Predicting infectivity of Arbuscular Mycorrhizal fungi from soil variables using Generalized Additive Models and Generalized Linear Models

IRNANDA AIKO FIFI DJUUNA1,2,* , LYNETTE K. ABBOTT2, KIMBERLY VAN NIEL3
1Department of Soil Sciences, Faculty of Agriculture and Agriculture Technology, The State University of Papua (UNIPA), Manokwari, Gunung Salju St. Amban Manokwari 98314, West Papua, Indonesia. Tel. +62-0986-211830, Fax: +62-0986-213513, *e-mail: irnanda_d@yahoo.com.au
2School of Earth and Geographical Science, The University of Western Australia, Crawley 6009, WA
3School of Earth and Geographical Science, The University of Western Australia, Crawley 6009, WA


ABSTRACT

Djuuna IAF, Abbott LK, Van Niel K (2010) Predicting infectivity of Arbuscular Mycorrhizal fungi from soil variables using Generalized Additive Models and Generalized Linear Models. Biodiversitas 11: 145-150. The objective of this study was to predict the infectivity of arbuscular mycorrhizal fungi (AM fungi), from field soil based on soil properties and land use history using generalized additive models (GAMs) and generalized linear models (GLMs). A total of 291 soil samples from a farm in Western Australia near Wickepin were collected and used in this study. Nine soil properties, including elevation, pH, EC, total C, total N, P, K, microbial biomass carbon, and soil texture, and land use history of the farm were used as independent variables, while the percentage of root length colonized (%RLC) was used as the dependent variable. GAMs parameterized for the percent of root length colonized suggested skewed quadratic responses to soil pH and microbial biomass carbon; cubic responses to elevation and soil K; and linear responses to soil P, EC and total C. The strength of the relationship between percent root length colonized by AM fungi and environmental variables showed that only elevation, total C and microbial biomass carbon had strong relationships. In general, GAMs and GLMs models confirmed the strong relationship between infectivity of AM fungi (assessed in a glasshouse bioassay for soil collected in summer prior to the first rain of the season) and soil properties.

Key words: AM Fungi, GAM, GLM, land use.

INTRODUCTION

Most agricultural crops are colonized by arbuscular mycorrhizal fungi (AM fungi), which occur in almost all soils, but the variability of species abundance differs across soil types (Abbott and Robson 1991). These fungi are claimed to be a main component of the soil micro biota in most agro-ecosystems. Some studies have been done on the benefit of indigenous AM fungi to the growth of plants in field soils, but it is difficult to assess the contribution of these fungi under the field conditions (Fitter 1985; Jakobsen 1994; Jakobsen et al. 2002).

The infectivity of AM fungi in soil has been found to be related to agricultural management practices such as (i) cropping systems (e.g. Thompson 1987; 1991; Bagayoko et al. 2000; Johnson et al. 1991; Hendrix et al. 1995), (ii) fertilizer application (e.g. Abbott and Robson 1984; Thomson et al. 1992; Gryndler et al. 1990; Liu et al. 2000; Joner 2000), (iii) cultivation (e.g. McGonigle and Miller 2000; Kabir et al. 1997; Johnson and Pfleger 1992; Douds et al. 1995), and (iv) land use intensity (Oehl et al. 2003; 2004). Agricultural practices are also known to reduce the abundance and diversity of mycorrhizal fungi (Boerner et al. 1996; Helgason et al. 1998). However, these relationships do not always hold.

Statistical modeling is commonly used to test relationships and to predict species distributions. With statistical techniques and GIS tools, the development of predictive habitat models has rapidly increased especially in ecological studies (Guisan and Zimmermann 2000). Statistical methods are based on correlation and often have as their purpose, the aim of prediction.

Soil properties and land use history are the most important variables which are directly related to the infectivity of AM fungi. However, the nature and strength of the relationships between AM fungal infectivity, soil properties and land use history has not been explored. If there is a strong relationship, it would be possible to spatially predict the infectivity of AM fungi in soil based on soil characteristics and land use variables using statistical modeling.

A range of data analysis methods can be applied to develop models for spatial prediction using environmental correlation. The most common method includes multiple regression models such as generalized linear modeling (GLM) (McCullagh and Nelder 1989) and generalized additive modeling (GAM) (Hastie and Tibshirani 1986; Yee and Mitchel 1991), which appear to be increasingly used for predicting species and habitat distribution. These models have been extensively reviewed by Franklin (1995), and Guisan and Zimmermann (2000).
GAMs and GLMs were selected in this study because these statistical models have been used as an exploratory tool in the analysis of species distribution with respect to the environmental factors. GAMs have been used in numerous studies of species-environment relationships (e.g., Bio et al. 1998; Austin 1999; Guisan and Zimmermann 2000). They are considered useful for exploring the shape of the response function because they do not assume any general shape of the response prior to estimation (Austin and Meyer 1996). Bio et al. (1998) concluded that GAMs are a useful and practical technique for improving current regression-based vegetation models by allowing for alternative complex response shapes. GLMs allow those response functions to be parameterized and their significance tested (Franklin 1998). GAMs and GLMs approaches have enabled biologists to model species responses to a wide range of environmental data types under a single theoretical and computational framework (Yee and Mitchell 1991).

The objectives of this study were: (i) to test the strength and nature of the relationships (using GAM) and, if possible, predict the infectivity of AM fungi in soil based on soil characteristic variables and land use history on a Wickepin farm, and (ii) to model the relative impacts of land use and soil characteristics at a farm scale.

**MATERIALS AND METHODS**

**Data sets**

Soils and land use history data of Wickepin (Djuuna 2006) were used. The 291 sampling points contains records of 15 data sets determined for each paddock. Each sampling point covers an area of 150 m². The infectivity records of AM fungi from bioassay data such as the percentage of root length mycorrhiza (%RLC) was used as response variable (dependent variable), and the following nine soil variables as predictors (independent variables): elevation, pH, EC, total C, total N, P, K, soil texture, and microbial biomass carbon. The selection of the predictor variables was based on important major factor influencing AM fungi generally in soil. Correlation coefficient tests were used to determine if there is any correlation among the predictor variables. The selected predictor variables show weak correlation with each other except for total C and total N. Because of this high correlation between total C and total N (r=0.91), one of these predictor variables cannot be used in the modeling. Total N was therefore only examined for shape and strength of the response in GAM modeling, but was not included in the stepwise model development and k-fold cross-validation analysis. Total N was selected for removal rather than total C, because two components of N (N-NH₄) and N-NO₃ were included.

**Methods of analysis**

**Generalized Linear Models and Generalized Additive Models**

GAMs were first parameterized using soil characteristics and land use variables. A smoothing spline term was used to explore the shape of the response curve and the strength of the relationship between response variable (dependent variable) such as %RLC and predictors (independent variables) i.e. soil characteristics (pH, EC, P etc). GLMs were parameterized for all variables using the response functions suggested by the GAMs (Franklin 1998). The interactions of variables were evaluated for significance in the GLMs. All variables were tested for significance using forward and backward stepwise selection, with response curves at different levels of complexity for each variable. The stepwise model selection procedure was started with the full model, and then at each step, one independent variable was tested for omitting and re-introducing to the model (Pearce and Ferrier 2000). Variables in the GLMs were tested for significance and deleted if not significance (backward elimination). The backward elimination is one of the procedures to be used in the model which is most powerful for fitting models to designed experimental situations (Nicholls 1989). In addition, the forward stepwise procedure is useful for exploratory model building. The model that had the lowest value for Akaike’s Information Criterion (AIC) was kept. The selection stopped when there was no independent variable addition or omission that would lower the AIC value. Statistical analyses were carried out using the S-Plus version 6.2 for Windows.

**Goodness of fit**

The model fit and significance of the variables were evaluated using the residual deviance (analogous to the residual sum of squares in the linear model). The residual deviance can be compared with analysis of variance using a χ²-test.

**Model evaluation**

The GAM and GLM models obtained from stepwise selection for infectivity of AM fungi were compared and evaluated in terms of discrimination using the Area Under the receiver operator Curve (AUC). This method provides a threshold-independent evaluation of the predictive performance of models than traditional comparison of relative error (Bradley 1997; Duda et al. 2001).

**Cross validation**

Model power and stability was evaluated using a k-fold cross-validation method (Fielding and Bell 1997). This method provides insight to the predictive power of the model, by measuring how well models developed from different data segments predict the omitted data. It gives a reasonable estimation of how well the model would perform on new data and can point to model instability. Data were randomly split into 10 equal-size groups. At each iteration, 9 of the 10 groups were used to build the model, and the other group was used as an independent validation set to evaluate the performance. This method was repeated 10 times. As a measure of how well the model predicts unknown data, the mean prediction error and the variance in the prediction error were also calculated across all iterations. Mean prediction error provides information on how well all models performed, while the variance gives insight to model instability.
RESULTS AND DISCUSSION

GAMs and GLMs: Variable selection and response curves

Regarding the different response shapes between soil variables and the percentage of root length colonized by AM fungi, GAMs parameterized for root length mycorrhizal suggested skewed quadratic responses to soil pH, NH$_4$-N, NO$_3$-N and microbial biomass carbon; cubic responses to soil K, soil P and EC; and linear responses to elevation, total carbon, C/N ratio, clay and silt as well as the land use history (pasture). These results show that there is a tendency of a strong relationship between some soil variables and the root length colonized by AM fungi, especially for the linear response shapes. For example, elevation and total carbon showed negative linear relationships with percentage root length colonized by AM fungi. This indicates that the infectivity of AM fungi varies with elevation and total carbon distribution in the soil. However, percent clay and silt showed positive linear relationship to infectivity. Based on the distribution of soil types on the farm, sandy soils were more common at higher elevation, and this corresponded with lower infectivity. The different response shapes between soil variables and the total mycorrhizal root length, GAMs parameterized for total mycorrhizal root length suggested skewed quadratic responses to NH$_4$-N, soil pH, EC, C/N ratio and microbial biomass carbon; cubic response to elevation; and linear responses to NO$_3$-N, total carbon, soil P and soil K. In general, the strength of the relationships between percentage root length colonized by AM fungi and soil properties on this farm showed that some soil properties had little influence on the infectivity of AM fungi. In general, the GAM models demonstrated that total mycorrhizal root length was not influenced by some of the soil properties, including NH$_4$, NO$_3$, P, pH, K and EC. This result contradicts that from greenhouse studies which have found strong explanatory relationships between P and the infectivity of AM fungi (e.g. Abbott and Robson 1982; Bolan 1991; Marschner and Dell 1994; Ryan et al. 2004; Thomson et al. 1992; Vejsadova et al. 1989). However, the past studies have focused on a small range of P at low levels (0-60 mg/kg) compared to those found at the Wickepin farm (11-500 mg/kg). The shape of the response curve within the range of previous studies for P and %RLC is in agreement with these findings, displaying a reasonably tight fit around these levels of P with decreasing AM fungi infectivity with increasing P. In addition, there was a reasonably strong negative relationship between elevation and percentage root length. This relationship may be explained by higher soil moisture levels at lower elevations as well as a deeper soil profile. Percentage and total mycorrhizal root length also increased with increase in total N but decreased with increase in C.

The summary of the shape of response functions for each soil variable and land use history for %RLC is presented in Table 1. Among the response variables, elevation, microbial biomass carbon and total C showed significant correlations (p<0.05) with %RLC for the GAM model. This was also similar to %RLC of the GLM model. The significant relationships for %RLC, for example with elevation, were generally linear, although some of the weak relationships were skewed quadratic (e.g. pH).

<table>
<thead>
<tr>
<th>Response function</th>
<th>Elevation</th>
<th>Land use</th>
<th>pH</th>
<th>EC</th>
<th>Total C</th>
<th>P</th>
<th>K</th>
<th>MBC</th>
</tr>
</thead>
<tbody>
<tr>
<td>%RLC (n=291)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GAMs</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linear</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Piecewise linear</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Skewed quadratic</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Bimodal</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>No relationship</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Significant</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Note: 1 = Presence, 0 = Absence

The deviance from the GAM and GLM models obtained by stepwise analysis is summarized in Table 2. Both the GAM and GLM models had similar values of residual deviance and the null deviance. However, the residual deviance of GAM models for %RLC were lower than the GLM models, showing that the GAM models had slightly better model fits. The GAM models explained 14% of the variation in %RLC while the GLM models only explained 6.5%. These outcomes show that although others have found strong relationships with soil properties and mycorrhizal colonization, these relationships may not be evident in field data.

The final models from the stepwise selection method and the model evaluation are shown in Table 3. The GAM model for %RLC had the best AUC value (0.61) while all models had the same kappa (pk) value (0.50). The explanatory variables selected for the final GAM and GLM models were very similar with only minor differences. The most frequently selected explanatory variables were elevation, land use, total carbon, soil P, soil K, microbial biomass carbon and soil pH. However, the weakness of the relationship with P, as discussed earlier, along with other evident but weak relationships (e.g. elevation) is demonstrated in the weak predictive power of all four models.

Results based on the 10-fold cross-validation are presented in Table 4. The mean prediction error for %RLC differed for GAM and GLM. However, mean prediction error of the GLM models was lower than that of the GAM model. This may indicate that although GAMs had a better fit, across all iterations they may be over fitted and unable to generalize to unknown data. However, the AUC values for the GAM model for %RLC indicate that the final GAM model selected was reasonably good at predicting unknown data.

In general, the regression models of GAM and GLM developed in this study provide insight to the nature of the relationship between AM fungi and environmental factors in the field. However, they were not good predictors of AM fungi infectivity. These prediction models mainly depend on environmental variables such as soil properties and land...
use features. As some soil properties had low correlation with the infectivity of AM fungi, the GAM and GLM models were influenced by these relationships. This is supported by some studies which investigated how the variation in response variable would influence the outcome of the model (Pearce and Ferrier 2000; Pearce et al. 2001).

In the GAM response curve, there was a linear relationship between the infectivity of AM fungi and environmental and soil properties such as elevation, total carbon, C/N ratio, soil P, EC, soil K and NO₃. Among these predictor variables, soil P is only the major nutrient that was expected to directly influence AM fungi status in the soil.

It is generally accepted that the beneficial effects of AM fungi decrease as the supply of P increases (Abbott and Robson 1984). Previous studies have found that very high and very low levels of soil P reduced AM fungi colonization (e.g. Koide and Li 1990). However, these studies have investigated a smaller range of P, with much lower values than found at this site. There is some evidence in the GAM response curves that the relationship found by others between %RLC and P may be identified within the known range.

Generalized additive modeling is a powerful tool to facilitate choice of a possible response shape without having to assume any particular relationships between the dependent and independent variables (Shipley and Hunt 1996). However, most of the relationships were found to be linear. The linear response curves of some soil properties were not followed with the significant correlation with the infectivity of AM fungi. Elevation, total carbon and soil microbial biomass carbon showed significantly correlations (p<0.05) with %RLC in the GAM and GLM models. These results indicated that these three factors were better predictors of the infectivity of AM fungi in the soil than other soil properties.

Elevation was an important factor in this field study because it predicted the infectivity of AM fungi better than other soil properties. This result might be correlated with the soil moisture depth of the soil profile, as at low elevations the infectivity of AM fungi was higher. Different elevation can cause a different moisture content of soils, because soil water has a characteristic of potential gravity. That is mean the water tend to move to the lower part or low elevation. Consequently, the moisture content of the upper elevation is more dry than on the lower elevation. There have been no other agricultural studies which have identified the influence of elevation on the status of AM fungi in the soil. However, in tropical rainforest, Meyer (1973) hypothesized that lower elevation of forest trees are predominantly by AM fungi, while at higher elevations mostly found the ectomycorrhizal. As noted by Read (2002), tropical grasslands are predominantly by AM fungi, while in the deserts areas are dominated by AM fungi plants with occasional ectomycorrhizal trees. In addition, in more temperate regions, at low elevations, grasslands are also dominated by AM fungi with some individual of ectomycorrhizal trees. These broader ecological studies may have little relevance to the local landscape at the farm at Wickepin however, where low agricultural plant diversity occurs.

The final GAM model was chosen from the stepwise-selected models based on their performance assessed through cross-validation. More than a half of the soil

---

**Table 2. Summary of GAM and GLM final models of the infectivity of AM fungi from soil characteristics and land use history.**

<table>
<thead>
<tr>
<th>Infectivity of AM Fungi</th>
<th>GAM</th>
<th>GLM</th>
</tr>
</thead>
<tbody>
<tr>
<td>%RLC</td>
<td></td>
<td></td>
</tr>
<tr>
<td>n of observation</td>
<td>291</td>
<td>289.16</td>
</tr>
<tr>
<td>Null deviance</td>
<td>284.64</td>
<td>255</td>
</tr>
<tr>
<td>Residual deviance</td>
<td>270.26</td>
<td>272</td>
</tr>
<tr>
<td>Residual degree of freedom</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>n of variables</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 3. The final models and their model evaluation parameters**

<table>
<thead>
<tr>
<th>Infectivity of AM Fungi</th>
<th>Model</th>
<th>Terms</th>
<th>AUC</th>
<th>pk</th>
</tr>
</thead>
<tbody>
<tr>
<td>%RLC</td>
<td>GAM</td>
<td>Land use + s (elevation)+s (NO₃)+s (TotalC)+s (P)+s (K)+s (MBC)</td>
<td>0.61</td>
<td>0.50</td>
</tr>
<tr>
<td></td>
<td>GLM</td>
<td>Elevation + Total-C + C/N + P + MBC + I (totalC^2)+I (C/N^2)</td>
<td>0.55</td>
<td>0.50</td>
</tr>
</tbody>
</table>

**Table 4. Cross validation results of GAM and GLM models of the infectivity of AM fungi from soil characteristics and land use history (K=10 replicates; SQRT= Square root)**

<table>
<thead>
<tr>
<th>Infectivity of AM fungi</th>
<th>n of observation</th>
<th>Mean prediction error</th>
<th>SQRT variance prediction error</th>
<th>Mean log prediction error</th>
<th>SQRT log variance prediction error</th>
<th>Mean prediction error</th>
<th>Mean log prediction error</th>
<th>SQRT log variance prediction error</th>
</tr>
</thead>
<tbody>
<tr>
<td>%RLC</td>
<td>291</td>
<td>1593.35</td>
<td>19.03</td>
<td>1.69</td>
<td>0.02</td>
<td>211.88</td>
<td>16.40</td>
<td>0.80</td>
</tr>
</tbody>
</table>
properties were able to predict the %RLC. However, the ROC analysis, which was used to assess the expected performance of the GAM model, was rather low for %RLC as the AUC value was lower than 0.7%. Similar results were also shown for the GLM model with a low AUC value.

The infectivity of AM fungi has previously been found to depend partly on soil characteristics and the environment. This study, however, has shown either a lack of relationship (e.g. with P) or a contradictory relationship (e.g. with C) to those found in glasshouse experiments. There may be various reasons for this for the marked differences between this field study and that of previous glasshouse studies. The outcomes in this study may be due primarily to driving factors which have not been considered or included in these models. Field experiments do not allow for control of a number of factors, including unknown environmental and historical (including climatic and weather) influences. Certain findings (such as the relationship with P) may have been affected by the inclusion of fertilizer in field samples. However, it is also possible that clear relationships in glasshouse studies may not be directly applicable to the field, where competing influences may interact and impact both the strength and nature of relationships.

This study suggests that some soil properties and land use variables could be used as predictor variables to predict the infectivity of AM fungi. Further investigation is required to develop models with good predictive power, including the selection of explanatory variables. However, this study has demonstrated that predictive modeling using standard statistical regression procedures can be applied to studies of AM fungi.

CONCLUSION

This study found that soil properties (i.e. total carbon, phosphorus, potassium, nitrate, soil pH, and microbial biomass carbon) and land use tended to have weak linear relationships with %RLC on a sheep-cropping farm at Wickepin in Western Australia. The relationship with elevation was again supported in this analysis. Further study is necessary into the potential direct drivers of infectivity in paddocks, for which elevation was a weak surrogate. This study demonstrated that a generalized additive model and a generalized linear model are useful tools to study the nature and strength of the relationships between environmental and land use variables and the infectivity of AM fungi.

ACKNOWLEDGMENT

The authors would like to thank Australian Government through provision of an Australian Development Scholarship, ADS-AusAID for the Postgraduate Scholarship, and Dr Bervin Turlach (The University of Western Australia) for the training in use of the GAM and GLM models.

REFERENCES


