Publ. Astron. Obs. Belgrade No. 101 (2021), 209 – 215 *Poster*
DOI: 10.5281/zenodo.5645712

A LINEAR REGRESSION MODEL FOR SOIL SALINITY PREDICTION IN THE GREAT HUNGARIAN PLAIN USING SENTINEL 2 DATA

G. SAHBENI

Department of Geophysics and Space Science, Eötvös Loránd University,
Budapest, Pázmány Péter stny. 1/A, 1117, Hungary
E-mail: gsahbeni@caesar.elte.hu

Abstract: Salts occur naturally within soil and water. When exceeding the thresholds, salinity becomes a severe threat, damaging agricultural productivity, water and soil quality, biodiversity, and infrastructures. Multispectral data retrieved from Sentinel-2 MSI sensor were used in this study to predict soil salinity in the Great Hungarian Plain. For this purpose, samples were collected from the upper layer of soil between mid-September and mid-October in the Hungarian Soil Monitoring System framework. The application of multiple linear regression analysis between salt content (g/kg) and remotely sensed data revealed a highly moderate correlation with a coefficient of determination R^2 equals 0.52, a p-value equals 0.001198, and an RMSE equals 0.194 g/kg. The model can be employed to highlight soil salinity levels in the study area and understand the efficiency of land management strategies, considering its moderate predictive power.

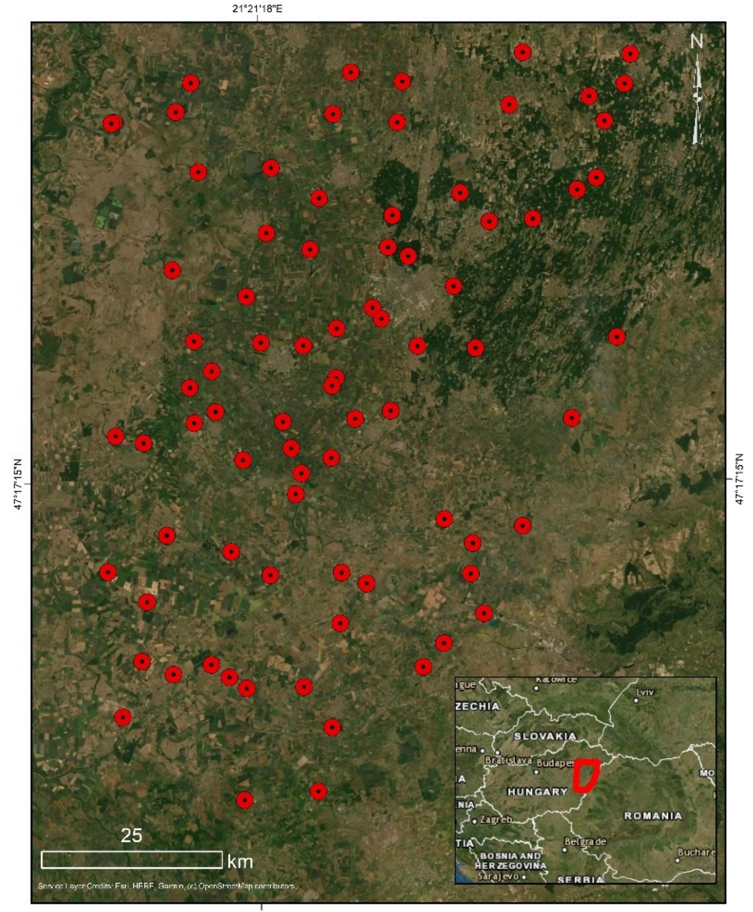
Keywords: Soil Salinity, Sentinel 2, Multiple Linear Regression.

1. INTRODUCTION

Salinization is a widespread land degradation form induced naturally by parent material weathering or artificially due to irrigation with saline water (Ondrasek and Rengel 2021). It occurs in dry regions where water balance is negative. Globally, 397 million ha are affected by salinization (FAO 2000), with 3.8 million ha of saline soils in Europe (Stanners 1995). In Hungary, salt-affected soils cover 13% of the total area and exhibit the most natural continental salinization features (Tóth 2009). Using spaceborne and airborne products coupled with adequate methods for salinity prediction has become a valuable alternative to map salt behavior in the subsoil and maintain its levels under control. In this context, many scholars have explored the efficiency of multispectral, hyperspectral, and radar sensors in salinization monitoring (Weng et al. 2010, Bannari et al. 2018, Szatmári et al. 2020, Sahbeni 2021a). This study aims to examine the importance of multispectral sensors, notably Sentinel-2 MSI, in predicting soil salinity with lower costs and acceptable accuracy.

2. STUDY AREA

The study area covers 6903.5 km^2 (Figure 1), at an average elevation of 89m above sea level. It is characterized by a mean yearly temperature of 11°C (Tóth et al. 2014), a mean precipitation yearly rate of 560 mm, and a mean evaporation rate of 900 mm (Hungarian Meteorological Service 2018).



Author: G.SAHBENI
Data source: RISSAC, 2019

Figure 1: Location of the study area and sampling sites.

3. MATERIALS AND METHODS

3.1. SOIL SAMPLES

Eighty-one soil samples were collected in the Hungarian Soil Monitoring System (SIMS) framework. SIMS is a national soil monitoring program that collects soil data from around 1235 sites and generates Hungary's most unified, thematically detailed, and up-to-date soil database (Bakacsi et al. 2019). An average sample is taken from 9 drillings from the 0-30 soil layer in a 50 m diameter circle (Berényi-Üveges 2015). Salt content values are measured from the saturated paste extract according to the Hungarian Standard MSZ-08-0206/2-1978 (MSZ 1978).

3.2. REMOTELY SENSED DATA

Once the Sentinel-2 MSI image was downloaded from the European Space Agency (ESA) Copernicus portal, atmospheric and radiometric calibrations were applied using Sentinel-2 Toolbox. Then, spectral indices (Table 1) were computed using ENVI IDL 5.3. Additionally, we acquired an SRTM digital elevation model provided by the OpenTopography facility to explore potential associations between salinity levels and elevation. The digital elevation model was reprojected to the Universal Transverse Mercator (UTM) coordinate system using WGS 1984 datum assigned to north UTM Zone 34. Corresponding values to field data were retrieved using ArcMap 10.3, and a database including remotely sensed data and salt content values was developed.

Table 1: Spectral indices and their mathematical expressions.

Index	Expression
NDVI	(NIR − R) / (NIR + R) (Rouse et al. 1974)
NDSI	(R − NIR) / (R + NIR) (Khan et al. 2005)
VSSI	2 * G – 5 * (R + NIR) (Dehni and Lounis. 2012)
BI	$\sqrt{(R^2 + NIR^2)}$ (Khan et al. 2005)
SI	(R * G) / B (Allbed et al. 2014)
SI_1	$\sqrt{(G * R)}$ (Douaoui et al. 2006)
SI_2	$\sqrt{(R * NIR)}$ (Dehni and Lounis 2012)
SI_3	$\sqrt{(G^2 + R^2 + NIR^2)}$ (Douaoui et al. 2006)
SI_4	$\sqrt{(G^2 + R^2)}$ (Yahiaoui et al. 2015)
RVI	R / NIR (Krtalic et al. 2019)
DVI	NIR − R (Tucker. 1979)
Int_1	(G + R) / 2 (Bouaziz et al. 2011)
Int_2	(G + R + NIR) / 2 (Bouaziz et al. 2011)
SR	(R − NIR) / (G + NIR) (Dehni and Lounis 2012)
SAVI	(1 + L) ∗ (NIR − R) / (NIR + R + L) (Huete 1988)
$SSSI_1$	$SWIR_1 - SWIR_2$ (Bannari et al. 2008)
$SSSI_2$	$(SWIR_1 * SWIR_2 - SWIR_2 * SWIR_2) / SWIR_1$ (Bannari et al. 2008)

3.3. REGRESSION ANALYSIS

A multiple linear regression analysis was conducted via RStudio to define the statistical significance of independent variables in relationship with soil salinity variation. In this context, we employed the R squared model selection to extract only significant variables. R squared represents the proportion of variance for a dependent variable which independent variables can explain. Thus, a model with a larger R-squared (equation 1) value can explain a more significant percentage of data variance (Romero 2007).

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_\mathbf{i})^2}{\sum(y_i - \bar{y})^2} \qquad (1)$$

Where $\hat{y}_i$ is the estimated value, y_i is the actual value, and $\bar{y}$ is the mean value.

4. RESULTS

Data distribution is positively skewed, according to the descriptive analysis report. The mean equals 0.49 g/kg, whereas the median equals 0.3 g/kg. Besides, a spatial variability was found due to the discard between the minimum (= 0 g/kg) and the maximum (= 5.6 g/kg). Table 2 summarizes the main statistical parameters of field data.

Table 2: Descriptive statistics of salt content samples.

	Minimum	1st quantile	Median	Mean	3rd quantile	Maximum
Salt content (g/kg of soil)	0	0.2	0.3	0.49	0.6	5.6

The final model’s main characteristics are presented in Table 3.

Table 3: Characteristics of the final model.

R^2	p-value	RMSE	Significant Variables
0.52	0.001198	0.1942	NDVI, SAVI, RVI, DVI, BI, VSSI, SI, SI1, SI2, Int1, B2, B11, and B12

Figure 2 shows the relationship between measured and estimated salinity values using the linear regression model. We split the dataset into two parts: a training set (70%) used to tune the model, and a test set (30%) used to check its statistical significance. Overall, the model yielded acceptable results with a coefficient of determination equal to 0.52, showing a highly moderate correlation and a p-value close to zero (< 5%), revealing a strong statistical significance. Nevertheless, a quite high prediction error was produced due to data redundancy.

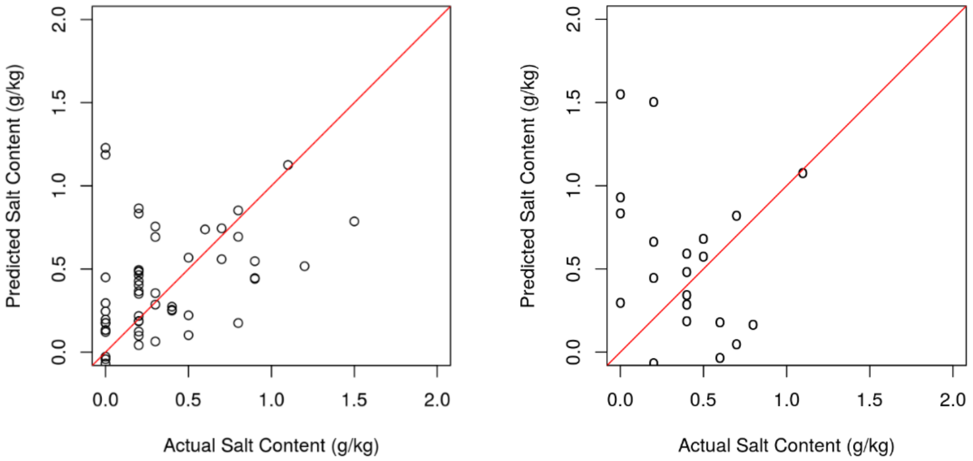


Figure 2: Relationship between actual and predicted salt content values (g/kg); (a) Training set (70%) and (b) Test set (30%).

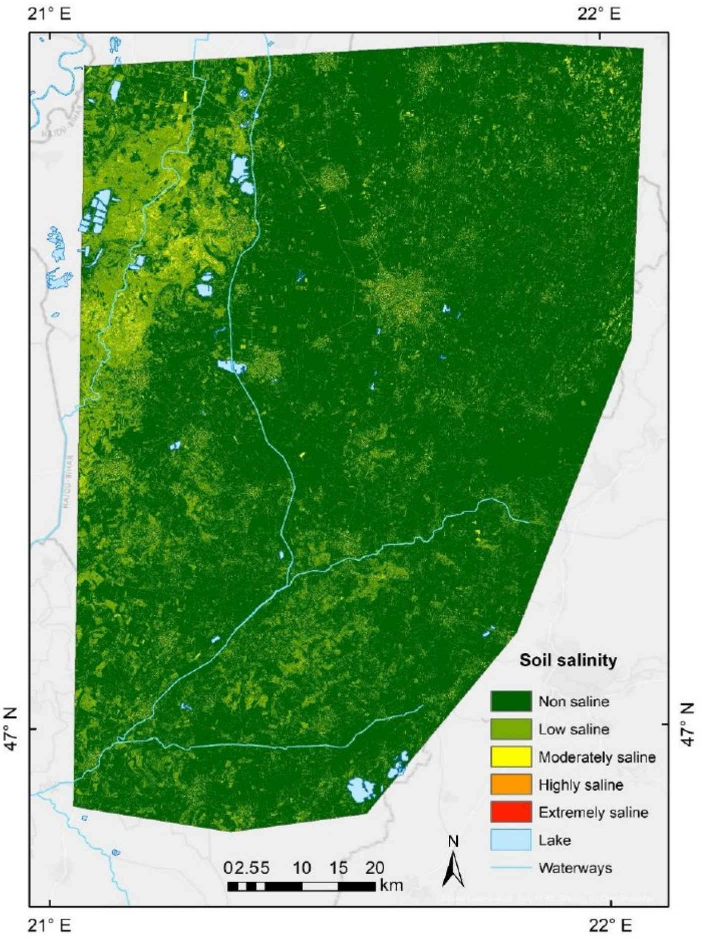


Figure 3: Soil salinity prediction map using the final model.

Around 4% of the total pixels were assigned negative values. This can be explained by the residual noise caused after atmospheric correction (Weng et al. 2010), which will be investigated in future studies. Based on Figure 3, 80% of pixels are classified as non-saline soils and 18% as low saline. This distribution of classes was expected due to the dominance of non-saline samples in the database, followed by low saline ones.

5. CONCLUSIONS

This study demonstrates the efficiency of Sentinel-2 MSI data in predicting soil salinity with acceptable accuracy. Hence, regression analysis offers a reliable approach for soil salinity assessment with affordable costs. The model explains 52% of the data spatial variance, with an RMSE equals 0.1942 g/kg of soil. Overall, remote sensing depicts a valuable alternative for conventional methods when coupled with representative field data. Yet, further research will be conducted to reduce prediction errors and overcome the issue of data multicollinearity.

An improved version of this research can be found in Sahbeni (2021b).

Acknowledgements

The author expresses her gratitude to Prof. L. Pásztor from the Research Institute for Soil Sciences and Agricultural Chemistry for providing field data[1].

The Sentinel 2A image was downloaded from the Copernicus Open Access Hub[2].

SRTM elevation model was downloaded in GeoTIFF format from the OpenTopography Facility with support from the National Science Foundation under NSF Award Numbers 1833703, 1833643, and 18336323[3].

References

Allbed, A., Kumar, L., Sinha, P.: 2014, *Remote Sensing*, **6**, 1137.

Bakacsi, Z., Tóth, T., Makó, A., Barna, G., Laborczi, A., Szabó, J., et al.: 2019, *Hung Geogr Bull.*, **68**, 141.

Bannari, A., El-Battay, A., Bannari, R., Rhinane, H.: 2018, *Remote Sens.*, **10**, 855.

Bannari, A., Guedon, A. M., El Harti, A., Cherkaoui, F. Z., El Ghmari, A.: 2008, *Communications in Soil Science and Plant Analysis*, **39**, 2795.

Berényi-Üveges, J.: 2015, *Workshop to develop 250000 soil databases for Danube Basin using eSOTER methodology*.

[1] https://www.mta-taki.hu/en

[2] https://scihub.copernicus.eu

[3] https://opentopography.org

Bouaziz, M., Matschullat, J., Gloaguen, R.: 2011, *Comptes Rendus Geoscience*, **343**, 795.

Dehni, A., Lounis, M.: 2012, *Procedia Engineering*, **33**, 188.

Douaoui, A.E.K., Nicolas, H., Walter, C.: 2006, *Geoderma*, **134**, 217.

FAO: 2000, *Extend and causes of salt-affected soils in participating countries, Global Network on Integrated Soil Management for Sustainable Use of Salt affected soils*.

Huete, A.R.:1988, *Remote Sensing of Environment*, **25**, 295.

Hungarian Meteorological Service: 2018. *Precipitation conditions of Hungary*.

Khan, N.M., Rastoskuev, V.V., Sato, Y., Shiozawa, S.: 2005, *Agric. Water Manage.*, **77**, 96.

Krtalic, A., Prodan, A., Racetin, I.: 2019, *19th SGEM International Multidisciplinary Scientific Geo Conference Proceedings*, **19**, 449.

MSZ 1978: 1978, Hungarian Standard no. MSZ 08–0206–2:1978. *Hungarian Standards Institution, Budapest (in Hungarian)*.

Ondrasek, G., Rengel, Z.: 2021, *The Science of the total environment*, **754**.

Romero, A.: 2007, *A note on the use of r-squared in model selection, Department of Economics, College of William and Mary*.

Rouse, J.W., Haas, R.H., Schell, J.A., Deering, D.W.: 1974, *Third Earth Resources Technology Satellite 1 Symposium*, **I**, 309.

Sahbeni, G.: 2021a, *SN Appl. Sci.*, **587**, 1.

Sahbeni, G.: 2021b, *Open Geosciences*, **13**, 977.

Stanners, D.: 1995, *Europe's Environment: The Dobris Assessment. Office for Official Publication of the European Communities, Luxembourg*.

Szatmári, G., Bakacsi, Z., Laborczi, A. Petrik, O., Pataki, R., Tóth, T., et al.: 2020, *Remote Sens.*, **12**, 4073.

Tóth, T.: 2009, *World Soil Resources Report*, **104**, 201.

Tóth, T., Balog, K., Szabo, A., Pásztor, L., Jobbágy, E. G., Nosetto, M. D., Gribovszki, Z.: 2014, *AoB PLANTS*, **6**, plu054.

Tucker, C.J.: 1979, *Remote Sensing of Environment*, **8**, 127.

Weng, Y.L., Gong, P., Zhu, Z.L.: 2010, *Pedosphere*, **20**, 378.

Yahiaoui, I., Douaoui, A., Zhang, Q., Ziane, A.: 2015, *J Arid Land.*, **7**, 794.