Development of New Correlation Fines Content, NSPT and CPT Using Neural Network Approach and Multilinear Regression to Support Liquefaction Hazard Analysis

Akhmad Muktaf Haifani1, Hadi Suntoko2, Adi Gunawan Muhammad3, Siti Alimah4

Abstract

Identification and characterization of constituent soil types in the form of Fines Content (FC) values are essential in analyzing the potential of soils to be liquefied. Multiple Linear Regression is one of the fundamental statistical models used to determine the causality between target and predictor geotechnical parameters. The study used multilinear regression approaches and artificial neural networks to get optimal results from FC predictions. The study considers the correlation between the SBT Index and FC and several other parameters such as NSPT, Depth, Totally Overburden Stress, Initially Overburden Stress, and Sleeve Friction. The coefficient of determination resulting from the regression process shows a reasonably strong relationship between the independent and target parameters, as much as 61.4%. In comparison, the Neural Network is 96.928%, which indicates a nonlinear influence.

Keywords: Linear Regression; Neural Network; Fines Content; NSPT

1. INTRODUCTION

Liquefaction occurs in an area characterized as water-saturated sandy soil that experiences earthquakes. Liquefaction events describe the process of transforming a granular material from solid to liquid due to increased pore pressure and reduced effective stress (Division, 1978). Liquefaction is more likely to occur in water-saturated, unconsolidated soils with low porosity, such as sandy loam, sand, and fine gravel. During an earthquake, the unconsolidated sand layer tends to experience volume shrinkage (Ni & Fan, 2002) states that several factors could affect the liquefaction potential, such as relative density (or void ratio), adequate confining pressure, soil characteristics, and fines content. (Uyeno, 1976) Mentioned that muddy sand is resistant to liquefaction due to increased FC by 20%. (Cetin et al., 2004) used FC, resistance corrected SPT 60 % ((N1)_60), cyclic stress ratio (CSReq), moment magnitude (Mw), and initially effective overburden stress (\[\sigma^v\]) to calculate Probabilistic Liquefaction Hazard Analysis.

Geotechnical investigation in the form of a Sieve Test is a critical procedure for characterizing the types of constituent soil in FC values based on the soil grain of the soil sample (ASTM, n.d.). Soil type is determined established on the percentage of grain size.

---

1 Research Centre for Nuclear Reactor Technology- National Research and Innovation Agency (BRIN), Bld 80, KST BJ Habibie, Serpong, South Tangerang, Indonesia, https://orcid.org/0000-0002-7652-6303, akhm020@brin.go.id
2 Research Centre for Nuclear Reactor Technology- National Research and Innovation Agency (BRIN), Bld 80, KST BJ Habibie, Serpong, South Tangerang, Indonesia. hadi004@brin.go.id
3 Research Centre for Nuclear Fuel Cycle and Radioactive Waste Technology, National Research and Innovation Agency (BRIN), Bld 20, KST BJ Habibie, Serpong, South Tangerang, Indonesia, adig001@brin.go.id
4 Research Centre for Nuclear Reactor Technology- National Research and Innovation Agency (BRIN), Bld 80, KST BJ Habibie, Serpong, South Tangerang, Indonesia, siti012@brin.go.id
(FC) retained by sieve No. 200. (Boulanger & Idriss, 2014) Considered FC as a parameter that could trigger liquefaction created by the correlation between FC and the Soil Behavior Type Index (SBT-Ic). This index explains a parameter determined from the end resistance of the Cone Penetration Test (CPT) (Qc) and sleeve friction (fs).

Researchers currently use a linear regression approach by considering various variables to FC. Researchers also often rely on Bayesian classification (BC), Artificial Neural Networks (ANN), and adaptive neuro-fuzzy inference systems (ANFIS) to predict an independent parameter (Šapina, 2016). These methods are an alternative because it takes into account the non-linearity of the soil (Abdipour et al., 2018); (Farjam et al., 2014); (Mansouri et al., 2016); (Sabzalian et al., 2014). Several researchers have formulated the nonlinear approaches have a higher level of accuracy than other methods. This relationship can predict better than the linear-based method (Khairunniza-Bejo et al., 2014) (Safa et al., 2016). The ANN method can demonstrate this phenomenon, which provides less error rate and is more proficient than Multiple Linear Regression (MLR) or other liner-based models to find the best variable (Odabas et al., 2014).

The liquefaction probability analysis incorporates the error factor into the uncertainty in the model parameter (ε). The analysis considers the term "load" and the uncertainty of the factor parameters of the NSPT correction factor (N1.60), Cyclic Stress Ratio (CSR), Grain Size (FC), Moment Magnitude (Mw), and Effective Stress (σ′v) (Cetin, 2000). The analysis can then be estimated by summing the probabilities of all potential sequences of parameters that determine the liquefaction, which involves combination in the liquefaction domain. The grain size parameter taken from the CPT data is important to determine accomplishment in calculating the liquefaction probability (Tabrizi-Zarrinhabaei et al., 2019). This study uses the MLR and ANN approaches to obtain the good quality assessment of the FC value from several available geotechnical parameters. This study aims to produce a new equation based on a statistical gradation approach (Haifani, 2021) using the MLR method to predict FC values through a combination of NSPT and CPT data. The results of this formulation were then validated using the ANN method.

2. METHODS
In the following, the development of FC model is presented using both MLR and ANN methods. The FC model was developed based on the NSPT-CPT dataset of geotechnical investigation.

2.1. Geotechnical Investigation of the NSPT-CPT Method
The procedure to identify geotechnical procedures applied in in-situ site investigations are The Cone Penetration Test (CPT) and Standard Penetration Test (NSPT) which describe geotechnical foundation designs and establish soil properties. CPT is trustworthy for carrying out field investigations and carrying out geotechnical designs. CPT examination expresses soil stratigraphy and records rapid and constant parameters such as cone end resistance (QC) and sleeve friction (fs) (Guen, 2014). The advantages of CPT are the process of repeatability, incessant measurement, and ease of usage (Robertson et al., 1983). CPT analyses explain the soil characteristics of bearing capacity, pore water pressure, and sleeve friction ((Peixoto et al., 2000); (Guen, 2014) ; (Lingwanda et al., 2015)). NSPT is used to obtain the desired soil profile per depth, soil properties, identify soil stratigraphy, and its lateral circulation and find untenable soil profiles due to liquefaction ((Guen, 2014); (Tsukamoto et al., 2004)). NSPT is broadly employed in determining soil parameters based on empirical correlation (Davidson et al., 1999). NSPT signifies the influence of historical stresses and strains, soil structure, effective horizontal stresses, and combined relative and vertical stresses (Seed, 1979). NSPT provides an overview of in-situ soil characteristics under earthquake stress conditions, is exciting to
conceal under laboratory examinations and envisages natural soil liquefaction for upcoming earthquakes (Tokimatsu & Yoshimi, 1983). The combination of CPT and NSPT can properly characterize soil grain size in the form of Fines Content (FC) in the site area.

Geotechnical investigation used dataset covers 44 boreholes in locations where there is evidence of liquefaction (Figure 1-2). The depth of boreholes was relatively uniform, ranging from 4 to 20 m depth, with the groundwater depth between 0.3 to 7 m, hence relatively shallow (TRS-PVMBG, 2012). NSPT analysis provides information on soil classification, index parameters such as water content, density, and some geotechnical applications through semi-empirical procedures, such as evaluating shallow foundation settlement and the pile's bearing capacity and evaluating the potential for sand liquefaction. The SPT examination used a split tube dropped from a height of 75 cm, using a 63.5 kg hammer, to press down the pipe to a total depth of 45 cm. NSPT data estimates every 15 cm step, with no more than 50 hammer blows for every 15 cm step (Akin et al., 2011).

Subsurface information, such as geological profiles, NSPT, Vs., and CPT (Figure 2), were obtained from the borehole data and seismic reflection investigations. It is necessary to develop a geotechnical subsurface model using bore logs representing typical soil profiles in the area. Based on the soils' nature, they have been classified into general groups to identify the respective layers. The study uses a frequency distribution approach to tell how each soil type frequencies. As with all methods, SPT investigations have uncertainties that will affect the analysis results. Experts agree that there is a tolerance for NSPT ranging from at best 1.4% and at worst 100% (Kulhawy & Mayne, 1990), (Kulhawy & Trautmann, 1996), (Schmertmann, 1975), (Youd & Idriss, 2001).

![Fig. 1. Boreholes locations of NSPT, CPT, and seismic measurements in the vicinity of the study area (Muktaf & TAMPUBOLON, 2022)](image-url)
Development of New Correlation Fines Content, NSPT and CPT Using Neural Network Approach and Multilinear Regression to Support Liquefaction Hazard Analysis

2.2. Multi Linear Regression (MLR)

Multiple Linear Regression (MLR) study is an analysis that has more than one predictor variable. Multiple Linear Regression analysis techniques determine the significant effect of two or more independent variables (X1, X2, X3, Xk) on the target variable (Y). In MLR analysis, the least square method is used, including estimating the regression coefficient, the Sum Square Error (SSE), and Sum Square Total (SST). The regression reduces the quantity of the squared upright spaces from each data position on the line to get the best fit. The regression coefficient describes each predictor variable’s unrelated contribution in predicting the target variable in multi linear regression analyses. In contrast to simple linear regression, a description of each independent variable’s degree of interaction or correlation is necessary.

The least-squares method, commonly known as Least Square, estimates model parameters in multi-ple regression analyses. This method’s concept is that the regression coefficient \( \beta \) has a minimum error (Weisberg, 2005).

2.3. Artificial Neural Network (ANN)

Artificial Neural Network (ANN) is a processing data system that imitates the human neural net-work system. ANN consists of connected contribution and yield. Each connection has a unique weight to get the best prediction estimation results. The ANN arrangement includes Input Layer as the initial layer that connects data sources with the processing network. The obscure layer becomes the connecting layer for input variables to produce yield, the Yield Layer, by describing ANN pro-cessing’s yield. This yield is generated based on the weight, several concealed layers, and the activa-tion function. The Initiation function is used to get the yield and input value. The equation used in the Neural Network is given in Equation 1 as follows:

\[
y = \emptyset \left( \sum_{i=1}^{n} W_i X_i + b \right)
\]

Where are: \( y \) – Output, \( \emptyset \) – initiation function, \( \sum_{i=1}^{n} W_i X_i \) – Input accumulation with weight, \( b \) – refraction.

The backpropagation algorithm repeatedly performs learning on a multilayer feed-forward Neural Network to study the weighting for predictions on the label class in a data set or attribute representing an entity (Han et al., 2012). A multilayer feed-forward neural Network comprises one contribution layer, one or more concealed layers, and one yield
layer.

Feed-Forward Backpropagation occurs because the input unit goes to the obscure layer and later goes to the yield layer, producing yield. However, if the yield does not match expectations, the contribution will spread backward on the obscure layer and be forwarded to the yield layer. In the backpropagation process for each training, the analysis will modify the weight to reduce the mean square error between the estimated response quantity \( \hat{y}_i \) and the observed response quantity \( y_i \).

The following steps describe an artificial neural network's modelling utilizing the backpropagation algorithm: The first step is to initiate weighting and bias on the Network. The weights and biases are random numbers between -1 to one or -0.5 to 0.5. The second stage is to spread the input forward. Every \( j \)th output is the input value for each obscure layer or output layer. It can be written as \( O_j = I_j \). Then the net input is calculated from unit \( j \) to the previous layer \( i \) can be formulated given in Equation 2 as follows:

\[
I_j = \sum_i^n w_{ij} O_i + \theta_j \tag{2}
\]

Where are:

- \( w_{ij} \) – the connecting weight of the \( i \)th unit in the previous \( j \)th unit layer,
- \( O_i \) – the \( i \)th unit output of the previous layer,
- \( \theta_j \) – the \( j \)th unit's bias.

Then the output is calculated, namely \( O_j \) using the logistic, sigmoid, or function. This function is a squashing function to map input with a large domain into a smaller range, from 0 to 1 using the Equation 3 as follows:

\[
O_i = \frac{1}{1 + e^{-I_j}} \tag{3}
\]

The third stage is the Back propagate error process, namely the weighting and bias update process, to get the best error in the network prediction results. This process is performed using the Equation 4-7 as follows:

\[
Err_j = O_j (1 - O_j) \sum_i^n Err_i w_{ij} \tag{4}
\]

\[
\Delta w_{ij} = (l) Err_j O_j \tag{5}
\]

\[
w_{ij} = w_{ij} + \Delta w_{ij} \tag{6}
\]

\[
\Delta \theta = (l) Err_j \tag{7}
\]

\[
\theta_j = \theta_j + \Delta \theta \tag{8}
\]

Where are:

- \( Err_j \) – an error in the next layer,
- \( \Delta w_{ij} \) – weight addition,
- \( w_{ij} \) – weight update results,
- \( l \) – the percentage of learning between 0 and 1,
- \( \Delta \theta \) – added bias,
- \( \theta_j \) – updated bias.
3. RESULTS

Based on the previous study of (Yi, 2014), (Suzuki, 1998) and (Boulanger & Idriss, 2014), we developed FC linear model using three independent variables shown in Equation 9-12 (Haifani, 2021) as follows:

\[ X_1 = (\text{NSPT} + Q_c) \]  
\[ X_2 = \left( \frac{f_s}{Q_c} \right) \]  
\[ X_3 = \text{(Depth)} \]  
\[ X_4 = \left( \frac{\sigma_0}{\sigma'_0} \right) \]

Where:

- \( Q_c \) – measured tip resistance,
- \( f_s \) – sleeve friction,
- \( \sigma_0 \) – total overburden stress,
- \( \sigma'_0 \) – effective overburden stress.

3.1 Multiple Linear Regression Analysis

For the hypothesis test we used the coincident assessment (F-assessment) and partial assessment (t-test). The F-assessment is applied to simultaneously assessment the predictor variable's effect on the target variable (Ghozali, 2016). This assessment’s hypothesis is \( H_0 \): No predictor variables significantly affect the target variable, and \( H_1 \): At least one independent variable significantly affects the target variable. Table 1 shows the results of the F-assessment using the R Studio software. Based on the results in Table 1, the model produces a decision to reject \( H_0 \) which indicated by the p-value of the F-assessment of 1.42e-114, which is still lower than the significance probability 0.05. It means there was one independent variable that significantly affects the target variable.

Table 1. The result of F-assessment

<table>
<thead>
<tr>
<th>Model</th>
<th>p-value F Assessment</th>
<th>Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1</td>
<td>1.42e-114</td>
<td>Reject ( H_0 )</td>
</tr>
</tbody>
</table>

Afterward, we performed the t-assessment to determine the result of every unconstraint variable on the target variable. Table 2 shows the outcome of the t-assessment. From the results of Table 2, with a substantial level of 5%, it can be concluded that the variables substantially influence the target variable are \( X_1 \) and \( X_2 \). Based on the t-assessment result, this study formulated empirical formulation for FC as shown in Equation 13.

Table 2. The result of t-assessment

<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Estimate</th>
<th>Standard Error (SE)</th>
<th>tStat</th>
<th>P-Value</th>
<th>Decision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1.10520</td>
<td>0.025766</td>
<td>42.896</td>
<td>2.9282e-179</td>
<td>Reject ( H_0 )</td>
</tr>
<tr>
<td>( X_1(\text{NSPT} + Q_c) )</td>
<td>-0.24105</td>
<td>0.009786</td>
<td>-24.631</td>
<td>2.336e-91</td>
<td>Reject ( H_0 )</td>
</tr>
<tr>
<td>( X_2(\frac{f_s}{Q_c}) )</td>
<td>-0.04046</td>
<td>0.005262</td>
<td>-7.6898</td>
<td>6.6475e-14</td>
<td>Reject ( H_0 )</td>
</tr>
<tr>
<td>( X_3(\text{Depth}) )</td>
<td>0.026618</td>
<td>0.007836</td>
<td>3.3970</td>
<td>0.00072956</td>
<td>Reject ( H_0 )</td>
</tr>
<tr>
<td>( X_4(\frac{\sigma_0}{\sigma'_0}) )</td>
<td>-0.03171</td>
<td>0.010663</td>
<td>-3.7386</td>
<td>0.00020403</td>
<td>Reject ( H_0 )</td>
</tr>
</tbody>
</table>
\[ FC = 1.1052 - 0.241 \times \log(\text{NSPT} + Qc) - 0.040 \times \log \left( \frac{f_s}{Qc} \right) + 0.027 \times \log(\text{Depth}) - 0.031 \left( \frac{\sigma_0}{\sigma'_{0}} \right) \]  

Equation 13 describes that FC will be worth 1.1052, assuming no influence from other variables. Every one-unit increase of the variable \( X_1 \) (\( \log(\text{NSPT} + Qc) \)) will decrease the FC value by 0.241, assuming the other variables are considered constant. Every one-unit increase of the \( X_2 \) variable will decrease the FC value by 0.04, assuming the other variables are considered constant. Every one-unit increase in the \( X_3 \) variable will increase the FC value by 0.027, assuming the other variables are considered constant. Every one-unit increase of the \( X_4 \) variable will decrease the FC value by 0.031, assuming the other variables are considered constant. This equation can be applied in an area with conditions including depth (0.2 - 20 m), \( Vs \) (112.7484-483.5294 m/sec), NSPT (3 - 56), and \( Qc \) (0.26-35.19 MPa).

From the calculation, we also gained the coefficient of determination, \( R^2 \), for the model which was 61.4%. It means that the independent variable can explain the target variable, FC, by 61.4%, and the rest is explained by other variables not included in the model.

3.2. Artificial Neural Network (ANN)

The algorithm employed in this study is the Multi-layer Neural Network, namely 64 obscure layers, one obscure layer processing, and 1000 epoch (iterations) learning process to form a linear equation model. The results of data processing using the Neural Network produced 385 weights factor.

The multi-layer neural network model produces an \( R^2 \) of 0.96928 (96.928%). It means the model explains the \( Y \) variable by 96.928%, and the balance is explained by other variables which are not part of the model. Meanwhile, the value of Adjusted \( R^2 \) Square is 0.96906 (96.906%).

3.3. Validity Assessment based on Performance Work

It is essential to estimate model prediction errors with accuracy because (a) it gives understanding into its precision, (b) permits comparison of numerous simulations, and (c) ones are used to determine threshold warnings (Salazar et al., 2017), (Swanepoel et al., 2016). To promote the evaluation of the ultimate computation results, the analysis implements distinct execution calculation purposes, namely, the mean absolute error \( (MAE) \), mean square error \( (MSE) \), maximum absolute error \( (S) \), and the coefficient of determination \( (RSq) \), using Equation 14-16. The analysis has to calculate approximately the indecision correlated with prototype simulation; the remainder of the prediction set is calculated and analysed (Kang et al., 2017). Table 3 shows the result of the performance evaluation functions of MAE, MSE, S and Rsq between MLR and ANN methods.

\[ MAE = \frac{1}{N} \sum_{i=1}^{N} |Y_{obs} - Y_{pred}| \]  

\[ MSE = \frac{1}{N} \sum_{i=1}^{N} (Y_{obs} - Y_{pred})^2 \]  

\[ S = \max |Y_{obs} - Y_{pred}| \]  

Where are:

\( Y_{obs} \) – Fines Content data from proposed equation 18 (Multiple Linear Regression)

\( Y_{pred} \) – Fines Content data from Artificial Neural Network calculation.
Table 3. Comparison between MAE, MSE, S, and R-Square between MLR and ANN

<table>
<thead>
<tr>
<th>Type Analysis</th>
<th>MAE</th>
<th>MSE</th>
<th>S</th>
<th>R-Square</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple Linear Regression</td>
<td>0.0824</td>
<td>0.0118</td>
<td>0.4558</td>
<td>0.614</td>
</tr>
<tr>
<td>Artificial Neural Network</td>
<td>0.0504</td>
<td>0.0071</td>
<td>0.4149</td>
<td>0.96928</td>
</tr>
</tbody>
</table>

The study uses graphs and box plots to show comparisons between predicted and actual FC values in order to understand the distribution of data and the ability of the selected model to predict FC. As described in Figures 3 and 4, the FC value predicted by the ANN model and by the MLR model has a tendency to resemble and great accuracy with the actual FC value.

Fig. 3. Performance of the (a), MLR for the Fines Content (b), ANN for the Fines Content (FC): fitting of observed FC values and predicted values

Fig. 4. Comparison of boxplot between FC observed, FC from MLR, and FC from ANN

4. DISCUSSION

In general, all ANN models have greater effectiveness in expecting FC than the multilinear regression models. In this study, the calculations show a substantial quantity of the relationship among the input parameter and FC because of a nonlinear relationship, which can be the primary justification for the differences between the ANN and MLR models' performance. Nevertheless, the variance in operation between the two models for predicting FC shows the significance of selecting the appropriate model. The high-level capability of the ANN modelling method associated to the MLR method to predict output variable because this method captures a particularly nonlinear and a complex correlation...
between output and input variables ((Gholipoor et al., 2013); (Khairunniza-Bejo et al., 2014), (Mansourian et al., 2017), (Singh et al., 2003)).

Based on the interpretation of the pack plot (Figure 4), it also shows relatively the same results, namely the maximum, minimum, Q1, median, and Q3 values between the predicted FC and the actual FC using the ANN method when compared with the FC produced by the MLR approach. The MAE, MSE, and S calculations for the FC ANN results show lower values than the FC results from MLR. Also included in this case, the value of the coefficient of determination shows that the R-Square value of ANN (96.928%) is higher than the MLR (61.4%).

5. CONCLUSIONS

The analysis results using two methods, linear regression with the Least Square approach and Artificial Neural Network, appear that the R-Square and Adjusted R-Square are 61.4% 61.1% 96.928%, 96.906%, respectively. Artificial Neural Network shows a better model for estimating FC based on comparing the results of the two calculation methods. Artificial Neural Networks in this study can produce the best results because Neural Networks can learn by themselves, learn from examples, and apply to the same events. Besides, the Neural Network's adaptive and flexible nature explains the soil's nonlinear characteristics from each regression parameter used in the study.

Acknowledgements

The funding for this research was provided by Ministry Research of Research Technology and Higher Education of Indonesia. This support is gratefully acknowledged. We are also thankful to PVMBG for providing the necessary data.

Conflict of Interest

The authors declare no conflict of interest.

References


Guen, O. (2014). Correlation between SPT and CPT. NTNU.


