257(88)90106-x
- Jafari, A., P. A. Finke, J. Vande Wauw, S. Ayoubi, and H. Khademi. 2012. Spatial prediction of USDA-great soil groups in the arid Zarand region, Iran: Comparing logistic regression approaches to predict diagnostic horizons and soil types. *European Journal of Soil Science* 63: 284–309. doi:10.1111/j.1365-2389.2012.01425.x
- Jafari, A., H. Khademi, P. A. Finke, J. Van de Wauw, and S. Ayoubi. 2014. Spatial prediction of soil great groups by boosted regression trees using a limited point dataset in an arid region, southeastern Iran. *Geoderma* 232–34: 148–63. doi:10.1016/j.geoderma.2014.04.029
- Johari, A., G. Habibagahi, and A. Ghahramani. 2006. Prediction of soil–water characteristic curve using genetic programming. *Journal of Geotechnical and Geoenvironmental Engineering* 132: 661–65. doi:10.1061/(asce)1090-0241(2006)132:5(661)
- Kovacevic, M., B. Bajat, and B. Gajic. 2010. Soil type classification and estimation of soil properties using support vector machines. *Geoderma* 154: 340–47. doi:10.1016/j.geoderma.2009.11.005
- Koza, J. 1992. *Genetic programming: On the programming of computers by means of natural selection*. Cambridge, MA: The MIT Press.
- Koza, J. R. 2010. Human-competitive results produced by genetic programming. *Genetic Programming and Evolvable Machines* 11(3–4): 251–84. doi:10.1007/s10710-010-9112-3
- Koza, J., H. Bennett, D. Andre, and M. Keane. 1999. *Genetic programming III: Darwinian invention and problem solving*. Burlington, MA: Morgan Kaufmann.
- Lesch, S. M., J. Herrero, and J. D. Rhoades. 1998. Monitoring for temporal changes in soil salinity using electromagnetic induction techniques. *Soil Science Society of America Journal* 62(1): 232–42. doi:10.2136/sssaj1998.03615995006200010030x
- Li, H. Y., Z. Shi, R. Webster, and J. Triantafilis. 2013. Mapping the three-dimensional variation of soil salinity in a rice-paddy soil. *Geoderma* 195–96: 31–41. doi:10.1016/j.geoderma.2012.11.005
- Li, M., J. Im, and C. Beier. 2013. Machine learning approaches for forest classification and change analysis using multi-temporal Landsat TM images over Huntington Wildlife Forest. *GIS Science and Remote Sensing* 50(4): 361–84.
- Malone, B. P., A. B. McBratney, B. Minasny, and G. M. Laslett. 2009. Mapping continuous depth functions of soil carbon storage and available water capacity. *Geoderma* 154: 138–52. doi:10.1016/j.geoderma.2009.10.007
- Mariappan, V. E. N. 2010. Soil salinity assessment using geospatial technology, perspectives, approaches and strategies. *Indian Cartography* 30: 25–30.
- Mathworks. 2010. *Matlab version 7.0*. Natick, MA: The Mathworks Inc.

- McBratney, A. B., M. L. Mendonça-Santos, and B. Minasny. 2003. On digital soil mapping. *Geoderma* 117(1–2): 3–52.
- Metternicht, G., and J. A. Zinck. 2008. Spectral behavior of salt types. In *Remote sensing of soil salinization: Impact on land management*, ed. G. Metternicht, and J. A. Zinck, 21–36. Boca Raton, FL: CRC Press.
- Minasny, B., A. B. McBratney, and B. M. Whelan. 1999. Vesper Version 1.0. Australian Centre for Precision Agriculture. The University of Sydney: NSW, Australia, 2006, <http://www.usyd.edu.au/su/agric/acpa> (accessed May 10, 2013).
- Mollazade, K., M. Omid, and A. Arefi. 2012. Comparing data mining classifiers for grading raisins based on visual features. *Computer and Electronic in Agriculture* 84: 124–31. doi:10.1016/j.compag.2012.03.004
- Moore, I. D., R. B. Grayson, and A. R. Ladson. 1991. Digital terrain modeling: Review of hydrological, geomorphological, and biological applications. *Hydrology Processing* 5: 3–30. doi:10.1002/hyp.3360050103
- National Cartographic Center. 2010. Research Institute of NCC: Tehran, Iran. [www.ncc.org.ir](http://www.ncc.org.ir) (accessed June 15, 2014).
- Nelson, R. E. 1982. Carbonate and gypsum. In *Methods of Soil Analysis: Part 2: Chemical Methods*, ed. A. L. Page, 181–97. Madison, WI: American Society of Agronomy and Soil Science Society of America.
- Nelson, D. W., and L. P. Sommers. 1986. Total carbon, organic carbon and organic matter. In *Methods of soil analysis: Part 2: Chemical methods*, ed. A. L. Page, 539–79. Madison, WI: American Society of Agronomy and Soil Science Society of America.
- Nemes, A., W. J. Rawls, and Y. A. Pachepsky. 2006. Use of the nonparametric nearest neighbor approach to estimate soil hydraulic properties. *Soil Science Society of America Journal* 70: 327–36. doi:10.2136/sssaj2005.0128
- Nemes, A., J. H. M. Wosten, A. Lilly, and J. H. Oude Voshaar. 1999. Evaluation of different procedures to interpolate particle-size distributions to achieve compatibility within soil databases. *Geoderma* 90: 187–202. doi:10.1016/s0016-7061(99)00014-2
- Nield, S. J., J. L. Boettinger, and R. D., Ramsey. 2007. Digitally mapping gypsum and nitric soil areas using Landsat ETM data. *Soil Science Society of America Journal* 71: 245–52. doi:10.2136/sssaj2006-0049
- Olaya, V. 2004. A Gentle Introduction to SAGA GIS. p. 216.
- Omid, M., A. Mahmoudi, and M. H. Omid. 2010. Development of pistachio sorting system using principal component analysis (PCA) assisted artificial neural network (ANN) of impact acoustics. *Expert System Application* 37: 7205–12. doi:10.1016/j.eswa.2010.04.008
- Padarian, J., B. Minasny, and A. McBratney. 2012. Using genetic programming to transform from Australian to USDA/FAO soil particle-size classification system. *Soil Research* 50: 443–46. doi:10.1071/sr12139
- Parasuraman, K., A. Elshorbagy, and B. C. Si. 2007. Estimating saturated hydraulic conductivity using genetic programming. *Soil Science Society of America Journal* 71: 1676–84. doi:10.2136/sssaj2006.0396
- Pearson, R. L., and L. D. Miller. 1972. Remote mapping of standing crop biomass for estimation of the productivity of the short-grass Prairie, Pawnee National Grassland, Colorado: 8th International Symposium on Remote Sensing of Environment, October 2–6. pp. 1357–81.
- Poli, R., W. B. Langdon, and N. F. McPhee. 2008. *A field guide to genetic programming* (with contributions by J. R. Koza). <http://www.gp-field-guide.org.uk> (accessed December 10, 2015).
- Puente, C., G. Olague, S. V. Smith, S. H. Bullock, A. Hinojosa-Corona, and M. A. González-Botello. 2011. A genetic programming approach to estimate vegetation cover in the context of soil erosion assessment. *Photogrammetric Engineering & Remote Sensing* 77(4): 363–75. doi:10.14358/pers.77.4.363
- Richards, L. 1954. Determination of the properties of saline and alkali soils. In *Diagnosis and improvement of saline and alkali soils, agriculture handbook*, no. 60, 7–33. Riverside, CA: US Regional Salinity Laboratory.
- Rouse, J. W., R. H. Hass, J. A. Schell, and D. W. Deering. 1973. Monitoring vegetation systems in the Great Plains with ERTS. In *NASA SP-351: Proc. Third Earth resources Tech. Satellite-Symp. Vol. 1:*

- Technical Presentations Sec. A*, ed. S. C. Freden, E. P. Mercanti, and M. A. Becker, 309–17. Washington, DC: NASA Science and Technology Information Office.
- Shamsi, F. R. S., Z. Sanaz, and A. S. Abtahi. 2013. Soil salinity characteristics using moderate resolution imaging spectroradiometer (MODIS) images and statistical analysis. *Archive of Agronomy and Soil Science* 59: 471–89. doi:10.1080/03650340.2011.646996
- Sheng, J., L. Ma, P. Jiang, B. Li, F. Huang, and H. Wu. 2010. Digital soil mapping to enable classification of the salt-affected soils in desert agro-ecological zones. *Agricultural Water Management* 97: 1944–51. doi:10.1016/j.agwat.2009.04.011
- Slavich, P. G. 1990. Determining ECa-depth profiles from electromagnetic induction measurements. *Australian Journal of Soil Research* 28: 453–63. doi:10.1071/sr9900443
- Sparks, D. L., A. L. Page, P. A. Helmke, R. H. Leppert, P. N. Soltanpour, M. A. Tabatabai, G. T. Johnston, and M. E. Summer. 1996. *Methods of soil analysis*. Madison, WI: Soil Science Society of America.
- Stum, A. K., J. L. Boettinger, M. A. White, and R. D. Ramsey. 2010. Random forests applied as a soil spatial predictive model in arid Utah. In *Digital soil mapping: Bridging research, environmental application, and operation. Progress in soil science*, ed. J. L. Boettinger, D. W. Howell, A. C. Moore, A. Hartemink, and E. S. Kienast-Brown, 179–89. Logan, UT: Springer.
- Taghizadeh-Mehrjardi, R., B. Minasny, F. Sarmadian, and B. P. Malone. 2014. Digital mapping of soil salinity in Ardakan region, central Iran. *Geoderma* 213: 15–28. doi:10.1016/j.geoderma.2013.07.020
- Urdanoz, V., and R. Aragüés. 2011. Pre- and post-irrigation mapping of soil salinity with electromagnetic induction techniques and relationships with drainage water salinity. *Soil Science Society of America Journal* 75(1): 207–15. doi:10.2136/sssaj2010.0041
- Yang, M. D. 2007. A genetic algorithm (GA) based automated classifier for remote sensing imager. *Remote Sensing* 33(3): 203–13. doi:10.5589/m07-020