Journal of the American Mosquito Control Association, 13(1):28–34, 1997

USE OF GENERALIZED REGRESSION TREE MODELS TO CHARACTERIZE VEGETATION FAVORING *ANOPHELES ALBIMANUS* BREEDING

J. E. HERNANDEZ,[1] L. D. EPSTEIN,[2] M. H. RODRIGUEZ,[1] A. D. RODRIGUEZ,[1]
E. REJMANKOVA[3] AND D. R. ROBERTS[4]

ABSTRACT. We propose the use of generalized tree models (GTMs) to analyze data from entomological field studies. Generalized tree models can be used to characterize environments with different mosquito breeding capacity. A GTM simultaneously analyzes a set of predictor variables (e.g., vegetation coverage) in relation to a response variable (e.g., counts of *Anopheles albimanus* larvae), and how it varies with respect to a set of criterion variables (e.g., presence of predators). The algorithm produces a treelike graphical display with its root at the top and 2 branches stemming down from each node. At each node, conditions on the value of predictors partition the observations into subgroups (environments) in which the relation between response and criterion variables is most homogeneous.

INTRODUCTION

Data sets from field studies to characterize environments favoring mosquito breeding are complex. These data sets include variables such as vegetation type and percent coverage, presence of predators and other organisms, as well as other environmental conditions. The analysis of such complicated data sets is often a challenge to the analyst. This article introduces generalized tree models (GTMs) (Breiman et al. 1984, Ciampi 1991) to the study of mosquito ecology, a novel application of this statistical methodology. Generalized tree models are useful for exploring and analyzing relations among vegetation types and other variables that characterize environments favoring mosquito breeding. To illustrate, we use a GTM to characterize vegetation favoring breeding of *Anopheles albimanus* Wiedemann in the coastal plains of Chiapas, Mexico. *Anopheles albimanus* is a malaria vector that is commonly present along the coastal plains of Mexico, Central America, and northern South America (Faran 1980). Regression trees may prove to be very useful for planning and implementing control programs that focus on the larval stages of these vectors.

Tree-based modeling is a technique for uncovering structure in data, so-called because the primary method of displaying the fit is a binary tree. The tree uses vegetation types to stratify observations according to larval counts. The stratification allows the researcher to easily identify vegetation types associated with high larval counts.

Regression trees are constructed by recursive partitioning, a data analysis technique that recently has received much attention. After the work of Sonquist and Morgan (1964), who developed automatic interaction detection (AID), and Breiman et al. (1984), who developed classification and regression trees (CART), other authors have generalized these methods. In particular, Ciampi et al. (1987) proposed a framework for constructing regression trees with generalized linear models (GLM). The most common use of regression trees is to formulate and test hypotheses in the presence of complex interactions. We apply this methodology to characterizing the vegetation favoring breeding by *An. albimanus.*

Regression trees use 4 elements: a response variable (termed criterion), a set of variables to partition the data set (termed predictor variables), conditions on the predictor variables (termed split-defining conditions), and a homogeneity criterion to create the strata.

MATERIALS AND METHODS

With the support of the National Aeronautics and Space Administration (NASA), the Centro de Investigacion de Paludismo (CIP), Mexico, conducted a 1-year field study in a region of the coastal plains of the state of Chiapas, Mexico (Rodriguez et al. 1993). During this period, field teams from the Center collected chemical, ecological, and environmental data, as well as counts of *Anopheles* larvae in selected sampling units. Sampling units were inland bodies of water, also called habitats.

The study started with a description of the study area. Using satellite imagery, 5 ecological zones or vegetation units were identified: mangrove swamp, transitional swamp, riparian, pasture, and annual crop. These vegetation units have distinct ecological and environmental characteristics. Within the 5 vegetation units, 14 study locations were selected according to their accessibility and distance to nearby villages. Within each study location, larval habitats were described and classified according to their size (area), hydrological type, and vegetation.

[1] Centro de Investigación de Paludismo, CISEI, Instituto Nacional de Salud Pública, Apartado Postal 537, Tapachula, Chiapas 30700 México.

[2] Battelle Memorial Institute, 2115 East Jefferson Street, Suite 400, Rockville, MD 20852.

[3] Division of Environmental Studies, University of California Davis, Davis, CA 95616.

[4] Department of Preventive Medicine and Biometrics, Uniformed Services University of the Health Sciences, 4031 Jones Bridge Road, Bethesda, MD 20814.

Table 1. Vegetation variables used as predictors.

Variable name[1]	Most common genera/species[2]	Growth form
BROE	*Ammania coccinea* *Crinum erubences* *Egletes viscosa* *Pontederia sagittata* *Ludwigia octovalvis* *Cuphea cf. calophylla* *Verbesina* sp.	Broad-leaved emergent
BROH	*Salicornia bigelovii* *Phyloxerus vermicularis* *Batis maritima* *Portulaca oleracea*	Broad-leaved emergent halophytes
CYNO	*Cynodon dactylon*	Emergent perennial grass
CYPE	*Cyperus* sp. *Scirpus cf. cubensis*	Annual or perennial emergent graminoids (shorter than FIMB)
EICH	*Eichhornia crassipes*	Floating
FIMB	*Fimbristylis spadicea*	Emergent graminoid
HYME	*Hymenachne amplexicaulis* *Brachiaria mutica* *Panicum purpurascens* *Paspalum* sp.	Emergent grasses
JOUV	*Jouvea straminea*	Emergent perennial maritime grass
MUHL	*Muhlenbergia* sp. *Echinochloa colonum* *Pennisetum purpureum*	Tall emergent grasses
PHYT	Planktonic algae	Phytoplankton
PIST	*Pistia stratiotes*	Floating
RHIZ	*Rhizophora mangle* *Avicennia germinans* *Conocarpus erectus*	Mangrove trees
SALV	*Salvinia* sp. *Nimphaea cf. conardi*	Floating leaves
SUBM	*Ceratophyllum demersum*	Perennial submerged
TYPH	*Typha domingensis*	Emergent
DETR	Detritus	

[1] Name of the variable as entered in the models.

[2] Most common plant species grouped in each vegetation type.

A total of 86 different plant species were identified at the sites. These were classified into 15 groups according to morphological similarity (Table 1). One hundred forty habitats were identified and visited 4,288 times.

During each visit, percent vegetation coverage was recorded and larvae were sampled using a standard 500-ml dipper. Larvae were classified by age and larval counts were recorded by stage of development. Water conductivity, pH, and other physical–chemical data as well as the presence of predators such as fish and other insects was recorded for each sample (Table 2). Rodriguez et al. (1993) provide additional information about the NASA-CIP study and details regarding the data collecting and sampling scheme.

Table 2. Description of the predator criterion variables.

Variable name	Description
FISH	Indicator for presence of fish in the habitat
COLEOP	Indicator for presence of water beetles (Coleoptera)
HEMIP	Indicator for presence of water bugs (Hemiptera)
ODONAT	Indicator for presence of dragonfly nymphs (Odonata)

The development of software to fit the GTM required combining algorithms that construct tree structures with programs that fit GLMs. We used a statistical programming language called S-plus (Chambers and Hastie 1993). S-plus provides native functions for classification and regression trees. We built on these functions, thus reducing the programming effort.

The procedure to fit a regression tree model takes place in 2 stages. In the first stage, the procedure uses an algorithm to partition the data set into a number of groups (strata), each one of them as homogeneous as possible with respect to the variable being modeled (e.g., larval counts). The resulting tree is usually referred to as the "large tree." Large trees are often too elaborate to interpret and are likely to overfit the data. The second stage of the procedure, called pruning, uses an algorithm to cut off excessive branches.

For the sake of clarity, we explain regression tree methods in the context of the NASA-CIP study. The purpose of the analysis is to investigate how the expected larval counts depend on the percent coverage of vegetation groups (Table 1). The response variable is the count of 4th-stage larvae. The predictor variables are the percent coverage of the vegetation groups. The split-defining conditions our analyses use are statements about the ranges of the predictor variables (e.g., "the percent coverage of BROH is less than 25%," see Table 1 for abbreviations used to designate vegetation types). The algorithm uses deviance as a measure of homogeneity and it stratifies the data according to the split-defining conditions and the homogeneity criterion.

The partitioning algorithm starts with the entire data set, also called the root node (Breiman et al. 1984). The algorithm formulates split-defining conditions for each possible value of the predictor variables to create candidate splits. After reviewing all possible split-defining conditions, the algorithm selects the candidate split that maximizes the homogeneity criterion and uses it to partition the population into 2 subgroups, also called child nodes. By convention, the algorithm assigns the observations satisfying a condition to the child node on the left branch and the observations not satisfying it to the child node on the right branch. The algorithm proceeds recursively with each of the new nodes until no partition yields nodes with more than 10 observations, the minimum node size we specified as stopping rule. The stopping rule usually reported in the literature is to set the minimum node size to a small number such as 5 or even 2. We set this number to 10, however, to ensure that there were enough observations in each node to perform a regression, an important issue in the generalization presented below.

A terminal node is a node that the algorithm cannot partition further. It is also called a leaf. The leaves define the most homogeneous groups or strata. Therefore, an estimate of the expected response (such as the mean) at each node appropriately summarizes the data in that node. Reporting such estimates allows the investigator to examine how the response varies with the predictor variables (vegetation groups).

Often, there is additional covariate information about the response variable. These covariates may be confounders or effect modifiers and are often referred to as criterion variables. The relationship between the criterion variables and the response variable may vary by stratum. To incorporate criterion variables into the analyses, one combines tree-based methods with GLMs (see McCullag and Nelder 1989).

In linear regression, one models the expected value of the observations y_i as a linear function of parameters b_0, b_1, . . ., B_p, that is:

$$\mu_i = E(y_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}; \quad i = 1, \quad , n,$$

where $(x_i, \ldots, x_{ip})$ are p covariates for observation i, and n is the number of observations. Generalized linear models expand the traditional regression techniques to other distributions with the use of a link function $g(\cdot)$. This link function depends on the distribution one uses to model the data. In the application that concerns us, the observations are counts of larvae, which we model with the Poisson distribution. For the Poisson distribution, one expresses the log of the expected counts as a linear function of the parameters. Thus, if y_i denotes the larval count at habitat i and if $\mu_i = E(y_i)$, then

$$\log \mu_i = \beta_0 + \beta_1 x_i + \ldots + \beta_p x_{ip}; \quad i = 1, \quad , n.$$

Thus for the Poisson distribution, $g(\mu_i) = \log \mu_i$. In general, the link-transformed mean values are expressed as a linear function of the b_i, that is

$$g(\mu_i) = \beta_0 + \beta_1 x_i + \ldots + \beta_p x_{ip}; \quad i = 1, \quad , n.$$

In our analyses, the presences and/or absences of potential predators are the criterion variables (Table 2). The partitioning algorithm incorporates the effect of the criterion variables on the response variable. To construct a tree with criterion variables, the partitioning algorithm fits a regression model to the data in each node before partitioning them. Then, for each candidate split, the algorithm fits a pair of regressions, one for each candidate child node. The algorithm selects the candidate split that maximizes the homogeneity criterion. The GLM–tree-based method uses the deviance as measure of homogeneity. The deviance of a node is twice the difference between the maximum values of the log-likelihood under the saturated model and under the submodel. The algorithm calculates the difference between the deviance of the node being partitioned (parent node) and the sum of the deviance of the resulting candidate child nodes. This quantity is the homogeneity criterion. The best split is the one that maximizes the homogeneity criterion.

The partition algorithm can be summarized as follows:

```
— Start with the entire data set, the root node.
— With a GLM regress the larvae counts on the
  presence/absence of predators.
— With the first predictor variable:
    — Calculate its coverage range:
    — For increments of 0.5% of the range:
        — Partition the data set according to the
          value of the predictor.
        — For each of the resulting candidate child
          nodes regress with a GLM the larvae
          counts on the presence/absence of potential predators. Calculate the deviance
          of each child node.
        — Select the split that maximizes the
```

change in deviance (the homogeneity criterion).

- Repeat for each predictor variable.
  - Select the predictor variable that achieves the highest reduction in deviance.
- Apply this process to each child node until all nodes cannot be partitioned further (i.e., meet stopping rule).

In the second stage, the pruning algorithm creates a decreasing sequence of nested trees. That is, each new tree in the sequence is a subtree of the previous one. The first tree in the sequence is the large tree itself. The pruning algorithm successively snips off branches of the large tree until it reaches the root node. To snip off branches, the algorithm uses a modified version of the error–complexity measure (Breiman et al. 1984) as a selection rule. The error–complexity measure is a linear combination of the error cost of the tree and the complexity parameter. In the GLM–tree method the error cost is the sum of the deviance of all terminal nodes or leaves in a tree or subtree. The complexity parameter is a constant. If we think of the complexity parameter as the cost per terminal node, then the error–complexity measure is formed by adding the error cost of the tree and its complexity cost. A large value for the complexity parameter yields a small-sized subtree. The size of a tree is the number of terminal nodes it has. For each possible value of the complexity parameter there is a subtree that minimizes the error–complexity function. The pruning algorithm builds the sequence of nested trees by letting the complexity cost vary between zero and a number large enough so that the resulting pruned tree is the root node itself. For each value of the complexity cost, the pruning algorithm finds the subtree that minimizes the total deviance and includes it in the sequence.

Once the tree sequence is completed, one must select the right-sized tree. Selection of the right-sized tree is a process analogous to variable selection in regression analyses. For selecting the right-sized tree or selecting a variable in regression analysis, Akaike's information criterion (AIC) is useful (Akaike 1974). The AIC is designed for statistical model identification when there are several competing models, such as a sequence of nested tree models. The models and the maximum likelihood estimates of the parameters are used to compute the AIC,

$$\text{AIC} = -2(\text{maximum log likelihood}) + 2(\text{number of parameters independently adjusted}).$$

The AIC adapts easily to GTMs. For these models the AIC takes the form

$$\begin{aligned}\text{AIC} = -2\ [&(\text{deviance of the subtree})\\ &\div (\text{deviance of the large tree}\\ &\quad - \text{size of the large tree})]\\ &+ 2\ (\text{number of observations}\\ &\quad - \text{size of the subtree}).\end{aligned}$$

We used AIC to select the final tree model. This criterion prescribes that the right-sized tree is the one with minimum AIC. The final tree model shows how counts of larvae depend on the vegetation groups. In addition, the tree model estimates the effect of potential predators on the expected larval counts and describes how they may vary with the environments that leaves of the tree describe.

The GLM–tree algorithm uses the effect of the criterion variables to select and make the split. In this context, each leaf contributes more information than just the observed mean larval counts. One can use the observed mean larval counts as an adjusted measure to explore larval productivity and to identify environments associated with high mean counts of larvae. The proper way to examine the results of the GLM–tree-based model, however, is to take advantage of the additional information contained in the regression coefficients associated with each leaf. The regression coefficients estimate the effect of the criterion variables (potential predators) on the expected larval counts.

A dendrogram is used as a graphical representation of the tree structure. In this representation, we incorporate the observed mean larval counts for each leaf as part of the information available for the nodes. The distance from one node to its child nodes is proportional to the reduction in deviance. From the dendrogram, one can easily see the importance of each predictor as well as an adjusted measure of the observed mean larval count at each leaf.

RESULTS

We fit a GLM–tree model to the data from each of the 5 vegetation units. For the sake of brevity we only present here the results for the transitional swamp unit. The final tree for this unit (Fig. 1) displays in its leaves the observed mean larval counts. The observed means may ignore the effect of criterion variables and, therefore, they should be used as an exploratory tool only. An interpretation that uses the regression coefficients is more appropriate as it incorporates the effects of potential predators on the expected larvae counts (Table 3). The tree also displays a node identification number for relating the leaves to the data in Table 3.

The GTM for the transitional swamp unit has 11 leaves (Fig. 1). Visually, we distinguished 4 cluster subtrees (circled). The observations (habitats) with HYME $\geq$ 7.5% define one subtree. The observed mean larval counts of these observations is the highest. Leaves 6 and 7, a partition induced by PHYT, show observed mean larval counts of 9.6 (PHYT $<$ 10.5%) and 15.8 (PHYT $\geq$ 10.5%). In the subtree where HYME $<$ 7.5%, CYNO $<$ 1.5%, and CYPE $<$ 4.5%; leaves 16 (DETR $<$ 1.5%) and 34 (DETR $\geq$ 1.5% and FIMB $<$ 0.5%) have some of the lowest observed mean larval counts (0.64 and 0.66, respectively). In this branch, the observed

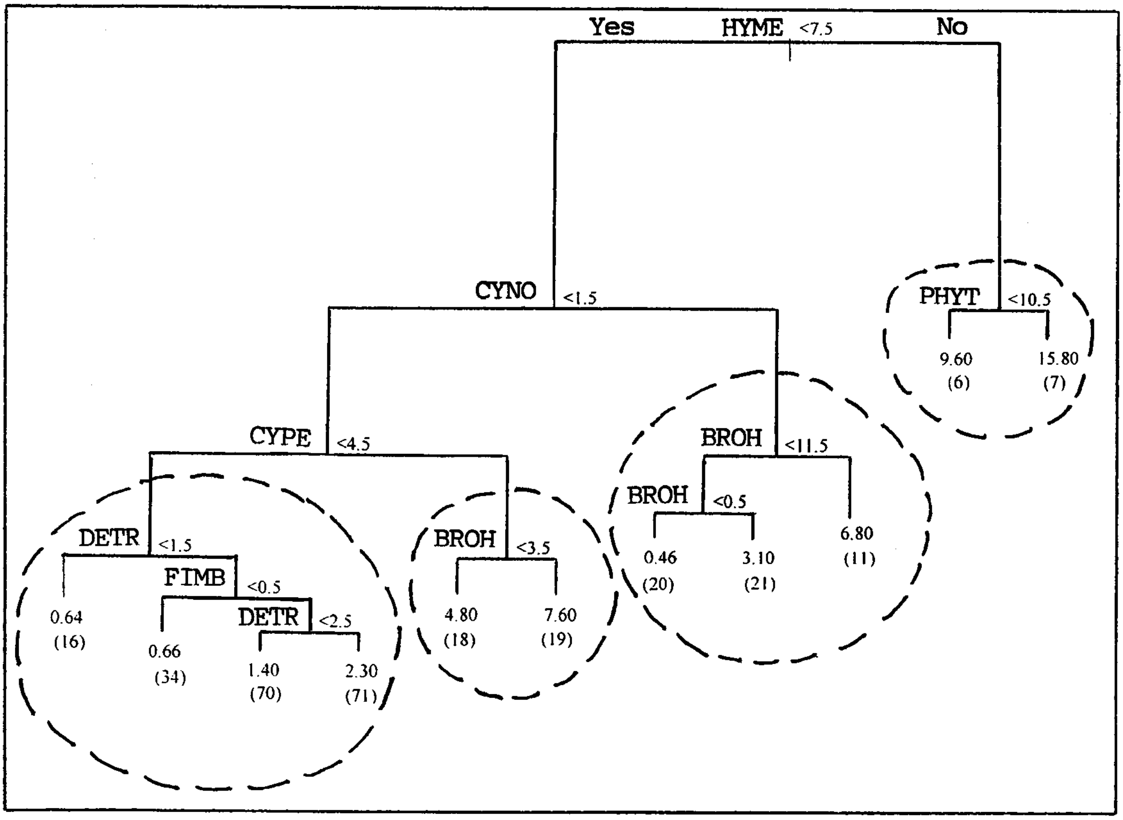

Fig. 1. Final tree model for the transitional swamp ecological zone. The model assigns the partition satisfying the condition to the child node on left and the partition not satisfying the condition to the child node on the right. The numbers below each leaf represent the observed mean larval counts and, enclosed in brackets, the leaf number.

means in leaves 70 (1.5% < DETR < 2.5% and FIMB ≥ 0.5%) and 71 (FIMB ≥ 0.5% and DETR ≥ 2.5%) are moderately low, 1.40 and 2.30, respectively. In the subtree where HYME < 7.5%, CYNO < 1.5%, and CYPE ≥ 4.5%, the observed mean larval counts are high. In leaf 18 (BROH < 3.5%), the observed mean larval count is 4.8 and in leaf 19 (BROH ≥ 3.5%) it is 7.6. In the subtree with observations where HYME < 7.5% and CYNO ≥ 1.5%, the leaves show that the observed means increase with increasing percent coverage of BROH. Leaves 20 (BROH < 0.5%), 21 (0.5% < BROH < 11.5), and 11 (BROH ≥ 11.5%) have observed mean larval counts of 0.46, 3.1, and 6.8, respectively.

Table 3 displays the estimates of regression coefficients for each leaf of the tree. The number under each coefficient is the ratio of the estimate to its estimated standard error. To assess the statistical significance of the coefficients one can compare these ratios to the quantiles of a normal distribution (Wald's test, McCullag and Nelder 1989). In 5 of the 11 leaves, none of the coefficients are statistically significant, whereas in the remaining 6, at least one coefficient estimate is significant. Coefficients of the criterion variables FISH, COLEOP, and HEMIP (predators, see Table 2 for definition) are negative in all leaves where they are significant, except in leaf 71. This indicates that, in this environment, FISH, COLEOP, and HEMIP have negative associations with the larval counts. The coefficient estimate for ODONAT is positive in every leaf where it is significant.

DISCUSSION

This discussion focuses on 2 main issues: the results of the NASA-CIP study and the use of generalized regression tree models.

According to the model depicted in Fig. 1, HYME is the group that most affects larval counts. One expects high densities of larvae in environments where the percent coverage of this vegetation group is equal to or greater than 7.5%. In addition, HYME interacts with PHYT. In fact, mean larval counts are higher when PHYT's coverage is equal to or greater than 10.5%. In these environments, the presence of fish or Coleoptera does not have a significant effect on the expected larval densities. In node 7 (HYME > 7.5, PHYT > 10.5), however, the presence of the order Hemiptera has a negative effect on the expected larval counts. In observa-

Table 3. Results of the regression tree for the transitional ecological zone.

Node ID number	Mean larval counts	*n*	Dispersion parameter	Regression coefficients			
				FISH	COLEOP	HEMIP	ODONAT
7	15.700	20	9.822	−0.260 (−0.570)[1]	0.338 (0.753)	−1.257 (−2.696)	1.001 (2.065)
6	9.580	24	9.709	−0.843 (−1.655)	−0.768 (−1.816)	−0.591 (−1.299)	−0.130 (−0.289)
11	6.760	34	10.440	−1.196 (−2.160)	0.640 (1.109)	−1.594 (−1.518)	−7.412 (−0.258)
21	3.080	24	6.116	−0.906 (−0.993)	9.594 (0.265)	0.136 (0.216)	−9.081 (−0.174)
20	0.462	26	1.595	−1.419 (−0.937)	−0.451 (−0.517)	1.571 (1.512)	0.949 (0.977)
19	7.620	21	3.646	0.501 (1.388)	−1.146 (−2.763)	−0.870 (−2.651)	−0.211 (−0.554)
18	4.850	20	6.727	−0.156 (−0.264)	1.327 (1.370)	1.077 (1.814)	0.227 (0.385)
71	2.320	157	4.923	0.805 (2.642)	0.097 (0.364)	1.059 (4.167)	0.754 (2.815)
70	1.410	44	3.407	0.369 (0.615)	−0.797 (−1.605)	−1.515 (−1.339)	−0.066 (−0.085)
34	0.661	109	2.796	−0.134 (−0.302)	0.171 (0.342)	0.259 (0.594)	0.437 (0.883)
16	0.636	176	4.626	−0.225 (−0.534)	−1.087 (−2.583)	0.326 (0.703)	1.854 (4.162)

[1] Numbers in parentheses represent the ratio of the estimated coefficient to its estimated standard error (Wald's test). Compare this quantity to the quantiles of a standard normal distribution to assess the statistical significance of the estimated coefficients. At the 0.05 alpha level and 2-sided hypothesis test, values greater than 1.96 or smaller than −1.96 are statistically significant.

tions where these water bugs are present, the expected counts are 80% less than in observations where they are absent.

CYNO and CYPE are the 2nd- and 3rd-most influential vegetation groups and their presence is associated with high larval counts. CYNO and CYPE interact with each other and both interact with HYME, BROH, FIMB, and DETR. Where the coverage of HYME is less than 7.5% and the coverage of CYNO is greater than 1.5%, increasing coverage of BROH is associated with increasing abundance of larvae (leaves 20, 21, and 11). In these environments, FISH is the only criterion variable with a significant effect on the expected larval counts, showing a negative effect in environments where coverage of BROH is greater than 11.5%.

If the coverages of HYME, CYNO, and CYPE are less than 7.5%, 1.5%, and 4.5%, respectively, we are likely to observe low larval counts, especially when the coverages DETR or FIMB are less than 1.5% and 0.5%, respectively. Slightly higher larval counts are likely when FIMB's coverage is equal to or greater than 0.5%.

Three potential predators (fish, Coleoptera, and Hemiptera) have a significant positive association with the larval counts in leaf 71 (HYME < 7.5, CYNO < 1.5, CYPE < 4.5, FIMB > 0.5, DETR > 2.5). Although this may contradict the hypothesis that these organisms are predators of *An. albimanus,* it probably means that floating detritus protects larvae so that the predator–prey interaction is minimized in environments described by this leaf. When statistically significant, organisms of the order Odonata have a positive association with larval counts, suggesting that these organisms are not high-impact predators and that *An. albimanus* habitats are equally favorable to the presence of Odonata naiads as they are for *An. albimanus* larvae or that Odonata are present only in high mosquito producing habitats.

Generalized tree models have several advantages over more traditional methods of analysis previously applied to the characterization of mosquito breeding sites. Traditional methods include analysis of variance (ANOVA) and linear regression. First, GTMs are easier to interpret. The tree describes the strata in terms of direct field variables. The use of GLMs to build trees increases the flexibility for using more appropriate distributions for the data (e.g., the Poisson distribution), which leads to superior inferences. Second, to simplify analyses, researchers often split the range of covariates to form dichotomous versions. Their choices for the split points do not follow any principles. This is not the case with GTMs, where the splits follow defined homogeneity criteria. Efficiency is an added advantage of GTMs, as the selection of the split points is automatic. Third, one of the most attractive properties of GTMs is their automatic detection and display of interactions among variables in the model. Traditional methods that attempt to use interactions quickly become unmanageable, especially if the number of variables in the model is large. Consider, for instance, a model with 15 variables such as the one in the Results section. To explore the main effects and all the possible interaction terms with a traditional ANOVA-type parameterization, the model would have 2^{15}-1 terms[5] and would require at least 10 times more observations. Such a model is not practical. Fourth, tree models are useful for summarizing complex data sets. Variables that do not contribute predictive information do not appear in the final tree structure, even if they are in the model. Hence, variables that have predictive power are separated by the GTM from variables that have no predictive power. Consider, for instance, the tree model in Fig. 1. Although all 16 vegetation groups in Table 1 are in this model, only HYME, CYNO, CYPE, BROH, DETR, and FIMB appear in the final tree structure.

Finally, generalized tree models add extra flexibility that allows adjusting for confounders or effect modifiers. These are entered into the model as criterion variables; they may be continuous or categorical. This adjustment takes place during the process of partitioning the observations. The only restriction is that the number of observations in the leaves must be sufficiently high, so that the algorithm can perform the regressions to adjust for them. The generalized tree structure shows the dependence of the response variable on the predictors while adjusting for the criterion variables. The coefficient estimates may also be used to address other scientific questions about the response variable.

At this time, software to fit GTMs is not readily available. There are some packages that already have modules for classification and regression trees such as CART (Breiman et al. 1984). S-plus (Chambers and Hastie 1993) also provides functions to grow and interact with tree models. Neither of these allow for the use of criterion variables or for the specification of the distribution of the data via a GLM. We used the S-plus programming language to develop the programs for our analyses. Fitting GTMs is computer intensive. For instance, to fit the model of Fig. 1 it took 4 h on a PC 486 running at 50 MHz and with 8 Mb of RAM. Work is now under way to make the program as efficient as possible. Copies of the programs are available upon request.

Although Akaike's criterion for model selection is a useful tool for determining the right-sized tree (Akaike 1974), this criterion is the subject of some controversy (Venables and Ripley 1994). Other techniques for model selection include cross-validation and bootstrap methods (Breiman et al. 1984, Chambers and Hastie 1993, Venables and Ripley 1994).

ACKNOWLEGMENTS

The Nasa-CIP project was funded partly by the NASA DI-Mod Project, a multidisciplinary effort supported by the following collaborating organizations: NASA Ames Research Center, Uniformed Services University of the Health Sciences, University of California (Davis), California State University (Fresno), University of Texas (Houston, El Paso), Stanford University, and Centro de Invesigacion de Paludismo (Mexico). This research project was funded by Centro de Investigacion de Paludismo and the Tropical Disease Research and Training Program (WHO/TDR).

REFERENCES CITED

Akaike, H. 1974. A new look at statistical model identification. IEEE Trans. Automatic Control 19:716–723.

Breiman, L., J. H. Friedman, R. Olshen and C. J. Stone. 1984. Classification and regression trees. Chapman and Hall, New York.

Chambers, J. M. and T. J. Hastie (editors). 1993. Statistical models. Chapman and Hall, New York.

Ciampi, A. 1991. Generalized regression trees. Computat. Stat. Data Anal. 12:57–78.

Ciampi, A., C.-H. Chang, S. Hogg and S. McKinney. 1987. Recursive partitioning: a versatile method for exploratory data analysis in biostatistics. D. Redeil Publishing, New York.

Faran, M. E. 1980. Mosquito studies (Diptera, Culicidae) XXXIV. A revision of the Albimanus Section of the subgenus *Nyssorhynchus* of *Anopheles*. Contrib. Am. Entomol. Inst. (Ann Arbor) 15:1–215.

McCullag, P. and J. A. Nelder. 1989. Generalized linear models, 2nd ed. Chapman and Hall, New York.

Rodriguez, A. D., M. H. Rodriguez, R. A. Meza, J. E. Hernandez, E. Rejmancova, D. Savage, R. Roberts , O. Pope and L. Legters. 1993. Dynamics of population densities and vegetation association of *Anopheles albimanus* larvae in a coastal area of southern Chiapas, Mexico. J. Am. Mosq. Control Assoc. 9:46–58.

Sonquist, J. A. and J. N. Morgan. 1964. The detection of interaction effects. Institute for Social Research, University of Michigan, Ann Arbor.

Venables, W. N. and B. D. Ripley. 1994. Modern applied statistics with S-plus. Springer-Verlag, New York.

[5] A 2-way ANOVA model in which the factors have only 2 levels needs 3 terms: one for each factor and one for the interaction, that is, 2^2-1. A 3-way ANOVA in which the factors have only 2 levels would use 7 terms: one for each factor, 3 for the 2-way interactions, and 1 for the 3-way interactions, that is, 2^3-1. A model comparable to the GTM presented in the results section would be a 15-way ANOVA and it would need as many as 2^{15}-1 terms.