RE-EM Trees For Predicting Temporal Changes in Soil Nitrate Levels

Jennifer Weeding
Department of Mathematical Sciences
Montana State University

May 7, 2010

A writing project submitted in partial fulfillment
of the requirements for the degree

Master of Science in Statistics

APPROVAL

of a writing project submitted by

Jennifer Weeding

This writing project has been read by the writing project director and has been found to be satisfactory regarding content, English usage, format, citations, bibliographic style, and consistency, and is ready for submission to the Statistics Faculty.

5/7/2010
Date

Mark Greenwood
Mark C. Greenwood
Writing Project Director and Advisor

Abstract

Regression trees are most appropriately used for modeling and prediction when observations are independent and identically distributed. We will be looking at an adjustment to regression trees, RE-EM Trees, which can be used when the observations are not independent. We suggest a new conditional pruning method, based on cross-validation for selecting an optimal tree size and discuss how to update terminal node estimates from the final mixed model and get standard errors of the node estimates. We compare the RE-EM Tree to a conventional mixed model in an application related to predicting fall to spring nitrate changes in farm fields in eight different locations in Montana, measured twice over two years.

Acknowledgements

I would like to thank Dr. Mark Greenwood for advising me on this project.

I would like to thank Dr. Clain Jones (LRES, MSU) for providing me with the data set that motivated this research and the Montana Fertilizer Advisory Committee (MFAC) that is funding me to complete the research.

I would like to thank Mr. Sydney Akapame for the many discussions we had about RE-EM Trees.

Contents

1 Introduction

Regression trees are nonparametric predictive modeling methods. We use trees for predictive modeling and/or when we are unsure about how the explanatory variables should enter the model, whether they are linearly related to the response, have complex interactions or any relationship to the response. The only assumptions that are attached to regression trees are the residuals should be independent, normally distributed, centered at zero, and have constant variance. When the observations are not independent, typical regression trees should not be used as they may misrepresent the true structure. We will look in depth at one method, RE-EM Trees, that adjust the typical regression tree to incorporate random effects. We then develop a way to prune the RE-EM Tree and get estimates and standard errors of the terminal nodes from the final mixed model. We will finish by comparing trees to the conventional mixed model.

I will explore these methods with a data set that is aimed at predicting the change in nitrate that occurs between fall and spring. The researchers would like to provide a simple way for farmers to predict this change. Farmers usually only have nitrate levels tested in the fall, therefore, being able to predict how much the nitrate level has changed over the winter is important. If farmers overfertilize, they will be spending more money than they originally needed, and if they underfertilize, the crop may not yield as much. Both of these conditions will impact how much the farmer profits and suggest the importance of building an accurate (and simple) predictive model.

1.1 Introduction to Data Set

The data are collected across eight different soil sampling sites which are located near research stations across Montana. At each site, eight fields are selected each year. The study takes

place over three years: 2007 - 2008, 2008 - 2009, and 2009 - 2010, although only two years are analyzed here. Each year different fields are selected, with a total of 24 fields selected per site over three years. The explanatory variables that are measured can be separated into two groups based on whether they were measured for each field or measured once for all fields in the site. Field level explanatory variables are those that vary from field to field and include August nitrate (lb N/acre), organic matter (%), crop (fallow, small grain, oilseed, and annual legume), and soil pH. Site level explanatory variables are measured at a local weather station near the site and include total precipitation (measured from September through February), average fall temperature (average of September, October, and November temperatures), and average winter temperature (average of December, January, and February temperatures). The response variable is the change in nitrate that occurs between April and August (April nitrate - August nitrate) in lb N/acre. Sampling sites and weather stations can be seen in Figure 1. We will assume that the sites in the study are representative of all sites in Montana as they are geographically dispersed based on the network of agriculture research stations.

The data set has some missing responses. After two years, the study should have 128 observations (8 sites * 8 fields per site * 2 years), and it only has 108. The changes in nitrate (April - August) can be seen in Figure 2, plotted based on the eight sampling locations. One outlier (change in nitrate of 194 lbs/acre) was removed from the analysis. While there may be important differences between the sites, that effect appears to be modest in Figure 2. There is a potential problem with the explanatory variables as they may be collinear with each other and also with the sites and years, potentially impacting the models we explore below.

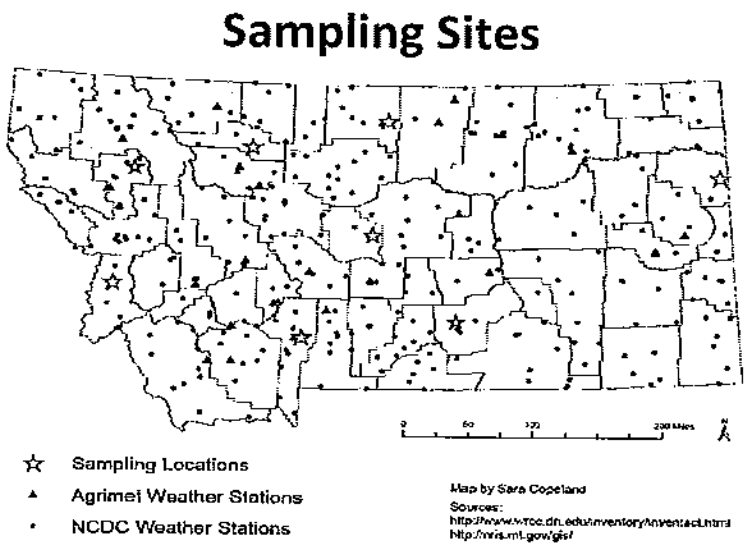

Figure 1: Sampling locations and weather stations.

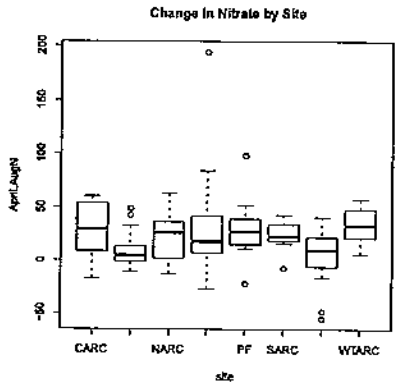

Figure 2: Side-by-side boxplots of the change in nitrate across the 8 sites.

2 Regression Trees

Regression trees are a useful exploratory data analysis tool, and are helpful when you have many explanatory variables, when those variables may have complex interactions, when you have

heavily skewed explanatory variables, or when some of the variables may have nonlinear relationships with the response. They have limited assumptions when compared to the typical multiple linear regression methods and can easily handle large data sets. They can even handle missing explanatory variables.

The general model for a regression tree is:

$$y_i = f(x_1, x_2, ..., x_K) + \epsilon_i$$

where $f(x_1, x_2, ..., x_K)$ is a hierarchical binary partitioning of observations based on values of the K explanatory variables (Maindonald and Braun, (2003), Venables and Ripley, (2002)). We assume that $\epsilon \sim N(0, \sigma^2 I)$, and also that the tree is able to approximate the "true model" with the variables available.

Regression trees are fit by a recursive partitioning algorithm in the R (R Development Core Team, (2009)) package *rpart* (Therneau & Atkinson, 2009). We begin by defining the residual sum of squares as $RSS = \sum_i (y_i - \bar{y})^2$, then split the full data set into two subsets by choosing the split that gives the maximum reduction in RSS. The split is chosen as the best split over all possible values of all the available explanatory variables. For the next split, each of the current cells is considered for splitting and the split is chosen that gives the maximum reduction in RSS. This pattern continues until there is no longer a split that will reduce the RSS.

Terminal nodes are defined as the set of cells that are not split. For classic regression trees, the model in each terminal node is a sample average of observations in that terminal node. We decide which terminal node an observation belongs to by asking a sequence of yes/no questions about the explanatory variables. If we answer "yes" to the question, we follow the branch to the left down to the next split and ask the next question. If we answer "no" to the question, we follow the branch

to the right down to the next split. This process continues until you reach a terminal node. For example, the regression tree in Figure 3 is the estimated tree for the nitrate data set. If we want to predict the change in nitrate that occurred over winter for an observation (#1), we begin by asking ourselves whether or not that observation had a value for August nitrate greater than or equal to 61. Suppose that answer is "no", therefore, we follow the branch to the right. The next question we ask is whether that observation had a value for total precipitation less than 3.415. If the answer is "yes", we would follow the branch to the left. We have reached a terminal node, and our prediction for the change in nitrate that occurred for that observation is 12.4 pounds per acre. For another observation (#29), we begin by asking whether or not it had a value for August nitrate greater than or equal to 61. The answer is "yes", so we follow the branch to the right. The next question we ask is whether it had a value for pH less than 7.7. The answer is "no", so we follow the branch to the right. We have reached a terminal node, and our prediction for the change in nitrate that occurred for that observation is 13.8 pounds per acre. The explanatory variable values and predictions are shown in Table 1.

Observation	AugN	OM	pH	Totalprecip	AveWinterTemp	$\hat{y}$
1	38	3	17	3.37	25.66	12.4
29	94	2	8.2	4.67	26.67	13.8

Table 1: Table showing 2 observations and their predictions.

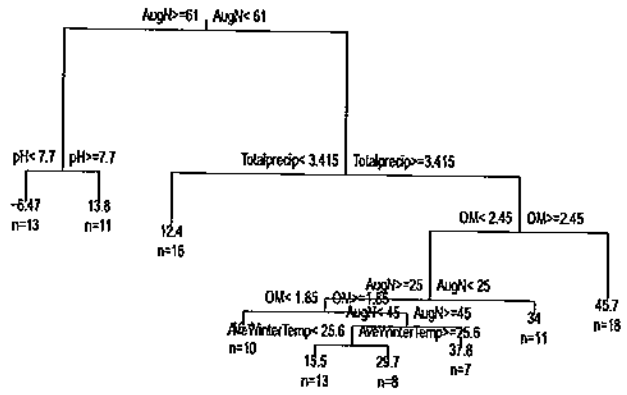

Figure 3: Example of a regression tree applied to the nitrate data set.

Regression trees are grown to the point of overfitting. One way to prune the tree involves using the cross-validated (CV) error. To calculate the cross-validated error for a tree of size m, you:

1. Randomly split the data into 10 equally sized subsets.

2. Withhold one subset of data at a time and fit the tree on the remaining data.

3. Predict $\hat{y}$ for each withheld subset of data, and calculate the prediction error (for each withheld subset of data).

4. Sum the squared prediction errors for each of the 10 subsets. This is the total CV prediction error.

5. Cross-validated error = $\dfrac{\text{total CV prediction error}}{\text{root node error}}$

There are two common rules for pruning the tree. One rule says to choose the tree with the smallest

cross-validated error, and the other says to choose the smallest tree whose cross-validated error is less than the minimum CV error + 1 SE. The minimum CV error + 1 SE rule is generally more conservative and picks a smaller tree, although sometimes the two rules will pick the same tree. If we apply pruning methods to the regression tree in Figure 3, the minimum CV rule tells us to prune the tree to where we only have three splits as seen in Table 2. The minimum CV error + 1 SE rule tells us to prune the tree to a zero split tree, which would be the root node (no tree). Figure 4 shows the same information, with a horizontal line at the value of the minimum CV error + 1 SE. We can control the size of the tree by specifying the value of the complexity parameter (c_p) as a condition in *rpart*. The c_p column in Table 2 shows the c_p values we would specify to get each size tree.

Splits	c_p	Relative Error	CV error	SE (CV)
0	0.177	1.000	1.022	0.173
1	0.078	0.823	1.016	0.158
3	0.039	0.666	0.983	0.150
4	0.017	0.627	1.083	0.169
5	0.017	0.610	1.112	0.164
7	0.016	0.576	1.123	0.163
8	0.010	0.560	1.112	0.163

Table 2: Table of cross-validated errors (CV error) and their standard errors (SE (CV)).

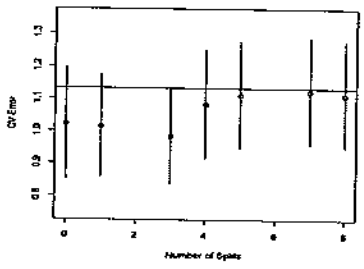

Figure 4: CV error plot

When comparing the two trees in Figure 5, we see that the pruned tree has a less "complicated" structure when compared to the unpruned tree. The value that the tree is splitting on stays the same between the two trees, but in the pruned tree we have taken out some of the explanation of "noise" that was occurring in the lower splits. The pruned tree suggests that AugN, Totalprecip, and OM are important explanatory variables when trying to explain the change in nitrate that occurs over the winter.

In typical multiple linear regression, if we have a missing explanatory variable for an observation it causes problems, specifically we have to throw the entire observation out. Regression trees can be fit to observed responses even if some explanatory variables are missing, with three different methods available for handling missing explanatory variables. To understand the methods, we first must define *surrogate variables*. At each split, *rpart* produces splits on other variables that are similar to the variable that it chose to split on. Surrogates are the splits that are next best to the split that was actually chosen (Therneau and Atkinson, 2009). The three options for dealing with a missing explanatory variable follow, with option 3 being the default in *rpart*.

1. An observation with a missing value for the primary split rule is not sent further down the tree. The prediction for that observation is a weighted average of the terminal nodes that fall

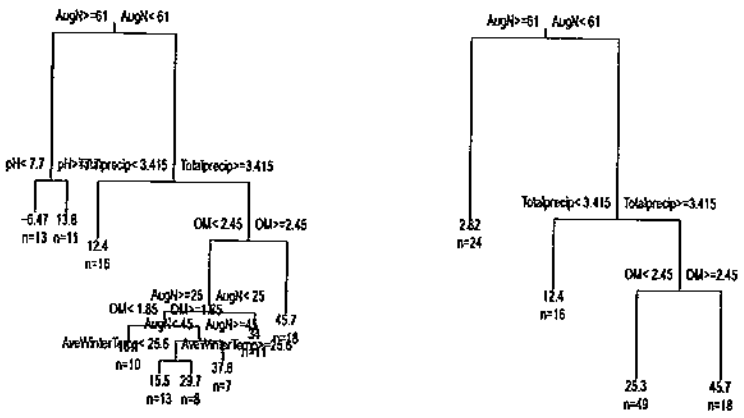

Figure 5: Regular regression tree (left) and the "pruned" regression tree (right). Pruning was based on the minimum cross-validated error rule.

under the primary split rule.

2. Use surrogates to split subjects missing the primary split variable. If all surrogates are missing, the observation is not split.

3. Use surrogates to split subjects missing the primary split variable. If all surrogates are missing, then the observation is sent in the majority direction.

Regression trees can be used to make predictions both easily and quickly, as explained above. They provide a nice visual tool that makes it easy to see which variables "may be" important when making predictions. Trees have limited assumptions on the fixed effects part of the model as compared to typical multiple linear regression models. They are, however, based on *iid* assumptions in the cross-validation methods and in tree growth. When using trees, observations that have missing

explanatory variables do not have to be discarded, and high leverage observations do not cause the same sort of problems as in linear models. Explanatory variables for regression trees can be quantitative, categorical, or ordinal. Ordinal variables are treated like they are quantitative, since location in the varialbe is all that matters. For categorical variables, the splits group levels of the variable together to provide the greatest reduction in RSS.

The typical regression tree should only be applied to independent observations. In 1992, Segal published the first paper in which trees were adapted to longitudinal data. In Segal's approach, all observations for one individual go to the same node. He assumes some parametric form of the covariance function that is assumed to hold over all subgroups. This approach can't handle missing data, requiring the same number of responses for all subjects.

In 2002, Abdolell introduced another method in which trees could handle longitudinal data. His method can be found in the R package, *longRPart*. Abdolell based his approach on a linear mixed effects model. Splits are likelihood based, and a new mixed effects model is estimated for each node. His approach requires a large sample size (default is to have a minimum sample size of 80 at each terminal node) and estimates unique mixed models for each node in the tree with no apparent penalty for the number of unique variance parameter estimates.

In 2009, Sela and Simonoff submitted a paper to *Machine Learning* in which they adjust the regular regression tree for random effects. Their method can be found in the R package, *REEMtree*. Sela and Simonoff's approach iterates between estimating the random effects and estimating a regression tree based on an "adjusted target value." They don't apply pruning methods when they are finished, nor do they go back and update the final estimates of the terminal nodes in the pruned tree, which can differ from those reported in their function. Their approach does allow random effects to cross tree nodes, allowing observations for a subject to end up in different nodes and

still have a common random subject effect. In 2008, Hajjem, Bellavance, and Larocque developed a method similar to Sela and Simonoff's, that adjusts trees for random effects. In the following section, we provide a thorough exploration of Sela and Simonoff's method in which we suggest a way to prune the fitted tree and update the final estimates of the terminal nodes. Finally, we apply their method to the nitrate data set.

3 RE-EM Trees

The Random Effects-Expectation Maximization (RE-EM) Tree is a typical regression tree adjusted to account for random effects. Consider the general mixed effects model with additive errors where we observe subjects $i = 1, 2, ..., I$ at times $t = 1, 2, ..., T_i$:

$$y_{it} = b_i + f(x_{it1}, x_{it2}, ..., x_{itK}) + \epsilon_{it}.$$

We assume the errors are normally distributed and independent between subjects, but can be correlated within subject, $\epsilon_i \sim N(0, R_i)$, where R_i is the subject-specific variance-covariance matrix. We also assume the random subject effects, b_i, are normally distributed, $b_i \sim N(0, \sigma_b^2)$. To fit an RE-EM Tree, one iterates between estimating the random effects part of the model (Pinheiro and Bates, 2000) and estimating the regression tree. Sela and Simonoff's estimation method is given as follows:

1. Initialize the estimated random effects, $\hat{b_i}$, to zero.

2. Iterate through the following steps until the mixed model likelihood converges:

   a. Estimate a regression tree approximating f, based on the adjusted target variable, $y_{it} - \hat{b_i}$, and predictors, $\mathbf{x_{it}} = (x_{it1}, x_{it2}, ..., x_{itK})$. Use this regression tree to create a set of

indicator variables, $I(x_{it} \in g_p)$, where g_p ranges over all the terminal nodes in the tree.

b. Fit the linear random effects model, $y_{it} = b_i + I(x_{it} \in g_p)\mu_p + \epsilon_{it}$. Extract $\hat{b_i}$ from the estimated model.

In this prcedure, estimated random effects ($\hat{b_i}$) are initialized to zero in step 1. We found better performance when we initialized the estimated random effects to a nonzero preliminary estimate. Sela and Simonoff state that the linear model with random effects in step 2b can be estimated using maximum likelihood (ML) or using restricted maximum likelihood (REML), and note using ML instead of REML had only a small effect on the resulting estimates. In their approach, they use REML. In our application, we observed large differences in the estimates we obtained when using ML and REML, possibly due to the small relative size of $\hat{\sigma_b^2}$. For the examples presented here, we estimate the mixed model with REML, to stay consistent with Sela and Simonoff.

The RE-EM Tree model with additive errors applied to the nitrate data set is

$$(AprilN - AugN)_{ij} = \alpha + f(\boldsymbol{X}) + b_i + \epsilon_{ij},$$

where $f(\boldsymbol{X})$ is the tree part of the model based on $\boldsymbol{X}$ = AugN, Treatment, OM, pH, Totalprecip, AveFallTemp, AveWinterTemp, b_i is the site random effect ($i = 1, 2, ..., 8$), and ϵ_{ij} is the random error for observation j in site i. The researchers are currently in the final year of data collection so we will ignore the random year effect in the present analysis.

Figure 6 contains the estimated full RE-EM tree part of the model. Comparison of this tree to the typical regression tree in Figure 3 shows many similarities. For instance, the first split in both trees is identical. Many of the next splits are also identical. In the *REEMtree* output, for each split, it also gives you other variables that provided a reduction in RSS that are comparable to the split on the variable that it chose. In Figure 6, if we follow the tree to the right from the first split,

we get to a split on organic matter. In the output, we were able to see that if the tree had split on total precipitation instead, it would have provided almost the same reduction in residual sums of squares. If the RE-EM Tree had split on total precipitation instead of organic matter, than the two trees would be almost identical.

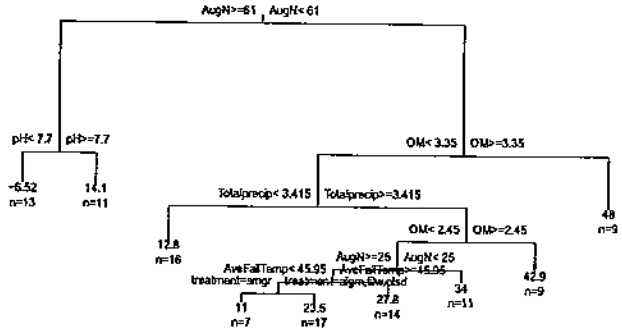

Figure 6: Example of a RE-EM Tree applied to the nitrate data set.

To prune the RE-EM Tree, we use cross-validated error. We begin by estimating the random effects, $\hat{b_i}$ for a tree of size m. Then we get the cross-validation error for $y_i{}^* = y_i - \hat{b_i}$ for a tree of size m. We select the tree with either the smallest cross-validated error or the smallest tree whose cross-validated error is less than the minimum CV error + 1 SE. After we fit the optimal pruned tree, we estimate the node means and random effects in the final mixed model, updating $\hat{\mu_p}$ from the last iteration of step 2a to 2b. In Sela and Simonoff's paper, the fitted tree is not updated with the estimates provided from the final mixed model. In typical regression trees, the same random splits are used for all tree sizes in the cross-validation process. We used different random splits of the data for each tree size because each tree cross-validation involved a new call

to *rpart*. This introduces more random variability in the CV process, making repeated CV runs even more important.

The pruning of a RE-EM Tree is a random process since different random data subsets are selected. We applied the pruning methods described above 30 times to the RE-EM tree, with the results provided in Table 3. From the table, we see that if we were using the minimum CV error pruning rule, 23 out of 30 times we would choose a two-split tree. If we were using the smallest tree whose CV error is less than the minimum CV error + 1 SE pruning rule, we would choose a zero-split (root node) tree 17 out of 30 times, and a two-split tree 11 out of 30 times. When there is a lot of noise present in the response variable compared to the size of the differences in the terminal nodes, cross-validation methods tend to favor simple trees. This happens because cross-validation methods have a difficult time distinguishing "significant" structure from noise if there is a lot of noise present in the data. A CV error plot for one of the 30 replications is shown in Figure 7. For this particular replication, the two pruning rules would both have chosen a two split tree. Figure 8 shows both the unpruned and pruned RE-EM trees.

	Splits					
Rule	0	2	3	4	6	8
minimum CV	1	23	3	0	2	1
min CV + 1 SE	17	11	1	1	0	0

Table 3: Results for 30 replications of the pruning process.

The final step in the pruning process involves estimating the node means and random effects in the final mixed model. These estimates are shown in Table 4 and Table 5. Intra-site correlation is calculated by taking the random site effect variance and dividing it by the sum of the random site

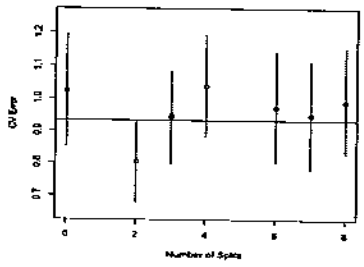

Figure 7: CV Error plot for one run of the pruning process.

variance and residual variance. The estimated intra-site correlation was 0.11, which indicates that we have small variation across the sites compared to the total variance in our data. The estimates from the final mixed model are slightly different than the estimates given on the pruned tree in Figure 8. The pruned *rpart* tree and the pruned RE-EM tree are shown in Figure 9. The pruned *rpart* tree has three splits, and the pruned *REEMtree* has two splits. The variables that may be considered "important" are August nitrate, total precipitation, and organic matter. We include total precipitation in the variables that we may consider "important," because we were able to see that if the tree had split on total precipitation instead of organic matter, it would have provided almost the same reduction in residual sums of squares.

Random Effect	Variance	95% Confidence Interval
Site SD	51.41	(9.30 , 282.91)
Residual SD	410.47	(309.76 , 543.36)

Table 4: Estimated Random Effects (REML).

Some advantages of using a RE-EM Tree are: they can handle missing explanatory variables, they are easy to understand since they clearly display which variables may be important, and we aren't as concerned with points that have high leverage. One disadvantages is that if a strong linear

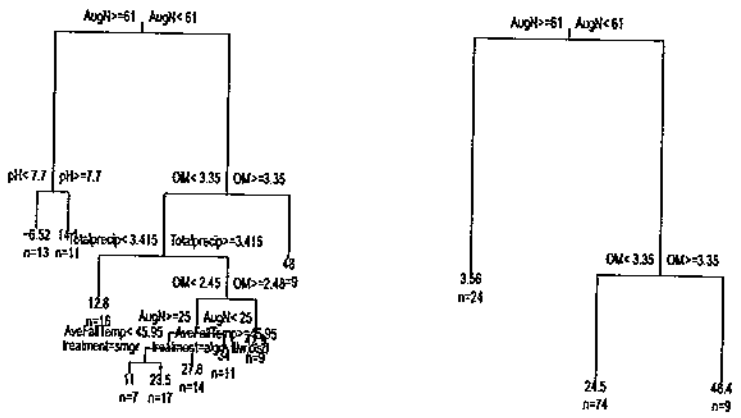

Figure 8: Unpruned RE-EM Tree (left) and pruned RE-EM Tree (right). Pruning was based on the minimum cross-validated error rule.

relationship is present, the trees are not as efficient as linear models. Additionally, if outliers are present, they greatly affect the trees performance just like in more conventional models.

Through the examples with the nitrate data set, we were able to explore both the regular *rpart* and RE-EM trees. We saw that the trees had similar performance, and that between site variation was minimal. We were able to see that August nitrate, total precipitation and organic matter are potentially "important" explanatory variables. We developed a way to prune the RE-EM Tree, which was not developed for *longRPart* or *REEMtree* previously. We updated the terminal nodes in the pruned RE-EM Tree with the estimates from the final mixed model, which is not part of the default in *REEMtree*. We were able to get standard errors for the terminal node estimates, which is also not typically a part of trees. The standard errors can be adjusted for the correlation structure and/or random effects, providing more accurate estimates of precision of node estimates than if

Terminal Node	Estimate	Standard Error	95% Confidence Interval
AugN$\geq$ 61	3.72	5.00	(-6.22,13.65)
AugN$<$ 61 & OM$<$ 3.35	23.98	3.51	(17.01,30.95)
AugN$<$ 61 & OM$\geq$ 3.35	51.82	8.12	(35.71,67.93)

Table 5: Estimated terminal nodes, along with their standard errors and a 95% confidence interval.

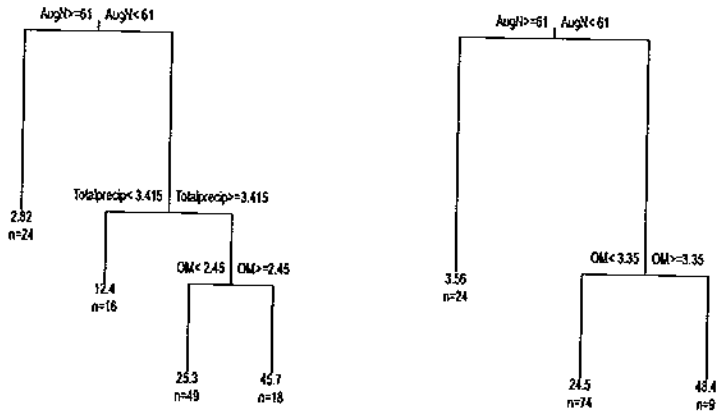

Figure 9: Pruned *rpart* tree (left) and pruned RE-EM tree (right). Pruning was based on the minimum cross-validation error rule.

important dependence is ignored. When the full data set is available, we will adjust the RE-EM Tree so that it can also handle crossed-random effects.

4 Conventional Mixed Model

The conventional mixed model is used when both linear fixed and random effects are present. The equation describing a mixed model is

$$y_i = X_i\beta + Z_i b_i + \epsilon_i,$$

where y_i is a vector of observations from the ith level of some random factor, β is the population average coefficient vector, b_i is a vector of random effects, and ϵ_i is a vector of errors from the ith level of the random factor, as discussed previously.

In our example, the model is

$$y_{ij} = X_i\beta + b_i + \epsilon_{ij},$$

where y_{ij} is the jth observation from the ith site, X_i is a matrix based on AugN, Treatment, OM, pH, Totalprecip, AveFallTemp, AveWinterTemp (which now enter the model linearly), β is the population average coefficient vector, b_i is the random site effect, and ϵ_{ij} is the error for the jth observation in site i. Two different models were selected based on two different model selection criteria. The first model, Model 1, was built based on including all of the explanatory variables present and also a three-way interaction between August nitrate, total precipitation, and the treatment. The reason for including the three-way interaction was based *a priori* on expected interactions of the researchers. A stepwise reduction testing procedure was applied to the full model. The analysis of variance (ANOVA) table for Model 1 is shown in Table 6, and diagnostic plots can be seen in Figure 10. From the ANOVA table, we can see that a fairly complicated model was chosen. The three-way interaction was significant, so all the two-way interactions and main effects related to that interaction remain in the model. Also, the main effect for organic matter was retained. Main effects for pH, average fall temperature, and average winter temperature were removed from

the model, suggesting they are not important after adjusting for the other effects. Random effect variances and 95% confidence intervals can be seen in Table 7. The estimated intra-site correlation for this model is 0.15. From the diagnostic plots in Figure 10, we see that the assumptions of constant variance and normality of the residuals appear to be valid. We have one observation that is "unusual", which can be seen in both the residual versus fitted value plot and also in the normal quantile plot.

Fixed Effects	numerator df	denominator df	F-value	p-value
Intercept	1	83	40.20523	<.0001
AugN	1	83	20.16172	<.0001
Totalprecip	1	83	2.50757	0.1171
treatment	3	83	1.39192	0.2510
OM	1	83	8.23462	0.0052
AugN:Totalprecip	1	83	1.94736	0.1666
AugN:treatment	3	83	1.95825	0.1266
Totalprecip:treatment	3	83	0.94734	0.4217
AugN:Totalprecip:treatment	3	83	2.61813	0.0563

Table 6: ANOVA table for Model 1.

Random Effect	Variance	95% Confidence Interval
Site	55.93	(9.68 , 323.34)
Residual	322.85	(242.95 , 429.03)

Table 7: Random effects for Model 1.

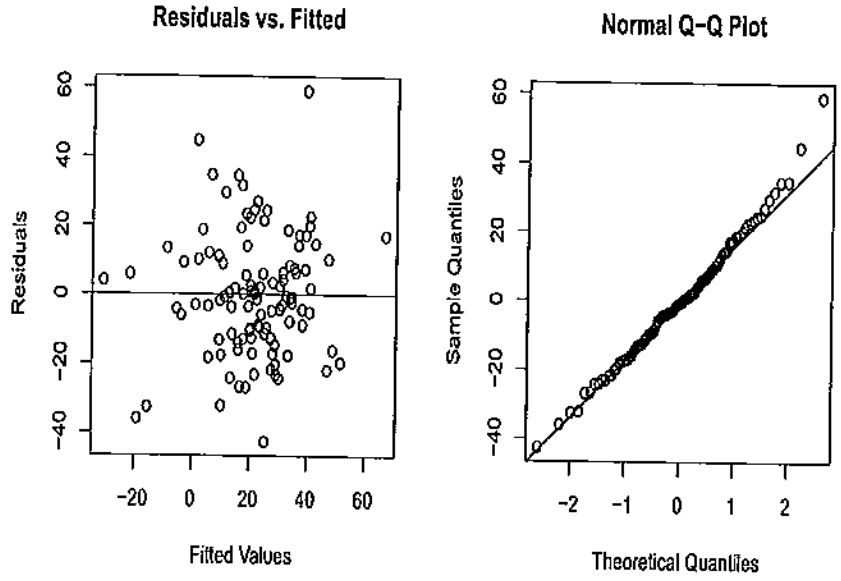

Figure 10: Model 1 diagnostic plots.

The second model, Model 2, was built based on Akaike's An Information Criterion, AIC (Akaike, 1973), and not including the three-way interaction discussed above. AIC is equal to: $-2 * logLikelihood + 2 * p$, where p = the number of parameters estimated in the model. AICs are used as a model selection tool, where preference is given to models with smaller AICs. Model 2 was built by including all of the explanatory variables as main effects, then applying a stepwise AIC approach which was allowed to both add and drop terms if the add/drop reduced the AIC. The ANOVA table for Model 2 is shown in Table 8, and diagnostic plots can be seen in Figure 11. From the ANOVA table, we see that this method chose a much simpler model than Model 1. The main effects for August nitrate, total precipitation, and organic matter remain in the model, as does the two-way interaction between August nitrate and total precipitation. The main effects for treatment, pH, average fall temperature, and average winter temperature were removed from the model and other two-way interactions were not incorporated. Random effect variances can be seen

in Table 9. Model 2 has a smaller random site effect variance than Model 1, but a larger residual variance, resulting in a smaller estimated intra-site correlation (0.05). From the diagnostic plots in Figure 11, we again see that our assumptions of constant variance and normality of the residuals look reasonable. The "unusual" observation remains, and can be seen in both the residual versus fitted value plot and also the normal quantile plot.

Fixed Effects	numerator df	denominator df	F-value	p-value
Intercept	1	95	68.32875	<.0001
AugN	1	95	20.51606	<.0001
Totalprecip	1	95	1.21796	0.2725
OM	1	95	6.80037	0.0106
AugN:Totalprecip	1	95	3.08138	0.0824

Table 8: ANOVA table for Model 2.

Random Effect	Variance	95% Confidence Interval
Site	22.10	(1.65 , 296.86)
Residual	413.84	(312.63 , 547.79)

Table 9: Random effects for Model 2.

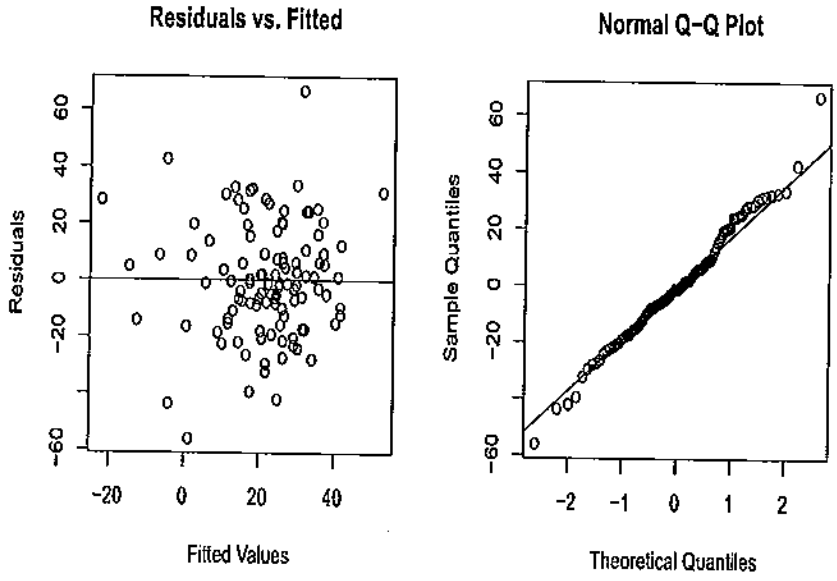

Figure 11: Model 2 diagnostic plots.

We re-estimated the mixed model of the pruned tree using maximum likelihood estimates to be able to compare AICs between the models. Table 10 shows the three models and their corresponding AIC and log-likelihood values. The tree method has the smallest AIC. Model 2 and the tree model have basically the same log-likelihood values, which indicates that they are similar fitting models but the tree uses two fewer degrees of freedom to achieve this fit. Random effect variances can be seen in Table 11. The tree model has a random site variance that is between the random site variance for Model 1 and Model 2. Also, the residual variance for the tree model is between the residual variance of Model 1 and Model 2. The estimated intra-site correlation is 0.09, which is slightly smaller than the estimated intra-site correlation when the tree was estimated using REML. Figure 12 contains diagnostic plots for the tree model. We see that our assumptions look reasonable, although the "unusual" observation is still present. When we analyze the final data set, we may re-consider removing this observation.

Model	df	AIC	Log-Likelihood
RE-EM Tree	5	962.19	-476.10
Model 2	7	966.68	-476.34
Model 1	19	969.37	-465.69

Table 10: AIC's and Log-Likelihoods for the 3 mixed models considered, sorted by AIC values.

Random Effect	Variance	95% Confidence Interval
Site	38.63	(6.54 , 228.19)
Residual	403.31	(305.21 , 532.92)

Table 11: Random effects for the tree model (ML).

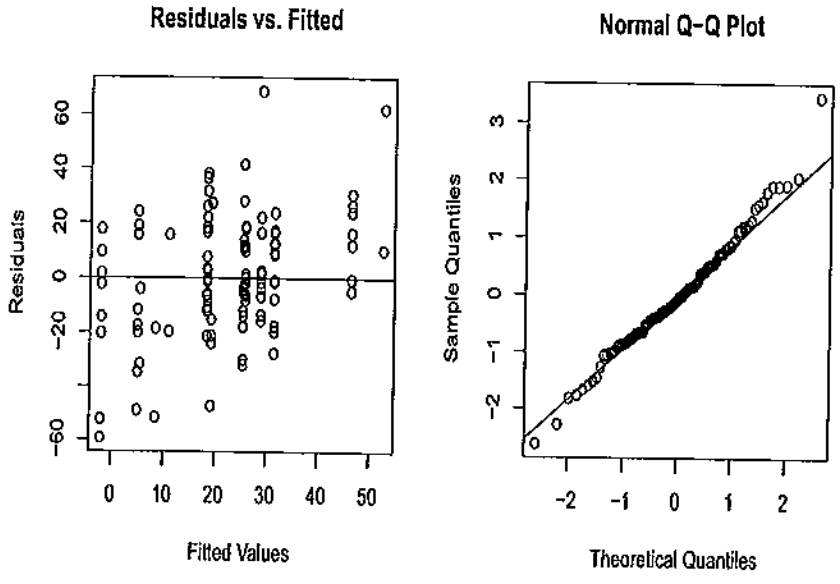

Figure 12: Pruned RE-EM tree diagnostics.

5 Conclusions

We looked at regular regression trees, RE-EM Trees, and the conventional mixed model, all applied to the nitrate data set. Regular regression trees cannot be used when observations are

not independent, therefore, RE-EM Trees were employed. We saw many similarities between the pruned regression tree fit and the pruned RE-EM Tree fit, which could be a result of the small intra-site correlation. Previously, methods for pruning the RE-EM Tree were not available, so we developed a novel method to prune the tree. We saw that outliers had a large impact on the tree building and pruning methods. We fit two different conventional mixed models, based on different model selection criteria. The model fit based on the stepwise reduction testing procedure, Model 1, was fairly complex with 19 parameters being estimated. Model 2, based on a stepwise AIC procedure and excluding the three-way interaction, had only seven parameters and a smaller random site variance. The tree model had only five parameters estimated and yielded the smallest AIC. The variables that were important in Model 2 were the same variables that showed up in the pruned RE-EM Tree. Also, the random site variance and residual variance of Model 2 were similar to that of the RE-EM Tree model. The assumptions of constant variance and normality appeared valid for all three of the models discussed.

A potential further application of the RE-EM Tree method consists of adjusting the RE-EM Tree to incorporate crossed-random effects. We anticipate the necessity of this adjustment when we analyze the final nitrate data set which will have eight sites measured over three years and at most eight fields per site per year. When we fit the RE-EM Tree, we saw large differences in the fitted tree when we used the estimation methods of restricted maximum likelihood (REML) compared to maximum likelihood (ML). Future research may explain why these differences in the fitted tree occurred. Also, when we pruned the RE-EM Tree, we used different random splits in each run of the cross-validation process. Future research may lead to a pruning procedure for RE-EM Trees that uses the same random splits in each run of the cross-validation process, making the process more similar to what is done in *rpart*.

References

[1] Abdolell, M. (2002) Binary partitioning for continuous longitudinal data; categorizing a prognostic variable, *Statistics in Medicine*, 21, 3395-3409.

[2] Hajjem, A., Bellavance, F., and D. Larocque. (2008) Mixed-Effects Regression Trees for Clustered Data. Les Cahiers du Gerad (discussion papers), www.gerad.ca/fichiers/cahiers/G-2008-57.pdf , 23 pages.

[3] Maindonald, J. & Braun, J. (2003) *Data Analysis and Graphics Using R - An Example-Based Approach*, Cambridge University Press.

[4] Pinheiro, JC and Bates, DM (2000), Mixed-Effects Models in S and S-PLUS, Springer,New York.

[5] Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., and the R Core team (2009). nlme: Linear and Nonlinear Mixed Effects Models. R package version 3.1-96.

[6] R Development Core Team (2009). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

[7] Segal, M. (1992) Tree-Structured Methods for Longitudinal Data, *JASA* , 87(418), 407-418.

[8] Sela, R. & Simonoff, J. (Submitted) RE-EM Trees: A New Data Mining Approach for Longitudinal Data, *Machine Learning*.

[9] Therneau, T. & Atkinson, B. (2009) R port by Brian Ripley, rpart: Recursive Partitioning, R package version 3.1-45, http://CRAN.R-project.org/package=rpart.

[10] Venables, W.N. & Ripley, B.D. (2002) *Modern Applied Statistics with S*, Springer.