Analysis of Edge Effect Transect Data Using Mixed-Effects Models.

Alan Swanson

Masters Writing Project

Dept. of Statistics, Montana State University

May 5, 2002

Introduction

Edge effects are the interactions between 2 adjacent, though abruptly separated, ecosystems (Murcia, 1995). In the Pacific Northwest the harvesting of trees has created a fragmented landscape, and there is interest in studying the edge effects that occur within oldgrowth forest adjacent to clearcuts.

The oldgrowth forest ecosystem, having evolved over millions of years, is very efficient at cycling nutrients. When a tree dies or needles fall to the ground, a diverse assemblage of microorganisms speeds decomposition and utilizes nearly all of the nutrients. Thus nutrients remain in the system and are eventually cycled back into living trees. This assemblage includes worms and mites, many species of bacteria, and mycorrhizal fungi which tap symbiotically into the roots of trees and funnel micronutrients into the trees' root system in exchange for photosynthetically fixed carbon. In an oldgrowth forest the air is cool and moist even on hot summer days, but in a clearcut, with no canopy of trees for insulation, temperatures soar and moisture levels drop. On clear nights, at any time of year, temperatures can drop radically in a clearcut due to radiation loss. It is thought that this variability in microclimate, coupled with the loss of living tree roots, harms the communities of

microorganisms that inhabit the soil, which, in turn, impacts the cycling of nutrients and causes nitrogen to be lost from the system.

Scientists interested in the ramifications of global climate change would like to predict how Northwest forests will respond to long-term changes in temperature. One possibility is that trees will grow faster and become a sink for carbon on a global scale. But nitrogen is a limiting nutrient in forest growth, so it is of interest to understand how much nitrogen is lost in a clearcut, and through which pathways this loss occurs. On a regional scale, due to the fragmentation of the forest, the edge environment may also contribute significantly to the loss of nitrogen. This paper describes the analysis of data from a study that is attempting to quantify the microclimate variability, it's impact on the below-ground community, and subsequent nitrogen loss present in the edge environment.

Researchers in ecology at UC Berkely are studying the impact of forest-clearcut edges on biogeochemical processes which affect nitrogen retention within fragmented oldgrowth Douglas-fir forests of the Pacific Northwest. The study site is at the Wind River Experimental Forest in southern Washington, where 9 transects have been established at edges separating oldgrowth forest from 9 separate clearcuts. Three of the edges are south facing, meaning that the edge is approximately east/west and the clearcut is south of the oldgrowth. Three additional edges are north facing, two face west, and the final edge is east facing. Within each transect, measurements have been made at 12 stations; one in the clearcut 120 meters from the edge, one at the edge, and the rest within the oldgrowth at 15, 30, 45, 60, 75, 90, 120, 150, 180 and 240 meters from the edge. The transects range in elevation from 500-700 meters, and are all within a 2 mile radius. Measurements have been collected on more than 100 variables in order to study edge effects on temperature and moisture by season, forest structure, litterfall by season, and nutrient availability in soil. The ultimate goal of the study is to model how edges affect nutrient cycling and forest health,

but the immediate goal is to study edge effects exhibited in the individual variables. Quantitatively, the major questions to be addressed are:

1. Is there an edge effect?
2. Does it change with the orientation of the edge?
3. How deep into the forest does it extend?

Previous edge effect studies have addressed the first two questions using a simple two-factor analysis of variance (ANOVA) model, and have addressed the last question using multiple comparisons. These studies have ignored the grouping of measurements into transects, treating each measurement as an independent observation. Ignoring this grouping of observations (into transects) doesn't necessarily affect the fit of a model, but it can lead to biased estimates of residual variance, and this can cause significance tests to be inaccurate (Neter et al, 1996). In addition to fitting a regular two-factor ANOVA model with distance and edge orientation (aspect) as factors, the following extensions to the ANOVA model will be described and compared to results from the two-factor ANOVA model:

1. Three-factor ANOVA with factors for distance, aspect and transect within aspect. This allows a separate mean for each transect.
2. Mixed effects model, in which the separate mean for each transect is treated as a random effect. This allows for broader inference.
3. Polynomial regression in which distance is treated as a continuous variable. This gives higher power and greater error degrees of freedom.

It will be shown that these extensions provide a better fit to the data and more reasonably estimate the significance of distance, aspect, and the interaction between aspect and distance.

Snow Cover Data

One of the variables measured is percent snow cover , and it will be the primary example for demonstrating the different models described above. Because snow cover was measured as the percentage of ground covered, it was transformed using the arcsin square root function, the recommended variance-stabilizing transformation for percent data (Neter et al, 1996). A plot of the untransformed data is shown in Figure 1. Note that the transects are grouped by aspect, with top row of plots showing edges with north aspects, the middle row shows south aspects, and the bottom row shows east/west aspects. Snow cover insulates the ground and stabilizes temperatures in the soil. Warmer daytime temperatures in the edge environment may cause reduced accumulations of snow. This lack of groundcover insulation may have a negative impact on microbial populations in the soil. Data from the clearcut station was excluded from this analysis because it is not relevant to examining the edge effect within the oldgrowth. In addition, it often showed problems with non-normality and heterogeneity of variance, and its inclusion caused problems in more complex models.

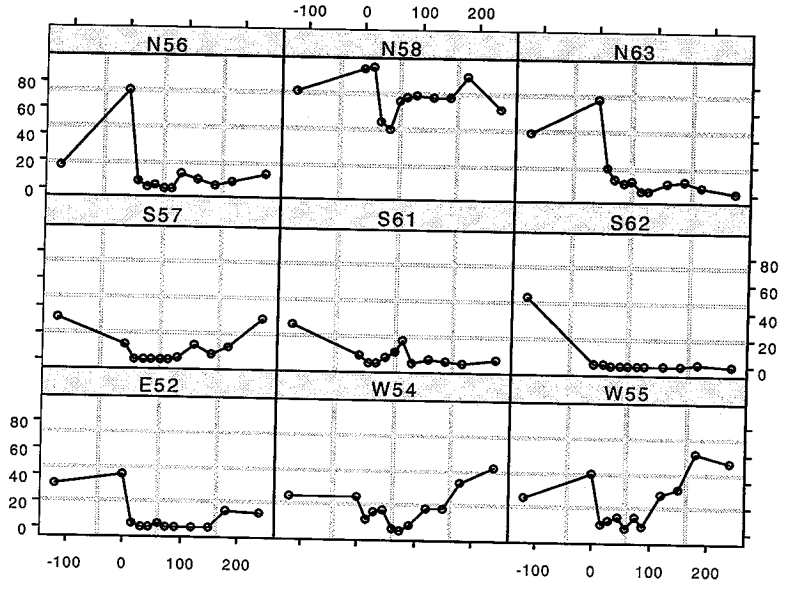


Figure 1. *Plot of percent snow cover by transect. Note that that the transects are grouped by aspect, with top row of plots showing edges with north aspects, the middle row shows south aspects, and the bottom row shows east/west aspects.*

Methods of Analysis

Software

The software package R (http://cran.r-project.org/) was used for all computations, and results were verified using SAS (http://www.sas.com/).

Two-factor ANOVA model

A two-factor ANOVA model was fit to the transformed data. This model fits a separate profile by aspect. For the *kth* observation taken at distance *j* from a transect with aspect *i*, the model is:

$$y_{ijk} = \mu_{...} + \theta_i + \tau_k + \gamma_{ik} + \varepsilon_{ijk}, \; i=1...3, j=1...3, \; k=1...11.$$

Where:

$\mu...$ is the overall mean.

θ_i is the effect of aspect *i*,

τ_k is the effect of distance *k*,

γ_{ik} is the interaction between distance and aspect,

and $\varepsilon_{ijk} \sim N(0, \sigma^2)$.

Such a model can be fit in R with the following code:

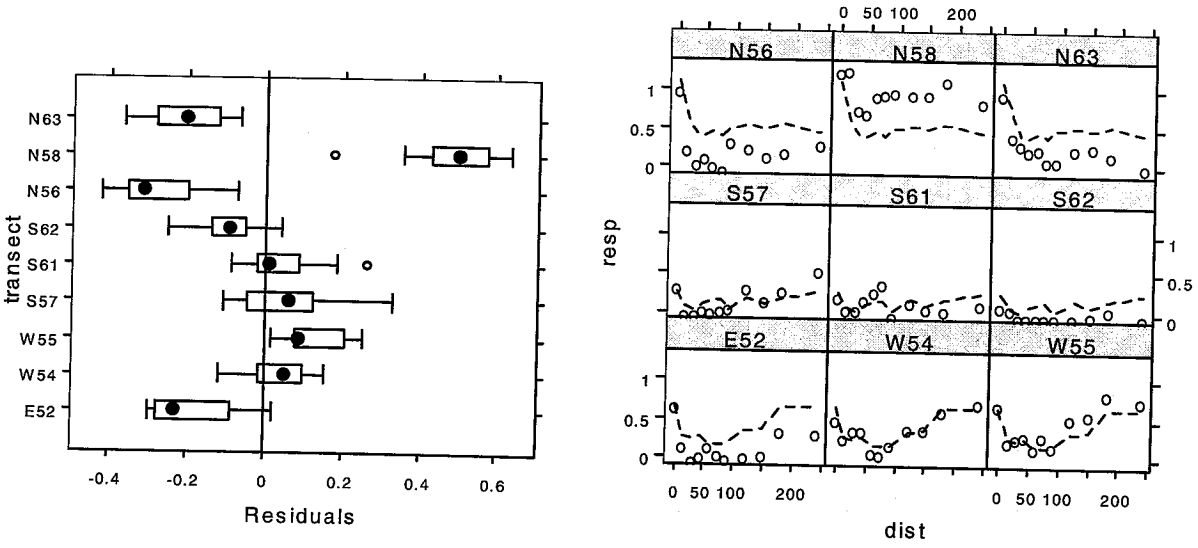
```
> anova.model <- lm(asin(sqrt(snow.cov/100)) ~ asp*factor(dist), data=snow)
```

Boxplots of residuals by transect (figure 2) show a serious problem with lack of fit. Within each aspect, the replicate transects differ in mean snow cover (the transect effect). In this simple model this variability associated with the transect effect is being absorbed by the residuals. This inflates residual variance and results in a misleading ANOVA table (below) in which the interaction between distance and aspect appears to be insignificant.

```
> anova(anova.model)
Analysis of Variance Table

Response: asin(sqrt(snow.cov/100))
                  Df Sum Sq Mean Sq F value    Pr(>F)
asp                2 2.6503  1.3252 15.5532 2.923e-06 ***
factor(dist)      10 1.6428  0.1643  1.9282   0.05659 .
asp:factor(dist) 20 0.9654  0.0483  0.5666   0.92182
Residuals         66 5.6233  0.0852
```

Figure 3 further illustrates the lack of fit, showing the data by transect with the fitted profile as a dashed line.



Figures 2 and 3. Figure 2 (left) shows boxplots of residuals from two-factor ANOVA model by transect. Figure 3 (right) shows the transformed data (circles) with fitted values as a dashed line.

Three-factor ANOVA model

One way this lack of fit can be corrected for is by fitting a separate mean for each transect. For the *kth* observation taken at distance *j* from a transect with aspect *i*, the model is:

$$y_{ijk} = \mu_{...} + \theta_i + \rho_{(i)j} + \tau_k + \gamma_{ik} + \varepsilon_{ijk}, \; i=1...3, j=1...3, \; k=1...11.$$

Where:

$\mu...$ is the overall mean.

θ_i is the effect of aspect *i*,

$\rho_{(i)j}$ is the effect of transect j nested within aspect i.

τ_k is the effect of distance *k*,

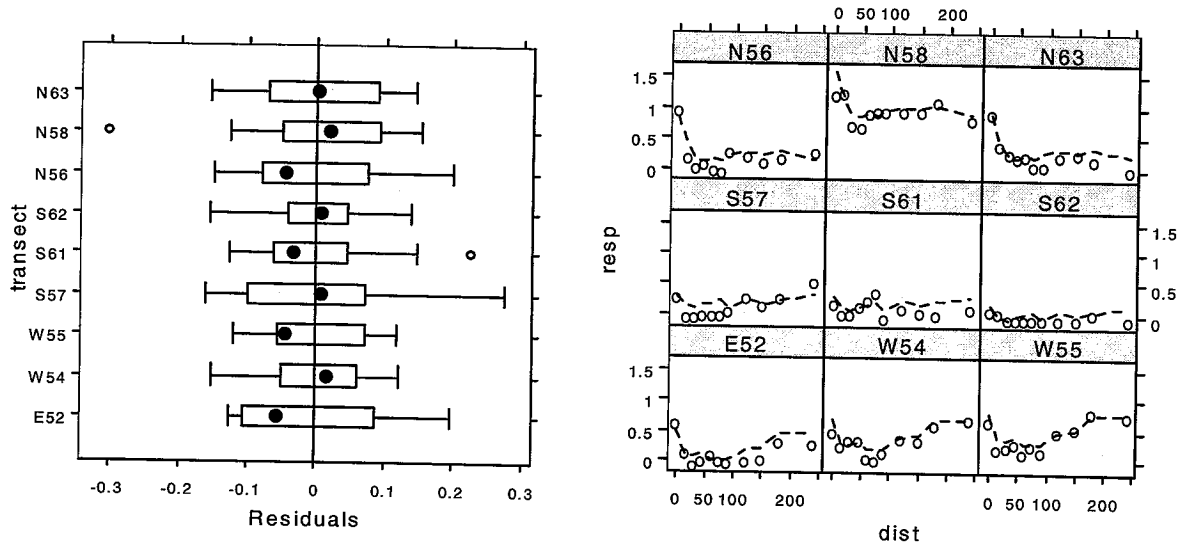
γ_{ik} is the interaction between distance and aspect,

and $\varepsilon_{ijk} \sim N(0, \sigma^2)$.

Such a model can be fit in R with the following code:

```
> anova.model2 <- lm(asin(sqrt(snow.cov/100)) ~ replicate %in% asp +
asp*factor(dist), data=snow)
```

Boxplots of residuals by transect from this model (figure 4) show that there is no longer a lack of fit problem, and a plot showing the data with fitted values (figure 5) confirms this.



Figures 4 and 5. Figure 4 (left) shows boxplots of residuals from three-factor ANOVA model by transect. Figure 5 (right) shows the transformed data (circles) with fitted values as a dashed line.

Diagnostic plots of this model are shown in figure 6. The plot of residuals versus fitted values (and the "scale-location" plot) show that the assumptions of constant variance has not been violated. The normal q-q plot shows that the residuals roughly follow a normal distribution. One outlier is indicated by this plot: observation 56, from the 0m station of transect N58. Cook's distance is a measure of the change in fit (summed over all observations), when an observation is removed from the model. This provides a good indication influential observations. The plot of Cook's distance shows that observation 56 is influential, as are observations 55 (240m station of transect S57) and 72 (75m station of transect S61). The typical cutoff value for Cook's distance is 0.2 so none of these observations were removed from the model.

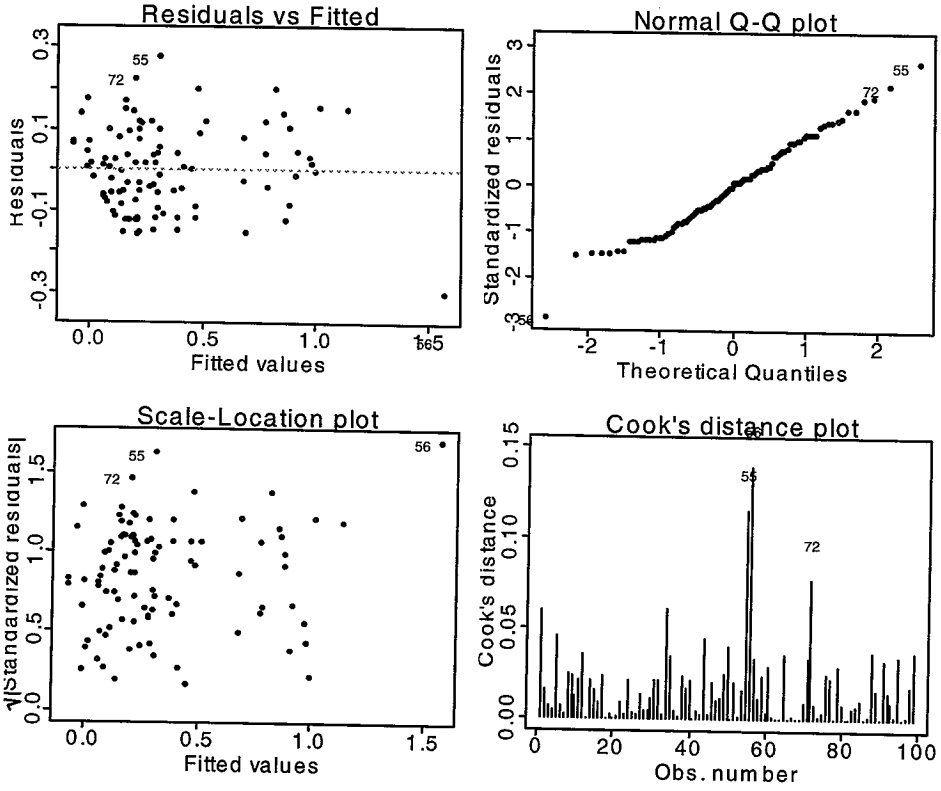


figure 6. Diagnostic plots from the 3-factor ANOVA model.

The models described thus far assume uncorrelated residual error. This assumption can be examined with an autocorrelation function (ACF) plot. An ACF plot shows the average correlation of residuals within the same transect as a function of their separation. Separation is measured in terms of lags, so such a plot is intended for equally spaced data. The transect data is not equally spaced so ACF plots are technically not appropriate, but they do give a reasonable picture of the correlation present in the data. An ACF plot of the residuals from the two-factor ANOVA is shown in figure 7. This plot shows some negative correlation at 5, 6 and 9 lags, but this is more indicative of lack of fit than autocorrelation.

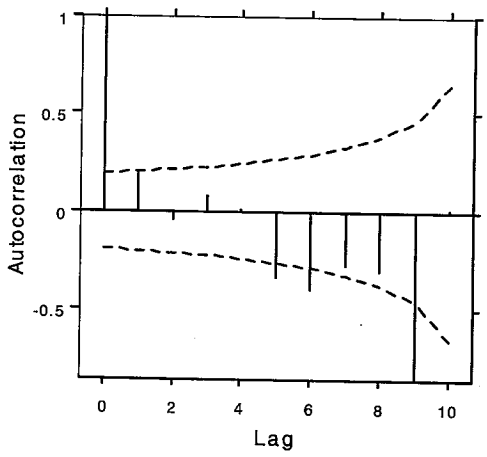


Figure 7. Autocorrelation function (ACF) plot of residuals from the two-factor ANOVA model. The line shows an alpha=.05 confidence region for the null hypothesis of no autocorrelation.

This model gives quite a different ANOVA table (below). It shows that the difference in means between transects with the same aspect is highly significant ($p<2.2e-16$), as is the distance:aspect interaction. In addition, the estimated residual standard deviation ($\hat{\sigma}$) drops from 0.2919 to 0.1346 with this model.

```
> anova(anova.model2);
Analysis of Variance Table

Response: asin(sqrt(snow.cov/100))
                 Df Sum Sq Mean Sq F value    Pr(>F)
asp               2 2.6503  1.3252 73.1853 < 2.2e-16 ***
factor(dist)     10 1.6428  0.1643  9.0730 6.619e-09 ***
replicate:asp     6 4.5369  0.7561 41.7602 < 2.2e-16 ***
asp:factor(dist) 20 0.9654  0.0483  2.6659  0.001799 **
Residuals        60 1.0864  0.0181
```

Mixed-effects models

The three-factor ANOVA model provides an adequate fit to the data, but is limited in that it only models the specific sample of transects that were measured. If the transects measured can be assumed to be a random sample from some larger population, we can model the transect effect as a *random effect.* That is, we can model it as random variation around some population mean. This allows us to make inference about the population of forest edges our sample was drawn from, rather than just those edges measured. In this case, we are not so concerned with making inference about the transect means, so a mixed-effects model is not neccesary. If there were no distance:aspect interaction, we might be interested in the effect of aspect on mean snow cover, but because of the significant interaction, this effect is of no interest. For the *kth* observation taken at distance *j* from a transect with aspect *i*, the model is:

$$y_{ijk} = \mu_{...} + \theta_i + b_{(i)j} + \tau_k + \gamma_{ik} + \varepsilon_{ijk}, \; i=1...3, j=1...3, \; k=1...11.$$

Where:

$\mu...$ is the overall mean.

θ_i is the effect of aspect *i*,

$b_{(i)j} \sim N(0, \sigma_b^2)$ is the random effect of transect *j* with aspect *i*,

τ_k is the effect of distance *k*,

γ_{ik} is the interaction between distance and aspect,

and $\varepsilon_{ijk} \sim N(0, \sigma^2)$.

This is called a mixed-effects model because it includes both fixed effects ($\mu, \theta, \tau, \gamma$), which are parameters associated with the entire population, and random effects (b), which are associated with the individual transects measured. Such a model can be fit in R with the following code:

```
> mixed.int <- lme(asin(sqrt(snow.cov/100)) ~ asp*factor(dist), random=~1|tran,
data=snow)
```

The command `summary(mixed.int)` gives the following output:

```
> summary(mixed.int)
```

```
Linear mixed-effects model fit by REML
 Data: snow
       AIC      BIC    logLik
  51.18885 127.8268 9.405575

Random effects:
 Formula: ~1 | tran
        (Intercept)  Residual
StdDev:   0.2590255 0.1345615
...
```

This shows that there is quite a bit of variability ($\ddot{\sigma}_b$ =0.259) in the mean response between transects of the same aspect.

As shown in a plot of fits versus distance by transect (figure 8), this model gives a fit virtually identical to that of the three-factor ANOVA model. The difference is that now we are not estimating the individual transect effects (although the individual transect effects are *predicted* by this model), but rather the variability of the transect effect (σ_b^2) is estimated. This reduces the number of parameters in the model and allows for broader inference.

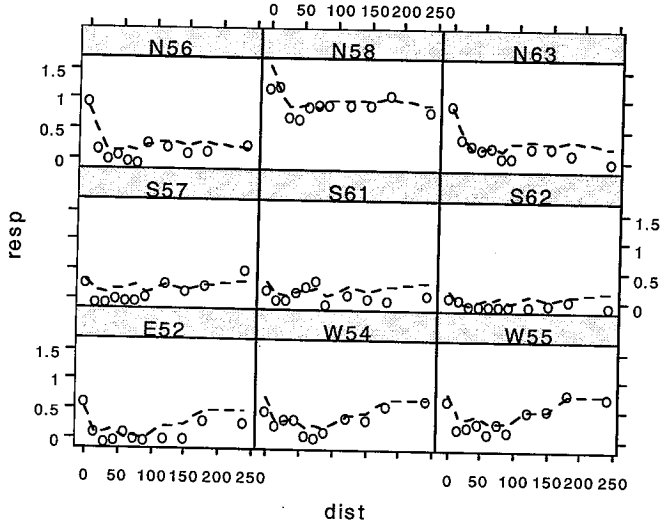


figure 8. Fit of mixed-effects ANOVA model by transect.

An ANOVA table from this model gives F-statistics for distance and distance:aspect which are identical to those from the three-factor ANOVA model.

```
> anova(mixed.int)
                  numDF denDF   F-value p-value
(Intercept)           1    60 15.140309  0.0003
asp                   2     6  1.752513  0.2515
factor(dist)         10    60  9.073008  <.0001
asp:factor(dist)     20    60  2.665940  0.0018
```

A plot of residuals versus distance by transect (figure 9) suggests that there remains some lack of fit with the mixed ANOVA model. In particular this plot shows a trend in the residuals for many of the transects. Transects N63, S57 and E52 stand out in this respect. By adding a random slope to our mixed-effects model this lack-of-fit problem can be resolved.

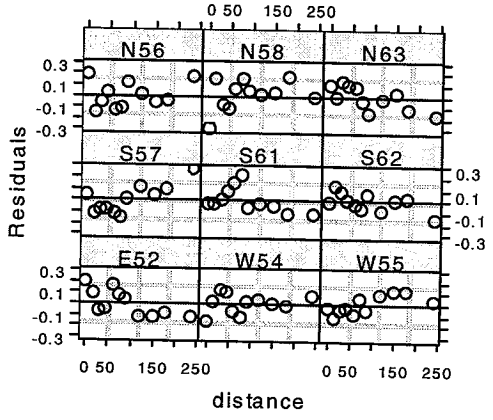


figure 9. Residuals versus distance by transect for the ANOVA model with random intercept.

A mixed-effect ANOVA model with random intercept and slope effects by transect can be fit in R with the following code:

```
> mixed.dist <- lme(asin(sqrt(snow.cov/100)) ~ asp*factor(dist),
random=~dist|tran, data=snow)
```

Models with different random effects can be compared using either the Akaike information criterion (AIC) or the Bayesian information criterion (BIC). AIC and BIC are both based on the log-

likelihood (logLik) of a model (a measure of fit), and both add a penalty for additional terms, because adding additional terms always results in a better fit. AIC and BIC differ in that BIC gives a higher penalty to additional terms. The model that gives the lower AIC or BIC is favored. The formulae for AIC and BIC are:

$$AIC = -2 \cdot logLik(model) + 2p, \text{ and } BIC = -2 \cdot logLik(model) + log(n) \cdot p$$

The mixed-model with random intercept gave an AIC value of 51.1 and a BIC value of 127.8, while the model with random slope and intercept gave values of 43.1 for AIC and 122.0 for BIC, so both criteria favor the random slope model. A normal q-q plot of residuals from this model (figure 10) shows that observation 56 is less of an outlier in this model.

For this (random slope and intercept mixed-ANOVA) model the issue of scope of inference is a concern. The slope is changing between replicate transects of the same aspect, so in order to make inference about forest edges other than those measured we must model the slope effect as random.

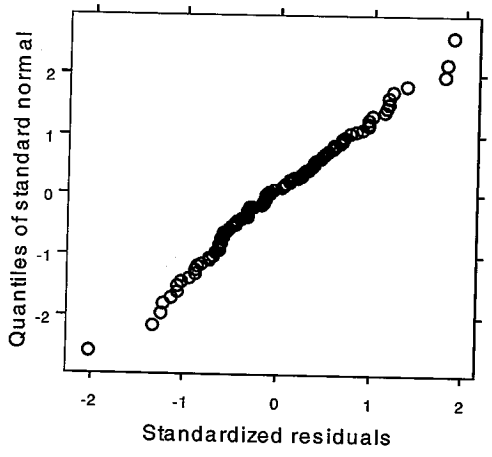


figure 10. Normal probability plot of residuals from random slope model.

An ANOVA table from this model gives smaller p-values for all effects. This is because the residual standard error (the denominator for the F-statistics) has dropped from 0.135 to 0.116 with the addition of the random slope.

```
> anova(mixed.dist)
                  numDF denDF   F-value p-value
(Intercept)           1    60 17.561937  0.0001
asp                   2     6  3.708520  0.0894
factor(dist)         10    60 12.058182  <.0001
asp:factor(dist)     20    60  2.248428  0.0083
```

A plot of residuals versus distance by transect (figure 11) shows no trend in residuals.

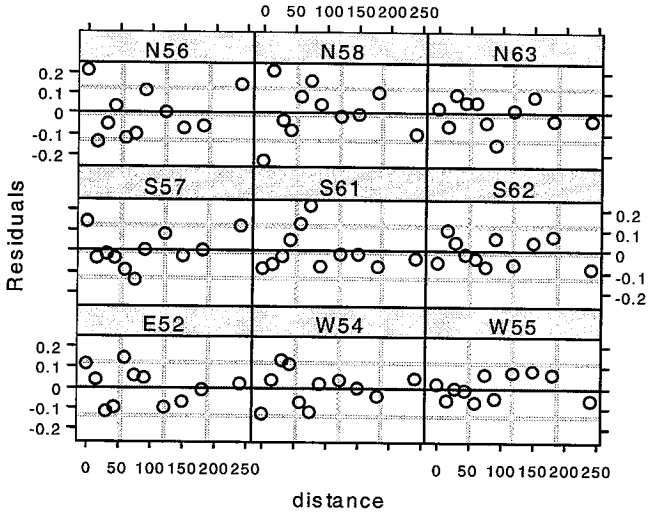


figure 11. Residuals versus distance by transect for the ANOVA model with random slope and intercept.

Polynomial Regression mixed-effects model.

The mixed effects model with random slope and intercept fits the data very well, but if we treat distance as a continuous variable rather than as a factor, the data can be fit nearly as well with fewer parameters. This is accomplished by modeling the distance effect, within each aspect, as an *nth* order polynomial. For the *kth* observation taken at distance *j* from a transect with aspect *i*, the model is:

$$y_{ijk} = \mu_{...} + \theta_i + b_{(i)j} + (\beta_{i1} + c_{(i)j}) \cdot dist_k + \beta_{i2} \cdot dist^2 + ... + \beta_{in} \cdot dist^n + \varepsilon_{ijk}, \; i=1...3, j=1...3, k=1...11.$$

Where:

$\mu...$ is the overall mean.

θ_i is the effect of aspect *i*,

$b_{(i)j} \sim N(0, \sigma_b^2)$ is the random intercept of transect *j* with aspect *i*,

$c_{(i)j} \sim N(0, \sigma_c^2)$ is the random slope of transect *j* with aspect *i*,

$\beta_{i1} ... \beta_{in}$ are the coefficients for an *nth* order polynomial describing the effect of distance for aspect *i*,

and $\varepsilon_{ijk} \sim N(0, \sigma^2)$.

Models with different fixed effects can also be compared with AIC and BIC, so long as the models are fit using the method of maximum likelihood (ML). This method of fitting gives biased estimates of variance so the restricted maximum likelihood (REML) method, which gives unbiased estimates of variance, is preferred for most situations. A series of polynomial models was fit with orders ranging from 1 to 10 (a 10^{th} order polynomial is equivalent to treating distance as a factor). In this case AIC favored a 5^{th} and BIC a 4^{th} order polynomial, both of which were favored over the ANOVA model.

The ANOVA table resulting from this model gives even smaller p-values for all effects. This is because the F-statistics are calculated as $F = \frac{\Delta_{SSE}/\Delta_{df}}{\hat{\sigma}^2_{full}} \sim F_{\Delta_{df}, df_{error}}$. In this case the change in error sums of squares (Δ_{SSE}) with additional terms is approximately the same as for the mixed-ANOVA model, but the change degrees of freedom (Δ_{df})with additional terms is half of that for the mixed-ANOVA model, so the F-statistics are approximately double.

```
> anova(mixed.poly)
Number of Observations: 99
Number of Groups: 9
                  numDF denDF   F-value p-value
(Intercept)           1    75 17.577020  0.0001
asp                   2     6  3.737010  0.0883
poly(dist, 5)         5    75 24.022035  <.0001
asp:poly(dist, 5)    10    75  4.011533  0.0002
```

A plot of fits versus distance by transect (figure 12) shows a very close fit, and diagnostic plots (figures 13 and 14) show normally distrubuted, uncorrelated residuals.

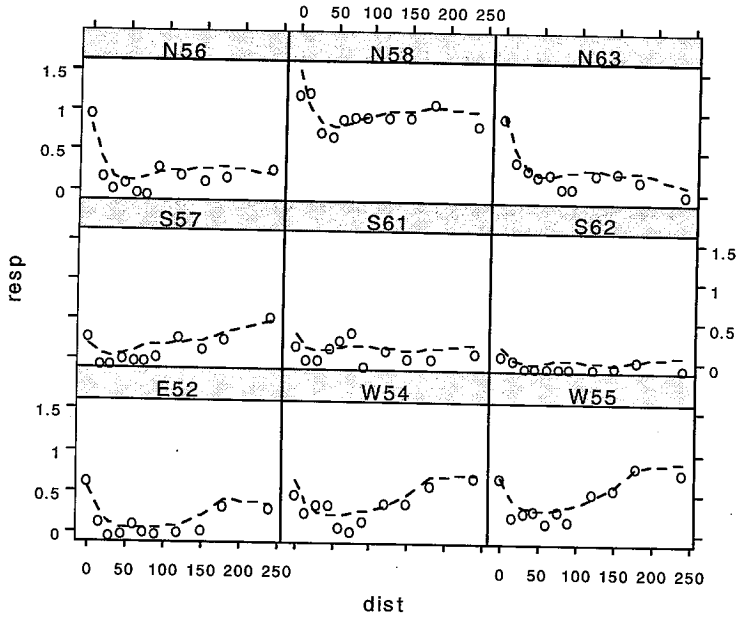
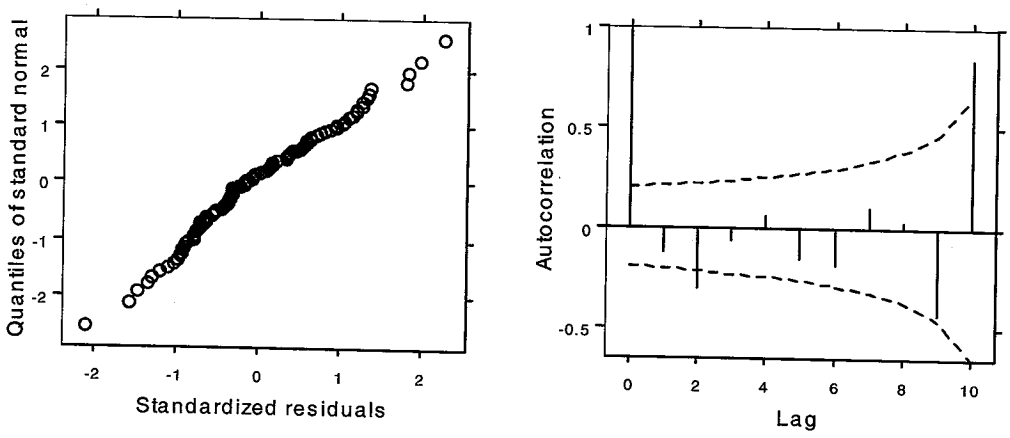


figure 12. Fits from the 5th order polynomial mixed-effects model



figures 13 and 14. Normal probability plot and ACF plot from the 5th order polynomial mixed-effects model

Discussion

The two-factor ANOVA model can be immediately discarded because, by ignoring the grouping of data into transects, it fits the data poorly. The three-factor ANOVA model fits the data much better but inference from this model is limited to the transects measured. The mixed-effects models allow broader inference while fitting the data as well or better. By modeling distance as a 5th order polynomial the number of parameters fit drops by 15, which improves the power while fitting the data quite well. Because it fits the data very well and gives the highest power, the mixed-effects model in which distance is modeled by a 5th order polynomial, with a random slope and intercept for each transect, is the preferred model for the snow cover data.

For the snow cover data, the effects associated with individual transects are small, so results of significance tests for distance and aspect effects are about the same for all of the models examined. For other variables, individual transect effects may be even greater, so the advantages of mixed-effects models will be greater as well.

Previous studies have measured the depth of influence (DEI) using multiple comparison procedures. This was done with the snow cover data using all of the models discussed, and with the exception of

the two-factor ANOVA, all gave roughly the same results. For variables with larger slope random effects, the results of multiple comparison procedures could be quite different. This is because random slope effects are incorporated when calculating the standard error of contrasts over distance (e.g. the contrast between mean response at 240m and the mean response at some other distance). However, multiple comparison procedures are not the ideal way to measure DEI. In all of the variables examined thus far, the edge effects appear to be occurring gradually over distance. If enough transects were measured, the contrast standard error would get small enough, and any contrast, no matter how insignificant biologically, could become statistically significant. A better way to measure DEI is with change-point methodology. In a change-point model we assume some steady state (constant mean and/or variance) at the old-growth end of a transect, then find the point (the change-point) beyond which some other model (a low-order polynomial function of distance) fits the data better.

Bibliography

Murcia, C. 1995. Edge effects in fragmented forest: implications for conservation. Trends in Ecology 10(2):58-62.

Neter, J., M. H. Kutner, C. J. Nachtsheim, W. Wasserman. *Applied Linear Statistical Models.* 4th ed. New York: McGraw Hill, 1996.

Pinheiro, J. C., and D. M. Bates. Mixed-Effects Models in S and S-PLUS. New York: Springer-Verlag, 2000.