



# Evaluation of Different Environmental Covariates Performance for Modeling Soil Salinity Using Digital Soil Mapping in a Susceptible Irrigated Rural Area

Judit Rodríguez-Fernández<sup>1</sup>(✉), Montserrat Ferrer-Juliá<sup>2</sup>, and Sara Alcalde-Aparicio<sup>2</sup>

<sup>1</sup> Facultad de Ciencias Biológicas y Ambientales, Universidad de León, Campus de Vegazana, 24071 León, Spain

[jrodrf07@estudiantes.unileon.es](mailto:jrodrf07@estudiantes.unileon.es)

<sup>2</sup> Grupo de Investigación Q-GEO, Universidad de León, Campus de Vegazana, 24071 León, Spain

**Abstract.** Soil is an indispensable resource for the development of the ecosystems, also working as a support for the human activities, being essential for the agricultural productivity. There are many soil degradation risks that cause a quality deterioration. One of the major risks is soil salinity, caused by the accumulation of salts both naturally and anthropically. For this reason, prevention measures are needed. To this end, soil properties inference and modelling result essential. Thus, the main objective of this research is to find the most useful environmental covariates for modeling soil salinity through the application of the Digital Soil Mapping (DSM) methodology in an irrigated rural area in Castile and León (Spain). For this purpose, 132 soil samples from two different laboratories were used, which contained electrical conductivity measured in saturated paste ( $EC_x$ ). In addition, several environmental covariates related to soil salinity were employed to perform a statistical analysis through the combination of multiple linear regression (MLR) and generalized linear models (GLM). Afterwards, the best prediction model and its explanatory covariates were selected. The MLR showed  $R^2$  values between 0.382 and 0.581 for the laboratories analyzed. In turn, all the models almost had the same main covariates, which were associated to remote sensing indices and topographic variables. Finally, it was concluded that the method is useful to determine the most important variables for modeling soil salinity, allowing more accurate predictions, identifying which susceptible areas need preventive measures and helping to achieve those SDGs targets that involve soil's conservation.

**Keywords:** Soil salinity · Electrical conductivity · Digital soil mapping

## 1 Introduction

Soil is an essential natural resource for the development of life in ecosystems, harboring large amount of biodiversity, functioning as a store and supply of water and nutrients and being the support for different human activities such as agriculture.

The importance of soil is even highlighted in the 2030 Agenda Sustainable Development Goals (SDGs). Several SDGs involve directly and indirectly the soil into their targets. In turn, Goals 12 “Responsible consumption and production” and 15 “Life on land” mention the use of sustainable production system and agricultural practices for improving soil quality, also preventing its pollution through proper management of chemical and waste. In this way, it also seeks to curb the causes of soil degradation, such as salinization [1].

Therefore, it is necessary to carry out preventive measures, such as the development of cartography and models to predict how salinity will evolve. To this end, the Digital Soil Mapping (DSM) methodology and its derived models, such as the *scorpan* model, are proposed, which are based on the statistical inference of soil properties by searching for the statistical relationship between these properties measured in the field with different auxiliary variables or environmental covariates (climate, lithology, land use, vegetation, topography, etc.), finally extrapolating these relationships to those data lacking areas [2–5].

Thus, the main objective of this research focuses on the application of the Digital Soil Mapping (DSM) method in an irrigated area of Castile and León (Spain) to determine which are the most useful and relevant covariates for modeling soil salinity.

## 2 Study Area

The study area is located between the provinces of León and Zamora (Spain) (Fig. 1). It covers an area of 1500 km<sup>2</sup>, with altitudes between 680 and 930 m and generally flat relief. The average annual temperature is between 10 and 13 °C, with rainfall between 400 and 500 mm and average annual ETP of up to 800 mm [6, 7].

The dominant lithologies correspond to the Pleistocene and Holocene, formed by alluvial deposits in terrace areas, and by sand, silt and clay in valley bottom areas and river plains. Among dominant soil types, Cambisols are found in terrace zones in the center and north of the study area, and Fluvisols in the floodplains. Finally, the main land use is associated to irrigated crops, with a predominance of maize (65,000 ha), with poplar plantations also standing out in the riparian areas [8].

## 3 Methodology

The DSM methodology was used for the study through the application of the *scorpan* model developed by McBratney et al. [2], which proposes the integration of different environmental variables into a function that allows, in this specific case, the prediction of soil salinity ( $S_s$ ) (Eq. 1).

$$S_s = f(s, c, o, r, p, a, n) \quad (1)$$

These environmental variables, also called “soil-forming factors”, comprise soil (s), climate (c), organisms (o), topography (r), parent material (p), age (a) and spatial position (n). In turn, these variables are defined by different environmental covariates [2].

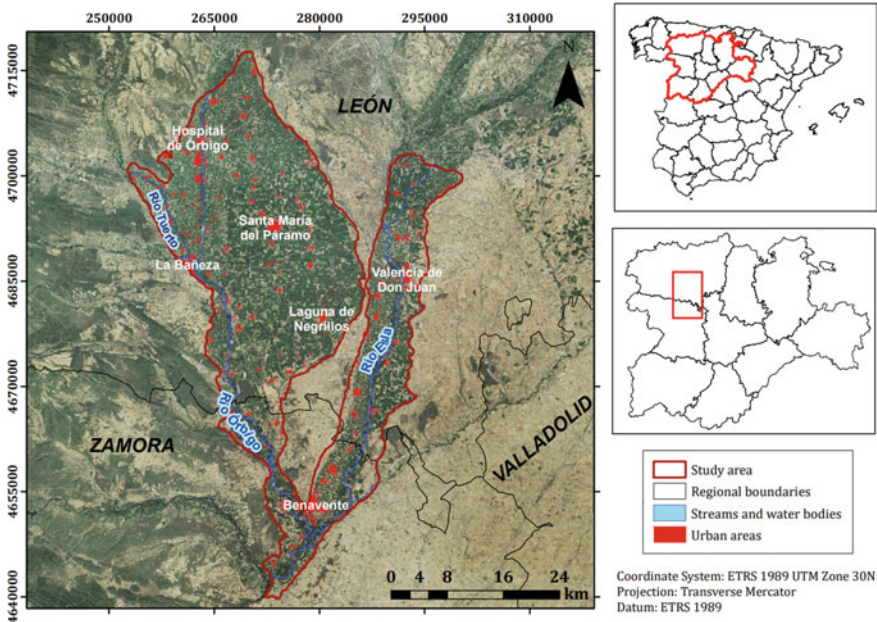
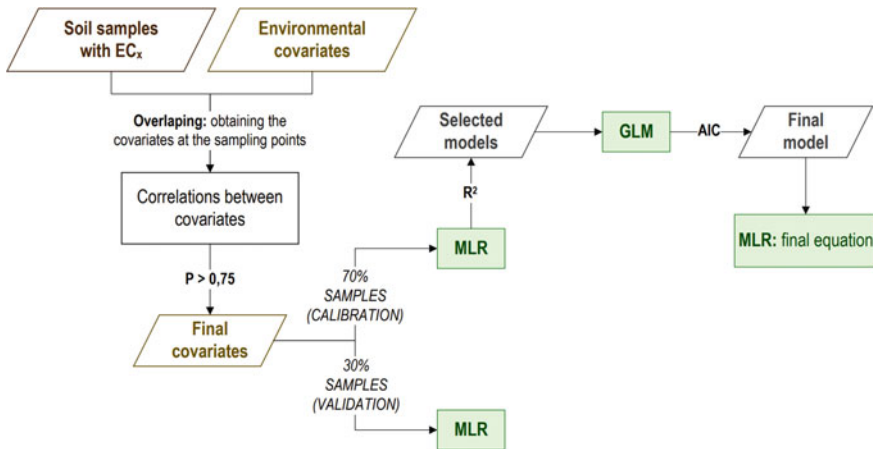


Fig. 1 Location of the irrigated area in León and Zamora, Spain

In this way, based on soil samples with salinity data and the different environmental covariates related to it, a relationship is established that allows estimating the values in unsampled areas. To do this, a spatial overlapping of the soil samples on the mapping of the different covariates was first applied, which allows obtaining at each sampling point both the value of the properties of that soil, as well as the values of each of the environmental covariates used. After obtaining these values, the relationships between the independent variables (environmental covariates) and the dependent variable can be modeled using different statistical methods. In this study, the dependent variable associated with salinity is the electrical conductivity measured in the saturated paste extract ( $EC_x$ ). In turn, the statistical technique that was applied was Multiple Linear Regression (MLR). The flow chart in Fig. 2 outlines the procedure followed.

### 3.1 Soil Data

The Soil Database of Castile and León, obtained from the Agrarian Technological Institute [9] was used to acquire soil data. This contains 914 surface samples from the first 25–30 cm of soil with measurements of electrical conductivity in the saturated paste extract ( $EC_x$ ). However, for the irrigated area the number of samples was lower (132), analyzed also according to the laboratories of origin (Table 1).



**Fig. 2** Applied methodology flow chart

**Table 1** Samples and EC<sub>x</sub> values range (minimum and maximum) for each laboratory analysed in the study area

Laboratory name	Number of samples	Mean EC <sub>x</sub>	EC <sub>x</sub> range (dS/m)
Análisis Integrales	77	0.30	0.2–0.70
APPLUS	55	0.26	0.2–0.69

### 3.2 Environmental Covariates

To obtain the environmental covariates, different data sources were used to derive a total of 24 covariates (Table 2). In the case of land use and lithology, as these are categorical variables, each of their classes was transformed into a binary numerical variable.

It should be noted that all covariates obtained are in raster format with a spatial resolution of 25 m × 25 m. There were covariates with high spatial resolution ( $\leq 25$  m) and others with very low resolution (those ones related to climate). It is always better to change from a high resolution towards a low resolution. For this reason, among those covariates with high resolution, it was decided to work with 25 m. The climate variables in general do not show such high spatial variability. Thus, it was concluded that although it is not the best solution to change from 500 to 25 m, the inherent error could be affordable with the DSM method.

### 3.3 Statistical Analysis

After overlaying and extracting the values of the 24 covariates in the sampling points at the study area, a correlation analysis was carried out to reduce the number of environmental covariates, given that the number of covariates was very high. In this way, those whose Pearson coefficient (P) was greater than 0.75 were eliminated from the analysis, as they give redundant information.

**Table 2** Data sources and environmental covariates used in the study for the different soil forming factors

Factors	Data sources	Spatial resolution	References	Covariates
Soil (s)	Sentinel-2 images	20 m × 20 m	[10]	Carbonates index (CAI) Clay index (CI) Brightness index (BI) Gypsum index (GI) Salinity ratio (SI)
Climate (c)	Assessment of natural regime water resources model (SIMPA)	500 m × 500 m	[11]	Mean annual temperature (°C) Mean annual rainfall (mm) Mean annual ETP (mm)
Organisms (o)	Sentinel-2 images CORINE Land cover (CLC) Castile and León crops and natural land map (MCSNCyL)	20 m × 20 m – 10 m × 10 m	[10] [12] [8]	Normalized Difference Vegetation Index (NDVI) Soil adjusted vegetation index (SAVI) Enhanced vegetation index (EVI) Land uses (2 classes)
Topography-relief (r)	Digital elevation model (DEM)	5 m × 5 m	[13]	Elevation (m) Slope (°) Aspect (rad) Standard curvature Profile curvature Planform curvature Topographic wetness index (TWI) Multi-resolution ridge-top flatness index (MrRTF) Multi-resolution valley bottom flatness index (MrVBF)

*(continued)*

**Table 2** (continued)

Factors	Data sources	Spatial resolution	References	Covariates
Parent material (p)	Castile and León geological map	–	[14]	Lithology (2 classes)

After this, Multiple Linear Regression (MLR) was applied using IBM SPSS Statistics 26 software. This technique is summarized by the following equation (Eq. 2), in which  $Y$  is the dependent variable,  $X_n$  are the predictors that explain the dependent variable,  $\beta_0$  is the intercept or origin,  $\beta_n$  are the coefficients that represent the weight and relationship of each environmental covariate with the dependent variable, and  $\varepsilon$  are the residual values that cannot be explained by the model [15].

$$Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \dots + \beta_n \cdot X_n + \varepsilon \quad (2)$$

Prior to its application, the dataset was segmented, with 70% of the samples being used for model calibration and the remaining 30% for subsequent validation. With the corresponding calibration samples, the MLR was applied using the “backward elimination” method. Although this method yields numerous models, only two models were selected, taking into account that the value of the coefficient of determination ( $R^2$ ) was adequate considering its significance value, and that the number of covariates was as small as possible.

Subsequently, for the selection of the final model variables, Generalised Linear Models (GLM) were estimated using the covariates of each model chosen, obtaining a value of the Akaike Information Criterion (AIC). In this way, the model with the lowest AIC value was selected or, in the case where the value was similar for both models, the one with the lowest number of covariates was selected.

Finally, once the covariates that best explained the electrical conductivity in each case were known, another MLR was applied again, forcing only these covariates to be used in order to obtain the final equation.

## 4 Results

Due to the low  $R^2$  coefficients achieved after working with all the available soil data in the study area, it was decided to work with two sample groups according to the analytical laboratory. The initial correlation analysis applied to the 132 soil samples showed similarities for both laboratories studied, discarding from the analysis the profile curvature and planform curvature variables, as well as a land use variable and a lithology variable.

Although the soil indices show high correlations, we worked with all of them since each one represents different edaphic characteristics. However, regarding the vegetation indices, NDVI and SAVI were discarded because their correlation with EVI showed a P value of 0.990. After this preliminary analysis, MLR was implemented.

- “Análisis Integrales” Laboratory

Two models were selected with  $R^2$  coefficients of 0.587 and 0.581, respectively. After performing the GLM with these models, an AIC value of  $-124.11$  was obtained for both, and the second model ( $R^2 = 0.581$ ) was finally chosen as it was the one that considered the fewest covariates. It can be seen that the indices, mainly soil indices, are the ones that best explain the  $EC_x$ , although topographic variables are also important (Eq. 3). This model showed a significance value of 0.000 ( $p < 0.001$ ), so the results are statistically representative.

$$\begin{aligned} EC_x = & -0.597 + \left( (4.682 \times 10^{-5}) \times BI \right) + (-0.036 \times LAND\_USE) \\ & + (0.053 \times CURVATURE) + (0.745 \times CAI) + (-0.043 \times TEMP) \\ & + (-0.004 \times ASPECT) + (0.037 \times LITHO) + (-0.461 \times CI) \\ & + (0.001 \times ELEVATION) + (0.566 \times EVI) \end{aligned} \quad (3)$$

- “APPLUS” Laboratory

Again, two models have been selected whose  $R^2$  values are 0.395 and 0.382, respectively. The GLM provides an AIC value of  $-115.66$  for both, with the second model ( $R^2 = 0.382$ ) being selected because it has fewer covariates. The covariates associated with the indices have higher weight, followed also by those corresponding to the topography factor (Eq. 4). In this case, the significance value is higher than 0.005 ( $p = 0.255$ ), so the results are not statistically representative.

$$\begin{aligned} EC_x = & -0.470 + (0.049 \times TEMP) + (-0.001 \times RAIN) \\ & + (-0.081 \times LAND\_USE) + (0.021 \times SLOPE) + (-0.005 \times ASPECT) \\ & + (-0.010 \times MRVBF) + (-0.080 \times CURVATURE) + (0.280 \times SI) \\ & + (0.208 \times EVI) + (0.260 \times CAI) + \left( (-3.036 \times 10^{-5}) \times BI \right) \end{aligned} \quad (4)$$

## 5 Discussion

The results obtained show that the DSM method is useful to know which variables best explain and model salinity, since the obtained  $R^2$  coefficients are quite good (0.581, 0.382). However, there is much uncertainty, which can be caused by the soil samples themselves, either by inhomogeneity of the measurements or by poor spatial distribution of the samples. It may also be due to the use of a high number of environmental covariates and to the error associated to them, as these come from several different data sources.

The indices derived from satellite images stand out, both those associated with the soil factor and the organism factor (Table 2), as well as the covariates corresponding to the topography factor. In the latter case, curvature is the most relevant.

Comparing these results with those obtained by other authors, they show similarities. Omuto et al. [16] applied MLR in a study area located in Lesotho, where they obtained an  $R^2$  value of 0.460, which does not differ much from those obtained in this study (0.581, 0.382). Mosleh et al. [17] also applied MLR, resulting in a worse  $R^2$  (0.110).

On the other hand, Taghizadeh-Mehrjardi et al. [18], although they used superlearning techniques, among all those statistical techniques was MLR which showed an  $R^2$  of 0.230, which is slightly worse than the results of this study.

In turn, these authors corroborate the importance of both satellite image-derived covariates and topographic covariates for spatial modelling of salinity [16, 18, 19]. It is also worth noting that Mousavi et al. [4] concluded that the best results are obtained when both types of covariates are used. In their case, by applying MLR using only satellite indices they obtained an  $R^2$  of 0.506, while using also topographic variables the  $R^2$  value increased to 0.660.

## 6 Conclusions

Following the analysis and discussion in this research, the DSM was concluded to be a useful methodology to obtain the most relevant covariates in the soil salinity modelling; highlighting those associated with the topography factor, as well as the variables corresponding to the indices calculated from satellite images.

Soil salinity is an emerging future challenge in agricultural areas, especially in those that are more susceptible. Salinity threatens soil quality and crops productivity, which means a reduction in supply to the population. Given the previous, further research on this topic is required, using methodologies such as DSM to identify those susceptible areas and allowing to apply preventive measures in order to achieve the SDGs targets.

**Acknowledgements.** This research has been funded by FEDER/Spanish Ministry of Science and Innovation—Agencia Estatal de Investigación/Projects ISGEOMIN (ESP2017-89045-R) and NSOURCES (PID2020-113912GB-I00) and the Ministry of Economy and Competitiveness Project PREVENT (CGL2015-66263-R).

## References

1. United Nations. <https://sdgs.un.org/>. Last accessed 11 Jul 2022
2. McBratney, A.B., Santos, M.M., Minasny, B.: On digital soil mapping. *Geoderma* **117**(1–2), 3–52 (2003)
3. Minasny, B., McBratney, A.B.: Digital soil mapping: a brief history and some lessons. *Geoderma* **264**, 301–311 (2016)
4. Mousavi, S.Z., Habibnejad, M., Kavian, A., Solaimani, K., Khormali, F.: Digital mapping of topsoil salinity using remote sensing indices in Agh-Ghala Plain, Iran. *Ecopersia* **5**(2), 1771–1786 (2017)
5. Zare, S., Abtahi, A., Shamsi, S.R.F., Lagacherie, P.: Combining laboratory measurements and proximal soil sensing data in digital soil mapping approaches. *CATENA* **207**, 105702 (2021)
6. de León Llamazares, A., Arriba Balenciaga, A., De La Plaza, M.C.: Caracterización agroclimática de la provincia de Zamora. Ministerio de Agricultura, Pesca y Alimentación, Madrid (1987)
7. de León Llamazares, A., Arriba Balenciaga, A., De La Plaza, M.C.: Caracterización agroclimática de la provincia de León. Ministerio de Agricultura, Pesca y Alimentación, Madrid (1991)
8. ITACYL. <https://mcsncyl.itacyl.es/>. Last accessed 12 June 2022



9. ITACYL. [https://suelos.itacyl.es/base\\_datos](https://suelos.itacyl.es/base_datos). Last accessed 09 Jan 2022
10. ESA. <https://scihub.copernicus.eu/dhus/#/home>. Last accessed 22 Feb 2022
11. MITECO. <https://www.miteco.gob.es/es/agua/temas/evaluacion-de-los-recursos-hidricos/evaluacion-recursos-hidricos-regimen-natural/>. Last accessed 20 Jan 2022
12. CNIG. <https://centrodedescargas.cnig.es/CentroDescargas/index.jsp>. Last accessed 22 Jan 2022
13. IDECYL. [https://opendata.jcyl.es/ficheros/cart/a2t01\\_elevaciones/](https://opendata.jcyl.es/ficheros/cart/a2t01_elevaciones/). Last accessed 20 Jan 2022
14. IDECYL. [https://idecyl.jcyl.es/geonetwork/srv/spa/catalog.search#/metadata/SPAGOBCYL\\_CITDTSGELIT](https://idecyl.jcyl.es/geonetwork/srv/spa/catalog.search#/metadata/SPAGOBCYL_CITDTSGELIT). Last accessed 07 Feb 2022
15. Triantafyllis, J., Lesch, S.M., La Lau, K., Buchanan, S.M.: Field level digital soil mapping of cation exchange capacity using electromagnetic induction and a hierarchical spatial regression model. *Soil Res.* **47**(7), 651–663 (2009)
16. Omuto, C.T., Vargas, R.R., Elmobarak, A.A., Mapeshoane, B.E., Koetlisi, K.A., Ahmadzai, H., Abdalla Mohamed, N.: Digital soil assessment in support of a soil information system for monitoring salinization and sodification in agricultural areas. *Land Degrad. Dev.* **33**(8), 1204–1218 (2022)
17. Mosleh, Z., Salehi, M.H., Jafari, A., Borujeni, I.E., Mehnatkesh, A.: The effectiveness of digital soil mapping to predict soil properties over low-relief areas. *Environ. Monit. Assess.* **188**(3), 1–13 (2016)
18. Taghizadeh-Mehrjardi, R., Hamzehpour, N., Hassanzadeh, M., Heung, B., Goydaragh, M.G., Schmidt, K., Scholten, T.: Enhancing the accuracy of machine learning models using the super learner technique in digital soil mapping. *Geoderma* **399**, 115108 (2021)
19. Nabiollahi, K., Taghizadeh-Mehrjardi, R., Shahabi, A., Heung, B., Amirian-Chakan, A., Davari, M., Scholten, T.: Assessing agricultural salt-affected land using digital soil mapping and hybridized random forests. *Geoderma* **385**, 114858 (2021)