Index swelling prediction of clayey soils

Amal Medjnoun [1], Zakaria Matougui [1], Mohamed Khiatine [2] and Ramdane Bahar [1]*

[1] Houari Boumediene University, Faculty of Civil Engineering, Algiers, Algeria
[2] Yahia Fares University of Medea, Faculty of Technology, Department of Civil Engineering, Medea, Algeria

Abstract. In civil engineering, statistical studies are widely used in risk studies of landslides, seismic, swelling-shrinking and other hazard studies. This study deals with a volumetric deformation of the clayey soil by water absorption, which is the phenomenon of swelling. It is a serious problem for lightweight structures such as highways. This phenomenon is characterized by two mechanical parameters measured by an oedometric test carried out in the laboratory, which takes a long time and is expensive, why scientists have always looked for other quick and less expensive alternatives to estimate these parameters. They developed models between classical parameters, easily determined in the laboratory, such as water content, dry density, percent of clay particles, plasticity index and mechanical properties of clay soils. The developed models are based on descriptive statistics, a neural network, or multiple regressions. To estimate the risk of swelling and to verify the applicability of the models of the literature on the Algerian soils, a database was collected from the geotechnical reports carried out on the sites of certain projects in Algeria. The objective of the present study consists of estimating the mechanical parameter of swelling, which is the swelling index from a set of physical tests carried out in the laboratory.

1 Introduction

Algeria experiences many disorders due to swelling-shrinking, during dry seasons as a result of soil moisture content evaporation. Shrinkage cracks appear and cause the degradation of light infrastructures [1] such as roads and boundary walls as shown in Figure 1.
Statistical studies make an important technical and economic contribution to geotechnical risk, helping to prevent risks and provide solutions before they occur. Predictive studies are based on the availability of data and the quality of the variables selected, and only those with satisfactory regression coefficients are retained. They have shown relationships between the physical parameters of the soil such as water content, dry density, liquidity limit and plasticity index, clay fraction, shrinkage limit, etc., with the parameters that measure the risk of swelling such as pressure, swelling rate, and swelling index [2], [3], [4] and [5].
Classifications and models are then established such as the classifications using the sensitive water parameters as the shrinking limit [6], the methylene blue value [7], the plasticity index [5], [8], and the swelling index [9]. Statistical software is developed with the evolution of computer tools, and new models have appeared with several parameters as noted by [4], [10], [11], [12], [13], etc.

* Corresponding author: medjnounamel@yahoo.fr
amedjnoun@usthb.dz

The research carried out has not yet found a mathematical prediction model applicable to all the swelling sites, despite the small number of parameters used in these models.
Previous studies have shown that the swelling pressure and rate are highly dependent on the initial state of the soil [1, 4, 14, 15 and 16]. Consequently, the models obtained are for local use and cannot be generalized to other sites. It is therefore interesting to establish a model based on the swelling index. This parameter is measured during the unloading phase of the oedometer test and is independent of soil dryness, wet density, and water content. This parameter is often used to predict the swelling potential of soils, but as it also takes time to determine at the beginning of the soil investigation, especially in the case of highly swelling soils, it is useful to predict it to save time, help guide the geotechnical testing, and speed up the remediation work. The classification of the risk of swelling based on the swelling index has already been established by [3].
Several studies have been carried out to predict the swelling index as [16 and 17]. Swelling index predicted by [17] using several machine learning models has found that the result of the Random Forest model was more effective than the empirical formulae, and the liquid limit has the most important effect on the prediction of the swelling index. The plasticity index has a moderate influence and ranked second. In addition, water content and dry density have little effect on swelling index.
This work is a contribution to the prediction of the swelling index, in the region of Medea and Tissemsilt located in the center north of Algeria. All the collected data in the database such as dry density, water content,

liquidity limit, and clay fraction, are subjected to descriptive statistical analysis. The problem of multicollinearity is considered, and only the factors presenting a good correlation coefficient are retained for the estimation of the swelling index. In the first part, two predictive models are run using Statistica software. Then, machine learning modelling is performed and the absolute mean error is calculated and discussed. Finally, all the prediction results are compared.

Fig. 1. Damage to boundary walls, footpaths, pavements, and hydraulic installations as a result of the shrink-swell cycle in the Medea region.

2 Materials and methods

The database of geotechnical parameters is collected from the construction project carried out on the slopes of Medea and from the Tissemsilt region. The two studied sites are located in the center of the high plateaus of Algeria. They are from the same geological formation which is the marine lower Miocene. The investigation tests are performed in the geotechnical laboratory. They describe the distribution of the soil grains size and give the percentage of clay fraction which is lower than 2 µm (C2), the soil water content value (W), the Atterberg limits which are liquid limit (Wl) and plasticity index (Ip), the dry and wet density respectively indexed (γd) and (γh). These tests are called physical tests and they don't take much time to obtain the result. This study is doing only on cohesive soil which has an amount of clay more than 35%

The mechanical tests are carried out in the laboratory, on a one-consolidation "oedometer" using French Standard AFNOR. This test is carried out on undisturbed soils, it consists of applying loads on the soil sample into its maximum strain, and then the loads are progressively decreased. At the unload stage, the swelling soil samples undergo free swelling by progressive water adsorption. The slope of this growth is measured and gives the swelling index.

First, data analysis is carried out using Statistica V7.1 software. Two methods of analysis are used: descriptive statistical analysis and multi-regression analysis. The latter is a generalization of simple regression. It is usually the determination of the relationship between a dependent variable and several independent variables through the fitting of a linear equation. This method makes it possible to assess the impact of several variables.

Secondly, data analysis is performed using open-source Python. Two machine learning algorithms are used to model the swelling potential, namely Random Forest (RF) and Gradient Boosting (GB) regressors. The choice of tree-based algorithms like Random Forest (RF) and Gradient Boosting (GB) is due to their flexibility in capturing both linear and nonlinear relationships between geotechnical parameters and swelling, their robustness to multicollinearity and their ability to provide insight into feature importance. Being non-parametric models, they make no assumptions about data distribution or variable relationships, enabling them to capture complex patterns without imposing constraints. Additionally, their ensemble learning nature helps mitigate overfitting and improve generalization performance, while their scalability makes them suitable for handling large geotechnical datasets efficiently.

The two algorithms are optimized through a trial-and-error process (GridSearchCV). The choice of these two models is due to their performance, reduced pre-processing, and the number of data samples (about 200 in the case of this study).

3 Result and discussion

3.1 Geotechnical analysis

The soil of Tissemsilt has the same geotechnical identification as the soil of Medea as stated by [16]. The analysis of the geotechnical parameters shows that the soils studied are moderately compact and moist in most of the two sites.

The Casagrande abacus, modified by [3], shows that the soils below line A are silty and those above are clayey. The horizontal and vertical bands indicate the degree of plasticity of clay and silt, as shown in Figure 2. The clays are plastic and moderately plastic in some parts of the sites.

The swelling index values ranged from 1.6 to 10%. The risk classification established by [9] shown in Table 1, indicates a high risk in a large part of the two sites and a certain risk of pathology in another part of the two studied sites.

Since the liquid limit and the soil swelling potential, both depend on the amount of water attempted to be absorbed [5], the relationship between the soil swelling potential and the plasticity index can be established as shown in Table 2. From this Figure and Table 3, the IP values of the studied soils are between 16 and 55% and mostly more than 20%, indicating a medium swelling

 https://doi.org/10.1051/matecconf/202439401002

potential in most of the two sites and a high swelling potential in the other part of the sites of Medea and Tissemsilt.

Table 1. Risk of pathology using the swelling coefficient S_{index}, established by [9]

Swelling index	Pathology risk
≥ 0,09	Certain
0,05 – 0,09	Very high
0,025 – 0,05	high
< 0,025	Less probable

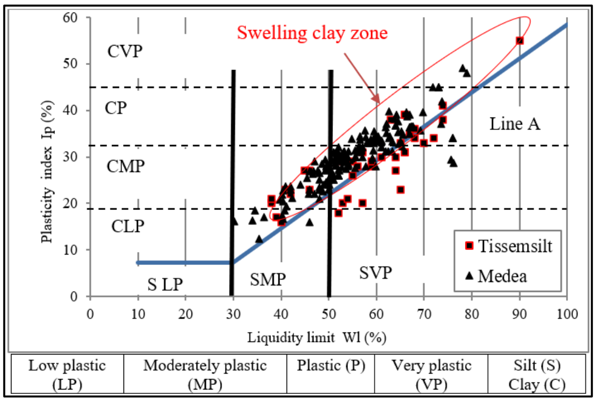

Fig. 2. Plasticity result on the modified Casagrande abacus by [3].

Since the liquid limit and the soil swelling potential, both depend on the amount of water attempted to be absorbed [5], the relationship between the soil swelling potential and the plasticity index can be established as shown in Table 2. From this Figure and Table 3, the IP values of the studied soils are between 16 and 55% and mostly more than 20%, indicating a medium swelling potential in most of the two sites and a high swelling potential in the other part of the sites of Medea and Tissemsilt.

Table 2. Swelling potential classification using plasticity index [5]

Swelling potential	Plasticity index
Low	0 - 15
Medium	10 - 35
High	20 - 55
Very high	35 and Above

William and Donaldson's swelling risk classification abacus, shown in Figure 3, based on the effect of the two geotechnical parameters: plasticity index and clay fraction, indicates a medium risk of swelling in most of the two studied sites. However, there is a high risk in some other parts of this sites.

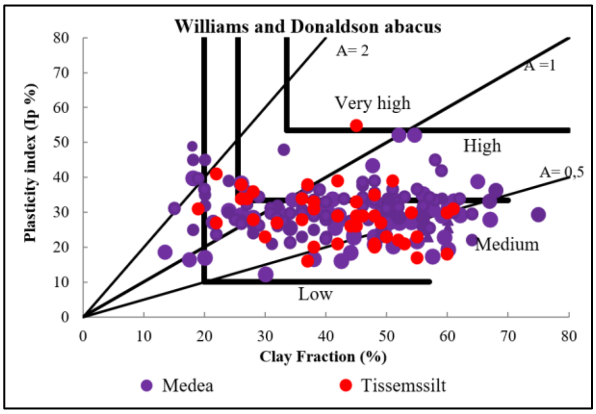

Fig. 3. Swelling potential prediction abacus established by [8]

3.2 Descriptive analysis

The risk of soil swelling shrinkage is a problem related to several soil parameters as the compactness or water content, plasticity and clay content.
To understand the relationships between variables, it is recommended to use all the geotechnical parameters in the predictive study. Statistical analysis of the studied site's data is presented in Table 3. This study requires an important number of samples, in this case, it is 221.

Table 3. Summary statistics of selected variables.

	count	mean	std	min	max
W	221	21.0	6.5	4.6	41.1
γd	221	17.1	1.2	14.5	19.8
Wl	221	55.9	9.6	38	90
IP	221	29.6	6.3	16	55
C2	221	51.1	6.9	40	68
CA	221	0.8	0.6	0.3	4.0
Sindex	221	0.065	2.7	0.019	0.131

The difference between the arrhythmic mean and the median is negligible for all factors according to the values of the standard deviation. From The first and third quartiles, and the diagram of normality, the variable distribution of the index (Sindex), plasticity index (Ip), clay fraction (C2), are normal as presented in the perplot of the investigated parameters Figure 4. The variable distribution of dry density and liquidity limit (Wl) are no-normal.

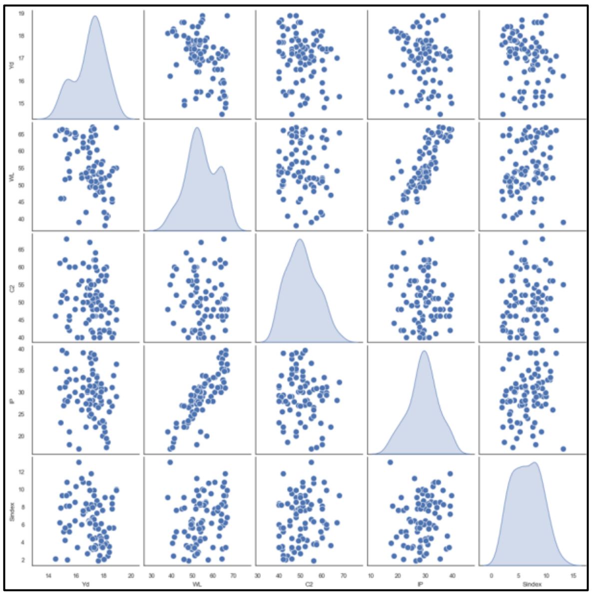

Fig. 4. Perplot of the geotechnical parameters.

Studies have shown that it depends on the compactness and water content of the soil and also on its plasticity and clay content. The study of this problem recommended the use of all the geotechnical parameters in the predictive study.

The correlation matrix is performed as presented in Figure 5. It presents the study of possible links between the geotechnical parameters of the studied soils.

The results suggest that there is a weak relationship between the swelling index and the liquid limit, plasticity index, and other geotechnical parameters. This means that an increase in the physical parameters tends to increase the swelling index proportionally, except for the dry density (γd) and water content which does not seem to have a negligible relationship with the swelling index.

The correlation matrix shows acceptable collinearity between the variables with high collinearity between IP and WL (R>0.84) due to the computational dependency between them, and shows a good correlation between water content and dry density (R=0.88), as explained by [4].

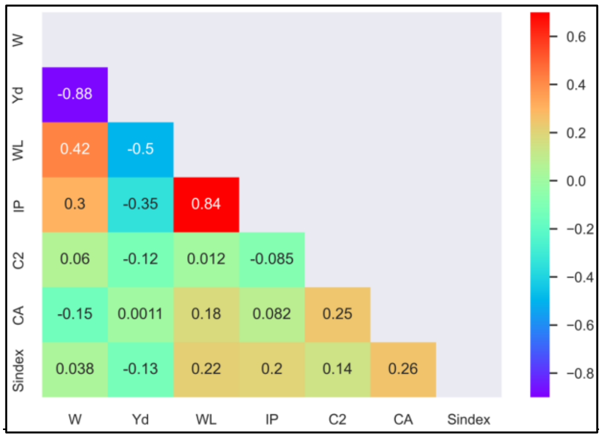

Fig 5. Correlation table of the used geotechnical factors.

3.3 Multiple regression analysis

From the result of the pairplot, the descriptive analysis of the 3D response surface analysis carried out with the Statistica software gives the models represented in Figures 6 and 7. Figure 6, shows the increase in swelling index proportionally to the increase of clay fraction and liquidity limit. This model shows that for the high values of clay fraction and liquidity limit, the swelling index is high, this later is proportional to clay fraction and liquidity limit. This model can be written as shown in equation (1):

$$\text{Sindex} = 0.436 + 0.0525\text{WL} + 0.0608\ \text{C2} \qquad (1)$$

The second model presented in Figure 7, shows that for the high values of clay fraction (C2) and plasticity index (IP) the swelling index is high, this later is also proportional to clay fraction and plasticity index. This model can be written as shown in equation (2):

$$\text{Sindex} = 0.6913 + 0.0607\ \text{IP} + 0.092\ \text{C2} \qquad (2)$$

The results of both Figures 6 and 7 are similar and indicate that the liquidity limit and the plasticity index have the same effect on the variation of the swelling index, This observation is similar to that investigated by [17].

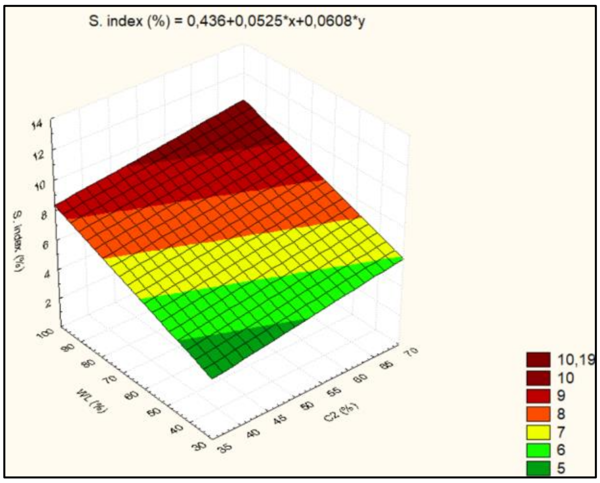

Fig. 6. Effect of clay fraction and plasticity index on swelling index

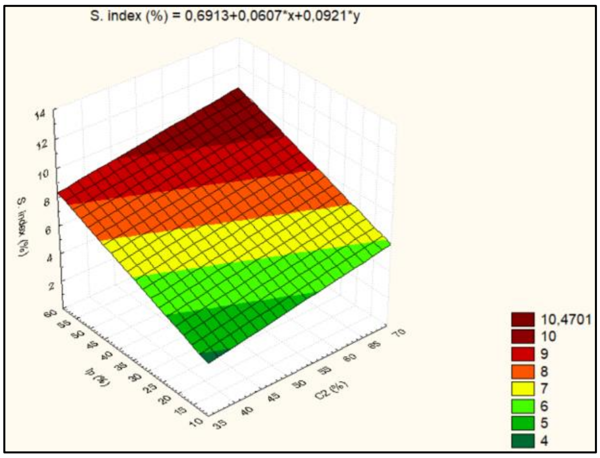

Fig. 7. Effect of clay fraction and plasticity index on swelling index

The calculation of the Mean Absolute Error is a metric commonly used to evaluate the performance of a predictive model. It measures the average absolute difference between the actual and predicted values.
For the first model cited in equation 1, the value of mean absolute error (MAE) is about 2.22, for the second model cited in equation 2, the value of mean absolute error is about 2.27. The two values are not much divergent and they are satisfactory for the prediction of soil swelling index.

3.4 Machine learning modelling

The same geotechnical variables are used in this study, the liquid limit is substituted by the plasticity index, because it is more used in geotechnical and gives the same influence on the swelling index based on the previous multiple regression study.
From Figure 8, the GB model shows that the plasticity index and clay activity are more significant variables than dry density and clay fraction. The standard deviation (std) of plasticity index (IP) and clay activity (CA) are relatively low for this model which indicates that the influence of these variables is stable than others. The std of clay fraction (C2) and dry density (γd) indicates that these variables are more sensitive (high std) and less relevant for Swelling index prediction.
The result of the Random Forest (RF) model is presented in Figure 9, shows that the clay activity and plasticity index (IP) are more significant variables than dry density and clay fraction. The standard deviation (std) of plasticity index (IP) and clay fraction (C2) are relatively low for this model which indicates that the influence of these variables is stable than others. This model is the opposite of the Gradient Boosting (GB); it shows that CA is more important than IP, and the γd is more sensitive than other parameters.

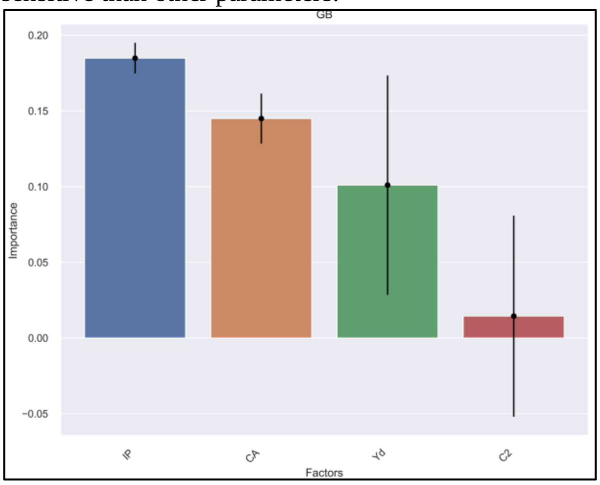

Fig 8. Gradient boosting models feature importance.

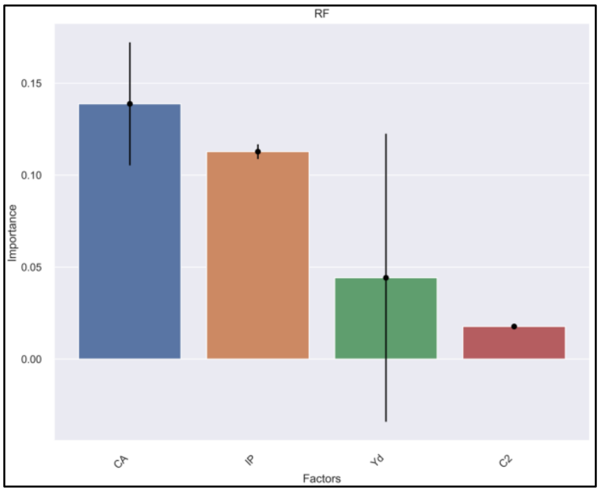

Fig 9. Random forest model feature importance.

The calculations of the values of mean absolute error (MAE) are presented Figures 10 and 11, which show the regression plot between the predicted values of the swelling index established by the two used models vs actual values. From these results, the models have estimated the swelling index with certain imperfections. The values of MAE are in the manner of regression coefficient and they are less than 1.5 % for the two models. The model of Gradient Boosting is more performed than Random Forest (GB R^2=0.74> RF R^2=0.53), that means that the GB is more suitable for the estimating swelling index.
Comparing the results of different methods used to predict the swelling index, traditional methods provide approximate assessments that give a general indication of the level of swelling risk. Meanwhile, approaches based on multiple regression analysis provide models that are also approximate but linked to various geotechnical factors. These models typically estimate the swelling index with a margin of error of less than 2.3.
The introduction of modern machine learning techniques represents a significant improvement, offering enhanced predictive capabilities and reducing the mean absolute error (MAE < 1.5) associated with the predictions. In particular, the Gradient Boosting model shows superior performance to the Random Forest model in estimating the swelling index. Despite the inherent limitations of regression-based approaches, both models provide estimates with satisfactory accuracy, with the Gradient Boosting model proving to be particularly good at estimating the swelling index.

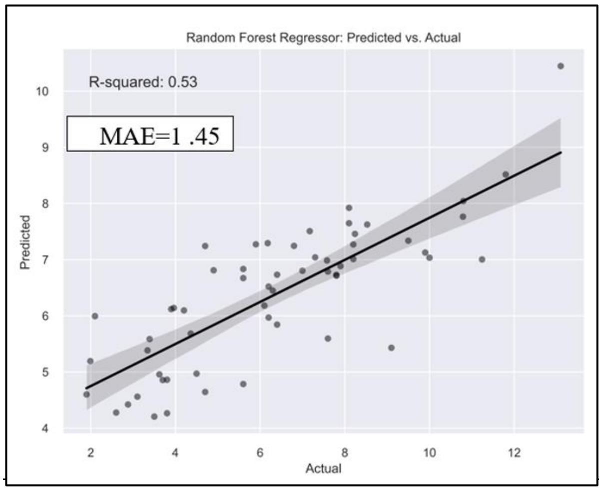

Fig 10. Estimated values of swelling index performed by Random Forest model versus actual values.

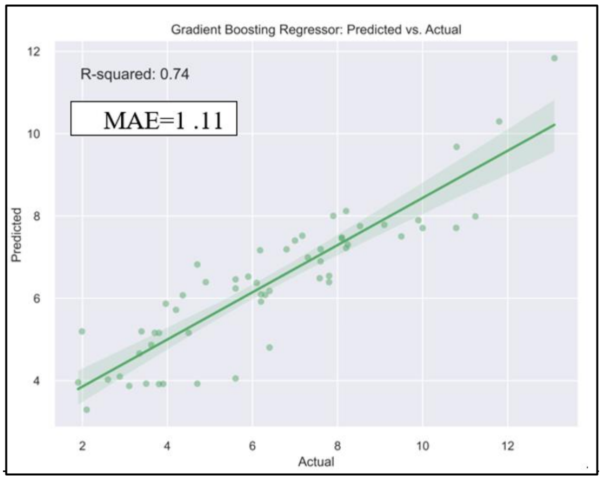

Fig 11. Estimated values of swelling index performed by Gradient boosting model versus actual values.

4 Conclusion

The swelling risk prediction is complex and is linked to several physical, chemical, mineralogical and structural parameters. Combining all these parameters results in mechanical behaviors that depend on the hydric state, the state of consolidation and the mineralogy.
The work was carried out on the same geological formation, in order to have the same mineralogical composition on soil with a clay content of more than 35%. The swelling classification methods, statistical models, and machine learning modeling showed a good relationship given by two geotechnical parameters, the plasticity index and the clay fraction. The plasticity is the amount of water adsorbed by soil, the adsorption is related to the mineralogy of clays that swell. The clay fraction is the amount of clay composed of swelling and no-swelling clay, this parameter was limited to more than 35% to have a significant amount of clay and showed the effect of clay plasticity.
All of the methods presented in this paper can be used to predict the risk of swelling. However, it is important to note that:

•William and Donaldson's swelling risk classification abacus gives just a degree of risk but uses a small number of data.

• The statistical model also works on small number of data and gives prediction values of the swelling index with an acceptable error less than 2.3.

• The Gradient Boosting model is more performed than Random Forest for estimating of the swelling index with a reduced error, but it needs a large data for its establishment.
The study of swelling prediction is an invaluable tool for decision making, guiding geotechnical investigations and managing the problem before it occurs.

References

1. A. Medjnoun, and R. Bahar, Shrinking-swelling of clay under the effect of hydric cycles. Innov. Infrastruct. Solut. **1**, (2016) https://doi.org/10.1007/s41062-016-0043-6.

2. H.B. Seed, R.J. Woodward, R. Lundgren, Prediction of swelling potential for compacted clays. J. of soil Mech. & Found., ASCE, **88**, 53, (1962)

3. D.E. Jones, W.G. Holtz, Expansive Soils – the Hidden Disaster. Civil. Eng. **43**, 49 (1973)

4. A. Medjnoun, and R. Bahar, Analytical and statistical study of the parameters of expansive soil. Inter J. Geol, Environ, Eng. **10**.10.2. (2016) https://doi.org/10.5281/zenodo.1123570

5. F. H. Chen, Foundations on expansive soils. **12**. 52 (Elsevier, New York, 2012)

6. W.T. Altmeyer, Discussion of engineering properties of expansive soils, J. of the Soil Mech and Found Div, American Soci Test Mat, **81**, 17. (1955)

7. A. T. Farid & M. Mosaid, Swelling potential prediction of expansive soils using blue methylene value. In Soil Behavior and Geomechanics 25 (2014) https://doi.org/10.1061/9780784413388.003

8. A.B. Williams. & G.W. Donaldson, Developments related to building on expansive soils in South Africa in Proceeding of 4th Int. Conf. On Expansive soils (1980), Denver 2, 834

9. D. Chassagneux, L. Stieljes, P. Mouroux, F. Menilliet, G.H. Ducreux, Cartographie de l'aléa retrait-gonflement des sols (sécheresse-pluie) à l'échelle départementale. Approche méthodologique dans les Alpes de Haute-Provence," Rapport BRGM R39218. 33. (1996)

10. B. Chittoori & A. J. Puppala, Quantitative estimation of clay mineralogy in fine-grained soils. J. Geot and Geoenv eng, **137**. 997. (2011). https://doi.org/10.1061/(ASCE)GT.1943-5606.0000521

11. I. Yilmaz, O. Kaynar, Multiple regression, ANN (RBF, MLP) and ANFIS models for prediction of swell potential of clayey soils. Expert Systems with

Applications, **38**, 5, 2011, https://doi.org/10.1016/j.eswa.2010.11.027

12. Y. Yukselen, A. Kaya, Suitability of the methylene blue test for surface area, cation exchange capacity and swell potential determination of clayey soils. Eng Geol, **102**. 1–2. (2008) https://doi.org/10.1016/j.enggeo.2008.07.002.

13. Y. Berrah, A. Boumezbeur, N. Kherici, et al., Application of dimensional analysis and regression tools to estimate swell pressure of expansive soil in Tebessa (Algeria). Bull. Eng. Geol. Environ., **77**. (2018) https://doi.org/10.1007/s10064-016-0973-4

14. N.S. ISık. Estimation of Swell Index of Fine Grained Soils Using Regression Equations and Artificial Neural Networks. Sci. Res. Essays, **4**, 1047. (2009)

15. A. Medjnoun, R. Bahar, & M. Khiatine, Caractérisation et estimation du gonflement des argiles algériennes, cas des argiles de Médéa. MATEC Web of Conferences. EDP Sciences. **11**, 03004. (2014). https://doi.org/10.1051/matecconf/20141103004

16. A. Medjnoun, M. Khiatine, & R. Bahar, Mineralogical and geotechnical characterization of swelling marly clays of the Medea region of Algeria. Bull. Eng. Geol. Environ **73**, 1259. (2014). https://doi.org/10.1007/s10064-014-0582-z

17. M. A Benbouras & A. I. Petrisor, Prediction of swelling index using advanced machine learning techniques for cohesive soils. App Sci, **11.** 536. (2021). https://doi.org/10.3390/app11020536