Application of Machine Learning for Prediction of Cone Penetration Test Data

E N Yabbarova\textsuperscript{1}, A A Zaikin\textsuperscript{1}, A I Latypov\textsuperscript{1}

\textsuperscript{1}Institute of Geology and Petroleum technologies, Kazan (Volga Region) Federal University, Kazan, Russia, Kremlevskaya str. 4/5, 420008

E-mail: yabbarova.en@mail.ru, kaskrin@gmail.com, airatlat@mail.ru

Abstract. The article discusses the use of machine learning to obtain predictive data of cone penetration tests of soils using the example of studying the geological environment of the Kazan city. The cone penetration test is one of the most popular and widely used types of field tests of soils. As a source material, a database of the physical and mechanical properties of soils and the database on cone penetration tests of soils of various genesis were used. The territory of the city of Kazan was considered separately for each terrace of the Volga river in accordance with the basic principles of engineering-geological zoning. For modeling of the dependence of parameters of cone penetration tests on soil properties was performed using machine learning. Mathematically, the problem was formulated as regressive - need to find functions that closely approximate field test data. As a result, it was found that the dependences obtained as functions of the cone resistance $q_c$ have very good convergence. This shows a good prospect of using machine learning to obtain real correlation dependencies, and therefore, to reduce the cost of engineering-geological surveys.

1. Introduction

The use of machine learning methods is relevant all over the world: in the oil and gas industry, medicine, agriculture, business, science, etc. The use of machine learning methods for solving practical problems of engineering and geological forecasting in the Russian literature is not numerous [1-11]. As for field tests, namely cone penetration test, the use of machine learning in the Russian literature is not found, despite the wide opportunities and simplification of engineering and geological surveys. Foreign researchers have extensive experience in using machine learning based on the results of cone penetration test.

For example, L. Junhai study of Jiangsu highway a General regression neural network for soil classification, since the existing tables for soil classification based on cone penetration test only provide an assessment of the behavior of the soil without specifying the expected physical and mechanical properties of soils [12].

P. Samui and other use the Minimax Probability Machine (MPM) artificial intelligence to reduce the risk of an earthquake on the example of the island of Taiwan and predict seismic soil liquefaction based on cone penetration test data [13]. The results of the study showed that the developed MPM model has a high level of predicting the potential of soil liquefaction based on drag and peak ground acceleration.
A. Mahgoub and other show the feasibility and effectiveness of using General Regression Neural Network to predict the internal friction angle $\phi$ and the modulus of soil deformation $E$ based on cone penetration test results using experimental data from the United Arab Emirates [14].

A. Mahgoub and other develop many empirical correlations between the SPT N-value, and other properties of soils. The principle objective of the current study is to demonstrate the feasibility and efficiency of using artificial neural networks (ANNs) to predict the soil angle of internal friction ($\Phi$), the soil modulus of elasticity ($E$) and tip resistance ($q_c$) of cone penetration test (CPT) results from SPT results considering the uncertainties and non-linearities of the soil [15].

Y. Erzin and other examine the results of field tests, including cone penetration test with pore pressure measurement in silty sands on the northern coast of the Izmir Gulf in Turkey. Based on the conducted research, a high convergence of forecast data obtained using artificial neural networks for determining the drag in silty sands was demonstrated [16].

B. Tarawneh developed a model of artificial neural network back propagation (ANN) for predicting the value of $N_{60}$ from CPT data. The data used in this study consisted of 109 SPT-SPT pairs for sandy, sandy-silty, and silty-sandy soils. The input variables of the ANN model are: CPT tip resistance ($q_c$), effective vertical stress, and CPT bushing friction ($f_s$). A different set of CPT-CPT data was used to test the reliability of the developed ANN model. It has been shown that the ANN model either underestimates the $N_{60}$ value by 7-16%, or overestimates it by 7-20%. It is concluded that reverse propagation neural networks are a good tool for predicting the $N_{60}$ value from CPT data with acceptable accuracy [17].

A. Ghaderi and other study the landslide-prone area of South-West Sweden, based on cone penetration data with pore pressure measurement. They describe the application of a new optimized multichannel generalized direct-link neural network (GFNN) for classifying different types of soil and predicting the distribution map of different types of soil at different depths, including finding complex boundaries between soils [18].

2. Objects and methods of research
Kazan city area is located in the east of the East-European platform on the left bank of Volga river and at the mouth of Kazanka river. The upper part of the geological section is composed of the Neogenic-Quaternary alluvium incisions of Volga and Kazanka. The soils composing the units of Quaternary alluvial terraces serve as the bases for the majority of engineering structures. They are represented by interbedding of clays, loams and silts with varying consistency as well as of sands with different grain-size composition and saturation degree. Clay soils, as a rule, predominate in the upper part of soil units.

The upper part of geological section in the historical part of the city and in low parts of the area is composed usually of reclaimed sands (aggravated and sand fills), more rare – of clay soils and heterogeneous reclaimed soils.

In this work uses archive data of surveys on cone penetration test and physical and mechanical properties of soils of various genesis in 2000-2020 on the territory of Kazan. For example, the collected database was used to model the dependence of cone penetration indicators – cone resistance $q_c$, MPa and total side friction resistance $Q_s$, kN on the parameters of properties of soils. The average values of properties of soils is given in Table 1. An example of the geological structure is shown in figure 1.
Table 1. Input variables of physical and mechanical properties of soils.

| Physical properties | solid particles density, g/cm³ | soil density, g/cm³ | dry soil density, g/cm³ | water content | plasticity index | liquidity index | void ratio | degree of saturation |
|---------------------|-------------------------------|--------------------|------------------------|---------------|-----------------|-----------------|-----------|-----------------------|
| Min                 | 2.07                          | 0.96               | 1                      | 0.1           | 0.06            | -0.1           | 0.49      | 0.29                  |
| Max                 | 2.74                          | 2.13               | 1.82                   | 0.56          | 0.72            | 1.5             | 2         | 0.93                  |
| Mean                | 2.65                          | 1.88               | 1.46                   | 0.37          | 0.24            | 0.24            | 0.95      | 1                     |

| Mechanical properties | modulus of deformation, MPa | undrained shear strength, kPa | angle of friction, degree |
|-----------------------|-------------------------------|-------------------------------|---------------------------|
| Min                   | 1.1                           | 2                             | 8                         |
| Max                   | 41                            | 78                            | 36                        |
| Mean                  | 14.4                          | 17.8                          | 22.7                      |

Figure 1. Location of the city Kazan and the points of cone penetration test (Zharkova, 2006 geomorphological map).
3. Mathematical model
Denote vector of all explanatory values \((W, I, P, \rho_S, \rho_d, e, S, E, c, \varphi)\) and IGE as \(X=(X_1, X_2, \ldots, X_m)\). We will use 1\(^{st}\) and 2\(^{nd}\) terraces to build the model (329 observations) and 3\(^{rd}\) and 5\(^{th}\) terraces to test it (251 observations).

The mathematical problem is formulated as a regression one: we need to find functions \(f_q(X)\) and \(f_Q(X)\) such that they approximate values \(q_c\) and \(Q_S\) as close as possible. As a fit score we will use mean squared errors:

\[
E \left( q - f_q(X) \right)^2, \quad E \left( Q - f_Q(X) \right)^2
\]

which need to be minimized (\(E\) is an expectation sign). In this setting two target functions \(f_q(X)\) and \(f_Q(X)\) can be modeled and fitted separately. From here and further a wildcard notation \(r\) can be replaced by both \(q_c\) and \(Q_S\).

As a model of choice for each of the targets we use additive model of the form

\[
f_r(X) = \beta_{r,0} + \sum_{m=1}^{M} f_{r,m}(x_m),
\]

where \(f_{r,m}\) are some smooth functions of one variable. Categorical variable IGE is perceived as an additive random effect.

Each of \(f_r, m\) is modeled as a penalized thin plate spline. These splines are parametric, so each of these splines is governed by a vector of parameters \(\beta_{r,m}\). The choice is dictated by a wide flexibility of such parametric representations.

The fitting is done by minimizing

\[
\sum_{n=1}^{N} \left( r_n - f_r(X_n) \right)^2 + \lambda \int \left( f_r'''(X) \right)^2 dX
\]

with respect to spline parameters \(\beta_{r,m}\) for some \(\lambda > 0\). The choice of lambda is dictated by minimizing the GCV (Generalized Cross-Validation) score \([19, 20]\).

Such models usually show a lot of flexibility in prediction problems. But we are also interested in significance of explanatory variables. This can be achieved by incorporating additional L2 penalty term to optimization target which can bring \(f_r, m\) and random effects to zero. Such penalization doesn’t significantly affect the prediction power, but improves interpretation.

As a measure of fit we utilize coefficient of determination \(R^2\) on test sets which is defined by

\[
R^2 = 1 - \frac{\sum (r_n - f_r(X_n))^2}{\sum (r_n - \bar{r})^2},
\]

where the summation is done over all test set indices. Values close to 1 indicate a good prediction power.

This model is a result of going through multiple regression models and fitting procedures including: linear regression, various neural networks structures, different penalization methods etc. We should note that this resulting model doesn’t differ from a simple linear regression. It can be seen from table 2, which is described below.

4. Estimation results
Test errors are given in table 1. As we can see, test sets determination coefficients are similar to the \(R^2\) in the training set, which is a good result. On the other hand, regression for \(Q_S\) seems to be not adequate, and it shouldn’t be used in practice.
Table 2. Coefficients of determination on training and test sets.

| Set        | $R^2_D$ | $R^2_Q$ |
|------------|---------|---------|
| Training set | 0.86    | 0.63    |
| Terrace 3  | 0.62    | 0.12    |
| Terrace 5  | 0.78    | -1.23   |

As for the functions $f_{r,m}$, summary is given in table 2. We present estimated degrees of freedom (EDF) for each smooth term, which is roughly a number of free parameters per variable. EDF equal to zero indicates that model doesn’t depend on this variate; EDF between 0 to 1 represents linear dependence. As we can see, most of dependencies are not far from linear regression.

Table 3. Smooth terms for regressions.

| Variable | Estimated degrees freedom, $q_c$ | Estimated degrees freedom, $Q_S$ |
|----------|----------------------------------|----------------------------------|
| W        | 0                                | 0                                |
| $I_P$    | 1.892                            | 1                                |
| $I_L$    | 0                                | 1.331                            |
| $e$      | 0                                | 0                                |
| $E$      | 2.962                            | 1.987                            |
| $\varphi$ | 5.756                           | 3.35                            |
| $c$      | 4.21                             | 0.063                            |
| $\rho$   | 0                                | 0.339                            |
| $\rho_S$ | 0                                | 0                                |
| $\rho_d$ | 3.291                            | 0                                |
| $S_r$    | 5.867                            | 0.251                            |

The random effect corresponding to lithotype seems to be not very significant for $q_c$: test on hypothesis of no effect has p-value around 0.9, and $R^2$ coefficient decreases only by 0.02. On the other hand, model for $Q_S$ significantly depends on lithotype (p-value around zero, and dramatical change of $R^2$ on the training set).

5. Conclusion

Thus, the analysis of the results shows that the use of machine learning allows us to obtain realistic parameters of cone penetration test based on the physical and mechanical properties of soils of various genesis, determined using laboratory studies. In addition, the use of machine learning for the analysis of the geological environment allows you to perform it quickly and with less human resources and economic costs, which makes it a convenient operational tool for conducting engineering and geological surveys.

6. References

[1] Lisenkov A B and Popov E V 2003 Metodika resheniya ekologo-gidrogeologicheskikh zadach pri pomoshchi apparata teorii informacii Uchebnoe posobie. MGGRU 23 p (in Russian)
[2] Strokova L A 2009 Ispol'zovaniye algoritmov Data Mining dlya resheniya prognoznykh zadach pri stroitel'stve metropolitena Izvestiya TPU vol 315 pp 82–86 (in Russian)
[3] Strokova L A 2009 Prognoz osedaniya zemnoj poverhnosti pri stroitel'stve gorodskikh tunnelej Vestnik MGSU pp 238–241 (in Russian)
[4] Kuzin A A 2013 The isolation of landslide-prone territory using the neural network method pp
The reported study was funded by RFBR, project number 19-35-90089.

Acknowledgments
The reported study was funded by RFBR, project number 19-35-90089.