Classification of soil quality using K-Nearest Neighbors methods

I D Ratih¹, S M Retnaningsih¹, and V M Dewi¹

¹Faculty of Vocational, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia

Email : iisdewiratih@gmail.com

Abstract. The damage or deterioration of the material due to reaction with the environment is called corrosion. The corrosive of soil can cause natural disasters such as landslides. Soil corrosivity is a serious problem in industrial and infrastructure activities. Soil corrosivity is also a problem at PT. IPMOMI, because of the large number of metal pipes that are planted underground which functions to drain fluid in the form of seawater as the main material for steam production. PT. IPMOMI wants to install metal pipes underground, Therefore PT. IPMOMI requires information about the status of soil corrosivity at that location. To overcome soil corrosivity, PT. IPMOMI will install cathodic protection on the soil with high corrosivity. One of the efforts to deal with this problem is by mapping the subsurface corrosion zone. Before mapping the corrosion zone, it is necessary to know the level of soil corrosivity whether it is very high, high, medium, low, or very low. So that to determine the level of soil corrosivity whether very high, high, medium, low, or very low, a classification method is needed. In this study, the classification method used is the nonparamteric classification method, namely K Nearest Neighbors. The results of classification accuracy were 83.3% and variable chargeability had a higher contribution than the depth level in the classification model.

1. Introduction

Soil corrosivity can be interpreted as the ability of the soil to cause corrosion. The level of soil corrosivity can be grouped based on resistivity, soil resistivity is the magnitude of the characteristics of the soil as an electrolyte medium to conduct electric current that causes corrosion. Each soil has a different level of corrosivity. Soil resistivity is thought to be influenced by chargeability and the level of soil depth [1, 2, 3]. Chargeability is a measurement that is often used in the measurement of polarization induction using the method time domain, chargeability is obtained in milliseconds. The value is chargeability more accurate in showing the presence of minerals in the rock that have an effect on soil corrosivity. Whereas the soil depth level is the size of the soil with the depth level measured from the ground surface in meters [4]. The experimental results indicated that the soils could be rated as mildly corrosive to less corrosive groups to the buried galvanized-steel and cast iron pipes in the study areas. A good positive or negative correlation coefficient between resistivity, moisture, chloride and sulfate contents implies that these soil parameters have an equal contribution to the rating of soil corrosivity [5]. Soil resistivity influences the corrosion of metals installed underground and can serve as an indicator of soil corrosiveness. From corrosion engineering perspective, the lower the resistivity, the higher the corrosivity and vice versa. Soil resistivity is most often measured using Wenner four-electrode method and resistance meter [6]. One of the cases of corrosion on metal pipes caused by soil corrosivity is at PT IPMOMI who is the company of steam power plant (PLTU) company located in Paiton, Probolinggo, East Java. PT IPMOMI has an underground pipe that functions to drain fluid in the form of sea water as the main material for steam production. The corrosion cases at PT IPMOMI is
the occurrence of corrosion in underground metal pipes, which will cause damage and deterioration of metal pipes due to reacting with the environment. The environment in question is soil, because metal pipes are buried underground, the soil corrosivity will affect the corrosion of metal pipes. Based on the description above, in this study, the soil quality classification at the PT IPMOMI pipe installation site was carried out based on the level of soil corrosivity using the algorithm in K-Nearest Neighbors in an effort to obtain classification results with a higher level of accuracy than previous research which would later serve as the basis for PT IPMOMI to install cathodic protection in soil parts with high and very high corrosivity. The K-Nearest Neighbor (K-NN) is one of the supervised method in Data Mining with the algorithm is classification method for a set of data based on learning data that has been classified previously. Included in supervised learning, where the results of new query instances are classified based on the majority of the proximity of the existing categories in K-NN. K-Nearest Neighbor (KNN) is one of the most popular algorithms for pattern recognition. Many researchers have found that the KNN algorithm accomplishes very good performance on different data sets [7, 8, 9, 10,11,12]. KNN has several main advantages, including simplicity and robustness to noisy [13]. The data set has been labeled, that is to say the selected neighbors of test sample have been classified correctly and there is no explicit training phase or it is very minimal [14,15] and [16] shows that KNN provides the highest classification accuracy compared to naive Bayes and decision trees.

This research is a continuation of the previous research about Dummy Regression Adjusted Control Chart has conducted, which is to map soil corrosivity zones by classifying soil corrosivity levels with chargebility and soil depth levels as predictor variables and soil resistivity as response variables. The classification of soil corrosivity is determined based on the soil resistivity value. The results show that chargebility and soil depth have a significant effect on soil resistivity and the dummy regression model provides a good model between 78.9%-92.1% in predicting soil resistivity. The soil resistivity prediction results obtained are used to classify soil corrosivity. Further research is needed because the corrosive nature of the soil can cause natural disasters such as landslides. Soil corrosion is also a serious problem in industrial and infrastructure activities, the piping system in industrial and infrastructure activities tends to increase. This piping system is considered to have a high level of integrity and is more effective and efficient than other transportation systems. In this piping system most types of pipe material used are metal types [6], besides its advantages metal also has many disadvantages when compared to other elements due to the nature of the metal which is easily corroded. Metal pipes that are buried under the ground surface have a high risk of corrosion, soil is the main cause of corrosion of these metal pipes.

2. Methods

2.1. Data Source
This research used secondary data that obtained from a survey conducted by [17], students from Department of Geophysical Engineering, Sepuluh Nopember Institute of Technology in 2017 at PT IPMOMI, Probolinggo. Research of data is the result of measurement in units 7 and 8 PT IPMOMI at location 1 with varying lengths of location, then obtained the value of chargebility, depth, and resistivity.

2.2. Variables
Variabel that used on this research included the response variable (Y), namely the level of soil corrosivity, then the predictor variable (X), namely chargebility and depth level based on measurement results in units 7 and 8 at PT IPMOMI, is described in Table 1 below.

| Variables | Information          | Scale    | Unit    |
|-----------|----------------------|----------|---------|
| Y         | Level of Soil Corrosivity | Ordinal  |         |
| X₁        | Chargebility         | Ratio    | Msec    |
| X₂        | Depth Level          | Nominal  | Location|
The explanation for each research variable is explained as follows.
1. Soil corrosivity is the ability of the soil to cause corrosion. The level of soil corrosivity can be grouped based on resistivity, soil resistivity is the magnitude of the characteristics of the soil as an electrolyte medium to conduct electric current which causes corrosion. The level of corrosivity based on the resistivity value is described in Table 2 below.

Table 2. Soil Corrosivity Level Based on Resistivity Value (Center for Metalurgi-LIPI, 1987)

| Soil Resistivity (ohm) | Corrosivity     |
|-----------------------|-----------------|
| <7                    | Very High       |
| 7-20                  | High            |
| 20-50                 | Medium          |
| 50-100                | Low             |
| >100                  | Very Lowmoderate|

2. Chargeability is a measurement that is often used in the polarization induction using the method time domain, chargeability is obtained in milliseconds. The value is chargeability more accurate in showing the presence of minerals in the rock that have an effect on soil corrosivity [4].
3. The level of soil depth is the measure of the soil with the depth level measured from the ground in units of meters. This research was conducted measurements at 1 location in units 7 and 8 PT IPMOMI which has 6 categories of soil depth, namely 0.17; 0.53; 0.92; 1.36; 1.84; 2.36 meters.

2.3. Analysis Step
The steps of analysis in this research are as follows:
1. Categorized the level of soil corrosivity based on the resistivity value.
2. Divided the data into 2 parts, namely data training (75%) and data testing (25%).
3. Forming a soil quality model using K-Nearest Neighbors with data training.
   a. Determines k positive integers based on the availability of learning data.
   b. Choose the closest neighbor from the new data as much as k.
   c. Determine the most general classification in step (b), using the highest frequency.
   d. Output classification from new sample data.
4. Classifying with data testing.
5. Get the Confusion Matrix.
6. Get the Importance Variable.

3. Results and Discussions
3.1 Data Classification
The method of analysis used in this study is the k-nearset neighbor (k-NN). In the k-NN method, the training dataset will be separated from the testing dataset. 75% of the total training dataset is taken from a total of 46 sample data. Whereas for testing the dataset is the rest of the data that will be tested for accuracy as in Table 3. K-NN is done by looking for k objects in the training data that are closest (similar) to the objects in the testing data.

Table 3. Classification Data set

| Count | Training Set | Testing Set |
|-------|--------------|-------------|
| 46    | 34           | 12          |

Soil quality classification results of PT. IPMOMI with k-Nearest Neighbor (k-NN) using values k 1 to 30 can be seen in Table 4.
Table 4. Prediction Accuracy Based on k Value on k Nearest Neighbor (k-NN)

| Parameter | Accuracy | Parameter | Accuracy | Parameter | Accuracy |
|-----------|----------|-----------|----------|-----------|----------|
| k=1       | 0.5      | k=11      | 0.75     | k=21      | 0.667    |
| k=2       | 0.667    | k=12      | 0.75     | k=22      | 0.667    |
| k=3       | 0.667    | k=13      | 0.833    | k=23      | 0.583    |
| k=4       | 0.75     | k=14      | 0.75     | k=24      | 0.667    |
| k=5       | 0.75     | k=15      | 0.75     | k=25      | 0.583    |
| k=6       | 0.75     | k=16      | 0.75     | k=26      | 0.583    |
| k=7       | 0.75     | k=17      | 0.75     | k=27      | 0.583    |
| k=8       | 0.75     | k=18      | 0.75     | k=28      | 0.583    |
| k=9       | 0.75     | k=19      | 0.667    | k=29      | 0.583    |
| k=10      | 0.75     | k=20      | 0.667    | k=30      | 0.583    |

Table 4 shows that the accuracy of the prediction results of k-Nearset Neighbor (k-NN) is the highest with a value of k = 13 because the instance can predict the class higher than the others, the accuracy reaches 0.833 or 83.3% and the calculation shows at the accuracy in Confusion Matrix. These results are also shown visually in Figure 1 below.

![Accuracy Chart](image)

**Figure 1** Prediction Result k-Nearset Neighbor (k-NN)

3.2 Confusion Matrix

Table 5 shows that there are four categories of soil quality in PT. IPMOMI is viewed from the level of soil corrosivity, namely the soil corrosivity level is Very High, High, Medium, Low, And Very Low. In the Confusion Matrix (Figure 1), there are a total of 12 predictions obtained by analyzing the testing dataset to see the suitability of the soil corrosivity level at PT. IPMOMI. From the 12 samples above, the classifier precisely predicts the level of soil corrosivity "Very High" is 7 times, then the prediction level of soil corrosivity is "High" is 2 times and the prediction of the level of soil corrosivity is "Medium" is 1 times, meanwhile the classifier does not predict the level of soil corrosivity "Low" and “Very Low”.

Whereas in the actual data there are 9 locations with a "Very High" level of corrosivity, while the other 4 are locations with a "High" category of corrosivity and the last 1 is a location with a "Medium" level of corrosivity and no locations with a "Low" and “Very Low” category. In the table can be seen that there is a error prediction at 2 locations that should be predict with "High" category but incorrectly predict "Very High" The accuracy calculation is as follows:

\[
\text{Accuracy} = \frac{TP}{\text{Total Data Set}} = \frac{10}{12} = 0.833
\]
The results of k-NN analysis using k = 13, the results show that k-NN will correctly classify 10 samples from a total of 12 sample testing datasets which can be seen from the confusion matrix table as in Table 6 with an accuracy value of 0.833 or 83.3%.

The confusion matrix between the predictions and the actual from the testing dataset to see the suitability of the soil corrosivity level is described at Table 5.

**Table 5. Confusion Matrix of k-NN**

| Prediction | Actual | Very High | High | Medium | Low |
|------------|--------|-----------|------|--------|-----|
| Very High  | 7      | 2         | 0    | 0      | 0   |
| High       | 0      | 2         | 0    | 0      | 0   |
| Medium     | 0      | 0         | 1    | 0      | 0   |
| Low        | 0      | 0         | 0    | 0      | 0   |

3.3 Importance Variable

Importance variable is used to select predictor variables that do not have a significant contribution to the model. Figure 2 shows that chargeability has a higher contribution than depth level to classification formation.

![Figure 2 Importance Variable Diagram](image)

4. Conclusions

In conclusion, the accuracy of the classification of soil corrosivity at one of the locations of PT. IPMOMI using the KNN method was 83.3%. Chargeability, has a higher contribution than the depth level in the classification model.

Acknowledgments

Thank you for Business Statistics Department, Institut Teknologi Sepuluh Nopember who has provided financial support

Reference

[1] Fon A N, Che V B and Suh C E 2012 Application of electrical resistivity and chargeability data on a gis platform in delineating auriferous structures in a deeply weathered lateritic terrain, eastern cameroon *International Journal of Geosciences* 3 960-71

[2] Liu H W, Dai Y and Cheng F 2020 Corrosion of underground pipelines in clay soil with varied soil layer thicknesses and aerations *Arabian Journal of Chemistry* 13 3601-14

[3] Melchers R E 2018 Correlation between soil electrical resistivity, polarisation resistance and
The International Journal of Corrosion Processes and Corrosion Control 53 524-30

[4] Cramer S D. and Jr. Covino B S 2003 ASM Handbook Volume 13A Corrosion: Fundamentals, Testing, and Protection, ASM International

[5] Poudel, A. et.all 2020. A Classification Approach for Corrosion Rating of Soil to Buried Water Pipelines: A Case Study in Budhanilkantha-Maharajganj Roadway Areas of Nepal. World Journal of Applied Chemistry 5 47-56.

[6] Lim K S, Yahaya N, and Ihman S R 2013 The relationship between soil resistivity and corrosion growth in tropical region Journal of Corrosion Science and Engineering 16 1-11

[7] Li C, Zhang S, Zhang H, Pang L, Lam K, Hui C, Zhang S 2012 Using the K-nearest neighbor algorithm for the classification of lymph node metastasis in gastric cancer. Computational and Mathematical Methods in Medicine 12 1-11

[8] Suguna N and Tanushkodi K 2010 An Improved k-Nearest Neighbor Classification Using Genetic Algorithm. International Journal of Computer Science 7 18-21

[9] Sanjaya S, Pura M L, Gusti S K, Yanto F, Syafria F 2019. K-Nearest Neighbor for Classification of Tomato Maturity Level Based on Hue, Saturation, and Value Colors. Indonesian Journal of Artificial Intelligence and Data Mining 2 101-106

[10] Yan X, Li W, Chen W, Luo W, Zhang C, Wu Q and Liu H 2013 Weighted K-Nearest Neighbor Classification Algorithm Based on Genetic Algorithm. TELKOMNIKA Indonesian Journal of Electrical Engineering 11 6173-6178

[11] Safri Y F, Arifudin R and Muslim M A 2018 K-Nearest Neighbor and Naive Bayes Classifier Algorithm in Determining The Classification of Healthy Card Indonesia Giving to The Poor. Scientific Journal of Informatic 5 9-18

[12] Khuman, M B 2013 Classification of remote sensing data using KNN method. Journal Inform. Knowledge Res. Electron. Commun. Eng 2 817-821

[13] Parvin H, Alizadeh H and Minati B 2010 A Modification on K-Nearest Neighbor Classifier. Global Journal of Computer Science and Technolog 10 37-41

[14] Zhao M and Chen J 2016 Improvement and Comparison of Weighted k Nearest Neighbors Classifiers for Model Selection. Journal of Software Engineering 10 109-118.

[15] Imandoust S B and Bolandraftar M 2013 Application of K-nearest neighbor (KNN) approach for predicting economic events theoretical background. International Journal of Engineering Research and Applications 3 605-610

[16] Jadhav S D and Channe H P 2013 Comparative Study of K-NN, Naive Bayes and Decision Tree Classification Techniques. International Journal of Science and Research 5 1842-1845

[17] Dewi, Kiki Kartika (2017). Pemetaan Zona Korosivitas Tanah Berdasarkan Nilai Chargeability Menggunakan Metode Time Domain Induced Polarization Konfigurasi Dipole-Dipole. Case Study : PT. IPMOMI.