International Journal of Research in Agriculture and Forestry
Volume 7, Issue 3, 2020, PP 21-34
ISSN 2394-5907 (Print) & ISSN 2394-5915 (Online)

Evaluating Linear and Nonlinear Regression Models in Mapping Soil Salinity

Luis A. Garcia[1*] and Ahmed A. Eldeiry[2]

[1]Luis A. Garcia, Professor, Department of Civil and Environmental Engineering, College of Engineering and Mathematical Sciences, University of Vermont, Burlington, USA
[2]Ahmed A. Eldeiry, Ph.D. Former Research Scientist, Department of Civil and Environmental Engineering, Colorado State University, Fort Collins, USA

***Corresponding Author:** Luis A. Garcia, Luis A. Garcia, Professor, Department of Civil and Environmental Engineering, College of Engineering and Mathematical Sciences, University of Vermont, Burlington, USA, Email: Luis.Garcia@UVM.edu*

ABSTRACT

In this study a number of Linear and Nonlinear Regression models are evaluated in the context of mapping soil salinity. The linear regression models evaluated were Ordinary Least Squares (OLS) and the Generalized Linear Model(GLM)while the nonlinear regression models were Multivariate Adaptive Regression Spline (MARS) and Artificial Neural Networks (ANN).These models were applied to six soil salinity datasets collected in a study area in the Lower Arkansas River Basin in Colorado. The dataset consisted of three fields where alfalfa was growing and three fields where corn was growing. Multispectral IKONOS satellite images, which covered both the corn and alfalfa fields, provided remote sensing data. The objectives of this study were: a) evaluate different regression models and determine the best one to use in mapping soil salinity; b)assess the contribution from remote sensing data in mapping soil salinity; c)compare the contribution of remote sensing data to improving the mapping of soil salinity infields planted with alfalfa vs fields planted with corn; and d) investigate the impact that spatial patterns, autocorrelation, and the distribution of soil salinity samples collected in a field have on the performance of the different regression models. To evaluate the performance of the regression models the Nash-Sutcliffe Efficiency (NSE) and the Root Mean Square Error (RMSE) statistical methods were used. In this study the nonlinear regression models performed better than the linear regression models for mapping soil salinity and overall the ANN performed the best. In addition, this study found that remote sensing data (e.g. IKONOS satellite images) have the potential to contribute information to improve the mapping of soil salinity. The results showed that: 1) the remote sensing contribution was more significant in fields planted with cornas compared with fields planted with alfalfa; 2) the higher the range of soil salinity in the study area, the higher the autocorrelation; and3) the performance of the regression models improves the closer to a normal distribution the soil salinity data collected is.

Keywords: *Soil salinity; Mapping; Remote Sensing; Regression.*

INTRODUCTION

Salts decrease the amount of soil water that plants can extract due to an increase in the osmotic potential, this has an adverse effect on the plant's metabolism (Douaik et al. 2004). Therefore, it is imperative that in areas that have the potential to be affected by soil salinity plans are developed to manage it. One of the first steps in developing a soil salinity management plan is to map the soil salinity to quantify its spatial extent.

Regression models can be used to generate the spatial extent of soil salinity, if some predictive data is available (e.g. remote sensing).Two types of regression models can be used: linear and nonlinear. Linear regression models involve modeling the relationship between a dependent variable (e.g. soil salinity) and one or more independent variables (e.g. IKONOS satellite image bands). In the case that one independent variable is used (e.g. one band of an IKONOS satellite image), the process is called simple linear regression. For more than one independent variable (e.g. more than one band of an IKONOS satellite image), the process is called multiple linear regression (Freedman, 2009). In multivariate or multiple linear regressions, multiple correlated dependent variables are predicted, while in simple linear regression a single scalar variable is predicted (Rencher and Christensen, 2012).

The first linear regression model used in this study is an Ordinary Least Squares (OLS) model which is one of the most widely used regression models. OLS is based on studying the relationship between two or more variables (Gujarat, 2003; Upton et al, 2002). This empirical model assumes that the model's error term is normally, independently, and identically distributed (i.i.d.). OLS yields the most efficient unbiased estimators for the regression model's coefficients. However, when there is a certain degree of correlation between the residuals, the OLS model can be misleading.

The second linear regression model used in this study is the Generalized Linear Model (GLM), which is a flexible generalization of the OLS model and allows for response variables that have error distribution models other than a normal distribution. Generalized linear models were formulated as a way of unifying various other statistical models, including Linear, Logistic, and Poisson regressions (Nelder and Wedderburn, 1972).

Nonlinear regression differs from linear regression in that the least-squares estimators of their parameters are not:1) unbiased; 2) normally distributed; and3) minimum variance estimators. In nonlinear regression, the observed data is modeled by a function which is a nonlinear combination of model parameters and depends on one or more independent variables. A Multivariate Adaptive Regression Splines (MARS) model is evaluated in this study. The MARS model was introduced by Jerome H. Friedman in 1991 (Friedman, 1991). MARS is a generalization of recursive partitioning that uses spline fitting instead of other simple fitting functions (Breiman et al., 1984; and Lewis and Stevens 1991). MARS works by fitting a model in the form of an expansion in product spline basis functions whose predictors are chosen using a forward and backward recursive partition strategy. MARS produces continuous models for high dimensional data that can have multiple partitioning and predictor variable interactions (Lewis and Stevens, 1991).

This study also evaluated the performance of an Artificial Neural Network (ANN) to predict soil salinity. ANN's are a computational model typically organized in layers and made up of a number of interconnected nodes which contain an activation function. They are inspired by the biological neural networks that comprise animal brains. These models often consist of a large number of neurons, which are simple linear or nonlinear computing elements, frequently interconnected in complex ways and commonly organized into layers (Sarle 1994). The input layer communicates to one or more hidden layers where the actual processing is done via a system of weighted connections. The hidden layers link to an output layer where the answer is the output. The mathematical structure of ANN's is capable of identifying complex nonlinear relationships between input and output data sets and has been found to be useful and efficient, particularly in problems that are difficult to describe using physical equations (Hsu K. et al., 1995).The multilayer perception of ANN's can be considered as nonlinear regression and discriminant models can be implemented with standard statistical software (Sarle 1994).

Use of remote sensing of surface features to identify and map salt affected areas has been used extensively (Allbed, et al., 2014; Abbas, et al., 2013; Wu, et al., 2008; and Robbins and Wiegand 1990). McColl et al. (2012) mentioned that spatial resolution, clouds, surface roughness and vegetation cover affect the use of remote sensing data. Eldeiry and Garcia (2008, 2010, and 2011) stated that integrating geostatistical techniques with remote sensing data has great potential in estimating soil salinity. Wiegand et al. (1994) developed a procedure for using soil salinity, plant information, digitized color infrared aerial photography and videography to help determine soil salinity. Color and thermal infrared aerial photography as well as spectral image interpretation techniques have also been used for mapping surface land salinity (Abuelgasim and Ammad, 2019;Abbasm 2013; and Spies and Woodgate, 2004).

In order to determine field scale variation and heterogeneity, several Electromagnetic Induction (EMI) devices are currently utilized such as EM-31, EM-34, and EM-38, all of which were developed by Geonics Ltd. in Mississauga, ON, Canada. Robinson et al. (2009) used EMI to identify zones of water depletion and accumulation. Eldeiry and Garcia (2011, 2012a, and 2012b) used a variety of techniques to map soil salinity and manage crop yield. Zhu et al. (2010a, b) relied on repeated EMI surveys to detect the spatial variations of soil moisture, soil texture, soil type and subsurface flow paths. Saey et al. (2013) demonstrated the potential of multi-receiver EMI soil surveys to map and interpret the soil landscape and to discern archaeological as well as small and natural features.

It is important to analyze the spatial point pattern, autocorrelation, and distribution of the collected samples before any regression or interpolation of the data. Diggle (2003) classifies spatial point patterns in three main classes: aggregation (clustering), regularity (inhibition) and complete spatial randomness (CSR). Ripley's K-function is a spatial analysis method used to describe how point patterns occur over a given area of interest. This method estimates the expected number of random points within a distance (*r*) of a randomly chosen point within a plot and it is typically used to compare a given point distribution with a random distribution (Ripley, 1976).Another method to describe the pattern of the points are Moran scatter plots which provide a tool for visual exploration of spatial autocorrelation (Anselin 1966 and 2002). Anselin (2002) describes Moran scatter plots as the spatial lag (the average values of a location's neighbors) of the variable on the vertical axis and the original variable on the horizontal axis. In addition, histograms can be used as a quality control tool, since they provide a representation of the distribution of numerical data and an estimate of the probability distribution of a continuous variable (Tague, 2004; and Pearson, 1985).

The approach presented in this study involves integrating remotely sensed data, ground truth data (soil salinity data) and regression techniques to improve the mapping of soil salinity. The cross-correlation between soil salinity data and remote sensing data was established using reflection from crop cover as an indicator of soil salinity. Four regression models were used to regress the data derived from the IKONOS images with the soil salinity samples collected in a number of fields. Integration of remote sensing data with field data can minimize time and labor during the collection of field data and take advantage of information derived from the remote sensing data. Furthermore, this study compared relatively commonly used models, such as OLS and GLM to more sophisticated ones, such as MRAS and ANN's.

Materials and Methods

The Study Area

The study area is located in southeastern Colorado, near the town of La Junta (Figure 1). Fields in this area are planted with corn, alfalfa, wheat, cantaloupe, onions and other vegetables, and are irrigated by a multitude of irrigation methods including a mixture of border and basin, center pivots, and a few drip systems. Salinity levels in the irrigation canal systems along the river increase from 300 ppm total dissolved solids (TDS) near Pueblo to over 4,000 ppm at the Colorado-Kansas border (Gates et al. 2002). In this area, Colorado State University (CSU) has conducted an intensive field data collection effort that includes a depth to water table, irrigation amounts, crop yields, rainfall data, and soil salinity data. This study uses some of the soil salinity data that was collected in intensely monitored fields where in 2001 corn was planted and fields where in 2004 alfalfa was planted. IKONOS satellite images were acquired for both years to cover the area where the fields are located.

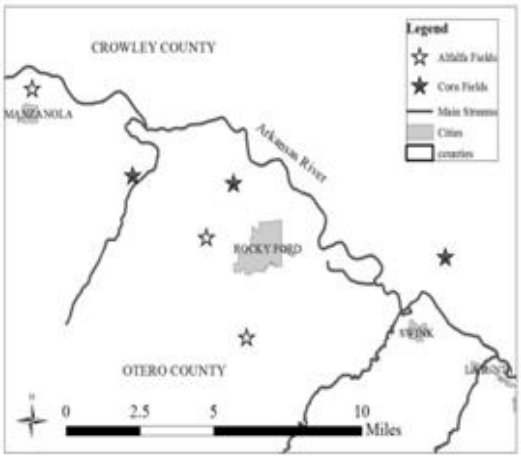

Figure1. *The study area with the location of the fields planted with alfalfa and corn*

Data Collection

Soil salinity was measured in the fields using an EM-38 electromagnetic probe. The EM-38 takes vertical and horizontal readings that can be converted to soil salinity estimates. The EM-38 has the ability to quickly cover large areas without ground electrodes and it provides depths of exploration of 1.5 meters in the vertical direction and 0.75 meters in the horizontal direction, respectively. Two soil salinity datasets were collected using EM-38 probes, the first data set consists of 257 points and the second consists of 181 points. Table 1 contains the summary of the data collected in the monitored fields that were used in this study. The table shows that the EM-38 soil salinity estimated for these fields ranges from 2.60 to 12.72 dS/m. The data also shows that the range of soil salinity in the alfalfa fields is higher than that in the corn fields.

Table1. *Description of the soil salinity data collected using an EM-38 in alfalfa and corn fields*

Dataset	# of samples	Average (dS/m)	STDEV (dS/m)	Min. (dS/m)	Max. (dS/m)
Alfalfa (2004)					
Field 04	71	9.4	6.0	2.7	20.7
Field 10	59	6.1	3.5	3.1	13.2
Field 14	46	4.7	1.0	3.1	6.8
Corn (2001)					
Field 09	108	3.0	0.2	2.6	3.7
Field 40	80	6.2	2.6	3.0	12.2
Field 80	68	5.1	2.0	2.7	11.7

The data in Table 1 shows a wide range of standard deviation (STDEV) from a low of 0.2 dS/m to a high of6.0 dS/m. In addition, the range of soil salinity varies from an arrow range of 2.6to3.7 dS/m to a wide range of 2.7 to 20.7 dS/m. The salinity content of the collected soil samples in the alfalfa fields show a higher range and standard deviation than the soil samples collected in the corn fields, which can be attributed to alfalfa's higher tolerance to soil salinity than corn. The number of soil salinity samples collected in each field was considered to be sufficient to properly run the regression and interpolation models.

The remote sensing images used in this research are IKONOS satellite images, which have three visible spectral bands (blue, green, and red), one near infrared (NIR) band and a panchromatic band. The spatial resolution of the blue, green, red, and NIR bands is 4 m, while the panchromatic band resolution is 1m.The spectral resolution of the blue, green, red, and NIR bands are as follows: 445-516, 508-595, 632-698, and 757-853 nm, respectively. The relatively high resolution of the IKONOS satellite images allows for more spatial detail in the study.

Exploratory Data Analysis

The spatial point pattern, autocorrelation, and distribution of the soil salinity data that was used in this study is discussed in this section. Figures 2 and 3 display the plots of Ripley's K, Moran's I and the Histograms of the soil salinity for the alfalfa and corn fields. The first row of each figure displays the locations of the points where soil salinity was collected, the second row displays the Riley's K-function, the third row displays the Moran's I, and the fourth row displays the Histograms. The summary of the plots for the Riley's K-function include the upper and lower simulation envelopes for 99 simulated realizations. The x-axis represents the radii (r) for the calculated simulations, while the y-axis displays the Riley's K-function. The dotted line is the expected value for a random pattern (CSR) and the solid black line is the observed count. If the solid black line, which represents the set of data points, extends below the lower envelope for a Poisson distribution (grey area on the plot), this suggests that the points are distributed regularly as opposed to randomly. For the alfalfa fields, the distribution plots of the Ripley's K-function are regular at the ranges of 50 ~ 150, 13 ~ 23, and 22 ~ 50 meters for field 04, field 10, and field 14, respectively. The distribution is random for the rest of the data ranges. The maximum deviation of the observed K-function from the complete spatial randomness (CSR) shows the distance where maximum regularity occurs; this occurs at 100 m for field 04, 18 m for field 10, and 35 m for field 14.For the corn fields, the plots of the Ripley's K-function demonstrate that the distributions are regular at the ranges of 22 ~ 60 m for field 09, 25 ~ 45 m for field 40, and 20 ~ 50 m for field 80. The patterns are random for the rest of data ranges(the dotted line is above the solid black line). The maximum regularity occurs at approximately 55 m for field 09, 37 m for field 40, and 42 m for field 80.From observing the Ripley's K-function results the conclusion is that the spatial pattern for both alfalfa and corn fields is not CSR and that regularity occurs in about half of the sample distance ranges.

In the Moran's I plots, the slope of the scatter plot can be used to determine the extent of linear associations between the values at a given location (x-axis) with values of the same variable at neighboring locations (y-axis). A positive slope correlates to a positive spatial autocorrelation; high values of the variable at location *i* tend to be clustered with high values of the same variable at locations that are neighbors of *i*, and vice versa. Depending on their position on the plot, the points in the Moran's I plot express the level of spatial association of each observation with its neighbors. The points in the upper right and lower left quadrants indicate positive spatial association of values that are higher or lower

than the sample mean, and the points with greater influence are displayed as darker asterisks. Of the Moran's I plots for the alfalfa fields, Field 10 shows the strongest autocorrelation followed by Field 14, while Field 04 shows the weakest autocorrelation. For the corn fields, the Moran's I plot points for Field 40 and Field 80exhibit a strong autocorrelation and Field 09 exhibits a weak autocorrelation.

The histogram plots at the bottom of both figures illustrate the distribution of the soil salinity data that was collected for both the alfalfa and corn fields. For the alfalfa fields, Field 14 demonstrates a distribution close to a normal distribution, while Field 04 and Field 10 are right skewed. For corn fields, Field 09 and Field 80 are close to a normal distribution while Field 40 is right skewed. It should be noted that in some cases the visual inspection of histograms might be unreliable, so the significant Shapiro-Wilk test was used in addition to the histograms to compare the sample distribution to the normal distribution to ascertain whether the data shows a serious deviation. The test shows a p-value of 0.372 which is larger than 0.05, indicating that the distribution of this field is not significantly different from the normal distribution. The Shapiro-Wilk test for the rest of the fields' p-values is less than 0.05, with the exception of Field 14, suggesting their distribution is not normal.

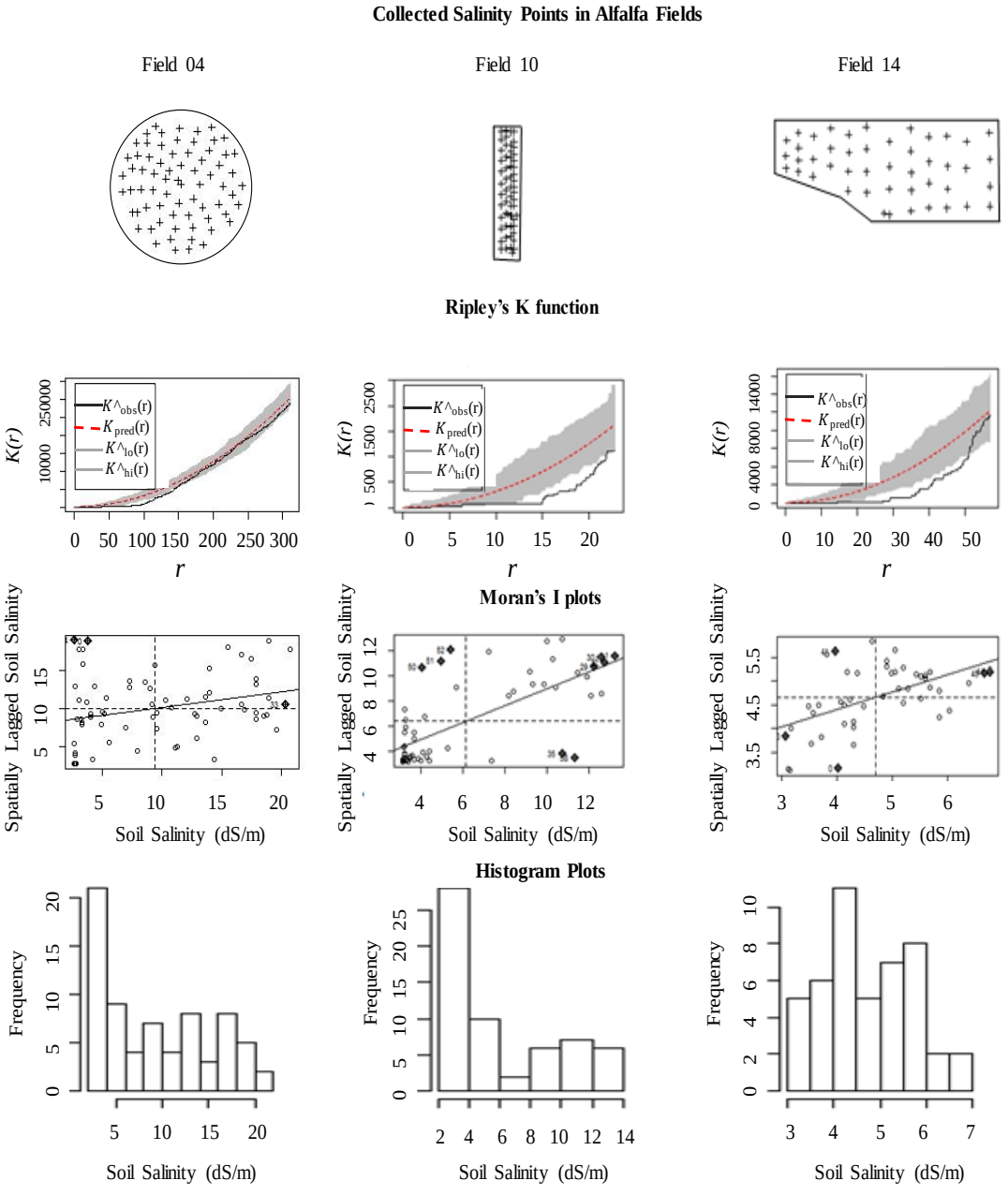

Figure2. *Map of soil sample locations, Ripley's K-function plots, Moran's I scatter plots, and Histogram plots of the three alfalfa fields.*

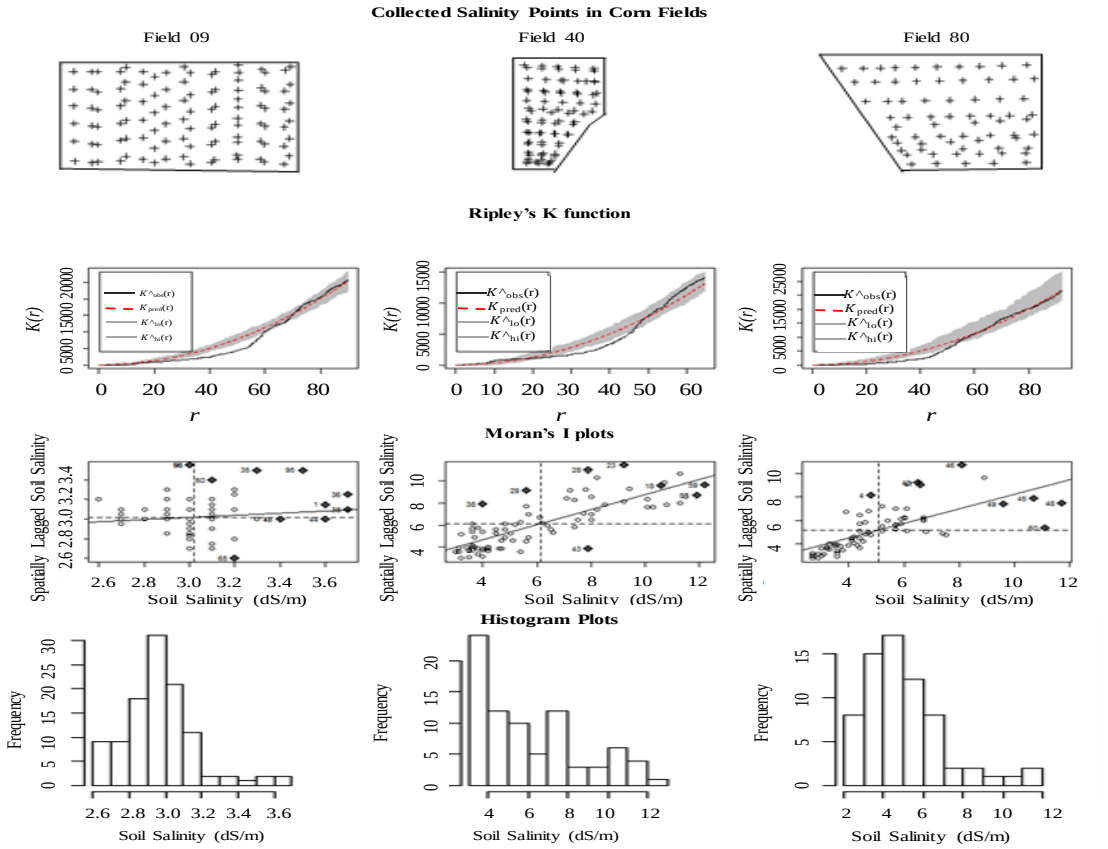

Figure3. *Map of soil sample locations, Ripley's K-function plots, Moran's I scatter plots, and Histogram plots of the three corn fields.*

Applying the Regression Models to the Data Sets

Linear Regression Models

The main assumptions of the linear regression models are that the relationship between the predictor (x) and the outcome (y) is linear and that the residual errors are normally distributed. There is an additional assumption that the residuals have constant variance (homoscedasticity), which is valid for the OLS model but is waved for the GLM model. One way to find the accuracy of the regression models is to check the residuals. In some cases, the data might contain some influential observations such as outliers that might affect the result of the regression. Therefore, there is a need to verify if removing any outliers will impact the results. The assumptions of the linear regression models can be confirmed by producing some diagnostic plots that visualize the residual errors.

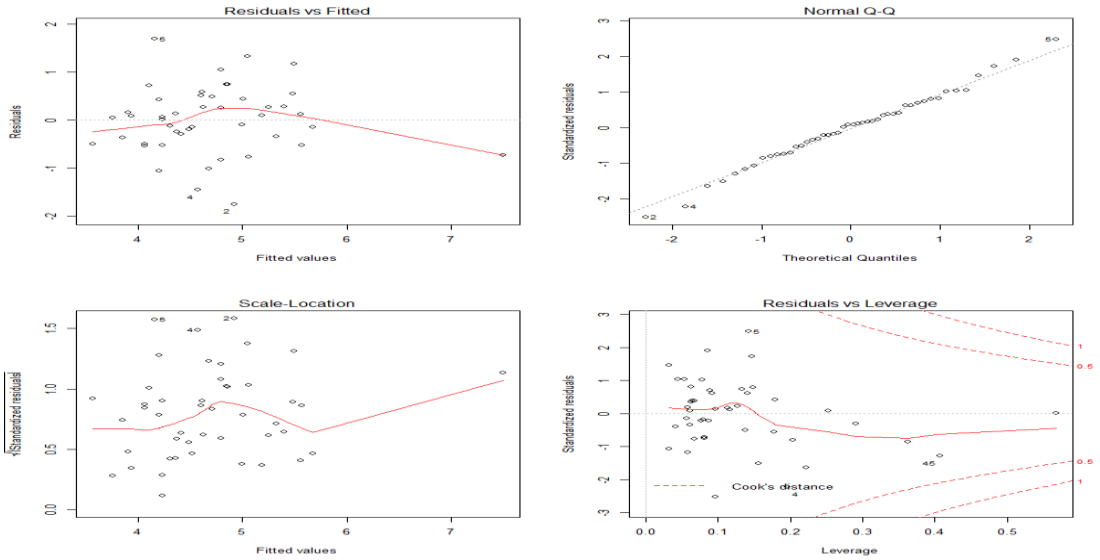

Figure4. *Diagnostic plots of the residuals of the OLS model*

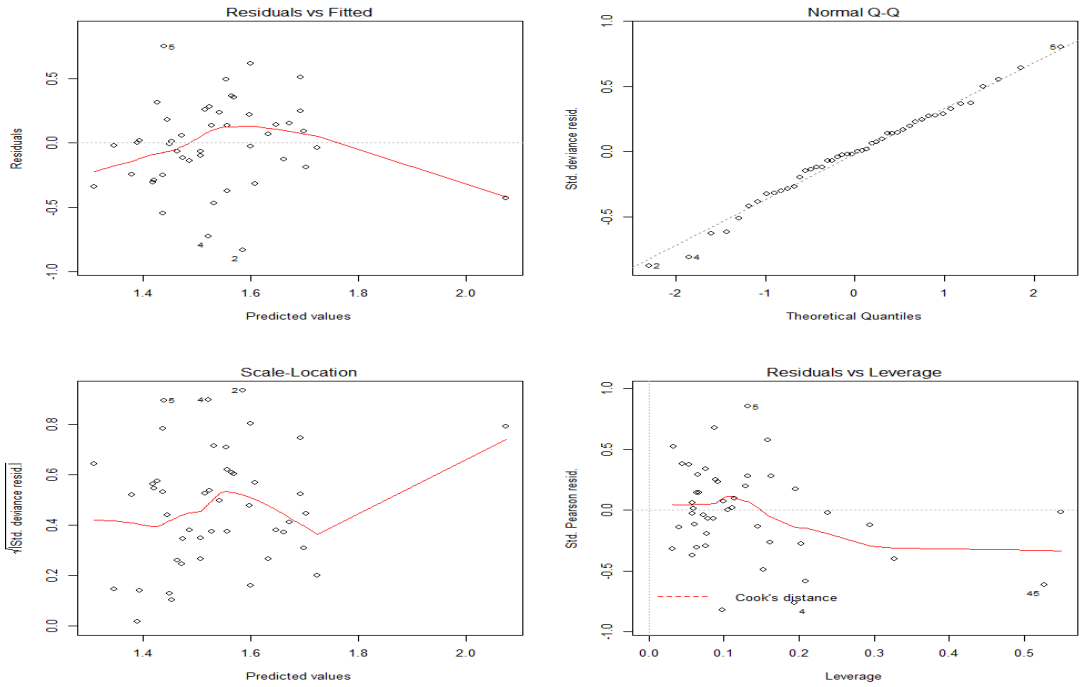

Figure5. *Diagnostic plots of the residuals of the GLM model*

Figures4and 5 show the diagnostic plots of the residuals of the OLS and GLM models. In each figure, the plot on the upper left shows the residuals versus the fitted values, which are used to check the linear relationship assumption. A horizontal red line and no distinct pattern among the points is an indication of a linear relationship, which is not present in these plots of residuals vs. fitted values for the two cases and therefore disproves the assumption of normal residuals. In both cases, there is no significant difference or improvement of the GLM over the OLS model. Also, the variation around the estimated regression line (red line) is not constant in both cases, which means the assumption of equal error variance is also disproven.

The plot on the upper right of both figures shows the normal probability plot (Normal Q-Q) of the residuals, which is used to check if the residuals are normally distributed. If the residuals follow a straight line, it is an indication of the normality of the residuals. The plots in both cases show that there are some points at the top and bottom that deviate from the straight-line pattern. These points do not severely deviate from the straight line, which implies that the residuals are close to a normal distribution.

The plot on the bottom left of both figures shows the scale or spread location. This plot is used to determine if the residuals are spread equally along the range of predictors (the homogeneity of the variance or homoscedasticity). A horizontal line with equally or randomly spread out points is an indication of a good model fit. The plot in both cases shows that the red line is not horizontal and there is also no equal or random spread of the points around the line, which means that the assumption of equal variance is also negated.

The plot on the bottom right of both figures shows the residuals versus leverage and is meant to assist in finding influential points that might impact the residuals, as not all outliers are influential in a linear regression analysis. Even though the data does have some extreme values, they might not be influential in the determination of a regression line. However, the collection of points in the upper and lower right corners can be influential against a regression line. The red dashed lines are called Cook's distance; when there are points that lie outside of Cook's distance; this is an indication that removing these points could influence the regression results. In other words, the regression results will be altered if these points are excluded. There are no points outside Cook's distance in the plot in both cases, which means there are no points that if removed would influence the results.

Overall, the analysis of the residuals for both the OLS and GLM models shows that both of them violated the assumptions of the models. This suggests that there might be some nonlinearity in the data and that both models were not able to handle it.

Linear Regression Models

Multivariate Adaptive Regression Spline Model (MARS)

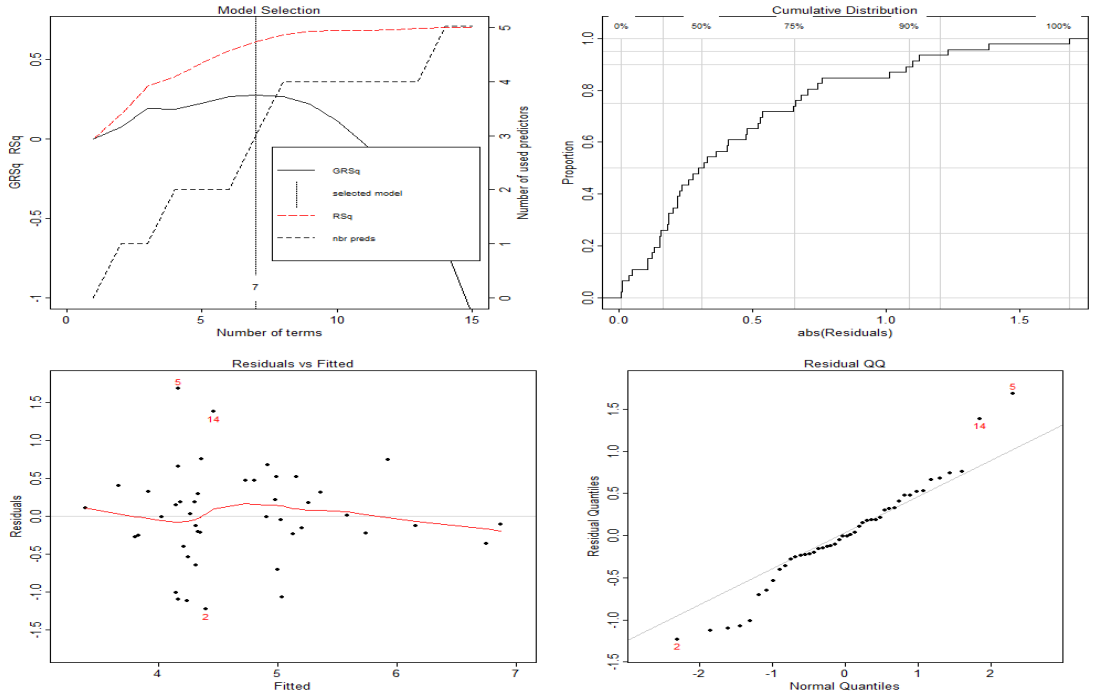

Figure6. *Diagnostic plots of the MARS model residuals*

Figure 6 shows the diagnostic plots of the MARS model residuals. The plot on the upper left shows the MARS model selection. The following two terms are related to the plot:1) RSq, which is represented by the red dotted line, normalizes the residual sum of squares (RSS). RSS varies from 0 (a model that consistently predicts the same value as the mean observed response value) to 1 (a model that predicts responses in the training data); and 2) *GRSq, which is represented by a solid black line,* normalizes the *Generalized Cross Validation* (GCV). Both the RSq and GRSq are measures of how well the model would predict values using data not included in the training set. The plot shows that RSq and GRSq initially run together and then diverge as the number of terms increases. The vertical dotted grey line, which is used to select the number of terms in the model, is positioned at the maximum GRSq and indicates that the best model has seven terms and uses all predictors. The word terms is related to the fact that MARS constructs a very large model by progressively adding basis functions or splines (interaction terms). The number of predictors is depicted by the black dotted line. The number of terms should generally be larger than the number of predictors (IKONOS satellite bands), which is set by the user when running the model.

The plot on the upper right shows the cumulative distribution of the absolute values of the residuals. The ideal model starts at 0 and rises quickly to 1. The value at the 50% vertical grey line represents the median absolute residual, which is approximately 0.35. The value at the 95% vertical grey line represents the absolute value of the residuals, which is less than 1.4. These values correspond to the training data, wherein the predicted values are within 1.4 units of the observed values 95% of the time. All absolute residual values estimated for soil salinity range between 0 and 1.5 dS/m.

The plot on the bottom left shows the scatter plot of the residuals and fitted values. The scales of the axes are intended to give an idea of the size of the residuals relative to the predicted values. Ideally, the residuals should show constant variance, meaning the residuals should remain evenly spread out as the fitted values increase. Measurement numbers 2, 5, and 14 (which are shown on the graph) can be considered as outliers that have observations that could potentially increase the residual variance in the MARS model.

The plot on the bottom right shows the residual Q-Q. There are some points in the upper and lower parts of the line which deviate from the straight line, therefore, the assumption of the normally distributed residuals is invalidated.

Artificial Neural Network Model (Ann)

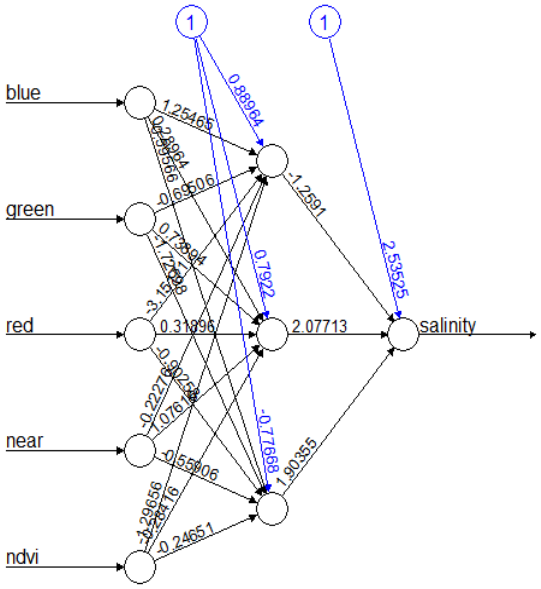

Figure7. *The architecture of a multi-layer Artificial Neural Network model (ANN)*

Figure 7 shows the architecture of a multi-layer Artificial Neural Network model (ANN). It shows the input, hidden, and output layers of an ANN. The data from IKONOS landsat image bands: blue, green, red, near, and NDVIon the left of the figure represent the input elements, also called the input vector. The hidden layers are the middle, while the output layer is the dependent variable (soil salinity) at the right side of the figure. The circles in blue represent the intercept or the constants. Every processing elementor neurode in a layer is connected to all processing elements in the next layer, with input neurodes connected to the hidden layer neurodes. This pattern continues until the neurodes in the last hidden layer are connected to the output layer neurodes. Each of these connections carry a value, commonly called a weight, which are the numbers displayed next to the lines.

Model Performance Evaluation

The following statistical parameters were used to evaluate the performance ofthe OLS, GLM, MARS, and ANN models used in this study:

- *Nash-Sutcliffe efficiency (NSE)* is a normalized statistic that determines the relative magnitude of the residual variance compared to the measured data variance (Nash and Sutcliffe, 1970). It is recommended as a performance measure by ASCE (1993) and Legates and McCabe (1999).NSE indicates how well a plot of observed versus predicted data fits a 1:1 line and is computed as shown inequation (1):

$$NSE = 1 - \left[\frac{\sum_{i=1}^{n}(Z_i^{obs} - Z_i^*)^2}{\sum_{i=1}^{n}(Z_i^{obs} - \bar{Z})^2}\right] \quad (1)$$

where Z_i^{obs} is the i^{th} observation for the constituent being evaluated,Z_i^* is the i^{th} predicted value for the constituent being evaluated,$\bar{Z}$ is the mean of the observed data for the constituent being evaluated, and n is the total number of observations. NSE values range between $-\infty$ and 1, with NSE=1 being the optimal value. Values between 0 and 1 are generally viewed as acceptable levels of performance, whereas values less than 0 indicate that the mean observed value is a better predictor than the simulated value and is regarded as unacceptable performance; and

- *Root mean square error (RMSE)*which is used to measure the prediction precision or model accuracy (Dobermann et. al., 2006; Triantafilis et. al., 2001) and is defined as shown in equation (2):

$$RMSE = 1 - \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Z_i^{obs} - Z_i^*)^2} \quad (2)$$

where Z_i^{obs} is the observed value of the i^{th} observation, Z_i^* is the predicted value of the i^{th} observation, and n is the number of points collected. The smaller the RMSE, the closer the prediction is to the measured values. The *RMSE* tends to place more emphasis on larger errors and, consequently, gives a more conservative measure of performance than the mean absolute error (MAE).

RESULTS

In this study graphical and analytical techniques are employed to evaluate the performance of the OLS, GLM, MARS, and ANN models used. Graphical techniques are used to perform an overview of the model performance and provide a visual comparison of the estimated and measured data (ASCE, 1993) and, according to Legates and McCabe Jr.(1999), graphical techniques are essential for appropriate model evaluation. In some cases, the graphical techniques might be misleading or do not provided enough information to interpret the results accurately, which is why analytical parameters are used for evaluation in addition to graphical parameters. The NSE and RMSE are used to evaluate how well the predicted data fitswith the observed data, the prediction precision, and the accuracy of the results.

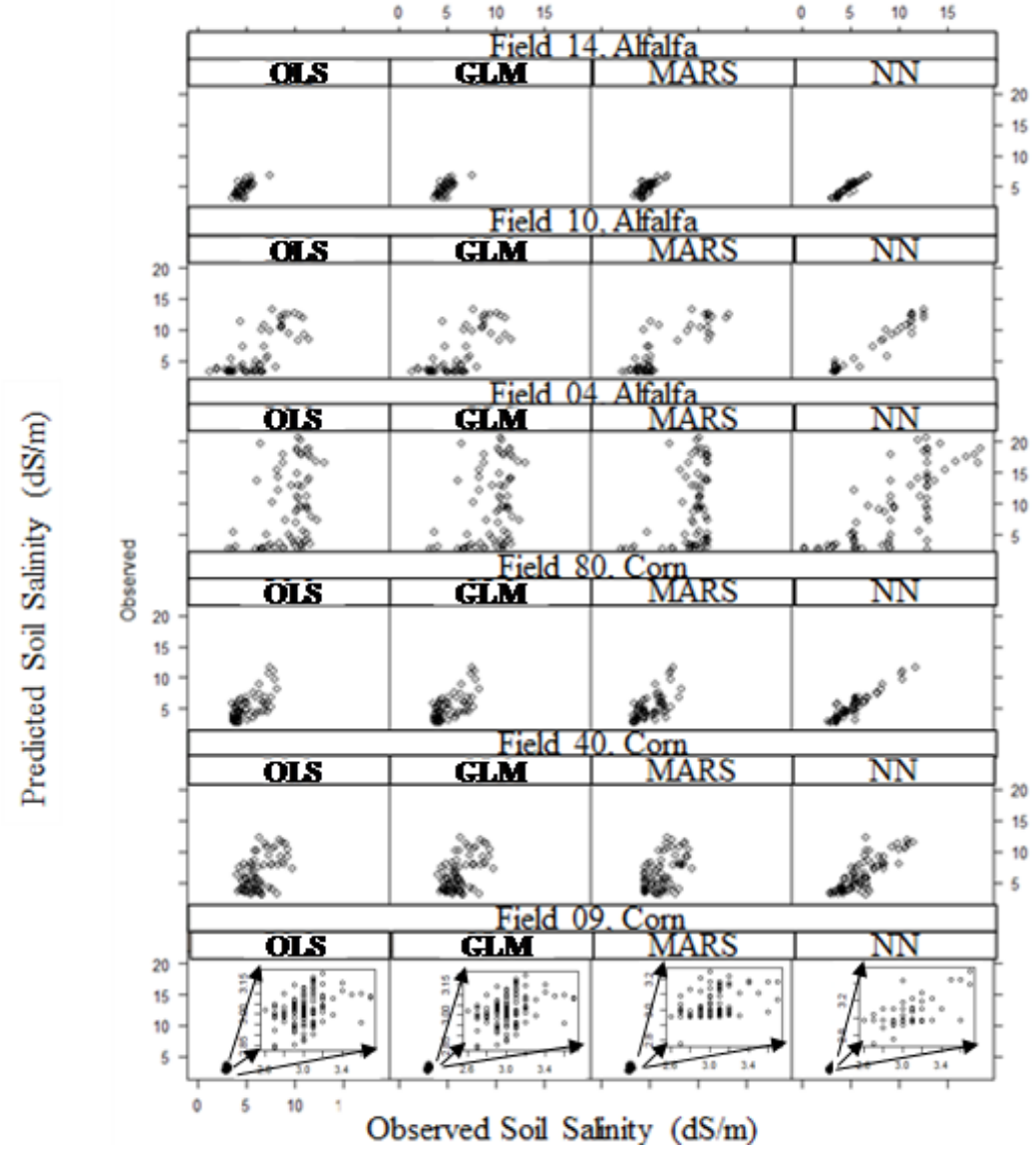

Figure8. *The scatter plots of the observed and predicted soil salinity of the alfalfa and corn fields when using the OLS, GLM, MARS, and ANN models*

Figure 8 displays the scatter plots of the observed and predicted soil salinity of the alfalfa and corn fields when using the OLS, GLM, MARS, and ANN models. Among the three alfalfa fields, Field 14 shows a strong trend between the observed and predicted data with minor differences among the four models. Field 10 shows some correlation between the observed and predicted data among the four models, while Field 04 shows poor performance among the four models. The ANN demonstrates the best performance of all four models, while the three other models show poor performance again with minor differences among them. Field 80 performed best out of the three corn fields, followed by Field 40 and lastly Field 09, which performed poorly. Of all four models, the ANN model exhibits the best overall performance; while the OLS, GLM, and GLM models exhibit overall poor performances and the differences among them are slight.

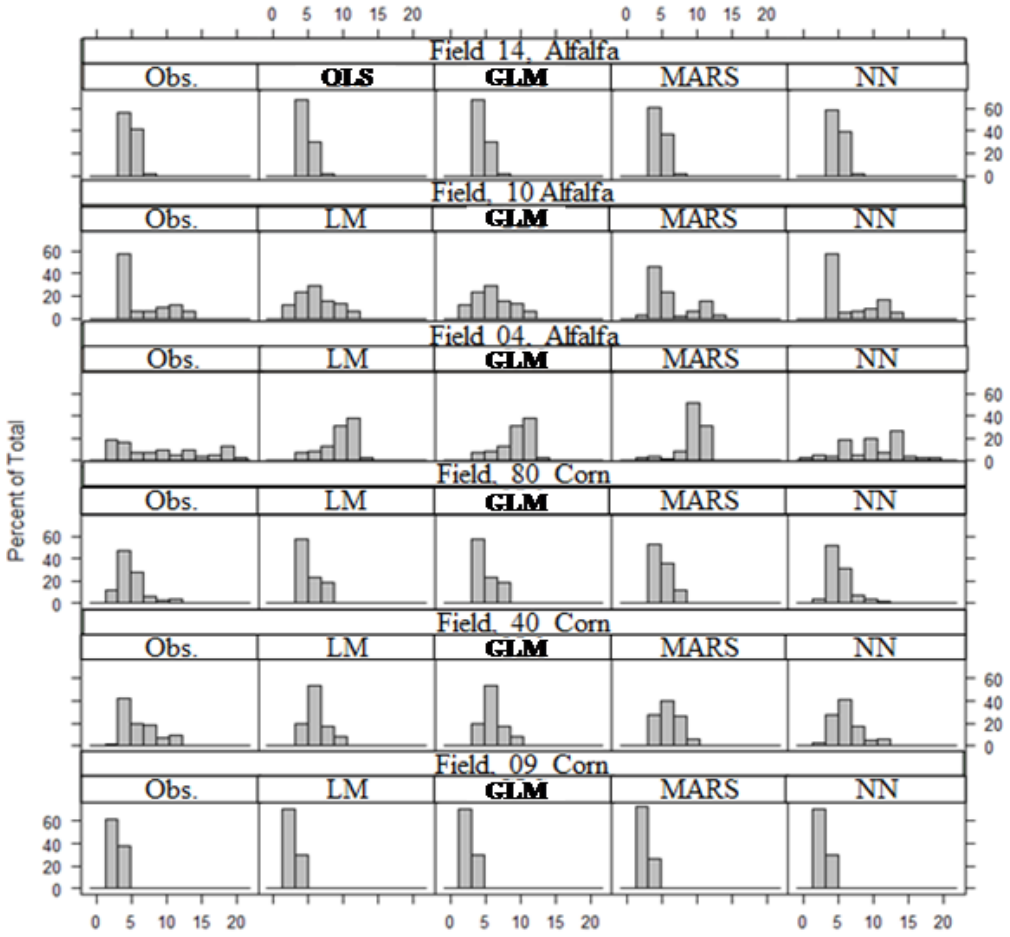

Figure9. *The histogram plots of the observed and the predicted soil salinity of the alfalfa and corn fields when using the OLS, GLM, MARS, and ANN models*

Figure 9displays the histogram plots of the observed and predicted soil salinity of the alfalfa and corn fields when using the OLS, GLM, MARS, and ANN models. The variation and shape of the distribution of the observed data set was compared to that of the predicted datasets of the different models. The alfalfa field's distributions show that there is no significant difference in the distributions between all the models for Field 14. The distribution of the predicted data using the ANN model is closest to the distribution of the observed data for the other two fields (Field 10 and Field 04). For these two fields, the distribution of the predicted data using the MARS model performs second best, while the OLS and GLM predicted data distribution is not a good match with the distribution of the observed data.

The distributions of the predicted data of the OLS and GLM models are close to the normal distribution for Field 10 and Field 04, while the observed data for these two fields are not. For the corn fields, there is no significant difference in the distribution of the predicted data using the OLS, GLM, MARS for the three fields. There is a slight improvement when using the ANN model over the other three models of these fields.

Table 2 shows the NSE and RMSE parameters that are used to evaluate the performance of the OLS, GLM, MARS, and ANN models when predicting soil salinity of the alfalfa and corn fields. The closest value of NSE to one is the best, positive values are acceptable, while negative values means that the model prediction is no better than using the mean value. Of the four models used, the ANN model shows the best NSE values, which are the closest to 1 in both the alfalfa and corn fields. For the alfalfa fields (Field 04, Field 10, and Field 14), the NSE values are 0.55, 0.96, and 0.90 respectively. Among the three alfalfa fields, both Field 10 and Field 14 have values close to 1 (0.96 and 0.90, respectively), while Field 04 has a value of 0.55. The MARS model reveals NSE values less than those of the ANN model but better than those of both the OLS and GLM models. Therefore, for the alfalfa fields, ANN performs the best followed by MARS, while both OLS and GLM have lower but still acceptable NSE values. The three corn fields, Field 09, Field 40, and Field 80, have NSE

values of 0.53, 0.67, and 0.89, respectively. Field 80 has the closest value to 1 (0.89), Field 40 has a value of 0.67, and Field 09 has a value of 0.53. The three other models (OLS, GLM, and MARS) have accepted values of the NSE parameter and they are all close to each other, so ANN performs best for the corn fields while OLS, GLM and MARS have acceptable values of NSE. Generally, the smaller the value of the RMSE, the better the performance of the model. For the alfalfa fields (Field 04, Field 10, and Field 14), the RMSE values for the ANN model are 4.01, 0.74, and 0.31, respectively. Of the four models, the ANN has the smallest values, followed by MARS, while both the OLS and GLM have the same performance levels with higher RMSE values. Of the three alfalfa fields, Field 14 has the smallest RMSE value at 0.31, followed by Field 10, and Field 04 has the greatest value of RMSE. For the corn fields (Field 09, Field 40, and Field 80), the RMSE values for the ANN model are 0.14, 1.47, and 0.68 respectively, which are smaller values in comparison to the other models. Among these three fields, Field 09 has the smallest value (0.14), and Field 40 has the highest (1.47), while Field 80 lies in between. The three other models (OLS, GLM, and MARS) show higher RMSE values and they are all close to each other in magnitude.

Table2. *The NSE and RMSE parameters that are used to evaluate the performance of the OLS, GLM, MARS, and ANN models when predicting soil salinity of the alfalfa and corn fields*

Alfalfa						
	Field 04		Field 10		Field 14	
	NSE	RMSE	NSE	RMSE	NSE	RMSE
OLS	0.15	5.49	0.52	2.42	0.49	0.69
GLM	0.15	5.49	0.52	2.42	0.49	0.69
MARS	0.11	5.60	0.73	1.82	0.61	0.60
ANN	0.55	4.01	0.96	0.74	0.90	0.31
Corn						
	Field 09		Field 40		Field 80	
OLS	0.13	0.19	0.29	2.16	0.46	1.47
GLM	0.13	0.19	0.29	2.16	0.46	1.47
MARS	0.17	0.19	0.30	2.16	0.46	1.48
ANN	0.53	0.14	0.67	1.47	0.89	0.68

SUMMARY AND CONCLUSIONS

In this study four regression models (OLS, GLM, MARS, and ANN) were evaluated using six datasets comprised of field data combined with remote sensing data, three of the fields had alfalfa growing and three had corn growing. These four models were selected because of their different abilities. The OLS model can handle normal errors and constant variance, while the GLM can handle non-normal errors and non-constant variance. The MARS model seeks to elaborate on individual behavior as a result of a combination of internal and external factors and influences.

ANN uses the processing of the brain as a basis to develop algorithms that can be used to model complex patterns and prediction problems. This study shows that among these four different regression models, the performance of the ANN model is the best overall, followed by MARS and then by OLS and GLM. Combining the graphical and analytical evaluations together, the regression models perform slightly better infields planted with corn than in fields planted with alfalfa. The model performance depends on the conditions of the data collected. It is clear from the results of this study that the performance of the ANN model was the best. However, the ANN performance was poor for some fields, so it is highly recommended that the condition of the field data should be evaluated before running any model. This study showed that the closer the pattern of the data used is to CSR, the higher the autocorrelation. Also, the closer to a normal distribution the soil salinity data collected is, the better the performance of the regression models. This study showed that integrating field data with remote sensing and choosing the appropriate regression model can reduce the amount of time and money spent by minimizing the data that needs to be collected from the fields. This study also determined that the selection of the evaluation parameters is very important since no single parameter is capable of evaluating the performance from all aspects and provide a credible and accurate assessment of the model's performance. Thus, it is strongly recommended not to lean in the direction of only evaluating one statistical parameter, since the use of several statistical parameters, if carefully selected, is

likely to result in a better assessment of the model performance.

REFERENCES

[1] Abbas, A., Khan, S., Hussain, N., Hanjra, M. and Akbar, S. *Characterizing soil salinity in irrigated agriculture using a remote sensing approach*, Physics and Chemistry of the Earth, 2013, 43-52.

[2] Abuelgasim, A. and Ammad, R. Mapping soil salinity in arid and semi-arid regions using Landsat 8 OLI satellite data, *Remote Sensing Applications: Society and Environment*, 2019, 13, 415-425.

[3] Allbed, A., Kumar, L., and Sinha, P. Mapping and Modelling Spatial Variation in Soil Salinity in the Al Hassa Oasis Based on *Remote Sensing Indicators and Regression Techniques, Remote Sensing*, 2014, 6(2), 1137-1157.

[4] ASCE Task Committee on Definition of Criteria for Evaluation of Watershed Models, Criteria for evaluation of watershed models. *ASCE Journal of Irrigation and Drainage Engineering*, 1993, 119, pp. 429–442.

[5] Anselin, L. The Moran Scatterplot as an ESDA Tool to Assess Local Instability in Spatial Association. In M. Fischer, H. Scholten, and D. Unwin (eds.), Spatial Analytical Perspectives on GIS. London: Taylor and Francis, 1996, pp. 111-125.

[6] Anselin, L. Exploring Spatial Data with DynES DA2. CSISS and Spatial Analysis Laboratory University of Illinois, Urbana-Champaign. September 12, 2002.

[7] Diggle, P. J. Statistical Analysis of Spatial Point Patterns New York. 2003, Oxford University Press Inc.

[8] Dobermann A., Witt C., Dawe D., Gines H.C., Nagarajan R., Satawathananont S., Son T.T., Tan P.S., Wang G.H., Chien N.V., Thoa V.T.K., Phung C.V., Stalin P., Muthukrishnan P., Ravi V., Babu M., Chatuporn S., Kongchum M., Sun Q., Fu R., Simbahan G.C. and Adviento M.A.A. Site-specific nutrient management for intensive rice cropping systems in Asia, *Field Crops Res*. 2002, 74, 37–66.

[9] Douaik, A., Van Meirvenne, M. and Toth, T. "Spatio-temporal kriging of soil salinity rescaled from bulk soil electrical conductivity." Quantitative Geology And Geostatistics, GeoEnv IV: 4th European Conf. on Geostatistics For Environmental Applications, X. Sanchez-Vila, J. Carrera, and J. Gomez-Hernandez, eds., Kluwer Academic, Dordrecht, The Netherlands, 2004, 13(8) 413–424.

[10] Eldeiry, A.A. and Garcia, L.A..."Detecting soil salinity in alfalfa fields using spatial modeling and remote sensing". *Soil Science Society of America Journal*, 2008, 72(1), 201-211.

[11] Eldeiry, A.A. and Garcia, L.A. "Comparison of ordinary kriging, regression kriging, and cokriging techniques to estimate soil salinity using LANDSAT images". *ASCE Journal of Irrigation and Drainage Engineering*, 2010, 136:355.

[12] Eldeiry, A. and Garcia, L. A. "Using Indicator Kriging Technique for Soil Salinity and Yield Management".*ASCE Journal of Irrigation and Drainage Engineering*, 2011,137(2), 82-93.

[13] Eldeiry, A. and Garcia, L.A. "Evaluating the performance of ordinary kriging in mapping soil salinity". *ASCE Journal of Irrigation and Drainage Engineering*, 2012a,138(12), https://doi.org/10.1061/(ASCE)IR.1943-4774.0000517.

[14] Eldeiry, A. and Garcia, L.A. "Using disjunctive kriging as a quantative approach to manage soil salinity and crop yield".*ASCE Journal of Irrigation and Drainage Engineering*, 2012b, 138(3), https://doi.org/10.1061/(ASCE)IR.1943-4774.0000392.

[15] Freedman, D.A. (2009). Statistical Models: Theory and Practice. Cambridge University Press. p. 26. A simple regression equation has on the right hand side an intercept and an explanatory variable with a slope coefficient. A multiple regression equation has two or more explanatory variables on the right hand side, each with its own slope coefficient.

[16] Friedman, J.H. "Multivariate Adaptive Regression Splines". The Annals of Statistics, 1991, 19: 1. Doi: 10.1214/aos/1176347963. JSTOR 2241837. MR. 1091842. ZB1 0765.62064.

[17] Gates, T.K., Burkhalter, J.P., Labadie, J.W., Valliant, J.C. and Broner, I."Monitoring and modeling flow and salt transport in a salinity-threatened irrigated valley". *Journal of Water Resources Planning and Management*, 2002, 128(2), 87-99.

[18] Gujarat D.N. Basic Econometrics. 2003, New Delhi; Tatar –McGraw-Hill.

[19] Hsu, K., Gupta, H.V. and Sorooshian, S. (1995): "Artificial neural network modeling of the rainfall-runoff process". *Water Resources Research*, 1995, 31(10), 2517-2530. https://doi.org/10.1029/95WR01955

[20] Legates, D.R. and McCabe Jr, G.J. "Evaluating the use of 'goodness-of-fit' measures in hydrologic and hydroclimatic model validation." *Water Resources Research*, 1999, 35(1), 233–241.

[21] Lewis, P.A.W. and Stevens, J.G. "Nonlinear Modeling of Time Series Using Multivariate Adaptive Regression Splines (MARS)". *Journal of the American Statistical Association*, 1991, 86:416, 864-877, DOI: 10.1080/01621459.1991.10475126.

[22] McBratney, A.B, Mendonca Santos, M.L. and Minasny, B., On digital soil mapping, *Geoderma*, 2003, 3-52.

[23] McColl, K.A., Ryu, D., Matic, V., Walker, J.P., Costelloe, J. and Rudiger, C. "Soil salinity impacts on l-band remote sensing of soil moisture". *IEEE Geoscience and Remote Sensing Letter*, 2012, 9, 262–266.

[24] Nelder, J., *and* Wedderburn, R. *"Generalized Linear Models"*. Journal of the Royal Statistical Society. *Series A (General). Blackwell Publishing. 1972, 135(3), 370–384. doi:10. 2307/2344614. JSTOR 2344614.*

[25] *Pearson, K. "*Contributions to the Mathematical Theory of Evolution. II. Skew Variation in Homogeneous Material". *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 1985, **186**: 343–414.

[26] *Qian, Shen-En. Ed. (2016). Optical Payloads for Space Missions. John Wiley & Sons. P. 824. ISBN 978-1-118-94514-8 – via Googlw Books.*

[27] Rencher, A.C., and Christensen, W.F. "Chapter 10, Multivariate regression – Section 10.1, Introduction", Methods of Multivariate Analysis, Wiley Series in Probability and Statistics, 2012, 709 (3rd ed.), John Wiley & Sons, p. 19, ISBN 9781118391679.

[28] Ripley, B.D. "The second-order analysis of stationary point processes". *Journal of Applied Probability*, 1976, 13, 255–266. doi:10.2307 /3212829.

[29] Robbins, C.W. and Wiegand, C.L. Field and laboratory measurements. Agricultural Salinity Assessment and Management, American Society of Civil Engineers, 1990, New York.

[30] Robinson, D.A., Lebron, I., Kocar, B., Phan, K., Sampson, M., Crook, N. and Fendorf, S. Time-lapse geophysical imaging of soil moisture dynamics in tropical deltaic soils: An aid to interpreting hydrological and geochemical processes. *Water Resources Research*, 2009, 32, doi: 10.1029/2008WR006984.

[31] Saey, T., Van Meirvenne, M., De Smedt, P., Neubauer, W., Trink, I., Verhoeven, G. and Seren, S. Integrating multi-receiver electromagnetic induction measurements into the interpretation of the soil landscape around the school of gladiators at Carnuntum. *European Journal of Soil Science*, 2013, 64,716-727.

[32] Sarle, W.S. "Neural network and statistical modeling". Proceedings of the Nineteenth Annual SAS users group international conference, 1994.

[33] Spies, B. and Woodgate, P. Salinity Mapping in the Australian context. Technical Report. Land and Water Australia. 2004, 153p.

[34] Triantafilis, J., Odeh, I. and McBratney, A. "Five geostatistical models to predict soil salinity from electromagnetic induction data across irrigated cotton." *Soil Sci. Soc. Am. J*, 2001, 65(3), 869–878.

[35] Upton G. and Cook I. Oxford Dictionary of Statistics. Great Britain; Oxford University Press, 2002.

[36] Wiegand, C. L., Rhoades, J.D., Escobar, D.E. and Everitt, J.H. Photographic and videographic observations for determining and mapping the response of cotton to soil salinity. *Remote Sensing of Environment*, 1994, 49:212-223.

[37] Wu, J., Vincent, B., Yang, J., Bouarfa, S., and Vidal, A. Remote Sensing Monitoring of Changes in Soil Salinity: A Case Study in Inner Mongolia, China. *Sensors*, 2008, 8, 7035-7049; DOI: 10.3390/s8117035

[38] Zhu, Q., Lin, H.S., & Doolittle, J.A. Repeated electromagnetic induction surveys for determining subsurface hydrologic dynamics in an agricultural landscape. *Soil Science Society of America Journal*, 2010a, 74, 1750-1762.

[39] Zhu, Q., Lin, H.S., and Doolittle, J.A. Repeated electromagnetic induction surveys for improved soil mapping in an agricultural landscape. *Soil Science Society of America Journal*, 2010b, 74, 1763-1774.

Citation: *Luis A. Garcia and Ahmed A. Eldeiry, "Evaluating Linear and Nonlinear Regression Models in Mapping Soil Salinity", International Journal of Research in Agriculture and Forestry, 7(3), 2020, pp 21-34.*