Open Journal of Statistics, 2020, 10, 239-251
https://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

Predicting the Underlying Structure for Phylogenetic Trees Using Neural Networks and Logistic Regression

Hassan W. Kayondo[1], Samuel Mwalili[2]

[1]Pan African University, Institute of Basic Sciences, Technology and Innovation, Nairobi, Kenya
[2]Department of Statistics and Actuarial Sciences, Jomo Kenyatta University of Agriculture and Technology, Nairobi, Kenya
Email: whkayondo@gmail.com, samuel.mwalili@gmail.com

How to cite this paper: Kayondo, H.W. and Mwalili, S. (2020) Predicting the Underlying Structure for Phylogenetic Trees Using Neural Networks and Logistic Regression. *Open Journal of Statistics*, **10**, 239-251.
https://doi.org/10.4236/ojs.2020.102017

Received: November 3, 2019
Accepted: March 31, 2020
Published: April 3, 2020

Abstract

Understanding an underlying structure for phylogenetic trees is very important as it informs on the methods that should be employed during phylogenetic inference. The methods used under a structured population differ from those needed when a population is not structured. In this paper, we compared two supervised machine learning techniques, that is artificial neural network (ANN) and logistic regression models for prediction of an underlying structure for phylogenetic trees. We carried out parameter tuning for the models to identify optimal models. We then performed 10-fold cross-validation on the optimal models for both logistic regression and ANN. We also performed a non-supervised technique called clustering to identify the number of clusters that could be identified from simulated phylogenetic trees. The trees were from both structured and non-structured populations. Clustering and prediction using classification techniques were done using tree statistics such as Colless, Sackin and cophenetic indices, among others. Results from 10-fold cross-validation revealed that both logistic regression and ANN models had comparable results, with both models having average accuracy rates of over 0.75. Most of the clustering indices used resulted in 2 or 3 as the optimal number of clusters.

Keywords

Artificial Neural Networks, Logistic Regression, Phylogenetic Tree, Tree Statistics, Classification, Clustering

1. Introduction

A phylogenetic tree is defined by [1] [2] as a tree that represents evolutionary relationships among species under consideration. Normally these trees are either sampled or simulated from populations. The populations are either structured or non-structured. A structured population is made up of types, called sub-populations as pointed out by [3] [4] [5]. A non-structured population does not have sub-populations and the disease dynamics are considered to be uniform or homogeneous. A phylogenetic tree can be studied by analysing the inferred tree shape as pointed out in [6]. The inferred tree shape can be summarised by tree statistics or indices as described by [7]. We employed tree statistics to predict the underlying population structure using classification techniques. These techniques were logistic regression and artificial neural network models.

Logistic regression is a special case of linear regression. Both linear and logistic regression have a dependent variable, say Y which is predicted using independent variables, say $X_1, X_2, \cdots, X_p$, in case where we have p independent variables. For linear regression, Y is a continuous variable, while Y is a categorical variable which takes on two values (dichotomous) for logistic regression, for example, logistic regression can be used to predict presence or absence of a certain symptom in patients, using variables like age, weight, race and others. Many studies have employed logistic regression to study various phenomena. For example, [8] used logistic regression to analyse 46 variable amino acid sites in reverse transcriptase for their effect on susceptibility. Another classification technique which we used was artificial neural network.

An artificial neural network (ANN) model consists of input neurons, hidden layers (with hidden neurons) and output neuron(s) as described in [9] [10] [11] [12]. For a classification problem, input neurons are features that are used during the learning process of the network. These are the input variables for the network as pointed out by [10]. Hidden layers and hidden neurons connect input neurons with output neurons. The output neurons are classification targets, for example, presence or absence of a disease. Layers and neurons in ANN models are connected by weights that are determined during the learning algorithm. ANN models are applicable in many fields, including financial management, manufacturing, pattern recognition, control systems, environmental science, among others as noted by [13]. For example, [9] used ANN models to predict five-year mortality for patients who were diagnosed with breast cancer. [13] applied ANN models to study rainfall-runoff patterns and forecasting floods. [10] applied ANN models for eutrophication prediction, where water quality indicators of a certain lake were predicted with reasonable accuracy. In other ANN applications, [14] used back-propagation neural network on classification of multi-spectral remote sensing data.

In this paper, we used logistic regression and ANN models for classification. The two classes were structured and non-structured populations. The independent variables were the tree statistics. We investigated the predictive ability of the logistic regression and ANN models. This was assessed using the average accuracy rates. We also performed unsupervised learning technique, called clustering. The aim was to obtain optimal number of clusters. Since we had structured and non-structured populations, average optimal number of clusters was expected to be two in order to consider clustering to have detected the underlying structure about the populations.

2. Methods

A linear regression model is given as:

$$\hat{Y} = \beta_0 + \sum_{j=1}^{p} X_j \beta_j. \tag{1}$$

where $\beta_j's$ are linear regression coefficients estimated using least squares method which minimises the residues $\sum_{i=1}^{n}\left(\hat{Y} - Y_i\right)^2$, given n observations, *i.e*, (x_i, y_i) for $i = 1, 2, \cdots, n$. Fitted values are denoted as $\hat{Y}$, while observed values for the dependent variable are the $Y_i's$. A given coefficient value (β_j) indicates the extent to which the mean of dependent variable changes when a unit shift in an independent variable (X_j) occurs, while keeping other variables in the model constant. Coefficient β_0 signifies a value the dependent value assumes when all independent variables in the model are equated to zero, or when they are all missing. Care has to be taken when interpreting β_0 as it might be meaningless in some regression models.

A logistic regression model is a special case of linear regression model for Bernoulli distributed dependent variable. The link function is the logit function which is defined as:

$$\operatorname{logit}\left(Pr\left(Y = 1\right)\right) = \ln\left(\frac{Pr\left(Y = 1\right)}{1 - Pr\left(Y = 1\right)}\right). \tag{2}$$

A logistic regression model is given below as:

$$\operatorname{logit}\left(Pr\left(Y = 1\right)\right) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j. \tag{3}$$

We simulated phylogenetic trees from both structured and non-structured populations. The structuring for the population was based on parameters for birth, death and migration. The choice for the values of the parameters in structured and non-structured populations were based on the work of [3] on estimation of binary character effect on speciation and extinction. In their work, birth, death and character change parameters were equal in the two sub-populations under a symmetrical scenario, while asymmetry was depicted by altering one of the three parameters at a time. In our simulation, the three parameters were changed at once to yield the asymmetry in all the three parameters. In our case, it is the asymmetry that introduced the population structure in the phylogenetic tree simulation. In disease epidemiology, a structured population means that there are two or more sub-populations with varying infectivity rates, hence the choice of the parameters in our simulated trees under structured and non-structured populations.

For parameters used in the simulation sets, the choice was based on a similar study done in [3]. We had three simulation sets, and in the first simulation set, we had: $n_1 = n_2 = 100$, $\lambda_1 = \lambda_2 = 0.5$, $\mu_1 = \mu_2 = 0.01$ and $m_{12} = m_{21} = 0.02$ for a non-structured population. Parameter λ_1 is the rate at which individuals in sub-population 1 give birth to new individuals (birth rate in sub-population 1), μ_1 is the rate at which individuals in sub-population 1 die (death rate) and m_{12} is the migration rate for individuals from sub-population 1 to 2. Parameters λ_2, μ_2 and m_{21} are analogously defined. Parameters n_1 and n_2 represent the number of leaves (tips) in a phylogenetic tree that belong to sub-population 1 and 2, respectively. The sum of n_1 and n_2 gives the total number of leaves in a phylogenetic tree. For a structured population, we had the parameters as: $n_1 = n_2 = 100$, $\lambda_1 = 1.5$, $\mu_1 = 0.03$, $m_{12} = 0.06$, while λ_2, μ_2 and m_{21} had the same values as those for non-structured. We simulated 1000 trees in total and therefore 500 for each of structured and non-structured populations. It should be noted that structured and non-structured population in this study correspond to asymmetry and symmetry models, respectively used by [3].

For the second simulation set, parameters for structured and non-structured remained the same as those in the first simulation set, but with only changes made on the number of leaves of trees. The total number of leaves was changed from 200 to 500. We therefore had $n_1 = n_2 = 250$. For the third simulation set, only the number of phylogenetic trees was doubled and we had 1000 for either structured or non-structured population, while other parameter values were the same as those for simulation set 1.

Using simulated trees obtained under structured and non-structured populations, we used eight tree statistics for classification and clustering. These included: number of cherries, Sackin, Colless and total cophenetic indices, ladder length, maximum depth, maximum width and maximum width over maximum depth. A cherry is defined as two leaves (tips) that are adjacent to a common ancestor node as described in [15]. A Sackin index index adds the number of internal nodes between each leaf and the root in a tree. This index was proposed by Sackin in 1972. For Colless index, the absolute difference between left and right hand leaves subtended at each internal node is computed. This is done over all the internal nodes and the sum gives Colless index. Details for Colless index can be obtained in [7]. The definition of total cophenetic index is given by [16]. Other definitions for ladder length, maximum depth of a tree, maximum width and maximum width over maximum depth can be found in [17]. The implementation of phylogenetic tree simulation and computation of tree statistics were implemented in Python software, version 3.7.3.

We then performed standardization for all the eight variables using a formula given by Equation (4).

$$z = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{4}$$

where z, $\min(x)$ and $\max(x)$ are standardized input, minimum and maximum for a particular variable, respectively. The standardization procedure used is explained in [18]. The standardized variables ranged in the interval $[0,1]$.

2.1. Training Artificial Neural Network and Logistic Models

With help of R package, neuralnet of [11], we first trained artificial neural network (ANN) models using all the standardized eight tree statistics as the input variables. These were: number of cherries, Sackin, Colless and total cophenetic indices; ladder length; maximum depth; maximum width, and width-to-depth ratio. We used generalized weights as described in [12] to identify four most influential input variables for each of the three simulation sets. We first used one hidden layer with one neuron to identify four most influential input variables. This was done to reduce input variables for ANN models. Reduced ANN models with few input variables converged faster.

For logistic regression, models were fitted using *glm* function of an R package called stats. The *glm* function fits generalized linear models. As pointed out by [19], these models comprise of a dependent variable (z), a set of independent variables ($x_1, x_2, \cdots, x_m$), predicted variable ($Y = \sum \beta_i x_i$) and a linking function ($\theta = f(Y)$). The linking function connects parameter θ of the distribution of z with the Y's of the linear model. We first fitted logistic models using all the eight variables. A summary of the output from *glm* function gives values of regression coefficients, z-values and $Pr(>|z|)$. For each of an input variable, z-values and $Pr(>|z|)$ were used to identify four most significant variables for each of the three simulation sets.

2.2. Parameter Tuning for Neural Network and Logistic Models

Using ANN model with one hidden neuron, we identified four most influential input variables using generalized weights for all the three simulation sets. [9] analysed contributions of covariates (input variables) for ANN models using generalized weights. They point out that the distribution of generalized weights for a particular covariate signifies whether the effects are linear (small variance) or non-linear (large variance). We plotted the generalized weights for all the eight inputs for each of the three simulation sets using the same range. Input variables that had a distribution of generalized weights close to zero were deemed to have less contribution in explaining the output variable as pointed out by [11]. Parameter tuning was then performed on reduced ANN models. The parameters that were tuned to obtain optimal models were the number of hidden layers and hidden neurons.

For each of the simulation set, having identified the four most influential input variables, we ran reduced ANN models with two hidden layers. In each of hidden layers, we varied number of hidden neurons between one and two. In the first reduced ANN model, we had one neuron for both hidden layers. For second reduced ANN model, we used two, and one neuron for first, and second hidden layers, respectively. A third reduced ANN mode had one and two neurons for first and second hidden layers, respectively. A fourth reduced ANN model had two neurons for both hidden layers. Models were compared using Akaike information criterion (AIC), Bayesian information criterion (BIC) and cross entropy (ce) error. [20] defined AIC and BIC in Equations (5) and (6), respectively.

$$\text{AIC} = -2\ln\left(L\left(\hat{\theta}\right)\right) + 2p \tag{5}$$

$$\text{BIC} = -2\ln\left(L\left(\hat{\theta}\right)\right) + p\ln\left(2\right) \tag{6}$$

where $L\left(\hat{\theta}\right)$ is the likelihood of estimated model. Parameter p and n are the total number of parameters estimated and sample size, respectively. For both AIC and BIC, smaller values imply better models. Cross entropy measures deviations of predicted outcome from the observed ones. The smaller the ce error, the better the model. Details of cross entropy can be found in [11] [12].

For logistic models, we tuned the number of input variables. We reduced the input variables from eight to four. We identified four most significant for easy comparison with ANN models since we had also reduced ANN models to four input variables.

2.3. Cross-Validation of Classification Results for ANN and Logistic Models

Having obtained optimal models for each of the simulation set for both ANN and logistic regression models, we performed 10-fold cross-validation for classification of simulated trees from both structured and non-structured populations. Measures used to compare classification results included: sensitivity, specificity, accuracy and area under the curve (AUC) for receiver operating characteristic (ROC) graphs. [21] defined sensitivity as a ratio of true positive to sum of true positive and false negative, while specificity as a ratio of true negative to sum of false positive and true negative. In classification problems, true positive and false positive are number of cases predicted as positive yet they were actually positive and cases predicted positive yet they were negative, respectively. True negative and false negative are analogously defined. This implies that false positive and false negative are considered as miss-classification cases. Accuracy is therefore a sum of true positive and true negative cases divided by the total number of cases classified. AUC for ROC graphs were defined in [22].

2.4. Clustering of Phylogenetic Trees Using Tree Statistics

Since ANN and logistic regression models are supervised learning techniques, we wanted to compare the two with unsupervised learning technique. We therefore did clustering by k-means. We were interested in finding out the optimal number of clusters that could be obtained from the tree simulated sets. We first used all the eight tree statistics and later reduced to four for easy comparison with ANN and logistic regression models. We used the exact four tree statistics that were used for reduced ANN and logistic regression models. Using R package NbClust of [23], we obtained optimal number of clusters for both full simulation sets (when all eight tree statistics used) and reduced simulation sets (when four tree statistics were used). NbClust gives optimal number of clusters for a given data set using thirty indices.

3. Results

3.1. Results for ANN and Logistic Regression Models

The visualization for a full ANN model for simulation set 1 is shown in Figure 1. The ANN model shown has one input layer with eight neurons. The entropy error was approximately 335 and it required 46934 steps to converge.

The corresponding generalized weights for ANN model in Figure 1 are shown in Figure 2. These generalized weights are for all the eight tree statistics for simulation set 1. From Figure 2, the four input variables with the largest variance, hence most influential in explaining the underlying structure for simulation set 1 are Colless and Sackin indices, maximum width and width-to-depth ratio. We also plotted the generalized plots for simulation sets 2 and 3. For these two simulation sets, the four most influential input variables were the same and these were Colless, Sackin, and total cophenetic indices and maximum depth.

We obtained optimal ANN models for each of the three simulation sets using AIC, BIC and entropy error. Results for simulation sets 1 and 2 are shown in Figure 3. The optimal model for simulation sets 1 and 3, had 2 neurons for the first hidden layer and 1 neuron for second hidden layer. For simulation set 2, the optimal model had 1 neuron for the first hidden layer and 2 neurons for the second hidden layer.

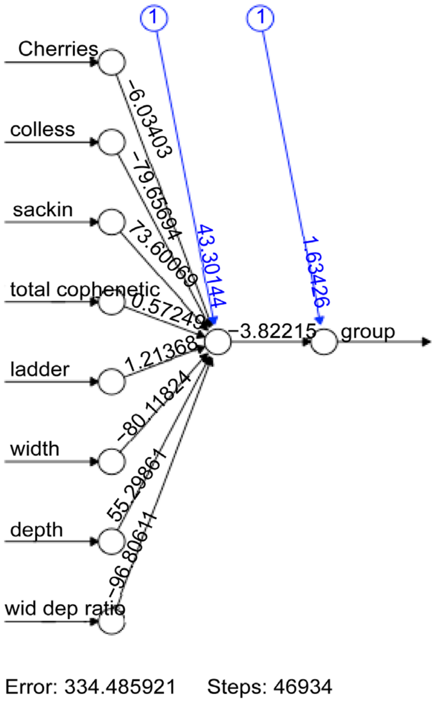

Figure 1. Aplot of a trained ANN model for simulation set 1 including synaptic weights, error and steps involved during the training.

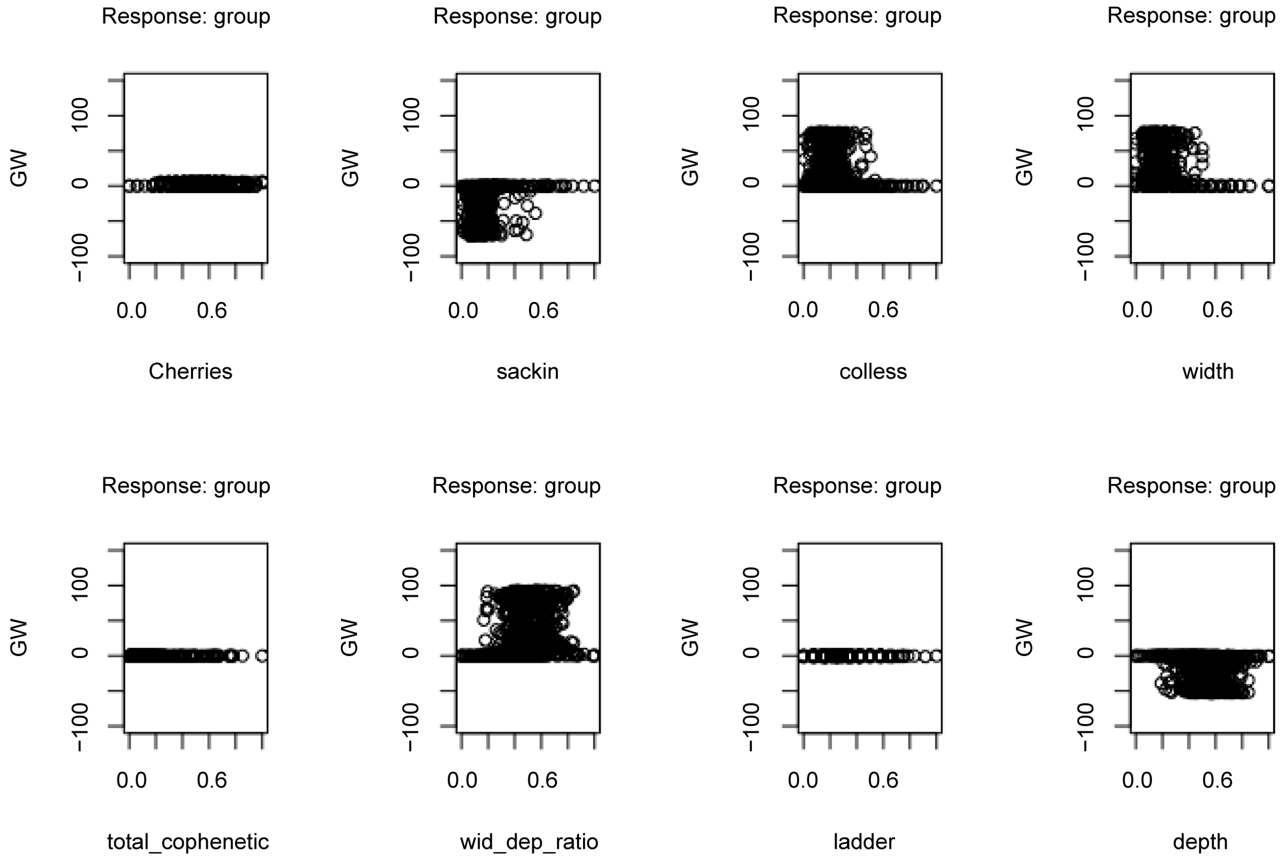

Figure 2. A plot of generalized weights with respect to each variable for general ANN model of simulation set 1.

For logistic regression models, the most significant variables for simulation sets 1 and 2 were: number of cherries, Colless, Sackin and total cophenetic indices. For simulation set 3, the four most significant variables were: number of cherries, total cophenetic index, maximum width and maximum depth.

3.2. Results for the 10-Fold Cross-Validation for ANN and Logistic Models

Having established the optimal models for both ANN and logistic regression models, we performed 10-fold cross-validation. Table 1 shows means for sensitivity, specificity, accuracy and AUC. Results for ANN and logistic regression are comparable, though in both models, simulation set 1 had the least mean values, but simulation set 3 had the best mean values for the measures used.

3.3. Results for Clustering

Figure 4 shows the optimal number of clustering using average silhouette width and gap statistic for simulation sets 1 and 2. For these two statistics, the optimal number of clusters was 2. We analysed both for full simulation sets (when all eight tree statistics used) and for reduced simulation sets (when only four tree statistics) were used. We had six reduced simulation sets, three according to reduced simulation sets used for ANN models and three according to reduced simulation

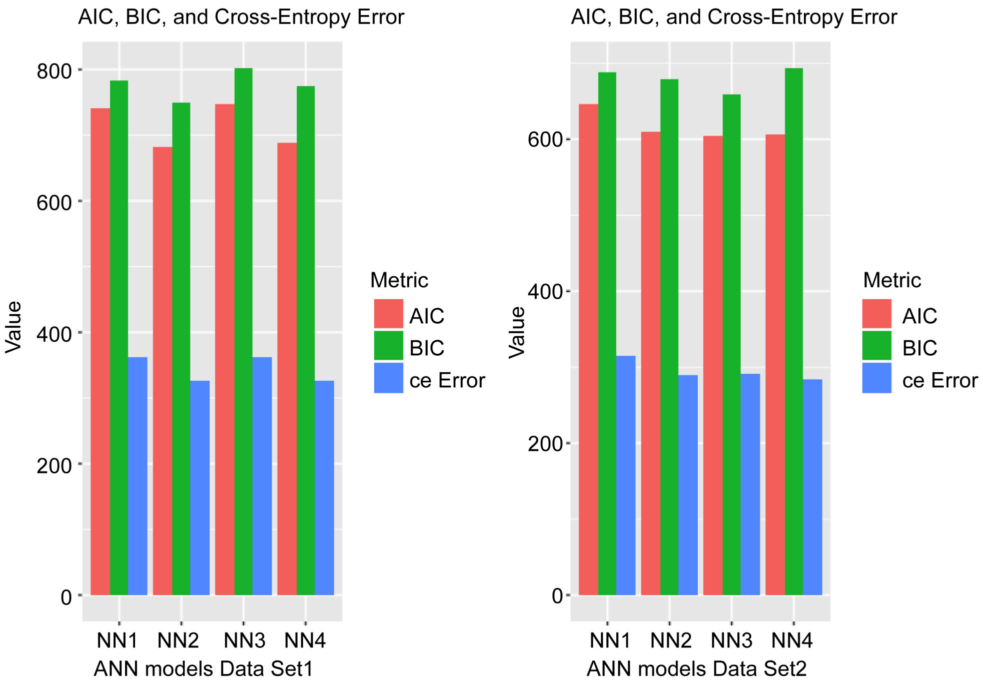

Figure 3. AIC, BIC and entropy error for reduced ANN models of simulation sets 1 and 2.

Table 1. Results of 10-fold cross-validated classification. The values shown are the average for the measures.

ANN 10-fold cross-validation for classification				
Dataset	sensitivity	specificity	accuracy	AUC
1	0.7322	0.7883	0.7560	0.7547
2	0.7954	0.8127	0.8010	0.8012
3	0.8371	0.8336	0.8355	0.8358
Logistic Regression 10-fold cross-validation for classification				
1	0.6976	0.6977	0.6960	0.6970
2	0.8111	0.8196	0.8150	0.8136
3	0.9650	0.9839	0.9740	0.9737

sets used for logistic regression models. This led to nine simulation sets (since we had three full simulation sets) that we investigated the optimal number clusters that was ideal for the data. Out of nine, five simulation sets resulted in the optimal number of clusters as two and the rest three.

4. Conclusions

From the results obtained, it was evident that ANN and logistic regression models had comparable performance. A comparison of reduced models with four input variables revealed that for any of the three simulation sets, at least two input variables in the reduced models for ANN and logistic regression were similar.

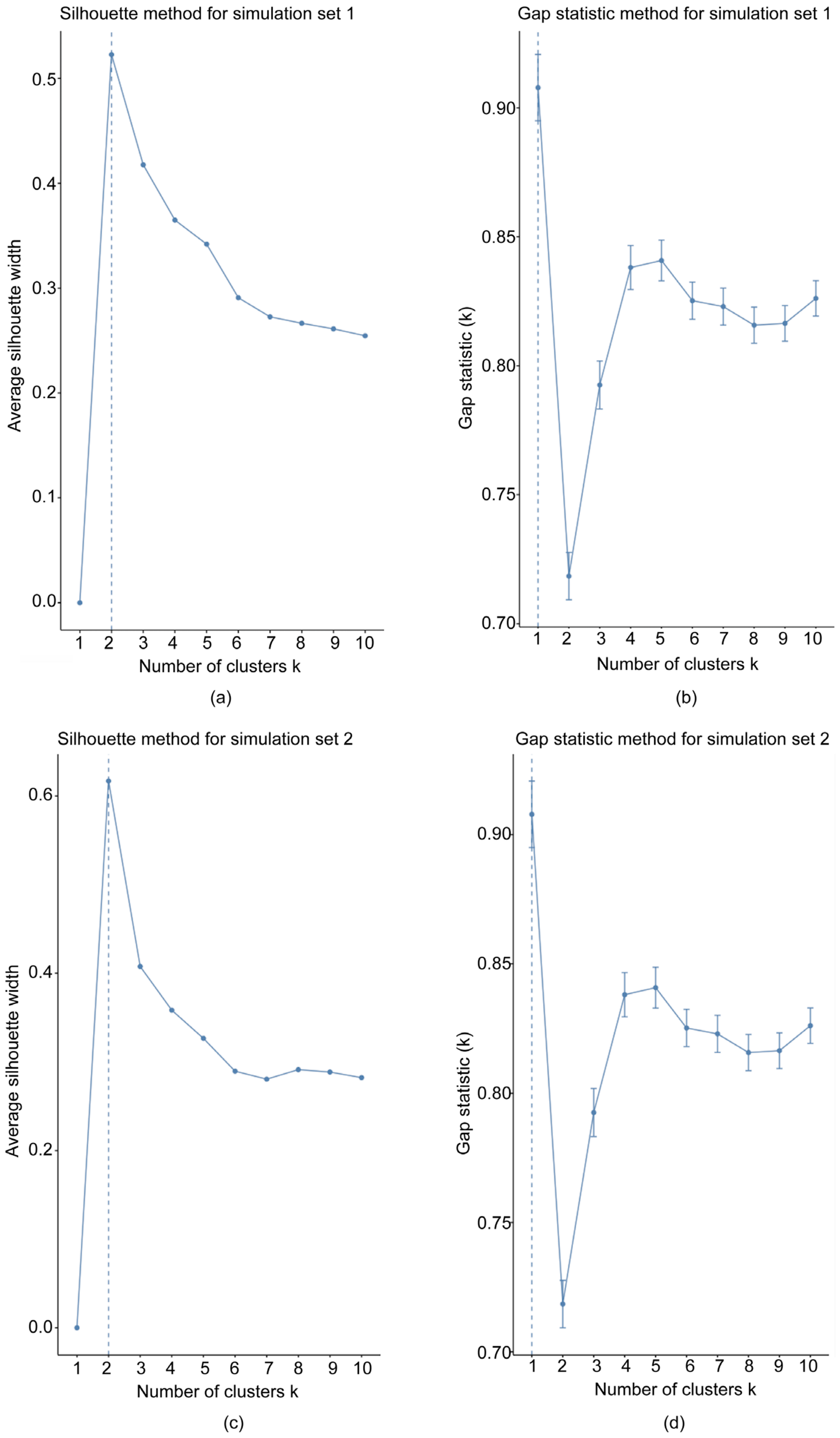

Figure 4. Average Silhouette width and Gap statistic indices that were used to detect the number of clusters in our simulation sets of data. (a) Average Silhouette width for simulation set 1; (b) Gap statistic for simulation set 1; (c) Average Silhouette width for simulation set 2; (d) Gap statistic for simulation set 2.

For simulation set 1, ANN models had the four most significant input variables as Colless and Sackin indices, maximum width and width to depth ratio. For logistic regression, four most significant variables were: number of cherries, Colless, Sackin and total cophenetic indices. For simulation set 2, three of the four significant input variables were common in both ANN and logistic regression models. These were: Colless, Sackin and total cophenetic indices. For simulation set 3, two input variables of the four most significant were common in both ANN and logistic models. These were: total cophenetic index and maximum depth.

For 10-fold cross-validation classification, in both ANN and logistic regression models, the mean values for sensitivity, specificity, accuracy and AUC were least for simulation set 1 and highest for simulation set 3, as shown in **Table 1**. The mean accuracy values for both ANN and logistic regression models were comparable with highest value of 0.974 for logistic regression for simulation set 3. The lowest was still for logistic regression of 0.696, and it was for simulation set 1. This was because phylogenetic trees simulated in set 3 had more leaves. This implied more information during the training of the classification models, hence better classification results for simulation set 3 compared to simulation set 1. We choose to compare logistic regression with ANN models because ANN models are considered as complex and whose internal mechanism is hard to understand, hence it is referred to as a black box classification technique in literature. Whereas logistic regression is one of the simplest regression models with only regression coefficients to be estimated during the model training. The fact that ANN models performed comparably with logistic regression models suggests that the tree statistics employed to predict the underlying population structure did well.

The results for clustering revealed that 2 or 3 clusters were optimal for most of the indices for clustering that were used. The unsupervised learning results reveal that structure was fairly detected by the clustering technique though not as accurate as expected since some indices were reporting 3 clusters. This is not surprising for clustering technique given the fact that it is a non-supervised technique.

Acknowledgements

We thank the Editor and reviewers for their comments which greatly improved this paper. The authors thank Pan-African University Institute of Basic Sciences, Technology and Innovation (PAUSTI) for funding this research.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Stadler, T. (2010) Sampling through Time in Birth-Death Trees. *Journal of Theoretical Biology*, **267**, 396-404. https://doi.org/10.1016/j.jtbi.2010.09.010

[2] Stadler, T. (2013) Recovering Speciation and Extinction Dynamics Based on Phylo-

genies. *Journal of Evolutionary Biology*, **26**, 1203-1219. https://doi.org/10.1111/jeb.12139

[3] Maddison, W.P., Midford, P.E. and Otto, S.P. (2007) Estimating a Binary Characteristics Effect on Speciation and Extinction. *Systematic Biology*, **56**, 701-710. https://doi.org/10.1080/10635150701607033

[4] Stadler, T. and Bonhoeffer, S. (2013) Uncovering Epidemiological Dynamics in Heterogeneous Host Populations Using Phylogenetic Methods. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **368**, Article ID: 20120198. https://doi.org/10.1098/rstb.2012.0198

[5] Volz, E.M. (2012) Complex Population Dynamics and the Coalescent under Neutrality. *Genetics*, **190**, 187-201. https://doi.org/10.1534/genetics.111.134627

[6] De Bruyn, A., Martin, D.P. and Lefeuvre, P. (2014) Phylogenetic Reconstruction Methods: An Overview. In: *Molecular Plant Taxonomy*, Humana Press, New York, 257-277. https://doi.org/10.1007/978-1-62703-767-9_13

[7] Blum, M.G., François, O. and Janson, S. (2006) The Mean, Variance and Limiting Distribution of Two Statistics Sensitive to Phylogenetic Tree Balance. *The Annals of Applied Probability*, **16**, 2195-2214. https://doi.org/10.1214/105051606000000547

[8] Brown, A.J.L., Precious, H.M., Whitcomb, J.M., Wong, J.K., Quigg, M., Huang, W., Daar, E.S., Richard, T.D., Keiser, P.H., Connick, E. and Hellmann, N.S. (2000) Reduced Susceptibility of Human Immunodeficiency Virus Type 1 (HIV-1) from Patients with Primary HIV Infection to Nonnucleoside Reverse Transcriptase Inhibitors Is Associated with Variation at Novel Amino Acid Sites. *Journal of Virology*, **74**, 10269-10273. https://doi.org/10.1128/JVI.74.22.10269-10273.2000

[9] Intrator, O. and Intrator, N. (2001) Interpreting Neural-Network Results: A Simulation Study. *Computational Statistics & Data Analysis*, **37**, 373-393. https://doi.org/10.1016/S0167-9473(01)00016-0

[10] Huo, S., He, Z., Su, J., Xi, B. and Zhu, C. (2013) Using Artificial Neural Network Models for Eutrophication Prediction. *Procedia Environmental Sciences*, **18**, 310-316. https://doi.org/10.1016/j.proenv.2013.04.040

[11] Günther, F. and Fritsch, S. (2010) NeuralNet: Training of Neural Networks. *The R Journal*, **2**, 30-38. https://doi.org/10.32614/RJ-2010-006

[12] Zhang, Z. (2016) Neural Networks: Further Insights into Error Function, Generalized Weights and Others. *Annals of Translational Medicine*, **4**, 300. https://doi.org/10.21037/atm.2016.05.37

[13] Dawson, C.W. and Wilby, R.L. (2001) Hydrological Modelling Using Artificial Neural Networks. *Progress in physical Geography*, **25**, 80-108. https://doi.org/10.1177/030913330102500104

[14] Heermann, P.D. and Khazenie, N. (1992) Classification of Multispectral Remote Sensing Data Using a Back-Propagation Neural Network. *IEEE Transactions on Geoscience and Remote Sensing*, **30**, 81-88. https://doi.org/10.1109/36.124218

[15] McKenzie, A. and Steel, M. (2000) Distributions of Cherries for Two Models of Trees. *Mathematical Biosciences*, **164**, 81-92. https://doi.org/10.1016/S0025-5564(99)00060-7

[16] Mir, A. and Rossello, F. (2013) A New Balance Index for Phylogenetic Trees. *Mathematical Biosciences*, **241**, 125-136. https://doi.org/10.1016/j.mbs.2012.10.005

[17] Colijn, C. and Gardy, J. (2014) Phylogenetic Tree Shapes Resolve Disease Transmission Patterns. *Evolution, Medicine and Public Health*, **2014**, 96-108. https://doi.org/10.1093/emph/eou018

[18] Milligan, G.W. and Cooper, M.C. (1988) A Study of Standardization of Variables in

Cluster Analysis. *Journal of Classification*, **5**, 181-204. https://doi.org/10.1007/BF01897163

[19] Nelder, J.A. and Wedderburn, R.W. (1972) Generalized Linear Models. *Journal of the Royal Statistical Society*: *Series A* (*General*), **135**, 370-384. https://doi.org/10.2307/2344614

[20] Aho, K., Derryberry, D. and Peterson, T. (2014) Model Selection for Ecologists: The Worldviews of AIC and BIC. *Ecology*, **95**, 631-636. https://doi.org/10.1890/13-1452.1

[21] Powers, D.M. (2011) Evaluation: From Precision, Recall and f-Measure to Roc, Informedness, Markedness and Correlation. *Journal of Machine Learning Technologies*, **2**, 37-63.

[22] Fawcett, T. (2006) An Introduction to Roc Analysis. *Pattern Recognition Letters*, **27**, 861-874. https://doi.org/10.1016/j.patrec.2005.10.010

[23] Charrad, M., Ghazzali, N., Boiteux, V. and Niknafs, A. (2014) Nbclust: An R Package for Determining the Relevant Number of Clusters in a Data Set. *Journal of Statistical Software*, **61**, 1-36. https://doi.org/10.18637/jss.v061.i06