Reprint ISSN 1923-7766 (Web Version)

International Journal of Experimental Agriculture

(Int. J. Expt. Agric.)

Volume: 10	Issue: 1	January 2020

Int. J. Expt. Agric. 10(1): 20-25 (January 2020)

IDENTIFICATION OF YIELD PREDICTORS OF WHEAT (*Triticum aestivum* L.) UNDER SALT STRESS USING RANDOM FOREST, MULTIPLE AND STEPWISE REGRESSION

M. HASANUZZAMAN, S.H.M.G. SARWAR AND M.S. ISLAM

An International Scientific Research Publisher

Green Global Foundation©

Web address: http://ggfjournals.com/e-journals archive
E-mails: editor@ggfjournals.com and editor.int.correspondence@ggfjournals.com

ISSN-1923-7766 (Online)
Int. J. Expt. Agric. 10(1):20-25(January 2020)

IDENTIFICATION OF YIELD PREDICTORS OF WHEAT (*Triticum aestivum* L.) UNDER SALT STRESS USING RANDOM FOREST, MULTIPLE AND STEPWISE REGRESSION

M. HASANUZZAMAN[1]*, S.H.M.G. SARWAR[2] AND M.S. ISLAM[3]

[1]Department of Genetics and Plant Breeding, Faculty of Agriculture, Hajee Mohammad Danesh Science and Technology University, Dinajpur, Bangladesh; [2]Central Farm, Hajee Mohammad Danesh Science and Technology University, Dinajpur, Bangladesh; [3]BRAC Agricultural Research and Development Centre (BARDC), CERDI Road, Joydevpur, Gazipur-1701.

*Corresponding author & address: Md. Hasanuzzaman, E-mail: mdhasanuzzaman@gmail.com
Accepted for publication on 7 January 2020

ABSTRACT

Hasanuzzaman M, Sarwar SHMG, Islam MS (2020) Identification of yield predictors of wheat (*Triticum aestivum* L.) under salt stress using random forest, multiple and stepwise regression. *Int. J. Expt. Agric.* 10(1), 20-25.

Salinity is one of the important limiting factors for the production of wheat in the southern coastal region of Bangladesh. The effectiveness of the selection of wheat genotypes depends on the perfectly-identified yield predictors' variables. The present study was conducted to assess the yield components under the salinity stress environment using the multiple regression, stepwise regression and random forest model. This research was conducted with ten wheat genotypes, grown in earthen pots with 10 dSm^{-1} salinity and control in consecutive two seasons of 2013-2014 to 2014-2015. All treatments were arranged in a complete randomized design (CRD) with three replications. Data were recorded on the shoot and root traits. The results showed that salinity treatment represses the development of roots causing grain yield loss of all wheat genotypes. Considering the predictors' variables, such as phenology: days to heading, days to maturity; yield attributes: effective tillers $plant^{-1}$, plant height, spike length, spikelets $spike^{-1}$, grains $spike^{-1}$; root traits: length, volume, fresh weight, dry weight, random forest, multiple linear regression and stepwise regression, all three methods have identified dry root weight and number of grains bearing tillers contributes to grain yield per plant under salt stress. Selection through these traits may be effective in a saline environment. The performance of random forest is superior to multiple linear regression and stepwise regression models showing the lowest MSE.

***Key words**: random forest, machine learning, salinity, wheat, root, effective tiller*

INTRODUCTION

Wheat is the most widely cultivated cereal in the world, a staple food for 40 percent of the world's population that contributes 20 percent of total dietary calories and proteins worldwide (Braun *et al.* 2010) and in Bangladesh most important cereal crops after rice in both economic and consumption (Hossain and da Teixeira Silva, 2013; BARI 2010).

Soil salinity severely constrains crop production in the southern part of Bangladesh and worldwide. The total global area of saline and sodic soils is estimated to be around 830 million hectares, more than 6% of the world's land and rising (Martinez-Beltranand and Manzur, 2005; Shrivastava and Kumar, 2015). It is estimated that over 50% of global arable land will be salinized by 2050 (Jamil *et al*. 2011). In Bangladesh, the coastal region covers about 20% area of the country (29,000 km^2) surrounded by more than 30% of the cultivable lands where 53% of the coastal areas are affected by salinity (Haque 2006). Salinity during the dry season is a major constraint to crop yield in southern Bangladesh, particularly in coastal zones (Dalgliesh and Poulton, 2011). Yield reductions of 50% in durum wheat under dryland salinity (James *et al*. 2012), 88% in bread wheat under high irrigation salinity (Jafari-Shabestari *et al*. 1995).

Machine learning, part of Artificial Intelligence, comprises of algorithms and statistical models that support the system to learn autonomously and improve from experience with simple programming (Wikipedia Contributors, 2019). Emerging technique ensemble learning produces a unique model from multiple predictions based on the same base algorithm (Friedl *et al.* 1999; Breiman 2001). Random forest is an example of ensemble learning. It is non-parametric and can utilize distribution-free data and correlated variables do not affect (Cutler *et al*. 2007). Autocorrelation and multicollinearity reduce the efficacy in linear regression (Draper and Smith, 1998).

Development of a new variety of Salinity tolerant is the only feasible way of improving yield in saline soils (Genc *et al.* 2007) and it is necessary to increase wheat production in by raising the wheat grain yield and the most efficient way to increase wheat yield in Bangladesh is to evolve the salt tolerance of wheat genotypes (Pervaiz *et al.* 2002; Shannon 1997). Limited attention has been given to develop a salt-tolerant variety of wheat in Bangladesh. As yield is low heritable, indirect selection may be effective. Identification of important primary traits that contribute to yield is essential for indirect selection. But, different tools identify different important primary traits. The effectiveness of selection depends on choosing the real primary traits depends on yield. Under this circumstance, this research was undertaken (1) to determine the important predictor/predictors for grain yield under salt stress in wheat and (2) to test the performance of the multiple regression, stepwise regression and random forest for yield prediction. This study is an application of statistical, and machine learning techniques to find out the primary trait which contributes most to yield per plant under saline condition.

MATERIALS AND METHODS

A pot experiment was conducted in the research field of Genetics and Plant Breeding, Hajee Mohammad Danesh Science and Technology University, Dinajpur, Bangladesh. Seeds of ten selected wheat genotypes

(Table 1) collected from Bangladesh Wheat and Maize Research Institute (BWMRI), Dinajpur, Bangladesh. The plants were grown in pots filled with a sandy loam soil having pH 7.5 and two treatments in the experiment i.e., non-saline (0 dS m^{-1}) and saline (10 dS m^{-1}). The experiment was replicated thrice in a completely randomized design and this study was continued in consecutive two seasons of 2013-2014 to 2014-2015.

Table 1. Name and pedigree of wheat genotypes used in the experiment

Entry	Name	Pedigree
1	Protiva	UP301/C306
		1187-1-1P-5P-5JO-OJO
2	Shatabdi	MRNG/BVC//BLO/PVN/3/PJB-81
		CM98472-1JO-0JO-0O-1JO-0JO-0R2DI
3	Prodip	G. 162/BL 1316//NL 297
		NC2055-4B-020B-020B-4B-0B
4	BARI Gom 25	ZSH 12/HLB 19//2*NL 297
5	BARI Gom 26	ICTAL123/3/RAWAL87//VEE/HD2285
		BD(JOY)86-0JO-3JE-010JE-010JE-HRDI-RC5DI
6	BAW-1151	SOURAV/KLAT/SOREN//PSN/3/BOW/4/VEE#5. 10/5/CNO
		67/MFD//MON/3/ SERI/6/NL297
		BD(DI)112S-0DI-030DI-030DI-030DI-9DI
7	BAW-1135	BAW-969/SHATABDI
		BD(DI)1319S-0DI-6DI-1DI-DIRC6
8	BAW-1168	BAW-923/BIJOY
		BD(DI) 1327S-0DI-3DI-1DI-DIRC4
9	BAW-1182	KAL/BB/YD/3/PASTOR
		CMSS99M00981S-0P0M-040SY-040M-040SY-16M-0ZTY-0M
10	BAW-1193	SOURAV/3/ALTAR84/AE.SQ.(224)//2*YACO/4/JUNCO//YD/PCI
		BD04JA178T-0DI-0DI-0JA-0JA-0JA

Source: Bangladesh Wheat and Maize Research Institute (BWMRI), Nashipur, Dinajpur

Each pot contained 10 kg of soil and filled with recommended doses of NPK fertilizers were used in both saline and non-saline treatments. The soil filled pots were irrigated with tap water fit for irrigation. Ten seeds were sown in each pot and after seedling emergence five plants were maintained in each pot by thinning and the plants were harvested at the maturity stage.

Data were recorded on root length, root volume, root fresh weight, root dry weight, spike length, number of effective tillers, plant height, days to heading, days to maturity, number of spikelets, number of grains per spike and grain yield per plant from ten different plants from each genotype in each replication. Data were partitioned into two parts: training data and testing data at 70:30 ratios randomly. Training data were used for model development and test data were used for evaluation of the model. Statistical tools, multiple linear regression, stepwise regression, and random forest were used to find the best predictors. R 3.5.0 (R Core Team, 2019) with randomForest (Liaw and Wiener, 2002) and caret (Kuhn 2008) packages were utilized to analyze the data.

RESULTS AND DISCUSSION

The number of grain bearing tillers per plant and dry weight of root significantly contributes to grain yield as per multiple regression (Table 4) and stepwise regression (Table 2).

Table 2. Showing coefficients of stepwise regression in wheat under salt stress

Variable	Estimate	SE	Pr(>|t|)
(Intercept)	24.49146	6.63262	0.000713
Grain bearing tillers per plant	0.91349	0.12367	8.71e-09
Days to maturity	-0.28311	0.06974	0.000244
Plant height	0.05312	0.03080	0.092947
Root dry weight	1.10770	0.24784	7.17e-05

Random forest uses two important parameters, m_{try} and ntree. Parameters m_{try} and ntree are the number of predictors used at each split and the number of trees in the forest respectively. It can be observed that the error rate becomes flat inbetween 400 to 500 trees (Fig. 1). This indicates beyond the value of 500, tree number have no significant impact on model accuracy. Therefore, for this study, the value for number of trees is kept constant at 500. The minimum out of bag error was observed against mtry 4 and higher score were observed below or upper of this value (Fig. 2).

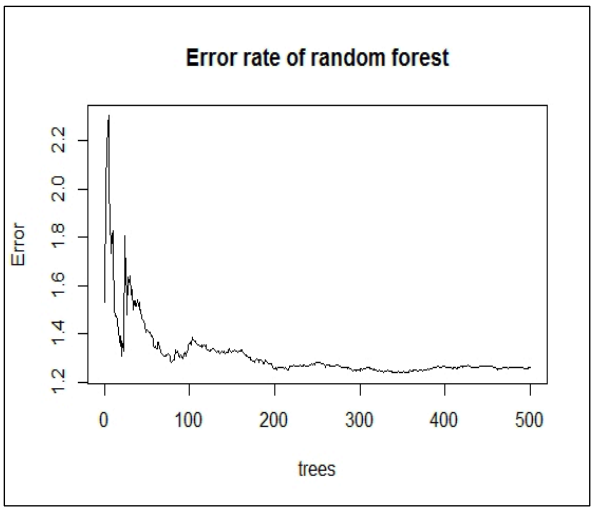

Fig. 1. Showing Error rate of random forest with different trees in wheat under salt stress

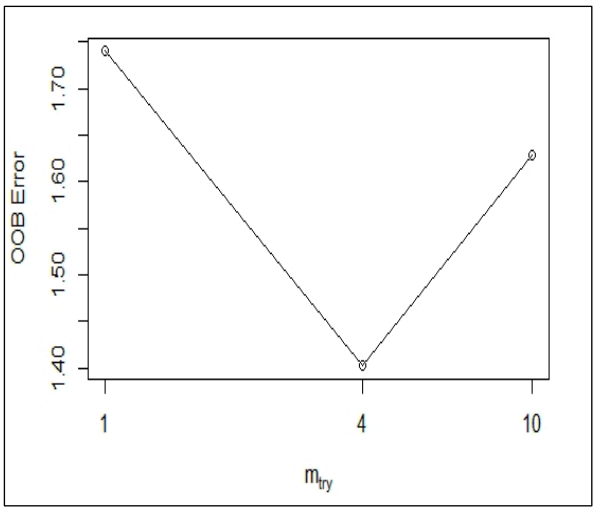

Fig. 2. Showing different mtry with Out of Bag Error in wheat under salt stress

The number of nodes varies from about 10 to 20 per tree and most of the trees have 13 to 14 nodes based on the random forest model (Fig. 3) indicates forest having trees with diversified number nodes gives a diversified decision, ultimately makes a single accurate decision. The highest value of mean decrease accuracy (%IncMSE) and Gini (IncNodePurity) were observed on root dry weight indicates that it is the most important primary trait which contributes maximum to grain yield per plant. The second important predictor was the effective tiller number per plant. Root traits volume, weight, and length also contribute to grain yield (Fig. 4).

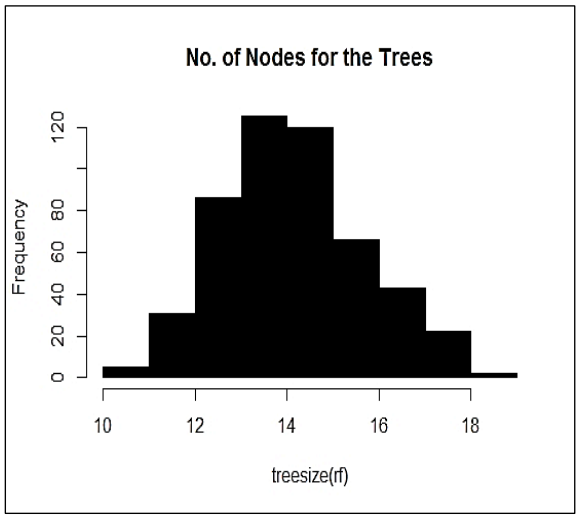

Fig. 3. Showing the number of nodes of the trees in wheat under salt stress

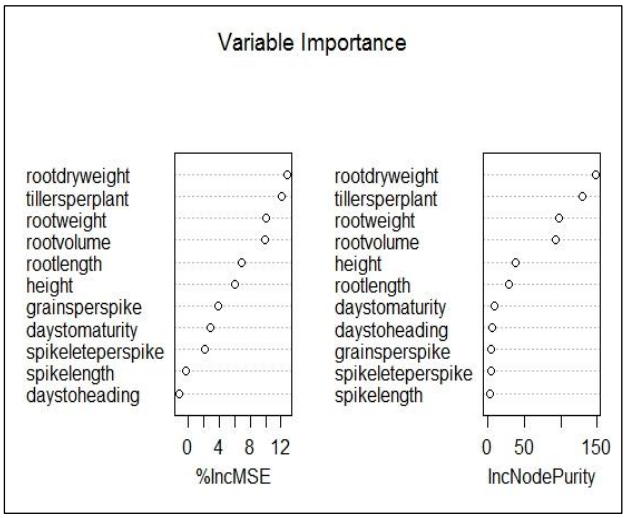

Fig. 4. Showing mean decrease accuracy (%IncMSE) and mean decrease Gini (IncNodePurity) of different variables in wheat under salt stress

Random forest model performance is further assessed using training (Fig. 5(a)) and testing (Fig. 5(b)) dataset respectively. (Fig. 5(a)) shows a decent fit between actual and predicted yield per plant based on the training dataset. But, the deviation is higher in the test data set (Fig. 5(b)). Relative MSE results of the three methods for grain yield per plant are presented in Table 3. Minimum MSE was observed in the random forest both in train and test data but explained variability was very close with multiple regression and stepwise regression.

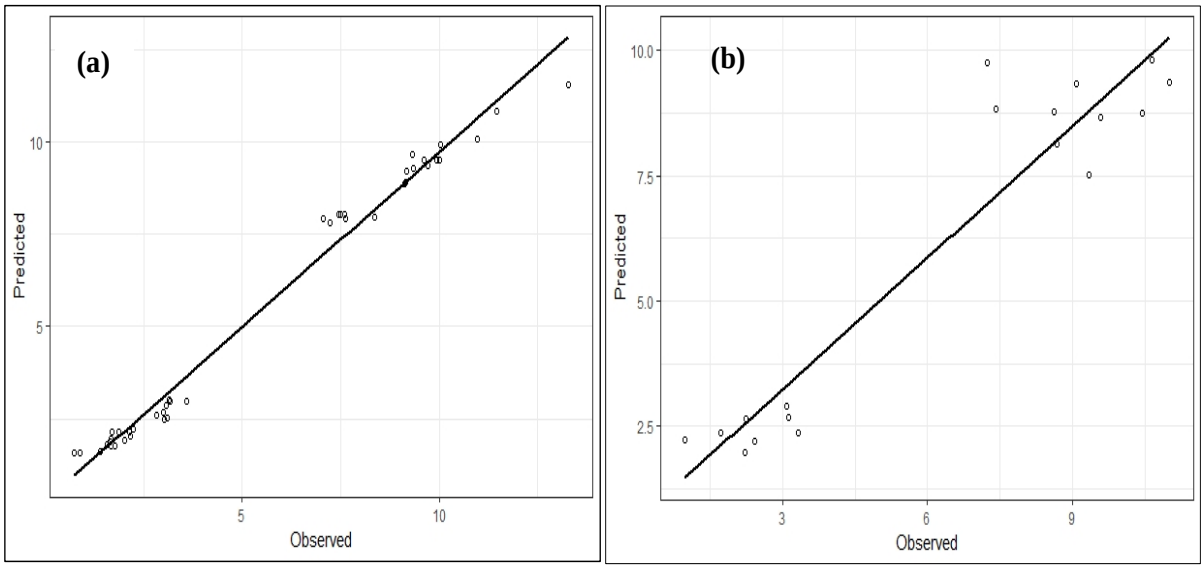

Fig. 5. Showing observed versus predicted yield per plant in wheat undersalt stress based on the random forest model (a) training data and (b) test data

Table 3. Showing Performance of statistical and machine learning tools used in wheat under salt stress environment

Tool	Train Data		Test Data	
	MSE	Variance Explained (R^2)	MSE	Variance Explained (R^2)
Multiple Regression	0.91	0.94	1.76	0.81
Stepwise Regression	0.83	0.95	1.92	0.84
Random Forest	0.25	0.90	1.35	0.83

This study is an application of statistical and machine learning techniques to find out the primary trait which contributes most to yield per plant under saline condition. A higher mean decrease in accuracy and Gini indicates higher variable importance. Based on random forest, grain yield per plant is depends on root traits: dry weight, volume, fresh weight, length and number of grain bearing tillers per plant (Fig. 4). But according to multiple and stepwise regression, grain yield depends on the dry weight of root and grain bearing tillers per plant (Table 2 and Table 4) under saline condition indicates salinity represses the development of roots causing yield loss. Maas *et al.* (1994) also reported that salt stress decreases the number of primary and secondary tillers causes yield reduction. Grain bearing tillers and dry weight of root should be selected for salt-tolerant high yielding wheat variety development. Selection based on root dry weight may be cumbersome. In that case, selection should be done based on the number of grain bearing tillers in each plant.

Table 4. Showing coefficients of multiple regression in wheat under salt stress

Variable	Estimate	Standard Error	Pr(>|t|)
(Intercept)	24.441883	8.475919	0.00733
Days to heading	-0.033938	0.081108	0.67872
Grain bearing tillers per plant	0.902166	0.260627	0.00169
Days to maturity	-0.266932	0.087958	0.00504
Height	0.040722	0.047476	0.39806
Spike length	0.122323	0.353266	0.73164
Spikelets per spike	0.020975	0.128771	0.87174
Grains per spike	-0.014877	0.036703	0.68820
Root length	0.023973	0.077861	0.76036
Root volume	0.024458	0.216878	0.91099
Root weight	-0.009495	0.221445	0.96609
Root dry weight	0.998763	0.407455	0.02050

Multicollinearity is a common limiting factor of regression models. MSE and R^2 evaluate the performance of a model, but the accuracy of the model does not depend always on the high value of R^2. As minimum MSE was observed in a random forest indicates the superiority over the other two statistical tools, multiple and stepwise regression. Random forest is robust, faster than bagging, distribution-free and gives variable importance

(Breiman 2001; Rodriguez-Galiano *et al*. 2014). Increase the number of trees reduces error makes the data fit (Rodriguez-Galiano *et al*. 2014). However, random forest acts as ‘‘black box’’ (Prasad *et al*. 2006). It does not compute regression coefficients and confidence intervals (Cutler *et al*. 2007). However, it measures variable importance and can be compared to other regression methods(Gromping 2009).

CONCLUSION

The potentiality of random forest techniques to model grain yield per plant under salt stress environment in wheat was explored in this study. It shows that the performance of random forest is superior to the multiple and stepwise regression. Saline environment reduces the root growth causes a decrease in grain yield. Grain bearing tillers and root dry weight contribute maximum to grain yield under salt stress. Selection through these primary traits can be effective for better grain yield under salt stress environment.

REFERENCES

BARI (2010) Wheat production in Bangladesh: a success story. http://www.bari.gov. bd/index.php?option=com_ simplest forum and view =postlist and topic= true and forumId=1andparentId=288 (4 November 2012).

Braun HJ, Atlin G, Payne T (2010) Multi-location testing as a tool to identify plant response to global climate change. In: Reynolds, CRP (ed) Climate change and crop production, CABI, London, UK.

Breiman L (2001) Random forests, Machine Learning, 45(1), pp. 5–32.doi: 10.1023/A:1010933404324.

Cutler DR, Edwards TC, Beard KH, Cutler A, Hess KT, Gibson J, Lawler JJ (2007) Random Forests for Classification in Ecology: Ecology, 88(11), 2783–2792.

Dalgliesh NP, Poulton PL (2011) Physical constraints to cropping in southern Bangladesh: Soil and Water. In Rawson H.M. (ed) sustainable intensification of Rabi cropping in southern Bangladesh using Wheat and mungbean. ACIAR Technical Report No. 78. Australian Center for International Agricultural Research: Canberra pp. 256.

Draper NR, Smith H (1998) Applied regression analysis, 3rd edn. New York: John Wiley & Sons.

Friedl MA, Brodley CE, Strahler AH (1999) Maximizing land cover classification accuracies produced by decision trees at continental to global scales, IEEE Transactions on Geoscience and Remote Sensing, 37(2 II), pp. 969–977. doi: 10.1109/36.752215.

Genc Y, McDonald GK, Tester M (2007) Reassessment of tissue Na^+ concentration as a criterion for salinity tolerance in bread wheat. *Plant cell Envirn.* 30(11), 1486-1498.

Gromping U (2009) ‘Variable Importance Assessment in Regression : Linear Regression versus Random Forest’, The American Statistician, 63(4), pp. 308–319.

Haque SA (2006) Salinity problems and crop production in coastal Regions of Bangladesh. Review article. *Pak. J. Bot.* 38(5),: 1359-1365.

Hossain A, da Teixeira Silva JA (2013) Wheat production in Bangladesh: its future in the light of global warming. AoBPLANTS5:pls042;doi:10.1093/aobpla/pls042.

Jafari-Shabestari J, Corke H, Qualset CO (1995) Field evaluatiuon to salinity stress in Iranian hexaploid wheat landrace accessions. Genet. Resour. Crop Evol. 42, 147–156. doi: 10.1007/BF02539518.

James RA, Blake C, Zwart AB, Hare RA, Rathjen AJ, Munns R (2012) Impact of ancestral wheat sodium exclsuion genes Nax1 and Nax2 on grain yield of durum wheat on saline soils. Funct. Plant Biol. 39, 609–618. doi: 10.1071/FP12121.

Jamil A, Riaz S, Ashraf M, Foolad MR (2011) Gene expression profiling of plants under salt stress. *Crit. Rev. Plant Sci.* 30, 435–458. doi:10.1080/07352689.2011.605739.

Kuhn M (2008) Building predictive models in R using the caret package. *Journal of statistical software*, 28(5), 1-26.

Liaw A, Wiener M (2002) Classification and Regression by randomForest. R News 2(3), 18-22.

Maas EV, Lesch SM, Francois LE, Grieve CM (1994) Tiller development in salt-stressed wheat. Crop Science, 34(6), pp. 1594–1603.doi: 10.2135/cropsci1994.0011183X003400060032x.

Martinez-Beltran J, Manzur CL (2005) Overview of salinity problems in the world and FAO strategies to address the problem, in *Proceedings of the international salinity forum* (Riverside), 311–313.

Pervaiz Z, Afzal M, Xi S, Xiaoe Y, Ancheng L (2002) Physiological parameters of salt tolerance in wheat. *Asian J. Plant. Sci*. 1: 478-481.

Prasad AM, Iverson LR, Liaw A (2006) Newer classification and regression tree techniques: Bagging and random forests for ecological prediction. Ecosystems, 9(2), pp. 181–199. doi: 10.1007/s10021-005-0054-1.

R Core Team (2019) 'R: A language and environment for statistical computing'. Vienna, Austria. Available at: https://www.r-project.org/.

Rodriguez-Galiano VF, Chica-Olmo M, Chica-Rivas M (2014) 'Predictive modelling of gold potential with the integration of multisource information based on random forest: a case study on the Rodalquilar area, Southern Spain', International Journal of Geographical Information Science, 28(7), pp. 1336–1354. doi: 10.1080/13658816.2014.885527.

Shannon MC (1997) Adaptation of plants to salinity. Adv. Agron. 60: 75–120.

Shrivastava S, Kumar R (2015) Soil salinity: A serious environmental issue and plant growth promoting bacteria as one of the tools for its alleviation. *Saudi J. Biol. Sci.* 22, 123–131. doi: 10.1016/j.sjbs.2014.12.001.

Wikipedia Contributors (2019) 'Machine learning', In Wikipedia, The Free Encyclopedia. Available at: https://en.wikipedia.org/w/index.php?title=Machine_learning&oldid=909989905 (Accessed: 9 August 2019).