Investigating the ironwood tree (*Casuarina equisetifolia*) decline on Guam using applied multinomial modeling

Karl A. Schlub[1], Brian D. Marx[1], Zelalem Mersha[2], and Robert L. Schlub[3*]

[1]Department of Experimental Statistics, Louisiana State University, Baton Rouge, LA 70803

[2]University of Florida, Tropical Research and Education Centre, Homestead, FL 33031

[3]University of Guam, Cooperative Extension Service, Mangilao, GU 96923

*Corresponding Author: rlschlub@uguam.uog.edu

ABSTRACT

The ironwood tree (*Casuarina equisetifolia*), a protector of coastlines of the sub-tropical and tropical Western Pacific, is in decline on the island of Guam where aggressive data collection and efforts to mitigate the problem are underway. For each sampled tree, the level of decline was measured on an ordinal scale consisting of five categories ranging from healthy to near dead. Several predictors were also measured including tree diameter, fire damage, typhoon damage, presence or absence of termites, presence or absence of basidiocarps, and various geographical or cultural factors. The five decline response levels can be viewed as categories of a multinomial distribution, where the multinomial probability profile depends on the levels of these various predictors. Such data structure is well-suited to a proportional odds model, thereby leading to odds ratios, involving cumulative probabilities which can be estimated and summarized using information from the predictor coefficient. Various modeling techniques were applied to address data set issues: reduced logistic models, spatial relationships of residuals using latitude and longitude coordinates, and correlation structure induced by the fact that trees were sampled in clusters at various sites. Among our findings, factors related to ironwood decline include basidiocarps, termites, and level of human management.

DATA COLLECTION

Data was provided from a 2009 University of Guam study on ironwood tree decline. There were 15 variables measured:

Response variable

Dec_sev (decline severity) is an ordered categorical variable from 0 to 4. If an observed tree is in perfect health, then it is recorded in the data as having dec_sev = 0 (the less healthy the tree, the higher the decline severity number.)

A series of ironwood tree images demonstrating the visual representation of the decline severity index

Dieback is a binary variable. If an observed tree was assigned a dec_sev greater or equal to 1, then the dieback value for that observation is equal to 1, otherwise the observation has a dieback value of 0.

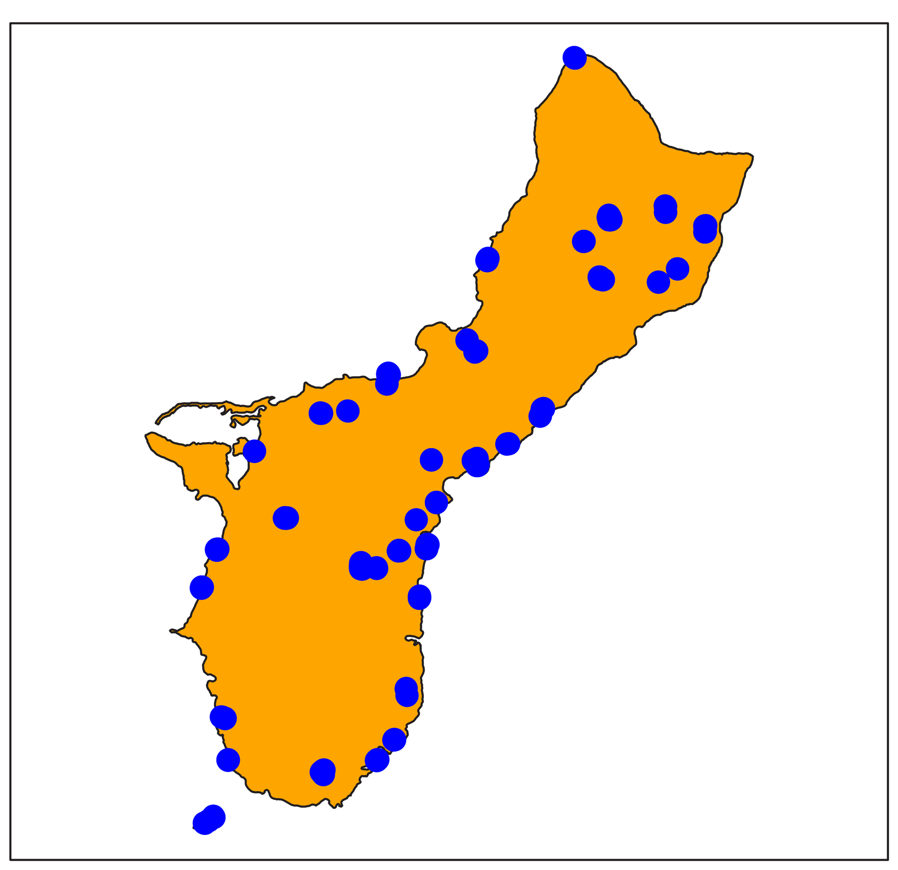
Explanatory variables

Structure variables: **num_stem** represents the number of tree main stems. **CBH** stands for circumference at breast height, 1.5 m above ground. **Density** is defined as the number of ironwood trees per square meter at a surveyed site.

Stress variables: **fire**, **conk**, **typhoons**, and **termites** are all binary categorical variables. A tree displaying any signs of stress was assigned a 1. Trees with no stress were assigned a 0.

Geographic variables: **lat** (latitude), **long** (longitude) and **altitude** were measured for each sampled tree using a GPS device. **Site** is an arbitrary number assigned for a location where the tree was sampled. If 30 sampled trees were located on one site, then all would be given the same site number. In total, there were 44 sites.

Miscellaneous variables: **human_mgmt** (0, 1, 2) assesses the level at which a sampled tree was managed [no management (0), intensive management (2)]. **Planted_natural** is a categorical variable and refers to tree establishment: planted (p) or natural (n).



A map of the island of Guam, blue dots indicate locations where ironwood trees were sampled.

STATISTICAL MODELS

Logistic Regression

Logistic regression is a form of analysis in which a binary response of 0's and 1's is fitted by a set of continuous and/or categorical explanatory variables. The model determines the probability that an experimental unit belongs to a group. In the context of the ironwood tree data, the probability of a tree experiencing dieback is generally unknown but can be estimated by a set of explanatory variables. The basic form of a logistic regression is:

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \alpha + \beta x$$

The formula on the right hand side resembles a standard linear regression setting, where alpha is an intercept term, and beta is a vector of coefficients for a vector of explanatory variables x. The left side of the equation is a log odds ratio. An odds ratio is a ratio of the probability of a tree having a dieback value of 1 over the probability of the complement, a tree not experiencing dieback.

Interpreting individual explanatory variables, in the context of a logistic regression, is determining which variables have a significant effect on the odds ratio. If an individual parameter ßi is equal to zero, this implies that it has no effect on the odds ratio. Confidence intervals for each explanatory variable can be constructed accordingly: $\beta i \pm z\alpha/2 \bullet S.E(\hat{\beta} i)$

Multinomial Regression

Although a logistic model is adequate for explaining differences between unhealthy vs. healthy trees, it is not adequate for evaluating several levels of tree health. Therefore, a multinomial model is needed, specifically one that uses a cumulative logit link function. A model which uses a cumulative logit function is required because the decline severity index has a natural ordering. The basic function resembles the logistic regression function, except that the odds ratio is replaced with a cumulative odds ratio. Interpretation of parameter estimates follows within the context of a cumulative odds ratio; and as in a logistic model, if a parameter estimate is equal to zero, it implies that the parameter has no effect on the cumulative odds ratio. Confidence intervals for parameter estimates are constructed similar to the parameters from the logistic regression.

RESULTS

After fitting a logistic model with all explanatory variables, except for the longitude and latitude variables (to account for spatial effects), and applying several model selection algorithms, the final logistic model was reduced down to six variables (Table 1). Pearson residuals were interpolated along a two-dimensional grid using the longitude and latitude coordinates. Since the interpolated residuals fall in the range of -1 to 2.5, it suggests that the logistic model has an adequate fit (Fig. 1). The results of the cumulative logit model are found in Table 2. Interpolating the Pearson residuals for the cumulative logit model resulted in extreme values; thereby, suggesting that the cumulative model may not be adequate (Fig. 2).

Table 1. Logistic modeling results

	Estimate	Std. Error	z value	$Pr(>\|z\|)$
(Intercept)	-417.5241	225.1959	-1.854	0.063732
conk	3.3060	0.2775	11.912	< 0.00001
human_mgmt	0.7571	0.1193	6.345	< 0.00001
termites	0.7468	0.1748	4.271	0.00001
density	-1.1633	1.1056	-1.052	0.292702
lat	-4.7615	1.2461	-3.821	0.000133
long	3.3110	1.6534	2.003	0.045229

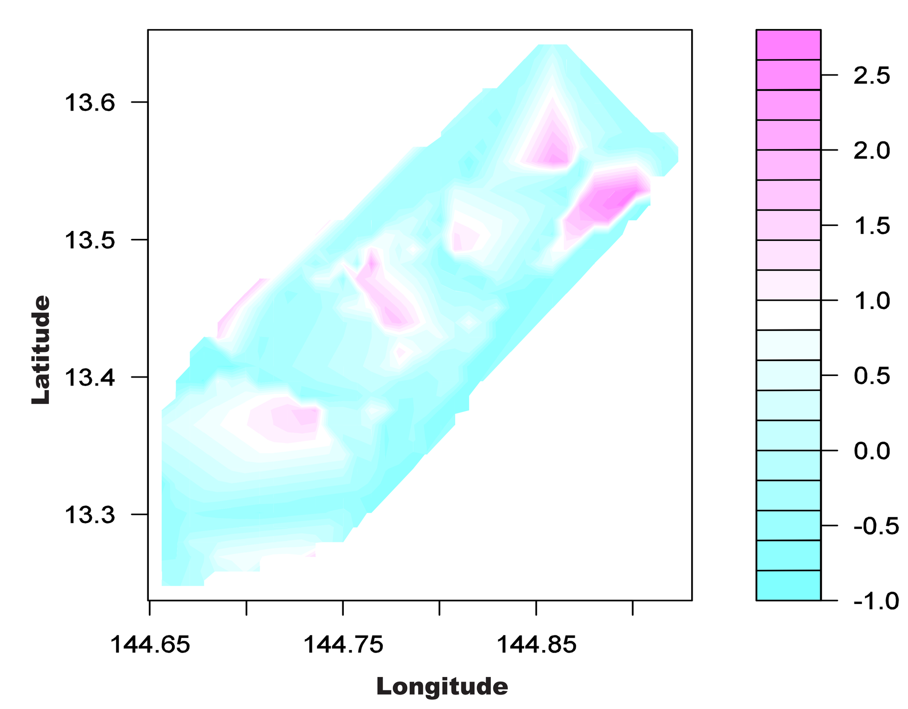


Figure 1. Interpolating Pearson residuals along a two-dimensional grid of the longitude and latitude coordinates.

Table 2. Cumulative logit model results

	Value	Std. Error	t value
(Intercept): 1	104.34581	200.95723	0.51924
(Intercept): 2	105.07729	200.95743	0.52288
(Intercept): 3	105.79070	200.95776	0.52643
(Intercept): 4	106.49698	200.95819	0.52995
density	-2.07296	0.82755	-2.50493
termites	-0.75484	0.15248	-4.95033
conk	-2.64156	0.17060	-15.48417
typhoon	-0.35112	0.17616	-1.99318
human_mgmt	-0.76839	0.10720	-7.16747
lat	3.78291	1.12303	3.36849
long	-1.05535	1.47362	-0.71616

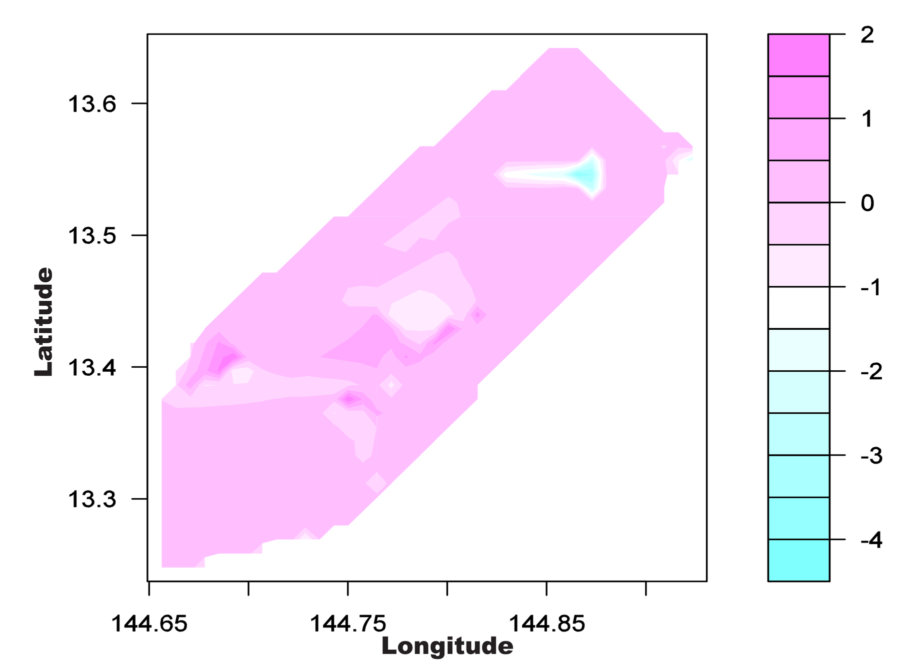


Figure 2. Interpolating Pearson residuals for the cumulative multinomial model.

CONCLUSIONS

The logistic model for the current data was the better model. From the results of the logistic model, the most significant variables that could explain the ironwood tree's state of health are the presence of basidiocarps, a high level of human management, or the presence of termites. The cumulative model, however, cannot explain trends in the data adequately. A reason could be that assumptions on proportional odds were not met. Possible ways to improve the cumulative model include adding more explanatory variables or moving away from the cumulative model to a multinomial model where the decline severity index is not considered ordinal.

ACKNOWLEDGEMENTS

Authors would like to acknowledge Dr. Bin Li of the Department of Experimental Statistics, Louisiana State University for suggestions and evaluation of model.