Soil Resistivity Prediction in South Africa: Using Decision Trees for Improving The Design Process of Overhead Power Line Tower Earth Electrodes

Ruan Pretorius
School of Computer Science and Applied Mathematics
University of the Witwatersrand
Johannesburg, South Africa

***Abstract*—Soil resistivity measurements are required for the earth electrode design of many structures like telecommunication towers, photovoltaic power plants, and overhead power lines to comply with design standards. For long lines of high voltage overhead power lines, this may require the laborious task of soil resistivity measurement at every tower site as well as a multi-step iterative design and construction process. As part of the initial design phase, time and money can be saved by having a model that predicts the soil resistivity based on freely available geospatial data. In this paper the feasibility of this soil resistivity prediction was investigated. This was done using descision trees trained on geospacial data such as bedrock depth, runoff, relief index, and biome type. The accuracy of the models produced in this study were between 48% and 54%. This relatively low accuracy was due to the models being trained on a small data set. Training similar models on a larger data set might lead to better results and forms part of the recommended future work.**

***Index Terms*—soil, resistivity, prediction, South Africa, machine learning, decision tree, design, power lines, earth electrodes**

I. Introduction

Soil resistivity is an electrical property of soil which describes to what extent it resists electric current flowing through it [1]. The earth electrode design of many structures like telecommunication towers [2], photovoltaic power plants [3], and overhead power lines [4], [5] are dependent on the soil resistivity at the location where they will be constructed.

In the case of high voltage overhead power transmission lines for example, the tower foot resistance must be below a specific value to comply with design standards. In South Africa the towers of a 400 kV power line must have a tower foot resisance of below 40 Ω [4]. This tower foot resistance depends highly on the local soil resistivity at the tower. Due to high soil resistivity at the location of construction of some of these towers, this design value is sometimes not met after the first design iteration [2], [4].

Towers that are found not to comply with the design standard have to be modified or retrofitted after the initial setup and construction to reduce the tower foot resistance to a satisfactory level. This requires a lot of extra time and money to fix. In some cases the towers need to be addressed on an individual basis and a unique re-design or updated design is required for each one [4].

The iterative design process just described includes many stages: These include the initial design and construction of the towers; the measurement of tower foot resistances for each tower after initial construction to see if it is below the design requirement; the re-design and modification of some of the towers that did not meet the requirement; and finally, after modifications are made, follow-up tower foot resistance measurements to see if the modifications were adequate [2], [4].

This process can be improved and the number of steps in this process can be reduced if reliable soil resistivity information was available without having to do physical measurements. The focus of this paper is to evaluate the feasibility of soil resistivity prediction in South Africa for this purpose. If the soil resistivity can be reliably predicted, the initial route layout and design of overhead power transmission lines can be adjusted so that no re-design and modification is required later on which will save time and money. For this study, prediction of soil resistivity was done by using freely available geospaial data.

The machine learning method selected for relating this geospatial data (input features) to soil resistivity (target variable) was the decision tree. Decision trees were used for this purpose mainly because of the human readability of their output. The hypothesis of a decision tree is represented as a series of if-then rules [9]. This is advantageous from an engineering design point of view since design decisions need to be transparent and clearly supported.

At the time of writing, no other methods were found in literature that attempted soil resistivity prediction with machine learning methods. Especially for imporving the design process of overhead power lines. As described above, the current design process is an iteritive process that involves time consuming physical soil resistivity measurements at the location of construction.

The rest of the paper is structured as follows. Section II describes the methodology followed to produce the results. This includes how the data set was constructed and a detailed description of the decision tree methods used. This is followed by Section III which contains the results and discussion

Funded by the DSI-NICIS National e-Science Postgraduate Teaching and Training Platform (NEPTTP)

and finally Section IV which contains the conclusions and suggested future work.

II. Methodology

The methodology followed to produce the results of this study is described here which includes the collection of raw data, data preprocessing, and the construction of a final usable data set. This section also includes a detailed description of the decision trees used and how the final data set was preprocessed for use by these models. Finally, a description of how the results of the decision trees were analysed is also given.

A. Data Collection

Two main sets of data were collected for the construction of the data set used in this study. These two main sets were geospatial data and soil resistivity data. The main aim of this study was to see to what extent freely available geospatial data could be related to soil resistivity data which is time-consuming to obtain. The geospatial data was used for the predictor variables or features of the decision tree models and the soil resistivity data was used for the target variables.

Soil resistivity depends on different parameters including moisture content, chemical composition, temperature, and porosity. [1]. Therefore, the geospatial data chosen for this study included biome type and runoff from Open Africa [6]. This data set was constructed by the Department of Agriculture, Fisheries and Forestry of South African Environmental GIS Data [6]. The reason behind choosing these variables was that the biome type might give an indication of climate and soil moisture content since different plants grow in different soil conditions. The runoff may also give an indication of soil moisture content. In addition to this, a geospatial data set with information on lithology, landform, soil type, relief index, and bedrock depth was also collected. This data set was created as part of the Soil and Terrain (SOTER) database programme by the International Soil Reference and Information Centre (ISRIC) [7]. The reason behind including this data was that it contained information about the parameters that influence soil resistivity as mentioned above. The geospatial data was available in a shapefile format which is a collection of polygons representing different regions with different properties on a map. Fig. 1 shows an example of this type of data representing the biome types in South Africa. The SOTER database contains a relational database with soil information. The soil information in this relational database is linked to polygons in a shapefile through a common ID column. To assemble the required data from this database in a format suited for use later on, an SQL query was created to return the required soil information contained in the database along with the corresponding information from the shapefile in one single dataframe.

The soil resistivity data was collected by physical measurements at different locations across South Africa. This was done using the four-point Wenner method. This method involves inserting four equally spaced electrodes into the earth and passing a known test current through the outer electrodes.

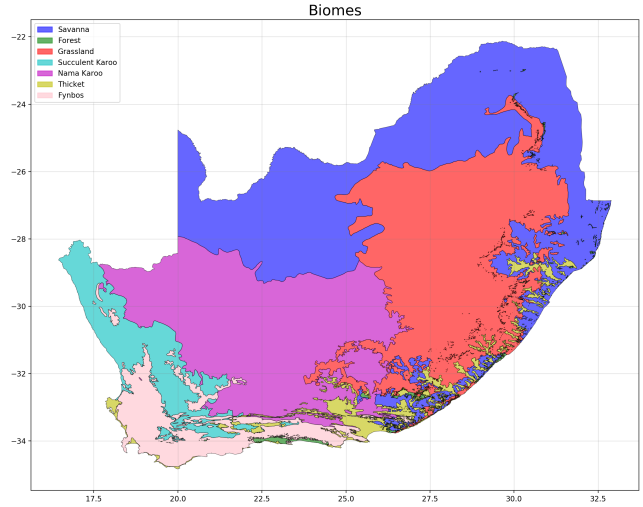


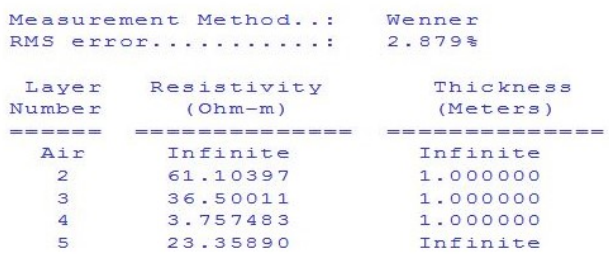
Fig. 1. Plot of geospatial data in shapefile format representing biome types of South Africa [6]

The potential is then measured between the inner electrodes. From here, the apparent soil resistivity can be calculated. The soil resistivity measured is roughly at the depth equal to the electrode spacing [8]. This process was repeated between eight and 15 times for different electrode spacings beween 0.5 meters and 16 meters to gather soil resistivity data at varying depths for each location measured. Fig. 2 shows a picture of the author collecting soil resistivity data using the four-point Wenner method.

Fig. 2. Photograph of author (R. Pretorius) taking soil resistivity measurements using the four-point Wenner method

These soil resistivity measurements were then imported into the RESAP module of SES CDEGS software to construct a three-layer soil model. The three layers were each one meter deep starting at the surface and reaching down to a depth of three meters. The reason for choosing three layers with a thickness of one meter was arbitrary so that a departure point for this study could be established. These depths can be varied

to suit more specific design considerations later on once this method is found to produce sufficiently accurate results. Fig. 3 shows a screenshot of the three-layer soil model output of SES CDEGS software (Note that the three one-meter thick layers are surrounded by two more layers of infinite thickness which are were required for modelling purposes but were not considered further in this study).



```
Measurement Method..:    Wenner
RMS error.........:    2.879%

 Layer    Resistivity        Thickness
Number      (Ohm-m)          (Meters)
======  ==============  ==============
  Air      Infinite          Infinite
   2       61.10397          1.000000
   3       36.50011          1.000000
   4       3.757483          1.000000
   5       23.35890          Infinite
```

Fig. 3. Example screenshot of three-layer soil model produced by SES CDEGS software using soil resistivity measurements (outer layers of infinite thickness were not of interest)

B. Data Set Construction

The locations of the soil resistivity measurements were recorded and therefore known. The GPS coordinates of these locations were then stored on a computer and the geospatial data corresponding to these coordinates could be retrieved. The geospatial data was then paired with the three-layer soil resistivity model values to create the final data set which was ready for preprocessing. The geospatial data was used as predictor variables or features and the soil resistivity data was used as target variables. Fig. 4 shows a diagram that indicates this process of data collection and data set construction. Note that at this point some of the features were numerical and some were categorical. A summary of the data set up to this point (before preprocessing) is shown in Table I. Because of the time consuming nature of collecting soil resistivity data, only 58 data points (58 locations) were used for this study.

TABLE I
DESCRIPTION OF DATA SET VARIABLES BEFORE PREPROCESSING

Variable	Use	Type	Note
Bedrock Depth	Predictor	Continuous	$\in [0.09, 1.04]$
Landform	Predictor	Categorical	5 Categories
Lithology	Predictor	Categorical	10 Categories
Soil Type	Predictor	Categorical	15 Categories
Relief Index	Predictor	Discrete	$\in [8, 89]$
Runoff	Predictor	Continuous	$\in [0.14, 1.79]$
Biome	Predictor	Categorical	7 Categories
ρ_1^{a}	Target	Continuous	$\in [4.14, 4462.9]$
ρ_2	Target	Continuous	$\in [0.60, 16038.4]$
ρ_3	Target	Continuous	$\in [1.50, 54005.6]$

[a]The three soil resistivity layers are labelled ρ_1, ρ_2, and ρ_3

C. Methods

Supervised machine learning was chosen for this study to relate the predictor variables (geospatial data) to the target variables (soil resistivity at one, two, and three meter depths). More specifically, decision trees were chosen for this purpose.

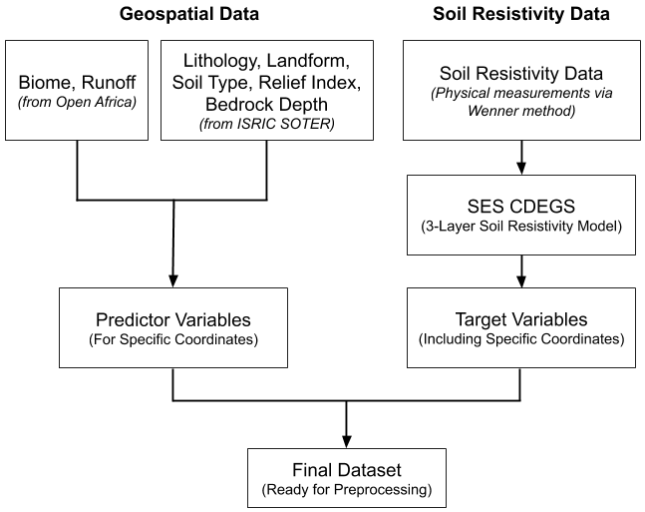


Fig. 4. Diagram indicating the process of data collection and data set construction

Decision trees are used for searching completely expressive hypothesis spaces and approximation of discrete-valued functions. Their output is easily human readable since they represent a hypothesis as a series of if-then rules [9].

The ID3 decision tree algorithm was selected for this study. This algorithm uses the statistical property called information gain to select which attributes are best for partitioning the data set in a way such that most information is gained from the target variables. Because of this, the resulting decision tree has the attributes which contain most classification power of the target variables closer to the root of the tree [9].

The information gain measures the reduction in entropy after partitioning the data by some attribute. Entropy is a measure of the impurity of a data set. Equations (1) and (2) describe how to calculate the information gain G and entropy H respectively [9].

$$G(S, A) = H(S) - \sum_{v \in A} \frac{|S_v|}{|S|} H(S) \tag{1}$$

$$H(S) = \sum_{i=1}^{c} -p_i \log_2 p_i \tag{2}$$

In the above equations, $G(S, A)$ is the information gain, S is a data set, A is an attribute of the data set, $H(S)$ is the entropy of the data set, S_v is a subset of S where A has the value v, p_i is the proportion of S belonging to class i, and c is the number of classes that attribute A can take on.

To avoid overfitting to the data set, pruning was applied to the decision trees after training. After trying different values, the maximum tree depth was set to five nodes. This produced the best overall results after applying 14-fold and shuffle-split cross-validation (see subsection F in this section for more details on accuracy analysis of decision trees used).

Since the soil resistivity of three different layers below ground were predicted, three different decision trees were

trained. One for each of the three layers between 0 meters and 3 meters below ground.

D. Data Preprocessing

Some preprocessing steps were necessary to make the data suitable for machine learning and more specifically for training decision trees.

Firstly, all categorical features had to be changed to numerical format. This was done by using dummy variables which split a categorical feature into the amount of possible categories it could assume. Each of these categories were converted into a new feature and was assigned a value of 1 if it belonged to that category and 0 otherwise. Because of the limited number of data points in the data set (58), the decision was made to reduce the amount of features in an attempt to make the hypothesis space completely expressive and to reduce the possibility of overfitting. Because the categories of the categorical features were transformed into features themselves, choosing only the categorical features with the least amount of unique categories would lead to a lower total amount of features. Therefore, landform and biome were kept and lithology and soil type were dropped from the data set.

To further ensure the hypothesis space was not restricted, the target variables were discretised into quartiles so they could only assume one of four values. To ensure that these bins were not categorical, they were assigned a discrete numerical value of 0 through 3. This was done because the quariles are ordinal categories.

E. Data Summary

After preprocessing the data, there were still 58 data points (58 different locations). The geographical distribution of which can be seen in Fig. 5.

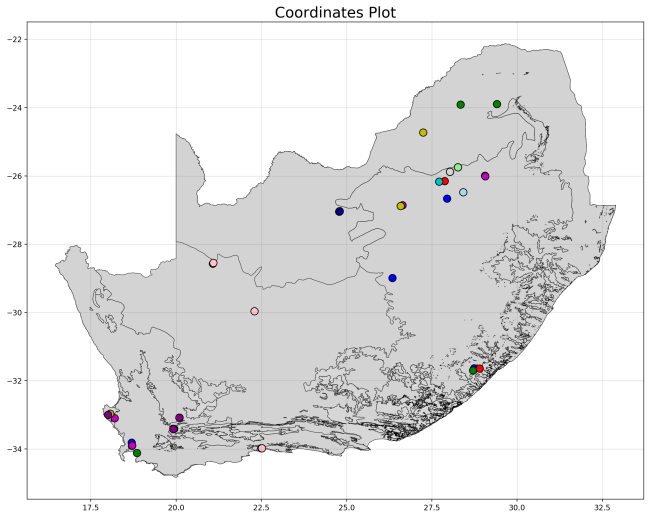


Fig. 5. Goegraphical distribution of the 58 data points used in this study (58 different locations)

The distribution of the three soil resistivity layers can be seen in the histograms of Fig. 6. The bins of these histograms correspond to the values given in Table II. From the soil resistivity distributions shown in the histograms of Fig. 6, it is clear that each bin is very close to equally represented. This is because the resistivity of each layer was discretised by quartiles.

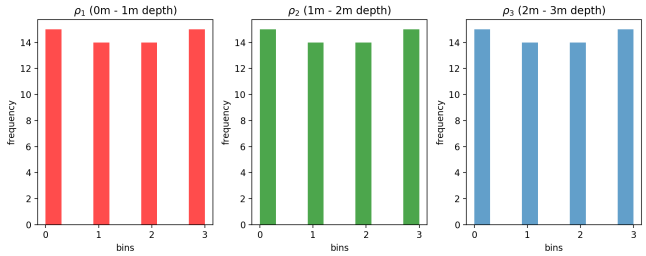


Fig. 6. Distribution of discretised soil resistivity data for all three layers between 1 meter and 3 meters deep

TABLE II
SOIL RESISTIVITY RANGES OF DISCRETISED BINS

Bin #	ρ_1 (0m - 1m)	ρ_2 (1m - 2m)	ρ_3 (2m - 3m)
0	4.14 - 70.95	0.60 - 19.74	1.50, 19.21
1	70.95 - 203.3	19.74 - 217.5	19.21, 138.7
2	203.3 - 908.0	217.5 - 882.3	138.7, 1637.8
3	908.0 - 4462.9	882.3 - 16038.4	1637.8, 54005.6

All soil resistivities are given in units of Ω-m

F. Analysis

To analyse the accuracy of the three decision trees trained in this study, 14-fold cross-validation as well as shuffle-fold cross validation was used. 14-fold cross validations was selected because it was high enough so that the training could be done on a significant portion of the data. Recall that there was only 58 data points in total. This meant that only four data points were held out for testing on each trained tree so that the tree could still be trained on the vast majority of data in order to have the best chance of capturing the characteristics of the data set. For the shuffle-fold cross-validation, a random set of four data points were held out for testing. For consistency, the shuffle-fold cross-validation was also repeated 14 times. Both methods yielded similar results as seen in Section III.

III. RESULTS & DISCUSSION

This section contains an exposition of the results followed by a discussion of these results.

A. Results

The results from both the 14-fold and shuffle-split cross-validation tests are summarised in Table III. The diagrams of all three trained decision trees are shown in Fig. 7, 8, and 9.

B. Discussion

From Table III it is clear that the 14-fold and shuffle-split cross-validation methods produced similar accuracy results. The average accuracy produced using 14-fold cross-validation

TABLE III
SUMMARY OF CROSS-VALIDATION ACCURACY SCORES FOR DECISION TREES PREDICTING ALL THREE SOIL LAYER DEPTHS

Tree #	Soil Layer	14-Fold CV (%)	Shuffle-Split CV (%)
1	0m - 1m (ρ_1)	53.21	39.29
2	1m - 2m (ρ_2)	47.14	53.57
3	2m - 3m (ρ_3)	51.07	51.79

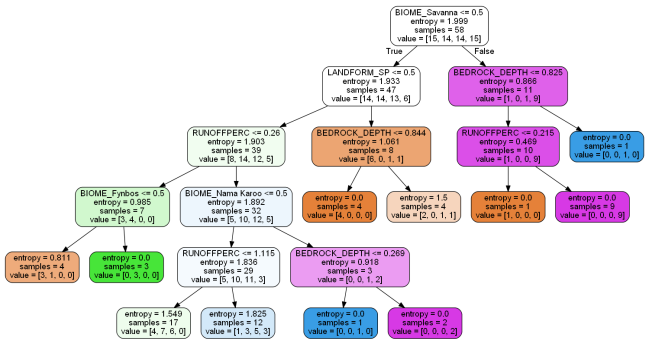


Fig. 7. Diagram representing the decision tree of predicting the soil resistivity of the first layer of soil ρ_1 (between 0 and 1 meters under ground) from geospatial data

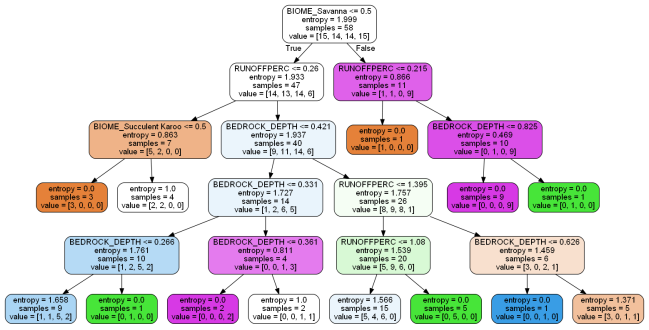


Fig. 8. Diagram representing the decision tree of predicting the soil resistivity of the second layer of soil ρ_2 (between 1 and 2 meters under ground) from geospatial data

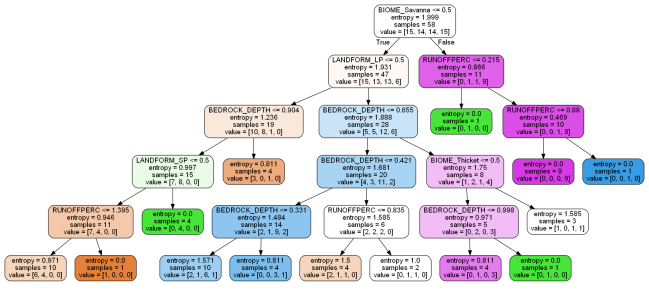


Fig. 9. Diagram representing the decision tree of predicting the soil resistivity of the third layer of soil ρ_3 (between 2 and 3 meters under ground) from geospatial data

for all three trees was 50.47% whereas for shuffle-split cross-validation it was 48.88%.

None of the trees could produce an accuracy above 53.57% using any type of cross-validation tested. This suggests that using this specific data set, the decision trees were not very successful at predicting soil resistivity between 0 and 3 meters below ground.

Looking more closely at the decision tree diagrams of Fig. 7, 8, and 9, it is clear that some of the leaf nodes on the trees do not have an entropy value of zero. This means that by following some of the routes down the tree there was no clear verdict reached as to what resistivity bin the data point should fall under. It is also worth nonting that this did not only happen at the fifth level of the tree where pruning was enforced. In all three trees this happened as early as level three of the tree. This suggests that it was not due to pruning.

Also worth noting from the three tree diagrams is that when branches did terminate in leaf nodes with entropy values of zero, they often had only a single data point in the verdict bin. Out of all three trees, the highest number of data points in a single bin of a leaf node with zero entropy was nine. This represents less than 16% of the entire data set which is much less than the roughly 25% of data that must fall under each bin of a leaf node for a tree that would give 100% accuracy (since the target variables were discretised by quartiles).

These results suggest that the data set used was not large enough. If the data set was larger, a larger proportion of the data set might have been present in the leaf nodes with zero entropy making the decision tree model seem ”more sure” of the final verdict. Further more, if the data set was larger, more features could have been included without restricting the hypothesis space. This could also have given the decision tree models more discrimination power leading to more accurate soil resistivity predictions.

IV. CONCLUSIONS & FUTURE WORK

This section contains the conclusions and suggestions for any future work that might be useful for developing this project.

A. Conclusions

The decision trees produced in this study were able to predict the accuracy of three soil layers between 0 and 3 meters under ground with an average accuracy of between 50.47% and 48.88%. None of these decision tree models were able to predict the soil resistivity with a greater accuracy than 53.57%.

The lack of accuracy was likely due to the limited size of the data set used. This led to the hypothesis space becoming too restricted for the decision trees to make accurate predictions and a need to reduce the amount of predictor variables used to train them.

B. Future Work

To improve on this study, the most important aspect is to increase the data set size used for training and testing

the decision trees. This would allow the addition of more predictor variables to the models which might lead to increased discrimination power and more accurate predictions.

Alternatively, other machine learning methods that might be better suited for use in restricted hypothesis spaces can be tested on this data set.

ACKNOWLEDGMENT

This work was funded by the DSI-NICIS National e-Science Postgraduate Teaching and Training Platform (NEPTTP). Some of the soil resistivity data which the CDEGS soil models were constructed from was kindly shared by: Barry Reid from Royal Haskoning DHV; Gary Thoresson from The Testing Guys; Ivan Grobbelaar from DEHN Africa; Johann Rossouw from EPCM Solutions; Dr. Pieter Pretorius from Terratech; Theunus Marais from Eskom; and Trevor Manas from LP Concepts.

REFERENCES

[1] M. Kižlo, A. Kanbergs, "The Causes of the Parameters Changes of Soil Resistivity," "The causes of the parameters changes of soil resistivity." Electrical, Control and Communication Engineering 25.25, 2009, pp. 43-46.

[2] L. W. Choun, C. Gomes, M. Z. A. Ab Kadir, W. F. Wan Ahmad, "Analysis of earth resistance of electrodes and soil resistivity at different environments," 2012 International Conference on Lightning Protection (ICLP), Vienna, 2012, pp. 1-9, doi: 10.1109/ICLP.2012.6344314.

[3] P. H. Pretorius, "Loss of equipotential during lightning ground potential rise on large earthing systems," 2018 IEEE International Symposium on Electromagnetic Compatibility and 2018 IEEE Asia-Pacific Symposium on Electromagnetic Compatibility (EMC/APEMC), Singapore, 2018, pp. 793-797, doi: 10.1109/ISEMC.2018.8393890.

[4] P. H. Pretorius, B. Ntshuntsha, S. Ramadhin, "Application of counter-poise in the reduction of tower footing resistance - low frequency design and case study," Cigre Symposium Cape Town - 2015, South Africa, Session 9 – Paper 7, 2015.

[5] Jinxi Ma, F. P. Dawalibi and W. Ruan, "Design Considerations of HVDC Grounding Electrodes," 2005 IEEE/PES Transmission & Distribution Conference & Exposition: Asia and Pacific, Dalian, 2005, pp. 1-6, doi: 10.1109/TDC.2005.1546811.

[6] Department of Agriculture, Fisheries and Forestry - South African Environmental GIS Data," Soil and Terrain Database (SOTER) for South Africa," March 17, 2020, Distributed by Open Africa. https://africaopendata.org/dataset/department-of-agriculture-fisheries-and-forestry-south-african-environmental-gis-data.

[7] International Soil Reference and Information Centre (ISRIC) - Soil and Terrain (SOTER) database programme," Environmental GIS Data," November 17, 2015, Distributed by ISRIC Data Hub. https://data.isric.org/geonetwork/srv/eng/catalog.search#/metadata/c3f7cfd5-1f25-4da1-bce9-cdcdd8c1a9a9.

[8] IEEE Guide for Measuring Earth Resistivity, Ground Impedance, and Earth Surface Potentials of a Grounding System," in IEEE Std 81-2012 (Revision of IEEE Std 81-1983), vol., no., pp.1-86, 28 Dec. 2012, doi: 10.1109/IEEESTD.2012.6392181.

[9] T. M. Mitchell, "Machine learning,", Singapore: McGraw-Hill, 1997, pp. 52-80.