Prediction of The Spread of *Acacia nilotica* Invasive Species Using Generalized Linear Models

Sahid Agustian Hudjimartsu, Yeni Herdiyeni Lili Budi Prasetyo, Agung Siswoyo

1Department of Computer Science, Faculty of Math and Natural Sciences, Bogor Agricultural University, Dramaga Campus, Bogor, Indonesia 16680
2Department of Forest Conservation and Ecotourism, Faculty of Forestry, Bogor Agricultural University, Dramaga Campus, Bogor, Indonesia, 16680
3Bromo Tengger Semeru National Park, The Ministry of Environment and Forestry, East Java, Indonesia, 65125

Received August 30, 2017/Accepted December 23, 2017

Abstract

*Acacia nilotica* is an invasive species in the Baluran National Park. The spread of *A. nilotica* is a serious threat that can disrupt the function of savanna ecosystems in providing food for herbivores and eventually led to a decrease in wildlife populations of herbivorous. The growth of *A. nilotica* invasive species is influenced by some environmental factors. In order to prevent the spread of *A. nilotica*, a model can be formed to predict the species distribution based on the environmental factors namely generalized linear models (GLM). This research used the data from Siswoyo's research in 2014, which consisted of environmental data and distribution of *A. nilotica*. The results showed that there were five environmental variables that significantly affect on the spread of *A. nilotica* invasive species. The variables included elevation, surface temperature, normalized difference vegetation index, normalized difference moisture index, and the distance from the river, with p-value less than 0.05. The accuracy of the prediction model on the spread of *A. nilotica* using GLM was 97.22% and the area under curve (AUC) value was 0.981.

Keywords: *Acacia nilotica*, Baluran National Park, GLM, invasive species, savanna

*Correspondence author, email: shudjimartsu@gmail.com, ph.: +62-81287920066

Introduction

The Baluran National Park (BNP) is renowned as an illustration of the savanna in Africa and also as a miniature forest because almost all types of Indonesian forests can be found in this national park. The park's natural diversity ranges from tropical rainforest to coral reefs that span from Bama Coast in the eastern region Baluran to Bilik Coast in the northern region Baluran (TN Baluran 2017). This condition makes the park a good habitat for biodiversity, especially for flora.

One of the flora in BNP is *A. nilotica*, which was originally brought in around the year 1969 as a natural firebreak (Sabarno 2002). The characteristics and the widespread distribution of *A. nilotica*, which was originally used as a protective savanna, make *A. nilotica* as an invasive species. The invasion of *A. nilotica* can disrupt the function of savanna ecosystems in providing food for herbivores and eventually led to a decrease in wildlife population of herbivorous (Caesariantika et al. 2011).

The spread of *A. nilotica* is influenced by some environmental factors (Kriticos et al. 2003). The relationship between the environmental factors and the species referred to as an ecology, which can be analyzed by using species distribution models (SDM). Other scholar use species niche model, ecological niche model or niche theory model terminology (Franklin 2009). The SDM provides a way to analyze the geographical influence of environmental factors on species (Segurado & Araújo 2004) by statistical or machine learning methods.

There are various techniques to develop SDM such as linear regression (Rahmat et al. 2012), multivariate (principal component analysis/PCA) (Prayogo et al. 2014; Favata et al. 2015; Prasetyo et al. 2017), maximum entropy (maxent) (Bowler 2014), or generalized linear models (GLM) (Guisan et al. 2002; Ulrich et al. 2016). The GLMs are alternative to the normal linear model, particularly when the independent variable does not follow normal distribution. GLM is an extensions of linear models that can cope with non-normal distributions of species (Venables & Ripley 1999). The GLM approach can also analyze how important the relationship between environment factors with the distribution of species (Guisan et al. 2002).

Several studies have considered GLMs in the prediction of species distribution. For example, Guisan et al. (1998) predict the distribution of plant species *Carex curvula* in the Swiss mountains. In the study, many predictors (environment factors) were used, and the authors then employ the PCA to reduce the dimensions before being
analyzed by GLM. Furthermore, Uguru et al. (2014), also consider GLM to predict the ingredient of protein in *A. nilotica* pods for goats feeding. Siswoyo (2014) also studies the habitat suitability of the species *A. nilotica* in the BNP using logistic regression.

The spread of invasive *A. nilotica*, which threatens the ecosystems of savannah, needs to be monitored by predicting its spread. The objective of this research is to predict spread of *A. nilotica* invasive species by analyzing the environment factors and how to know the correlation between the variable in the environment factors using GLM. For future work, this predictive modeling can be used as an initial step in maintaining existing ecosystem in BNP especially savanna ecosystems.

**Methods**

Several stages in predicting the spread of *A. nilotica* in BNP consists of: data collection, multicollinearity test, data division, GLM prediction modelling, model, and results testing. The tools used in this research are the application of R for statistical analysis by GLM and Quantum GIS for spatial analysis. The flow chart of this research can be seen in Figure 1.

**Data collection** This research used data of Siswoyo (2014) in BNP Situbondo, East Java Province. The data consists of environment data and distribution data of *A. nilotica*. There are six environmental data: elevation, slope, NDVI, NDMI, temperature, and distance from the river. These six factors are spatial data representing primary environmental regimes for the analysis of species distribution models (Franklin 2009) where as the distribution data are presence and absence of *A. nilotica*.

**Data pre-processing** Data pre-processing is divided into three following phases:

1 **Data extraction**

   To extract the value of environmental and distribution data, we use overlay analysis. The illustration of overlay analysis can be seen in Figure 2.

2 **Multicollinearity testing between predictor variables (environment data)**

   Multicollinearity is a condition where two or more predictor variables highly correlated with one and another. As a result, it is hard to find good estimates of their distinct effects on dependent variables (Midzi et al. 2010). Multicollinearity testing between predictor variables can be solved by examining the variance inflation factor (VIF). The VIF value is based on the calculation of tolerance in which if the tolerance value close to zero the variables have a high multicollinearity. Otherwise, if the tolerance value close to one the variables have a very small multicollinearity (Sunaryo et al. 2015).

   \[
   \text{Tolerance} = Tol = 1 - R^2 \tag{1}
   \]

   note: $R^2$ is the determinant coefficient of the predictor variables.

   \[
   \text{VIF} = \frac{1}{Tol} = \frac{1}{1 - R^2} \tag{2}
   \]

3 **Data division**

   The data is divided into training data and testing data with K-fold cross validation. Training data was used to analyze the model equations while the testing data was used to test the model equations. K-fold cross validation is a method used to evaluate the learning algorithm by dividing the data into K-fold, as K-1 fold is used as training data and 1 fold as the testing data (Liu & Özsu 2009). Data division with K-fold cross-validation can be illustrated as in Figure 3.

![Figure 1 Overlay and extraction data variable response and predictor.](image1)

![Figure 2 Research flow.](image2)
Generalized linear models (GLMs) Regression analysis is a modeling with a statistical approach that determines the relationship between variable response (influenced) and variable predictor (cause). The regression approach is theoretically capable of delivering good results when the response variables are normally distributed and the variance of the data is unchanged. One type of regression model is the GLM that can resolve when the response variable is not normally distributed and the variance of data changes (Guisan & Zimmermann 2000).

Ecological data sometimes violates the rules of the linear model assumptions. Venables and Ripley (2013) state that GLM is the development of linear regression, which can overcome non-distributed data on response variables and transform into linearity, which cannot be done by ordinary linear regression. The GLM can be expressed as Equation [3]:

$$g(\mu) = \beta_0 + \sum_{j=1}^{p} x_j \beta_j + \epsilon$$

note: $g = \text{link function}, \ E = \text{expected value}, \ Y = \text{response variable}, \ x = \text{predictor variable}, \ \beta = \text{vector of estimated coefficients (slope)}, \ \beta_0 = \text{estimated constant (intercept)}, \ \epsilon = \text{epsilon}.$

By GLM, the combination of the predictor variable produces linear predictor ($LP$) and the expected value of the $Y$ variable correlates with the $LP$ through the $g()$ as connecting function. The connecting function is described as the mean value of the $Y$ variable, based on the $LP$ or the expected exponential link from $Y$ to the predictor variable. For example, when $Y$ response is binary (0 and 1), then the distribution response ($Y$) is binomial and logit link is used for the connecting function. This connecting function is often used in species distribution modeling (SDM) because generally data of species emergence is binomial-distributed data. The GLM with binomial distribution is often called logistic regression (Franklin 2009). The connecting function on the binomial-distributed GLM, with the logit link connecting function is shown by Equation [4].

$$LP = \log \left( \frac{\mu}{1 - \mu} \right)$$

note: $LP = \text{Linear Predictor}, \ \mu = E(Y)$

In logistic regression we are often interested in the predicted probability of a class, where $\mu = E(Y)$ is the probability of class 1 (presence), and $(1 - \mu)$ is the probability of class 0 (absence), so by inverting the above equation we can solve for $\mu$ as shown in Equation [5].

$$\mu = \frac{e^{LP}}{1 + e^{LP}}$$

Model estimation In linear regression for estimating unknown parameter the method used is least square when variable $Y$ normal distributed. Unfortunately, the method cannot be applied to estimate unknown parameter because variable $Y$ is not normally distributed or dichotomous outcome. The estimation method that leads to the least squares function under the linear regression model is maximum likelihood. The method can provide the foundation for approach to estimate with regression logistic (Hosmer & Lemeshow 2000). It is required to maximizing the log-likelihood for estimating the parameters, because the parameter allegedly not linear and usually solved by an iteratively reweighted least squares (IRLS) algorithm (Franklin 2009).

Model selection Model selection can also be referred to as a selection variable is a model estimated to use a subset of different predictor variables and compared with some criteria (Franklin 2009). One of them uses the AIC value (Akaike Information Criterion), which is a statistical approach that is often used in model selection (Venables & Ripley 1999).

The equation of AIC is given by Equation [6]:

$$AIC = -2 \text{maximum log likelihood} + 2p$$

Where $p$ is the number of predictor variables, the smaller AIC value the better the model. There are several ways in the selection model, including:

1. backward elimination is removing the insignificant predictor variable in the model.
2. forward Selection is adding significant predictor variables to the model.
3. stepwise procedures are automatically using forward and backward in the predictor variable selection strategy.

Those ways are important for searching pormonious model that uses an appropriate subset all available predictor variables. Essentially, variety models are estimated using different subsets of predictors and compared with AIC, in terms of their fit to the data (Franklin 2009).

Model evaluation The classification table is most proper when classifying a purpose in the analysis (Table 1). Classification table is two-way frequency tables between the actual data and prediction data estimation results of the model based on testing data (Hosmer & Lemeshow 2000).

Receiver operating characteristic (ROC) is a method based on sensitivity and specificity. Sensitivity explains how well the model predicts presence, whereas the specificity explains how well the model predicts absence (Baldwin 2009). ROC curve is a plot between the probability of false positive (1-specificity) and true positive (sensitivity). The area under the ROC curve is called Area Under the Curve (AUC), ranging 0–1 and this value indicates the performance of model to classify the data. The AUC values...
of 0.5–0.7 are considered poor (poor model performance), values of 0.7–0.9 are considered to be good enough, and > 0.9 indicates excellent accuracy (Manel et al. 2001).

Results and Discussion

Model prediction for spread of Acacia nilotica invasive species in this study, prediction models were analyzed statistically using the GLM. Before the data were analyzed by GLM, we did data pre-processing on the existing data. The results of the preprocessing data such as data extraction between the response variable and predictor variables can be seen on Table 2.

After the extraction data, the next step is multicollinearity testing using VIF to determine whether there is collinearity between variables (Table 3). The results of multicollinearity test showed VIF value was lower than 10. If the VIF value is greater than 10, it indicates a problem multicollinearity between predictor variables (Sunaryo et al. 2015).

This research conducted two experiments with K values to obtain better results in the distribution of the training data and testing data. Based on the results of experiment K value on K-fold cross validation between K = 5 and K = 10, it was found that the number of K = 10 has better accuracy compared with the number of K = 5. The results of data division with K = 10 in cross-validation showed that there are 324 training data and 36 testing data. The training data were used as a data for modeling in GLM. The results GLM modelling are used to analyze the relationship between response (influenced) with the variable predictor (cause) and model selection as shown in Table 4.

Table 4 shows that there are several models in analyzing and selecting models using GLM.

Data analysis using GLM The result of predictive model showed that the distribution of A. nilotica with Model 4 resulted in significance and coefficient of each variable predictor (Table 5). As seen on Table 5, a total of 5 parameters were significant due to p-values less than 0.05, which means the model used has a confidence level of 95%. According to the result, these 5 predictor variables can affect the invasive distribution of A. nilotica species in TN Baluran. Furthermore, the probability equation of predictive model of the invasive species distribution of A. nilotica with GLM is shown in Equation [7],

\[ P = \frac{1}{1 + e^{(-0.853 + 0.024 \cdot x_1 - 25.434 \cdot x_2 - 0.002 \cdot x_4 + 0.478 \cdot x_5)}} \]

Table 1 Classification Table

<table>
<thead>
<tr>
<th>Predicted</th>
<th>Observed</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>True Positive (sensitivity)</td>
<td>False Positive</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>False Negative</td>
<td>True Negative (Specificity)</td>
<td></td>
</tr>
</tbody>
</table>

Table 4: Results of GLM modelling shows that there are several models in analyzing and selecting models using GLM. Model 4 showed that the distribution of A. nilotica with Model 4 resulted in significance and coefficient of each variable predictor (Table 5). As seen on Table 5, a total of 5 parameters were significant due to p-values less than 0.05, which means the model used has a confidence level of 95%. According to the result, these 5 predictor variables can affect the invasive distribution of A. nilotica species in TN Baluran. Furthermore, the probability equation of predictive model of the invasive species distribution of A. nilotica with GLM is shown in Equation [7],

\[ P = \frac{1}{1 + e^{(-0.853 + 0.024 \cdot x_1 - 25.434 \cdot x_2 - 0.002 \cdot x_4 + 0.478 \cdot x_5)}} \]

Note: \( x_1 \) = elevation, \( x_2 \) = normalized difference moisture index, \( x_4 \) = normalized difference vegetation index, \( x_5 \) = Distance from river, \( x_4 \) = surface temperature, \( e \) = exponential (2.718282).

By Equation [7], the coefficients of each predictor variables explain the relationship between predictor variable to the response variable. The negative coefficient has a negative relation to the response variable, thus the less predictor variable value, the higher probability of distribution of A. nilotica. Conversely, the positive value coefficient has a positive relationship to the response variable, if the value of the predictor variable increases, the higher probability of distribution of A. nilotica.

Distribution of invasive species A. nilotica based on elevation. Invasive distribution of A. nilotica species is spread at an elevation of 0–100 m above sea level (lowland) and a small portion is at an elevation of 100–200 m above sea level (Figure 4a). Based on the analysis, the high probability of invasive distribution of species A. nilotica located in the lowlands. According Sabarno (2002), the lowlands of BNP have alluvial soil and black soil types. Alluvial and black soil types are excellent for growing A. nilotica (Bargli & Bargli 2009).

The invasive spread of A. nilotica species based on NDMI NDMI is the index value derived from remote sensing data related to liquid water. NDMI index value greater than 0.1 is a signal of high humidity (moisture), while a low value (close to -1) explains the low humidity levels (Herbei et al. 2012). According to Figure 4b, it can be seen the distribution of A. nilotica by NDMI contained of the index value of 0.120–0.211. Its means the area has a moderate humidity. According to Kriticos et al. (1999), the growth of A. nilotica is influenced by moisture availability.

The invasive spread of A. nilotica species based on NDVI NDVI is the index value used to measure the greenness of vegetation. Higher NDVI values (close to 1), means a very good green vegetation canopy level, whereas an index value less than 0, is generally a body of water or area without vegetation (Prayogo et al. 2014). Map of distribution of A. nilotica based on NDVI can be seen in Figure 5a. Figure 5a
Table 2 Result of the extraction data

<table>
<thead>
<tr>
<th>Variable response (Y)</th>
<th>Elevation (masl)*</th>
<th>NDMI</th>
<th>NDVI</th>
<th>Distance from river (m)</th>
<th>Slope (°)</th>
<th>Surface Temperature (°C)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>27.00</td>
<td>0.11</td>
<td>0.33</td>
<td>201.25</td>
<td>4.62</td>
<td>24.79</td>
</tr>
<tr>
<td>1</td>
<td>9.00</td>
<td>0.11</td>
<td>0.30</td>
<td>494.77</td>
<td>2.92</td>
<td>24.85</td>
</tr>
<tr>
<td>1</td>
<td>59.00</td>
<td>0.11</td>
<td>0.32</td>
<td>180.00</td>
<td>9.16</td>
<td>27.20</td>
</tr>
<tr>
<td>1</td>
<td>23.00</td>
<td>0.17</td>
<td>0.28</td>
<td>127.28</td>
<td>10.00</td>
<td>22.02</td>
</tr>
<tr>
<td>1</td>
<td>17.00</td>
<td>0.21</td>
<td>0.37</td>
<td>757.17</td>
<td>5.40</td>
<td>21.99</td>
</tr>
<tr>
<td>0</td>
<td>239.00</td>
<td>0.24</td>
<td>0.39</td>
<td>270.00</td>
<td>5.28</td>
<td>21.52</td>
</tr>
<tr>
<td>0</td>
<td>537.00</td>
<td>0.16</td>
<td>0.16</td>
<td>30.00</td>
<td>14.16</td>
<td>19.10</td>
</tr>
<tr>
<td>0</td>
<td>582.00</td>
<td>0.16</td>
<td>0.16</td>
<td>42.43</td>
<td>24.10</td>
<td>18.05</td>
</tr>
<tr>
<td>0</td>
<td>639.00</td>
<td>0.15</td>
<td>0.16</td>
<td>335.41</td>
<td>19.46</td>
<td>16.13</td>
</tr>
<tr>
<td>0</td>
<td>786.00</td>
<td>0.09</td>
<td>0.08</td>
<td>488.36</td>
<td>36.05</td>
<td>13.41</td>
</tr>
</tbody>
</table>

Note: 1= Presence, 0 = absence

Table 3 Collinearity between variable predictor

<table>
<thead>
<tr>
<th>Variable predictor</th>
<th>Collinearity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Tolerance</td>
</tr>
<tr>
<td>Elevation</td>
<td>0.353</td>
</tr>
<tr>
<td>NDMI</td>
<td>0.291</td>
</tr>
<tr>
<td>NDVI</td>
<td>0.221</td>
</tr>
<tr>
<td>Distance from river</td>
<td>0.935</td>
</tr>
<tr>
<td>Surface temperature</td>
<td>0.505</td>
</tr>
<tr>
<td>Slope</td>
<td>0.573</td>
</tr>
</tbody>
</table>

Table 4 Result of GLM analysis and selection model

<table>
<thead>
<tr>
<th>Model</th>
<th>Total variable</th>
<th>AIC</th>
<th>Nagelkerke R²</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1</td>
<td>6</td>
<td>230.82</td>
<td>0.680</td>
</tr>
<tr>
<td>Model 2</td>
<td>5</td>
<td>228.85</td>
<td>0.680</td>
</tr>
<tr>
<td>Model 3</td>
<td>3</td>
<td>231.01</td>
<td>0.667</td>
</tr>
<tr>
<td>Model 4</td>
<td>5</td>
<td>228.85</td>
<td>0.680</td>
</tr>
</tbody>
</table>

Table 5 The coefficient and significance of each predictor variable

|                          | Estimate | Std. Error | z value | Pr(>|z|) |
|--------------------------|----------|------------|---------|---------|
| Intercept                | -1.853   | 2.583      | -0.717  | 0.473   |
| Elevation                | -0.024   | -0.004     | -6.706  | 0.000   ***
| NDMI                     | 17.462   | 8.593      | 2.032   | 0.042   *|
| NDVI                     | -25.434  | 10.069     | -2.379  | 0.017   *|
| Distance from river      | -0.002   | 0.000      | -2.606  | 0.009   **|
| Surface temperature      | 0.478    | -0.092     | 5.200   | 0.000   ***|

Signif. codes:  0 ‘***’  0.001 ‘**’  0.01 ‘*’  0.05 ‘.’  0.1 ‘ ’  1

Figure 4 Distribution map of *Acacia nilotica* based on elevation (a) and NDMI (b).
shows that the distribution of *A. nilotica* ranging 0.240–0.400. The NDVI value shows that area are savanna, shrubs, and secondary forest in BNP. The NDVI value indicates the high chance of invasive distribution of *A. nilotica* when grown in savanna and shrubs. In addition, based on the results of the survey (Siswoyo 2014) the distribution of *A. nilotica* is in some spot (spot) of savanna and shrub area.

**The invasive spread of *A. nilotica* species based on distance from rivers** Figure 5b shows *A. nilotica* can be found around riparian zones. Most of the *A. nilotica* were found in distance of 0–400 m from the river. According to (Bargali & Bargali 2009), *A. nilotica* can grow and develop properly while located near water sources. The results showed that the invasive distribution of *A. nilotica* increased as it grew or was near the border area of the river. This was reiterated by Radford *et al.* (2002) death of all plants, including *A. nilotica* is due to the location is out of riparian zones.

**The invasive spread of *A. nilotica* species based on surface temperature** Invasive species *A. nilotica* are uniformly distributed at surface temperatures of 18–27 °C as shown in Figure 6. The results show that the probability of distribution of *A. nilotica* species increases as they grow or are at higher surface temperatures. According to Kriticos *et al.* (2003) the overall species of *A. nilotica* can adapt and grow well at temperatures from 16–50 °C.

**Model evaluation** Evaluation predictive model of the spread

![Figure 5](image1)

**Figure 5** Distribution map of *Acacia nilotica* based on NDVI (a) and distance from river (b).

![Figure 6](image2)

**Figure 6** Distribution map of *Acacia nilotica* based on surface temperature.

![Table 6](image3)

**Table 6** Classification table based on GLM

<table>
<thead>
<tr>
<th>Prediction</th>
<th>Observed</th>
</tr>
</thead>
<tbody>
<tr>
<td>1(presence)</td>
<td>20</td>
</tr>
<tr>
<td>0 (absence)</td>
<td>15</td>
</tr>
</tbody>
</table>

![Figure 7](image4)

**Figure 7** ROC curve.
of A. nilotica species in BNP using GLM was done to find out how proper the model was able to predict the spread of A. nilotica. Model evaluation used data testing, started from data division with K-fold cross validation. Test results can be seen in the table of conformity classification (Table 6). Based on the classification table, it was obtained accuracy for class occurrence: a presence of 95.24%, an absence of 100%, and overall accuracy of the model was 97.22%. This proves that this model is very good, thus it can be used as a predictive model of the spread of invasive species in BNP.

Table 6: Classification Table

<table>
<thead>
<tr>
<th>Class Occurrence</th>
<th>Actual</th>
<th>Predicted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Presence</td>
<td>95.24%</td>
<td></td>
</tr>
<tr>
<td>Absence</td>
<td>100%</td>
<td></td>
</tr>
<tr>
<td>Overall Accuracy</td>
<td>97.22%</td>
<td></td>
</tr>
</tbody>
</table>

The area under ROC curve or AUC was used as a measure of performance of the classifier and the result AUC value of 0.981. This indicates that the model has a good performance in discriminating between the two categories. The discrimination of two categories is based on thresholding in which this research used spec sens. That was calculated from sum of the highest sensitivity and specificity.

The calculation of the spread A. nilotica invasive species using GLM prediction model revealed, that there are about an area of 7.759 ha which is predicted as a site to grow A. nilotica invasive species. A. nilotica invasive is allegedly spread from northern, eastern, and southern part of BNP. The model prediction map of the spread A. nilotica invasive species by GLM is shown in Figure 8.

Conclusion

Model predictions of the spread A. nilotica invasive species in BNP using GLM are very good with the accuracy of the model is 97.22%. Moreover, performance (AUC) value in discriminant between the two categories (presence and absence) of the occurrence of A. nilotica invasive species is 0.981. Environmental or predictor variables that significantly affect the response variable are: elevation, surface temperature, NDVI, NDMI, and the distance from the river. These 5 variables have p-value less than 0.05, which means that the model used has a confidence level of 95%.

References


