Ecology 190 | Lab 3: Land cover change modeling
Dahl Winters
November 15, 2005

1. **Write a 1/2 page description of the objectives of the lab and methods used to develop the predicted deforestation map.**

The objectives of the lab were to explore deforestation patterns in Central America between 1990 and 2000, to generate a CART model in Splus to explain the observed deforestation patterns, and finaly to use the CART model to generate a predicted deforestation map to see how well this model reproduces actual deforestation patterns. To do this, we first obtained GIS data provided by Dalia Conde of elevation, protected areas, villages, 1990 and 2000 deforestation, and roads in 2000, then mapped them in ArcInfo for an explorative look at how deforestation patterns compared to natural and anthropogenic features (elevation, villages, roads). However, for a more quantitative understanding of what factors predict deforestation, we needed to develop a CART model. We used ArcInfo to sample environmental data for 5000 random points within deforested areas and 5000 points over the entire area. This data was saved, then explored graphically in Splus to see how elevation, distance to road, distance to village, and protected area status varied between deforested sites and the overall landscape. We then ran a CART model script, specifying various numbers of nodes between 6 and 9 to see which number gave the least misclassification rate (and thus the best model for the observed deforestation patterns). Finally, we ran the DOCELL script to generate a map of predicted deforestation for ArcInfo using our CART model, and we looked at the residuals to see how different the predicted and actual deforestation maps were.

2. **We calculated distance from village using Euclidean distance. Why is this most likely the wrong distance measure to use (think about how people get from a village to the forest)? Look up definition of Euclidean distance to help you out. (1/4 page).**

Euclidean distance is the straight-line distance between two points. However, roads between villages and the forest likely wind around existing villages and other landforms (hills, lakes), and are not actually constructed in straight lines. Euclidean distance also doesn't account for elevation differences between villages and the forest. A better distance measurement would be the total path length.

3. **Why did we choose to compare random points of the landscape to deforested points (1/4 page)?**

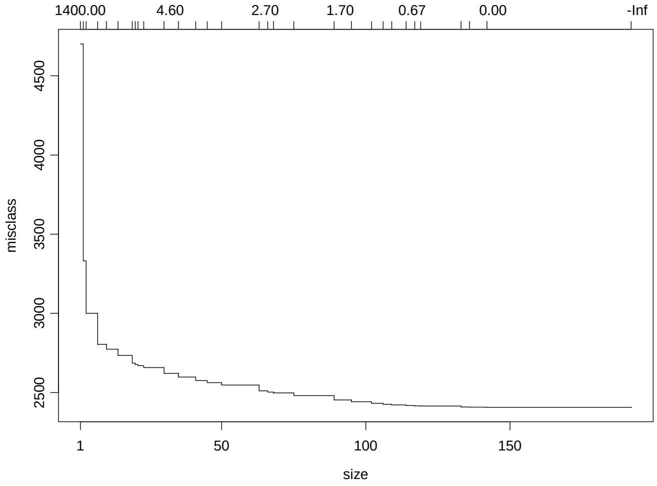
We did this to see patterns in the number of sites affected by a particular variable for both deforested areas and the overall landscape (i.e. to see which variables were important for predicting deforested areas).

- For elevation, we found that deforestation decreases with higher elevation in both deforested areas and the entire landscape, but that more deforested areas exist between 100-200 m elevation than would be expected from looking at the overall landscape. So elevation within that range is a good predictor variable.

- For distance to roads, all deforestation takes place within 10 km of a road, while landscape points extend up to 45 km from roads. So distance to roads < 10 km is another good predictor variable.
- There was a similar finding for distance to villages—all deforestation occurred within 10 km of villages, while there were random landscape points that extended up to 50 km from villages.
- Finally, for protected areas, there were more random landscape points in protected areas, and about an equal number of both unprotected and buffer areas. However, for deforested areas, the fewest number of points were found in protected areas, more in buffers, and the most in unprotected areas. This shows that much less deforestation takes place in protected areas compared to elsewhere in the landscape, which we would expect.

4a. What is the optimal number of nodes (leaves) based on the misclassification vs. node plot (1/4 page)?

By changing the number of nodes in the script, we found that the highest number of nodes possible in the tree are 192, and even then there are 2407 misclassifications out of 9545 (a 25% misclassification rate). However, if we just look at the range of 6-9 nodes we were asked to look at, 8 is the optimal number of nodes.



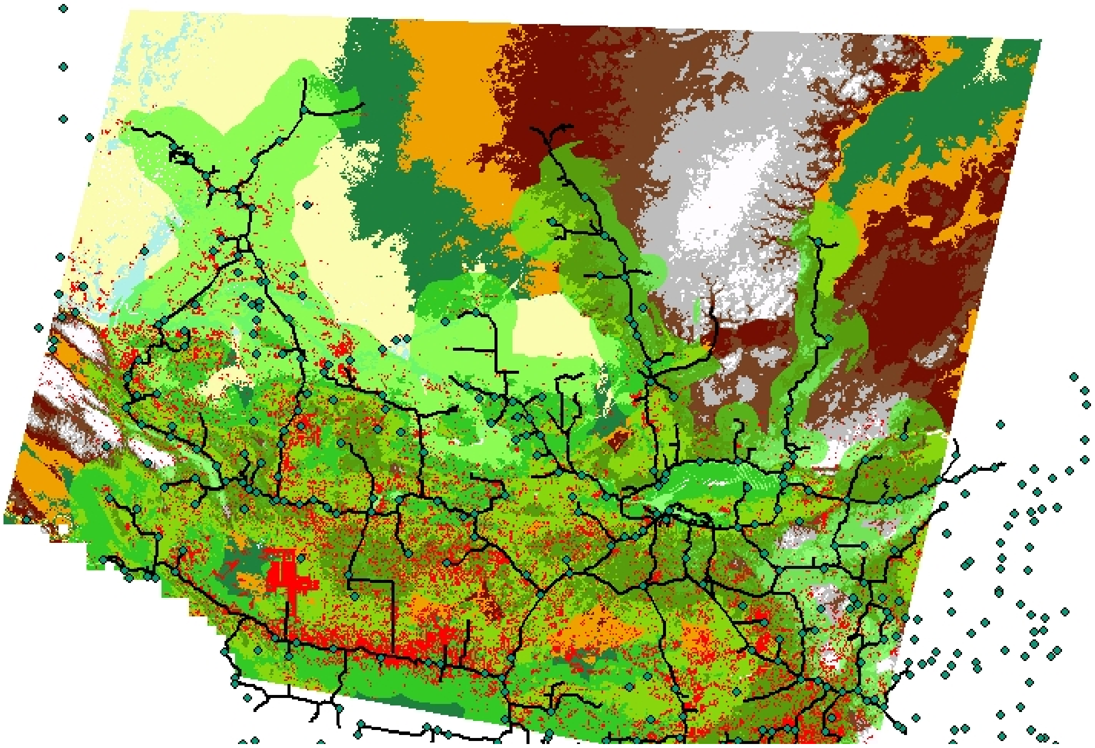
```
Numbers of nodes I tried, along with
misclassification rates:
6 or 7 nodes - 2805/9545 (unable to return 7 nodes)
8 or 9 nodes - 2774/9545 (unable to return 9 nodes)
50 nodes - 2548/9545
192 nodes (maximum tree size) - 2407/9545
```

4b. Do you agree with the predictor variables used in the model, should there be different ones used or less terms used (see line in Splus script that starts with deforest.tree to view model)? (1/4 pg)

I think the predictor variables used (`elevation + road.2000.d + village.dist + def.1990`) were good because they yield only a 25% misclassification rate for our model (meaning they account for 75% of the observed deforestation). However, 25% is still a bit high. This suggests we could add more variables to increase our model’s accuracy (see explanation in part d below).

4c. Look at the pred_defor map and compare to the observed deforestation. How is our model doing at predicting deforestation?

Looking at predictions.mxd, the def_2000 layer shows the actual deforestation in 2000. We find that the actual deforestation in this layer is much less in extent than the areas in the pred_defor layer, which came from our model. It looks like our model is overpredicting deforestation in many areas, especially in the northern half of the map. Interestingly, the predicted deforestation areas in the northern part of the map visually seemed to correspond more with being within a certain distance from roads (the black lines on the map), and less to distance to villages (the dots on the map), even though distance to villages was the first split in the CART model. Additionally, there were some areas to the west in the low-elevation region (colored yellow) that were deforested, but was not predicted by the model. The underprediction by our model was smaller in extent than the areas it overpredicted. A likely explanation for the model's overprediction and underprediction is its 25% error rate; there are probably more variables that affect deforestation that were not included in our model.



4d. What would improve the model best, adding new/different variables or pruning the tree differently? Think about what the misclassification plot is telling you (1/4 pg).

The misclassification plot tells us that pruning to leave fewer nodes greatly increases the misclassification rate (shown as the exponential increase on the left of the graph in question 4a), so we wouldn't want to do that. However, pruning the tree less to leave more nodes would only minimize the misclassification rate to at most 2407 misclassifications out of 9545 (the 25% error rate). Regardless of how the tree is pruned, we will not go below 25%. Thus, it looks like we need to add new variables to further refine our model. We are probably leaving out other important factors that, if included, would lower our misclassification rate below 25%.