PREDICTION OF SOIL PH FROM REMOTE SENSING DATA USING GRADIENT BOOSTED REGRESSION ANALYSIS

[1]V Anantha Natarajan, [2]M Sunil Kumar, [3]V Tamizhazhagan, [4]R M Chevdumoi
[1,2]School of Computing, Department of CSE, Mohan Babu University/
Department of Computer Science & Engineering, Sree Vidyanikethan Engineering College, AP, India.
[3]Department of Information Technology, Annamalai University, India.
[4]Chief Operating Officer, Sagri Bengaluru Private Limited, India.
Email: vananthanatarajan@vidyanikethan.edu
DOI: 10.47750/pnr.2022.13.S06.005

Abstract

The salt content of an agricultural field depends on the soil salinity and measuring its value precisely becomes essential. The dynamic changes of the factors are monitored using satellite data. This paper aims at estimating the value of soil pH based on the satellite data and the laboratory test results of pH values. Soil pH and salinity indices estimated from the satellite data has high correlation. A regression analysis is performed to explore the relationship between the salinity indices and the soil pH values over a study area. Salinity indices better represents the salt composition in the soil which is produced by reaction between the acidity and alkalinity as a chemical process. Based on the regression model the current soil pH value can be estimated with reference to the satellite data.

Keywords: soil salinity, soil pH, satellite data, regression analysis, salinity indices.

1. INTRODUCTION

The availability of certain nutrients and micronutrients, especially phosphorus and as well as biological activity, are influenced by soil pH. The soil pH or the salt composition is affected by the initial soil parent material, category of plant cultivated, the climatic condition especially the volume of rainfall, and the soil age. Soil pH is an important variable as a quality indicator as it controls different biological and chemical processes happening in soil. Soil pH measures the acidity or alkalinity which is critical in managing cropping process since it controls the nutrient availability for the crop. Thus the soil pH influences health of the crop and soil. Majority of the agricultural crops prefers a pH value between 5.4 and 7.4 range. The soil pH values can be grouped as follows [1];

Table 1. Category of soil-based pH value

Group	pH value range
Very High acidic	<3.5
Highly Acidic	3.5 – 4.4
Acidic	4.5 – 5.0
Moderate Acidic	5.6 – 6.0
Slightly Acidic	6.1 – 6.5
Neutral	6.6 – 7.3
Slightly alkaline	7.4 – 7.8
Moderately alkaline	7.9 – 8.4
Alkaline	8.5 – 9.0
High alkaline	>9.0

The mineral composition of the parent soil material, also the environmental factors affected the parent soil determines the soil pH value. The lateral or downward flow of water over the soil leaches the soil thus making the soil acidic in such as humid situation. But in dry and arid environments the soil is less insensitive to the weathering and leaching affects, soil will be neutral or alkaline [2]. Heavy rainfall, cultivated growth, volume of fertilizer, acidic rain, and oxidative condition are all factors that contribute to soil acidity. Rhizobium bacteria that fix nitrogen are impeded by low pH, while herbicides of the imidazolinone family breakdown slowly in acidic soil. Phosphorus and most micronutrient availability is reduced at high pH, and herbicides break down slowly.

Crop growth and development are affected by both acidic and alkaline soils. Crops cultivated in acid soils, for example, may be subjected to toxicity from Aluminium, and Manganese, as well as nutrient shortages in Calcium and Magnesium. When aluminium is present in the ionic Al3+ state, it causes aluminium poisoning, which is the most common concern in acid soils. At soil pH less than 5.0, the aluminium ion Al3+ is easily soluble in all of its form (acidic condition). Aluminum cannot be considered as a nutrient for plants, but rather its ionic form enters the root of crop through osmosis. Aluminum affects root development and growth through interfering with nutrient uptake and transport, division of cell, construction of cell wall, and enzyme function. High alkalinity in soil (sodic soils) on the other hand, makes poor infiltration, low conductivity (hydraulic), and limited retention of soil moisture, causing crop water stress.

For measuring the pH value of the soil the respective region's spectral reflectance data available in the visible NIR to SWIR band is used [3]. Soil salinity refers to a condition in which water soluble minerals in the crop roots zone obstruct crop growth. Soil testing, which determines the amount and kind of salts present, is used to determine the severity of the consequences and strategies for dealing with the problem. The salinity index depicts how much salt is contained in soils. Salinization of soil is a common types of soil degradation, particularly in dry and semi-arid areas where rather than evaporation the precipitation will be high. The Soil Remote Sensing Index (SRSI) can be considered as the transformation and synthesis index of the following spectral indices salinity index (SI) and the Normalized Difference Vegetation Indices (NDVI). The R and NIR bands are used to calculate the brightness index (BI). All the salinity indices and other spectral indices mentioned above including (NDSI, BI, and SRSI) have high correlation and hence in this study only soil salinity index or NDSI is used individually in predicting the soil pH values. The detailed methodology adopted in data collection, pre-processing, and model training are explained in the section 3 of this paper. The experimental procedure and the respective results were briefed in the section 4. Finally the section 5 concludes the paper with a detailed note on the scope for future work.

Table 2. Summary of spectral indices considered

SNo	Spectral Indices	Mathematical Equation
1	Soil salinity Index SI_1	$\sqrt{G \times R}$
2	Soil salinity Index SI_2	$\sqrt{G^2 + R^2 + NIR^2}$
3	Soil salinity Index SI_3	$\sqrt{G^2 + R^2}$
4	Normalized Difference Salinity Index (NDSI)	$\frac{R - NIR}{R + NIR}$
5	Brightness Index, BI	$\sqrt{R^2 + NIR^2}$
6	Soil Remote Sensing Index, SRSI	$\sqrt{(NDVI - 1)^2 + {SI_1}^2}$

2. LITERATURE SURVEY

Remote Sensing (RS) is commonly known as a low expenditure, quick, and repeatable method of obtaining quantitative and distributed (spatially) data on soil parameters. In soil research, the increasing capability of RS technologies, GIS, and spatial data models is opening up novel possibilities. Reflectance of the soil reacts to both static elements like soil type, mineralogy, and elevation, as well as time varying factors like soil moisture, tillage, texture or roughness, and residual of the crop, according to numerous studies. RS technology can be used to obtain both invariant and variant factors. Most studies, on the other hand, focus on a specific soil attribute, such as soil carbon, potassium, pH, or available water capacity, and do so exclusively at the very smaller study regions [7]. The goal of the research published in [8] was to see if field hyper-spectral reflectance, multi-spectral satellite imagery, and atmospheric variables could be used to forecast pH value changes in the soil of the croplands in the black soil region of Dehui in Northeast China. Only a few studies have attempted to quantify regional-scale changes in soil

qualities throughout time. Furthermore, the region or location specific characteristics of the correlations between RS variables measured and soil parameters is one of the most significant obstacles to widespread use of RS methods in soil research. Furthermore, due to the intimate association among both soil organic carbons (SOCs) and the spectral information, the majority of research has focused on SOCs, whereas models characterizing soil acidification have only been applied to soils infrequently. As a result, it was critical to evaluate the pH state of the soil and its temporal fluctuations at a broad scale. Overuse of chemical fertilizers has altered soil conditions and impacted crop yields across the Indian subcontinent's agricultural regions. Conversely, very few tests were carried to track and track variations in the soil pH [9][11]. As a result, precise assessment of the spatiotemporal variations of soil pH is critical for agricultural production sustainability and environmental preservation.

This study was aimed at utilizing spectral reflectance data extracted from the satellite information, and respective soil pH measured from laboratory tests, which was collected from the spatio-temporal data in soil pH of Nilgris region of TamilNadu from the period between September, 2020 and July, 2021. The objective is to i. examine the relationship between the soil pH obtained from the test reports and salinity indices derived from the multi-spectral data. ii. build predictive models based on the salinity indices and the soil pH as the independent and dependent variables respectively. iii. forecast the significant changes in soil pH over the region using the trained prediction model.

3. METHODOLOGY

The proposed approach aims at exploiting the relationship between the vegetation indices and the soil salinity for estimating the soil pH value similar to the work proposed in [10, 12]. An accurate regression model is required to uncover the correlation between the satellite bands and the soil pH.

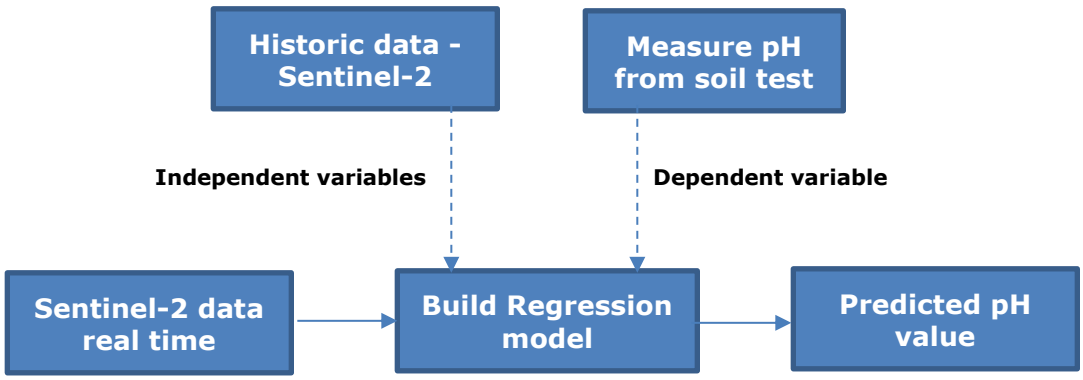

Fig.1 Process flow in prediction of soil pH

3.1 Data Collection

The geography of the study region is unusual due to its location in the Nilgiris Mountain range, which is site to the Doddabetta peak, which is regarded the tallest peak in South India. Nilgris is located in TamilNadu, India, at comparatively high latitudes (11° 08' to 11° 37' N, 76° 27' E to 77° 4' E), with an elevation peak of 2,637 m and a total size of 2552.50 km2. The district's high elevation causes low temperatures, which are exacerbated by the high moisture level in the atmosphere caused by vegetation exhalation. Lateritic soil, Red sandy soil, Red loam, black dirt, Alluvial soil, and Colluvial soil are the five principal soil types found in the Nilgiri district. Lateritic soil covers the majority of the district. Small patches of red sandy soil and red loams can be found. In the valleys, block soil develops, and water logging is widespread during the monsoon season. Along the lowlands and major river channels, alluvial and colluvial soils can be found. In total 1200 soil samples were collected from the study region and soil pH value was estimated based on laboratory tests. The corresponding satellite data was derived from the open source platforms over the study region. The multi-spectral satellite data was processed to remove the influence of different atmospheric conditions.

3.2 Atmospheric Correction

One of the most important aspects of result analysis is satellite data preprocessing. Because it might alter the final outcome, atmospheric correction is one of the most significant preprocessing stages. The primary goal of atmospheric correction is to correct satellite picture effects through the assessment of optical properties [4]. The conversion of top-of-atmosphere radiation obtained by sensors to surface reflectance is known as atmospheric correction. The procedure for calculation of Bottom of Atmosphere (BOA) reflectance from the satellite data (ToA – Top of Atmosphere reflectance) is presented in this section [5]. For converting the ToA to radiance it becomes essential to estimate the volume of solar radiation on top of the atmosphere, and divide the radiance by this number, taking into account the difference between radiance and irradiance as well as the angle of incident solar radiation. The relation between the ToA and radiance can be expressed

$$R_{ToA} = \frac{\pi * L}{\frac{1}{d^2} * I_{sun} * \cos\theta_{sun}} \quad \text{Eq. 1}$$

RTOA – ToA reflectance

L – Solar radiance

d – distance between sun and earth

ISun – average extra-terrestrial solar irradiance,

θSun – the angle between the direction toward the Sun and the normal of the Earth's surface.

As the distance between the Sun and the Earth fluctuates during the year, and the quantity of solar irradiance varies with it, compensate this by the factor 1/d2, where d is the precise distance between the Sun and the Earth at the time the image was taken. This would calculate the incoming irradiance per unit surface area if the Sun was at zenith (straight overhead). To calculate the irradiance per exposed unit surface area when the Sun is not at zenith, multiply this amount by the cosine of the solar zenith angle.

The reflected sunlight can be expressed mathematically as

$$L = \frac{\tau\rho\left(E_{dir} + E_{dif}\right)}{\pi} + L_p \quad \text{Eq. 2}$$

where L – radiance estimated at the sensor, τ represents the transmissivity, ρ represents the reflectance from the earth surface, and E_dir and E_dif denotes the direct and diffuse solar irradiance respectively. Thus the mathematical expression for estimating the surface reflectance from the solar radiance can be expresses as;

$$\rho = \pi(L - L_p)/\tau(E_{dir} + E_{dif}) \quad \text{Eq. 3}$$

3.3 Gradient Boosted Regression Analysis

It is a kind of additive model that performs prediction by combining results of a set of based models. This class of models can be expressed mathematically as;

$$g(x) = f_0(x) + f_1(x) + f_2(x) + \cdots$$

In boosted tree model decision trees are used as base classifiers. Combining multiple weak learners for yielding a better performance can be termed as model ensembling. In case of Random Forest algorithm the base classifiers are constructed independently by using subsample of the available dataset whereas the gradient boosting regression uses the gradient boosting as ensembling technique. Consider the least squares regression which aims at approximating a function F to estimate the values of y^=F(x) thus minimizing the error in estimation

$$\frac{1}{n}\sum_{i}(\hat{y}_i - y_i)^2 \quad \text{Eq. 4}$$

Where y^_i = predicted value F(x_i), y_i represents the actual target value, and n is the number of training samples.

Now consider the gradient boosting algorithm having M iterative stages. At each stage of the iteration some of the weak learner may yield y^_i= y¯ (y¯ – average of y). In order to improve the performance of the weak learner the gradient boosting algorithm adds new estimator h_m (x).

$$F_{m+1}(x) = F_m(x) + h_m(x) = y \quad \text{Eq. 5}$$

$$h_m(x) = y - F_m(x) \quad \text{Eq. 6}$$

Thus the gradient boosting algorithm will attempt to fit the function to the residual y- F_m (x). The residuals of a given model are proportional to the negative gradients of the loss function.

$$L_{MSE} = \frac{1}{n}(y - F(x))^2 \quad \text{Eq. 7}$$

$$-\frac{\partial L_{MSE}}{\partial F} = \frac{2}{n}\big(y - F(x)\big) = \frac{2}{n}h_m(x) \quad \text{Eq. 8}$$

4. EXPERIMENTS AND RESULTS

Soil pH prediction model was developed using Gradient boosting regression based on 1200 lab test reports from Nilgris region of TamilNadu and respective multispectral satellite data. The soil samples were split in to 75% (900) for the model training and 25% (300) for evaluating the performance of the model. Using cross validation procedure, error residual deviance and the Root Mean Square Error were determined for analyzing the performance of the model. The tuned prediction model was used to predict the pH value of the entire region. The Sentinel-2A satellite data was filtered based on the geometric or polygon coordinates and cloud mask. Then the multispectral data are processed to remove the effects of atmospheric conditions.

The changes in the soil pH map can be obtained by subtracting the current predicted pH values with historical soil pH map. Using modern software the changes in soil pH values can be obtained after a spatio-temporal analysis. The mean value of the soil pH observed in the soil samples was 6.027735, and the standard deviation was 0.925163. The maximum and minimum soil pH value was 8.860000 and 4.060000 respectively. The correlation between different salinity indices was analyzed using pearson's correlation coefficient which is calculated by computing the division of covariance of the two variables by their product.

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} \quad \text{Eq. 9}$$

Consider n number of samples represented as {(x_1,y_1),…..,(x_n,y_n)} where in the pearson's correlation coefficient, r_xy can be mathematically expressed as;

$$r_{x,y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad \text{Eq. 10}$$

where n is the number of samples; x_i,y_i denotes the individual sample; and $\bar{x}$ is the mean of all the samples.

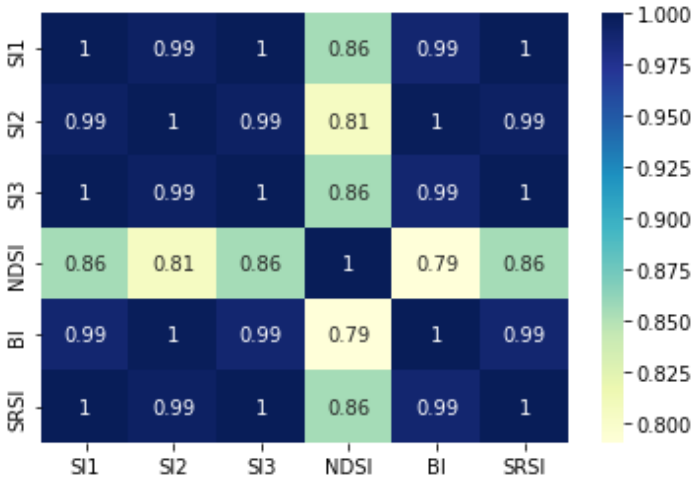

Fig. 2 Correlation among salinity indices

In the red and NIR bands, saline soils with a smooth and light salty crust surface had a higher spectral reflectance, whereas saline soils with a coarse dark puffy surface crust had a lower spectral reflectance. It confirms that salinity soil reflectance is determined by spectral features such as salt crust existence, soil color, and moisture content, all of which have a cumulative effect on the quantity of reflection. Fig. 3 shows the relation between the NDSI and the soil pH value.

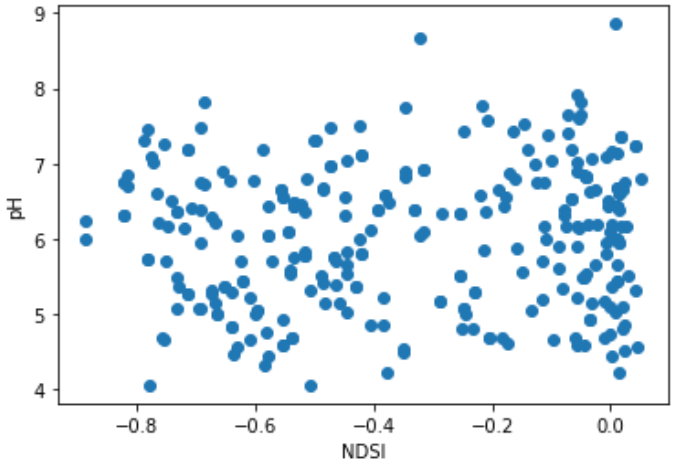

Fig. 3. Scatter plot of NDSI vs pH

The residual deviation indicates how effectively the intercept and inputs can predict our outcome. Smaller is preferable. The greater the difference between null and residual deviation, the better our input variables were at estimating the output variable. It is observed that the deviance increases when the iteration length is increased in gradient boosting.

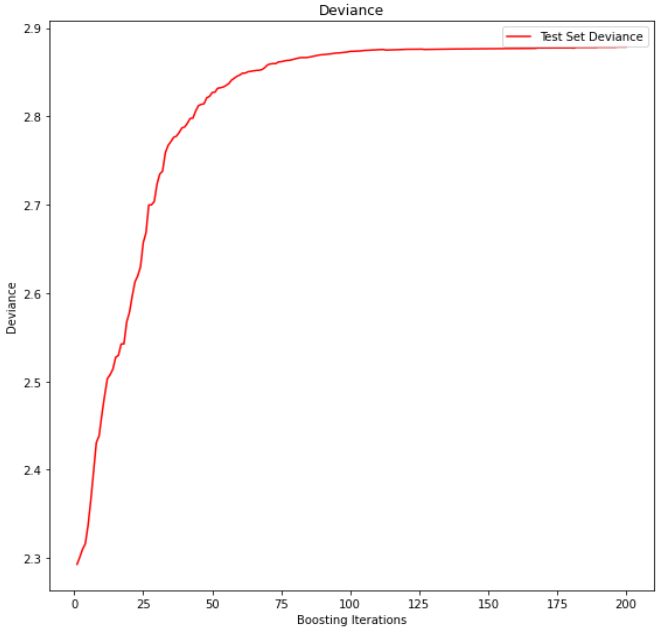

Fig. 4 Deviance of the Gradient Boosting regression

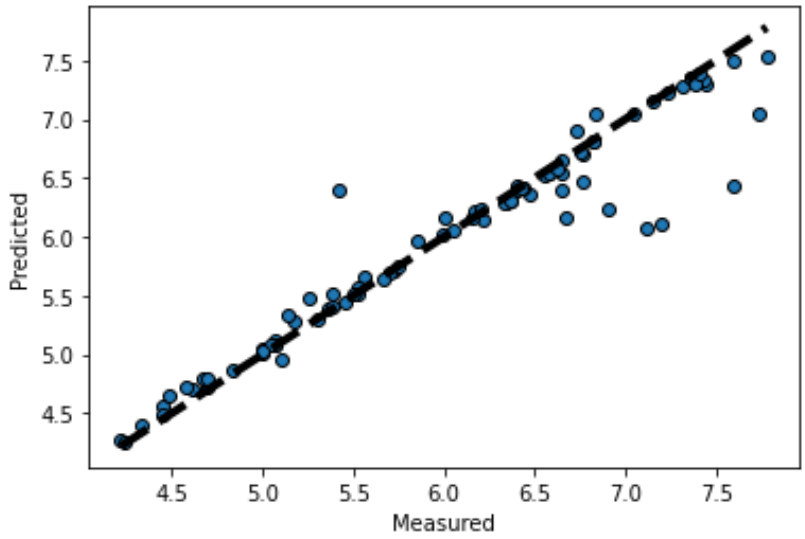

Fig. 5 Results of Gradient Boosting Regression

5. CONCLUSION

This paper describes the method for predicting the soil pH value from the satellite data. The efficiency of the model and comparison of the measured soil pH value with the predicted values proved to be useful for mapping the soil pH across a geographical region using any geospatial techniques. This type of prediction shall help geologist and agronomist in delineating the saline areas. The methodology and the approach followed for constructing the prediction model makes this as a resilient and promising mechanism for soil salinity forecasting. As a future work the same approach can be expanded in predicting other essential soil quality indicators. Also the model can be made robust by training with samples extracted from the different geographical regions.

REFERENCES

1. Soil Survey Staff. Soil survey laboratory methods manual. In: Burt R, editor. Soil Survey Investigations Report No. 42, Version 5.0. 5th ed. U.S. Department of Agriculture, Natural Resources Conservation Service. 2014. pp. 276-279.
2. Bloom PR, Skyllberg U. Soil pH and pH buffering. In: Huang PM, Li Y, Sumner ME, editors. Handbook of Soil Sciences: Properties and Processes. 2nd ed. Boca Raton, FL: CRC Press; 2012. pp. 19-14. ISBN: 9781439803059.
3. Ghazali MF, Wikantika K, Harto AB, Kondoh A (2020) Generating soil salinity, soil moisture, soil pH from satellite imagery and its analysis. Inf Process Agric 7(2):294–306
4. Chrysoulakis, N.; Abrams, M.; Feidas, H.; Arai, K. Comparison of atmospheric correction methods using ASTER data for the area of Crete, Greece. Int. J. Remote Sens. 2010, 31, 6347–6385.
5. R. Richter and Schläpfer, D.:Atmospheric/Topographic Correction for Satellite Imagery: ATCOR-2/3 UserGuide", DLR IB 565-01/11, Wessling, Germany, 2011.
6. Mayer, B. and Kylling, A.: Technical note: The libRadtran software package for radiative transfer calculations - description and examples of use, Atmos. Chem. Phys., 5, 1855-1877, 2005.
7. Xu, Y.M., Smith, S.E., Grunwald, S., Abd-Elrahman, A., Wani, S.P., 2017. Evaluating the effect of remote sensing image spatial resolution on soil exchangeable potassium prediction models in smallholder farm settings. J. Environ. Manage. 200, 423–433.
8. Zhang, Yue, et al. "Estimating temporal changes in soil pH in the black soil region of Northeast China using remote sensing." Computers and Electronics in Agriculture 154 (2018): 204-212.
9. Zhu, Q.C., de Vries, W., Liu, X.J., Hao, T.X., Zeng, M.F., Shen, J.B., Zhang, F.S., 2018. Enhanced acidification in Chinese croplands as derived from element budgets in the period 1980–2010. Sci. Total Environ. 618, 1497–1505.
10. Azabdaftari, A., and F. Sunarb. "Soil salinity mapping using multitemporal Landsat data." The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences 7 (2016): 3-9.
11. S. Shiva Prakash, "Educating and communicating with deaf learner's using CNN based Sign Language Prediction System", International Journal of Early Childhood Special Education (INT-JECSE) Vol 14, Issue 02, 2022.
12. Yu, Xinyang, et al. "Precise Monitoring of Soil Salinity in China's Yellow River Delta Using UAV-Borne Multispectral Imagery and a Soil Salinity Retrieval Index." Sensors 22.2 (2022): 546.