Analyzing Variation of Soil Salinity Content in the Agricultural Areas: A Factorial Analysis Based Random Forest Estimation Method

H Wang¹ and Y P Li¹,*

¹Environment and Energy Systems Engineering Research Center, School of Environment, Beijing Normal University, Beijing 100875, China

*yongping.li@iseis.org

Abstract. In this study, to identify the complex relationship between soil salinity content (SSC) and factors, and efficiently quantify the individual effects of factors on SSC, a random-forest-based factorial analysis (RF-FA) method by incorporating random forest (RF) and factorial analysis (FA) is developed. The RF-FA is applied in a case study. Results reveal that: (i) Compared with Back Propagation Neural Network (BPNN) and Support Vector Regression (SVR), RF is a more robust model for SSC simulation; (ii) The contributions of factors to the variation of SSC follow the order: Elevation (14.68%) > LST (11.69%) > Albedo_N (11.41%) > SI_2 (10.71%) > SAVI (10.21%), indicating topography and temperature are the macro factors controlling SSC; hence, providing sufficient irrigation water is necessary to mitigate soil salinization. The findings can help make effective strategies to relieve the soil salinization of the farmland and support the sustainable development of agriculture.

1. Introduction

Due to special climatic conditions that the lack of precipitation [1] and the strong evaporation, and the traditional agricultural management manners, Central Asia is seriously affected by salinization [2]. The proportion of farmland affected by salinization in five Central Asian countries is about 47.5% [3]. Therefore, the reliable method for monitoring the soil salinity content (SSC) over a large region is desired to ensure sustainable agriculture and to support the management strategies of protecting the ecological environment in Central Asia.

As well as the ease in development with smaller datasets obtained from the remote sensing and field sampling, statistical models can identify the complex relationship between SSC and factors via various algorithms, and been widely employed in SSC estimation and monitoring, such as random forest (RF) [4], artificial neural network (ANN) [5], and support vector regression (SVR) [6]. For example, Li, Li [7] employed models including RF to predict SSC of the irrigation oasis of Yinchuan China, results indicated that RF is more suitable for SSC estimation in the dry season, and topographic factors are the most important factors for SSC. Wang, Chen [8] used the sampled spectral data and five statistical models to estimate SSC of northwestern China; indicating that RF is more robust than other models for studies on hyperspectral simulation of SSC. Fu, Gan [5] coupled the particle swarm optimization and ANN to build SSC simulated models; the results indicated that the hybrid model has the satisfactory performance. The efficiency of simulating SSC for statistical models are proved by many studies. However, the statistical models are insufficient in identifying the quantitative impacts of factors on SSC,
resulting fail to analyze sensitivity for post-optimization. Factorial analysis (FA), is reliable method to quantify the effects of factors on the response variable (SSC in this study) [9]. Many previous studies reported the application of FA in many fields, such as hydrological simulation [10, 11], forest ecology [12]. However, few was conducted to identify the main and interactive effects of factors on SSC by incorporating statistical models (e.g., RF) with FA method.

Therefore, this study aims to develop a random-forest-based factorial analysis (RF-FA) method by incorporating random forest (RF) and factorial analysis (FA) into a framework. Then, the RF-FA method will then be applied to the irrigated farmland of the Syr Darya River Basin (SDRB) for quantitatively analyzing the effects of factors (e.g., vegetation, temperature, and topographical) on the SSC. The RF-FA method can (i) explore the complex relationship between SSC and factors, (ii) efficiently identify the main factors for SSC.

2. Methodology
RF is capable of capturing the relationship between predictors and predictands and has capacity to some extent avoid overfitting. RF has a low sensitivity to training set size and it was reported its efficiency in various applications in soil salt mapping [13-15]. The detailed steps are: (1) employing RF to capture the complex relationship between SSC and six kinds of factors (vegetation, temperature, evapotranspiration, underlying surface, topographic, and salinity index), ANN and SVR are used for comparing the simulation effectiveness of RF in SSC simulation; (2) employing FA to analyze the individual effects of factors based on the trained RF model.

3. Study area and data
SDRB is the sub-basins of the Aral Sea Basin in Central Asia. The main economic belts and urban agglomerations are located in the SDRB [16]. Unfortunately, SDRB possesses an arid and semi-arid climate and is threatened by soil salinization. Table 1 lists the information of the datasets. The datasets of factors (i.e.; NDVI) in the study were mainly collected from the National Aeronautics and Space Administration (https://modis.gsfc.nasa.gov/). The datasets of salinity Index 1 and 2 were calculated based on the MOD datasets. The soil samples of SDRB were obtained from the Central Asia Ecology and Environment Research Centre of the Chinese Academy of Sciences (CAS).

Table 1. The factors and datasets for case study

| Types                  | Name                                      | Datasets                          |
|------------------------|-------------------------------------------|-----------------------------------|
| Predictand             | Soil salinity content (SSC)               | Field sampling data in SDRB       |
| Vegetation factors     | Normalized difference vegetation index (NDVI) | MOD13Q1v6 / MYD13Q1v6 (Vegetation index) |
|                       | Soil adjusted vegetation index (SAVI)     | MOD09GQ (Surface Reflectance)    |
| Temperature factors    | Land surface temperature (LST)            | MOD11A2v4 (Land surface temperature) |
| Evapotranspiration factors | Evapotranspiration (max/min/mean ET) | MOD16A2v6 (Evapotranspiration)   |
| Underlying surface factors | Albedo in VIS (AIV)             | MCD43A3 (Albedo)                  |
|                        | Albedo in NIR (AIN)                     | SRTM v4 (DEM)                    |
|                        | Elevation                               |                                   |
Topographic factors | Slope factors
---|---
Salinity index factors | Salinity Index 1 (SI_1) | MOD09GQ (Surface Reflectance)
Salinity Index 2 (SI_2)

### 4. Result and Discussion

#### 4.1. Performance of random forest

The sampled SSC dataset (293 samples) was randomly partitioned into the calibration dataset (263 samples) and the validation dataset (30 samples). The calibration dataset was used to train the model, and the validation dataset was to test the generic ability of the models. Figure 1 shows the performance of SSC simulation of different models. The RF shows the best efficiency (calibration: $R^2 = 0.88$, RMSE = 1.31 g/kg, MAPE = 12.2%; validation: $R^2 = 0.74$, RMSE = 1.69 g/kg, MAPE = 14.3%) among the three models. The trained RF model will be used for the multi-level factorial analysis.

![RF, SVR and BPNN in (c) validation and (d) test period.](image)

**Figure 1.** Taylor diagram graphical for RF, SVR and BPNN in (c) validation and (d) test period.

#### 4.2. Main factors influencing SSC

The efficiency of the SSC simulated by RF is sensitive to the factors. The main effects of factors on SSC are quantified based on the multilevel FA. Design two levels (low and high) for each of the 13 factors, generating $2^{13}$ treatment combinations. SSC simulation with the trained RF model was then conducted for each of the treatment combinations. Figure 2 shows the simulated results. For the SSC of the SDRB, the highest, lowest and average streamflow of the $2^{13}$ treatments is 8.24 g/kg, 5.13 g/kg, and 2.24 g/kg, respectively. The large variation of SSC indicates the sensitivity of the RF-FA method.
Figure 2. Simulated SSC under different combinations of factor levels.

Figure 3 presents the main effects of factors on SSC. Take the plot of Elevation as an example. When Elevation is from the low level of 60m to a high level of 719.35m, SSC would decrease from 2.79 g/kg to 1.26 g/kg with; the slope of the line is negative, indicating that the effect of Elevation on SSC is negative. Overall, The LST and Albedo_V have a positive effect on SSC, and the other eleven factors have a negative effect on SSC. For the all factors, Elevation and SI has the steepest and gentlest slope, respectively, implying that Elevation and SI have the most and least significant main effect on SSC respectively.

Figure 4 presents the contributions of 13 factors to the variation of SSC. Results indicate that Elevation is the dominant factor to SSC with a contribution of 14.68%. This is because the flood irrigation is the main irrigation method for agricultural production in Central Asia, and elevation is the factor that determines the direction of water flow macroscopically; at high altitudes, salinity tends to follow irrigation water flow to the region with low elevations; at low altitudes, it is easy to accumulate salinity. LST is the second important factor with a contribution of 11.69%. it may be mainly attributed to that, the higher the temperature, the hotter and drier the farmland surface, and the salt accumulate on the surface with the loss of the surface water. Note that the four Evapotranspiration factors (ET_a, ET_max, ET_min, ET_mean) contribute less to SSC than the LST, this is mainly due to that the variation of ET are also limited by precipitation and irrigation; when evapotranspiration in different regions is similar, it is difficult to indicating the difference of the accumulated SSC caused by factors such as temperature. Generally, Topographic factors have the greatest impact on the variation of SSC with a total contribution of 21.59%; the second and third are underlying surface factors and vegetation factors, with the total contributions of 18.59% and 18.33%, respectively.
Figure 3. Main effects of factors on SSC in SDRB.
5. Conclusions
In this study, a random-forest-based factorial analysis (RF-FA) by coupling random forest (RF) and factorial analysis (FA) into a framework. The RF-FA models can not only identify the complex relationship between SSC and factors, but also efficiently quantify the effects of factors on SSC. Then, the developed model has been applied to the Syr Darya Basin of Central Asia, thirteen factors in six categories are select as the candidate predictors for analyzing the regional SSC variation. Main findings are as follows: (i) compared ANN and SVR, RF has the best performance during the calibration (R2 = 0.88, RMSE= 1.31 g/kg, MAPE= 12.2%) and validation period (R2 = 0.74, RMSE= 1.69 g/kg, MAPE= 14.3%). (ii) The contributions of factors to the variation of SSC follow the order: Elevation (14.68%) > LST (11.69%) > Albedo_N (11.41%) > SI_2 (10.71%) > SAVI (10.21%), indicating topography and temperature are the macro factors controlling SSC; hence, providing sufficient irrigation water is necessary to mitigate soil salinization. The findings are helpful for decision makers to formulate management strategies to relieve the soil salinization of the SDRB and protecting the ecological environment in Central Asia.

Acknowledgments
This research was supported by the Strategic Priority Research Program of Chinese Academy of Sciences (Grant No. XDA2006030202)
References
[1] Ma, Z., et al., Spatial and temporal precipitation patterns characterized by TRMM TMPA over the Qinghai-Tibetan plateau and surroundings. International journal of remote sensing, 2018. 39(12): p. 3891-3907.
[2] Koohafkan, P., Water and cereals in drylands. 2012: Routledge.
[3] Hamidov, A., K. Helming, and D. Balla, Impact of agricultural land use in Central Asia: a review. Agronomy for sustainable development, 2016. 36(1): p. 6.
[4] Wei, Y., et al., Updated information on soil salinity in a typical oasis agroecosystem and desert-oasis ecotone: Case study conducted along the Tarim River, China. Science of the Total Environment, 2020. 716: p. 17.
[5] Fu, C., et al., Determination of Soil Salt Content Using a Probability Neural Network Model Based on Particle Swarm Optimization in Areas Affected and Non-Affected by Human Activities. Remote Sensing, 2018. 10(9): p. 1387.
[6] Taghadosi, M.M., M. Hasanlou, and K. Eftekhari, Soil salinity mapping using dual-polarized SAR Sentinel-1 imagery. International journal of remote sensing, 2019. 40(1): p. 237-252.
[7] Li, Z., et al., Spatial Prediction of Soil Salinity in a SemiArid Oasis: Environmental Sensitive Variable Selection and Model Comparison. Chinese Geographical Science, 2019. 29(5): p. 784-797.
[8] Wang, S.J., et al., Performance Comparison of Machine Learning Algorithms for Estimating the Soil Salinity of Salt-Affected Soil Using Field Spectral Data. Remote Sensing, 2019. 11(22): p. 26.
[9] Montgomery, D.C., Design and analysis of experiments. 2017: John wiley & sons.
[10] Zhuang, X.W., et al., A hybrid factorial stepwise-cluster analysis method for streamflow simulation - a case study in northwestern China. Hydrological Sciences Journal-Journal Des Sciences Hydrologiques, 2016. 61(15): p. 2775-2788.
[11] Jia, Q., et al., Analyzing variation of inflow from the Syr Darya to the Aral Sea: A Bayesian-neural-network-based factorial analysis method. Journal of Hydrology, 2020: p. 124976.
[12] Wang, J., et al., Analyzing urban forest coverage variation in Guangzhou-Foshan region using factorial analysis based multivariate statistical prediction models. Forest Ecology and Management, 2019. 432: p. 121-131.
[13] Wang, J., et al., Quantitative estimation of soil salinity by means of different modeling methods and visible-near infrared (VIS–NIR) spectroscopy, Ebinur Lake Wetland, Northwest China. PeerJ, 2018. 6: p. e4703.
[14] Hoa, P.V., et al., Soil Salinity Mapping Using SAR Sentinel-1 Data and Advanced Machine Learning Algorithms: A Case Study at Ben Tre Province of the Mekong River Delta (Vietnam). Remote Sensing, 2019. 11(2): p. 128.
[15] Wu, W., et al., Soil salinity prediction and mapping by machine learning regression in Central M esopotamia, I raq. Land degradation & development, 2018. 29(11): p. 4005-4014.
[16] Wegerich, K., et al., Water security in the Syr Darya basin. Water, 2015. 7(9): p. 4657-4684.