GAMM-Trees for Understanding Northern Hemisphere Temperature Trends

Sydney Akapame

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

Sydney Akapame

This writing project has been read by the writing project advisor 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 Coordinator and Advisor

Contents

List of Figures

Acknowledgements

I would like to thank Dr. Mark Greenwood very much for his supervision throughout the course of this writing project. My appreciation also goes to Jennifer Weeding and Rory Peck. And most of all, thanks be to God.

Abstract

Trees, popularized by the work of Breiman et. al.(1984), have become an alternative statistical modeling approach. The appeal of trees perhaps is in the ease of interpretability and competitiveness with conventional linear models. However, regression trees, without any modification, cannot be used in a mixed modeling context where random effect estimation is also desired. Sela and Simonoff (2009) introduced Random Effects Expectation-Maximization (RE-EM) trees that are able to handle random effects. In some instances, the response may be related to one or more of the predictors through some smooth function(s) and trend estimation may be required in addition to the random effects. In this paper, we discuss Generalized Additive Mixed Models (GAMM) trees, a modification of RE-EM trees for cases where trend estimation is also of interest. We improve further on RE-EM trees by discussing a method for selecting an optimal tree in the GAMM-tree. Finally, we apply GAMM-trees to the Northern Hemisphere temperature anomaly series to investigate climate change.

1 Introduction: Climate Change

Climate change is a change in the statistical distribution of weather over periods of time that range from decades to millions of years. It can be a change in the average weather or a change in the distribution of weather events around an average (for example, greater or fewer extreme weather events). Climate change may be limited to a specific region, or may occur across the whole Earth. The most general definition of climate change is a change in the statistical properties of the climate system when considered over periods of decades or longer, regardless of cause. Accordingly, fluctuations on periods shorter than a few decades, such as El Nino, do not represent climate change. The term sometimes is used to refer specifically to climate change caused by human activity; for example, the United Nations Framework Convention on Climate Change defines climate change as "a change of climate which is attributed directly or indirectly to human activity that alters the composition of the global atmosphere and which is in addition to natural climate variability observed over comparable time periods". In the latter sense climate change is synonymous with global warming.

Factors that can shape climate are climate forcings. These include such processes as variations in solar radiation, deviations in the Earth's orbit, mountain-building and continental drift, changes in greenhouse gas concentrations and complex oscillations like El Nino Southern Oscillation (ENSO) and North Atlantic Oscillation (NAO). There are a variety of climate change feedbacks that can either amplify or diminish the initial forcing. Some parts of the climate system, such as the oceans and ice caps, respond slowly in reaction to climate forcing because of their large mass. Therefore, the climate system can take centuries or longer to fully respond to new external forcings (www.wikipedia.org/wiki/Climate_change).

1.1 Indicators of Climate Change

Evidence for warming of the climate system includes observed increases in global average air and ocean temperatures, widespread melting of snow and ice, and rising global average sea level. The most common measure of global warming is the trend in globally averaged temperature near the Earth's surface. The historical surface temperature dataset HadCRUT (Jones, 1994; Jones and

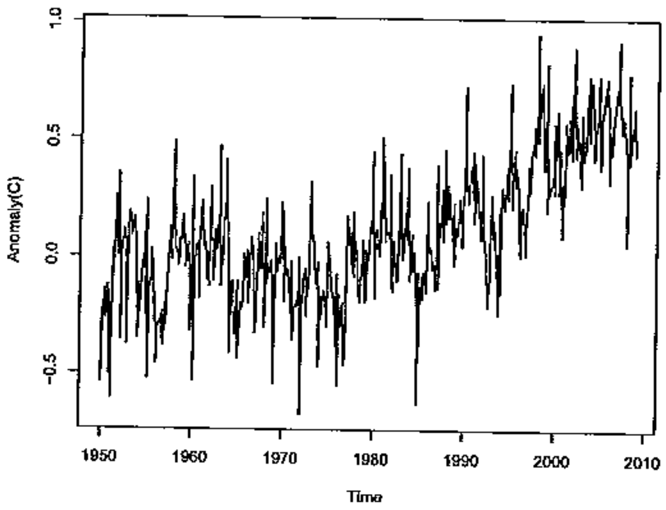
Moberg, 2003) has been extensively used as a source of information on surface temperature trends and variability (Houghton et al., 2001). Climate change can be investigated using this dataset. More importantly, we will be focusing on the Northern Hemisphere temperature anomaly series which dates back to 1850 and it is based on over 4394 stations around the world. The term "temperature anomaly" means a departure from a reference value or long-term average. A positive anomaly indicates that the observed temperature was warmer than the reference value, while a negative anomaly indicates that the observed temperature was cooler than the reference value. The anomalies are in $^{\circ}C$ (www.ncdc.noaa.gov/cmb-faq/anomalies.html) and are calculated by subtracting the station monthly mean (1961-1990) to generate the observed hemispheric mean anomaly. This removes the seasonal cycles from the series. The NH temperature anomaly series is displayed in Figure 1. It is evident from Figure 1 that the anomaly series shows an increasing trend. This trend is smooth, non-linear with some variability around it.

Why use temperature anomalies when absolute temperature measurements can be used? Absolute estimates of global average surface temperature are difficult to compile for several reasons. Some regions have few temperature measurement stations (e.g., the Sahara Desert) and interpolation must be made over large, data-sparse regions. In mountainous areas, most observations come from the inhabited valleys, so the effect of elevation on a regions average temperature must be considered as well. Using reference values computed on smaller (more local) scales over the same time period establishes a baseline from which anomalies are calculated. This effectively normalizes the data so they can be compared and combined to more accurately represent temperature patterns with respect to what is normal for different places within a region.

1.2 Modeling NH Temperature Anomalies

This paper will look at the Northern Hemisphere temperature anomaly series and attempt to explain variability in these observations using some associated climatic indices. Short-term fluctuations (years to a few decades) such as the El Nino Southern Oscillation (ENSO), the Pacific Decadal Oscillation (PDO), the North Atlantic oscillation (NAO), and the Arctic oscillation, represent climate variability rather than climate change. On longer time scales, alterations to ocean processes such

Figure 1: Northern Hemisphere Temperature Anomaly time series.



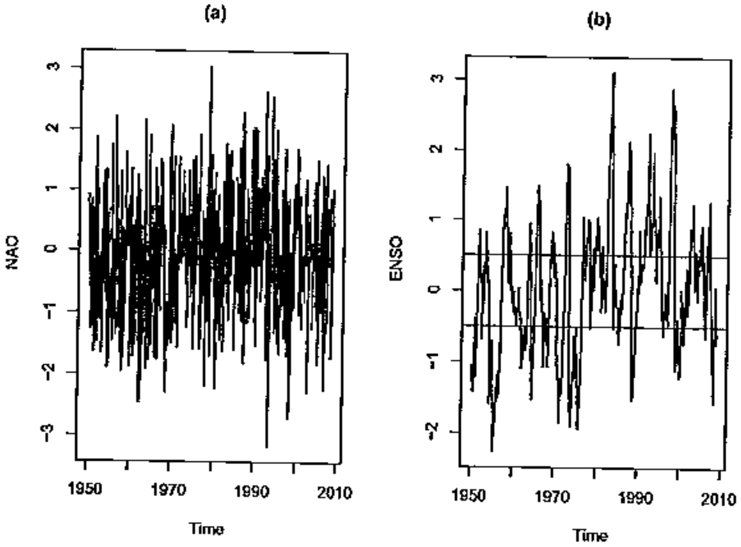
as thermohaline circulation play a key role in redistributing heat by carrying out a very slow and extremely deep movement of water, and the long-term redistribution of heat in the world's oceans. The climatic indices we will look at are ENSO and NAO.

ENSO is a Pacific ocean-based index. It cycles every 2 to 8 years and has two main phases, negative and positive phases (Jones 1989; Wigley 2000; Wolter et al., 1993). ENSO is composed of an oceanic component, called El Nino (or La Nina, depending on its phase), which is characterized by warming or cooling of surface waters in the tropical eastern Pacific Ocean, and an atmospheric component, the Southern Oscillation, which is characterized by changes in surface pressure in the tropical western Pacific. The two components are coupled: when the warm oceanic phase (known as El Nino or positive ENSO events, (> 0.5)) is in effect, surface pressures in the western Pacific are high, and when the cold phase is in effect (La Nina, negative ENSO events (< -0.5)), surface pressures in the western Pacific are low. Mechanisms that cause the oscillation remain under study.

NAO (www.cpc.ncep.noaa.gov) is an Atlantic ocean-based oscillation. It is a climatic phenomenon in the North Atlantic Ocean of fluctuations in the difference of atmospheric pressure at sea level between the Icelandic low and the Azores high. Through east-west oscillation motions of the Icelandic low and the Azores high, it controls the strength and direction of westerly winds and storm tracks across the North Atlantic. It is highly correlated with the Arctic oscillation, as it is a part of it. Strong positive phases tend to be associated with above-average temperatures. Negative phases lead to opposite conditions. The negative phase appeared just twice from 1979/80 through 1994/1995.

Monthly values of these indices are available as far back as 1950 and hence we will focus our attention on temperature anomalies from 1950 through 2008. We show ENSO and NAO in Figure 2. Positive ENSO events lead to abnormally warm Sea surface Temperatures, SST. La Nina, negative ENSO events, are associated with cold SST. ENSO is the more well-known of the two indices that may lead to year-to-year climate variability. ENSO is an east-west atmospheric pressure see-saw that directly affects tropical weather around the globe and indirectly impacts a larger area. It is associated with floods, droughts and other disturbances in a range of locations around the world. It is the most prominent known source of inter-annual variability in weather and climate around the world.

Figure 2: NAO (a) and ENSO (b) Monthly Time series.



ENSO and NAO are complex climatic indices such that the relationship between the indices and temperature anomaly may not be trivial. More importantly, these indices may have thresholds in them that may be key to explaining temperature anomalies. This may make using linear models suspect since they do not handle threshold relationships well. In the next section, we introduce the statistical concept of trees in detail, focusing on regression trees that might be useful here. We apply regression trees to the temperature anomaly series and discuss the results in the next section.

2 Trees

Statistical modeling almost always starts with the use of linear models. This is as a result of the ease of model interpretation and/or the approximate correctness of the linear model. This may not always be the case. Linear models may oversimplify an otherwise complex relationship between a response and some predictor(s). There are situations where although we can make distributional assumptions about our errors, modeling would require more than a linear model. In some datasets the linear model assumption may be inappropriate. This limits the extent to which we can explore the relationship between the response and the covariates. Even if we can specify models for very complex datasets, it may be impractical to develop inferential methods for them (Faraway, 2006). The relationship between the NH temperature anomalies and NAO and ENSO is an example of a potentially complex response-predictor(s) relationship.

Tree methodology is common in computer science literature. There are two general types of trees, classification trees and regression trees (Brieman et al., 1984). Regression trees are used for quantitative responses and classification trees for categorical responses. Trees have been applied to decision tree problems where there is no stochastic structure and a rule is required for making a decision (Faraway, 2006). Regression trees may be very competitive to linear models in situations where we have complex datasets. They are an algorithmically-based statistical method and they can be thought of as a cross between the linear model and the completely nonparametric approach. Regression trees are best understood through their underlying algorithms, where a recursive partitioning algorithm is used. For regression trees, this is described in the next few steps, for quantitative predictor(s), as in Faraway (2006), among others.

1. Consider all partitions of the region of the predictors into two regions where the division is parallel to one of the axes. That is, partitioning is done by choosing a point along the range of a particular predictor to make the split. We can have at most $(n-1)p$ splits. We can illustrate the number of splits in a very simple fashion. Suppose that we have a response y and p predictors. For a particular predictor x, with n ordered values, we can have splits of the form $x_1|x_2, x_3, ..., x_n$ where $|$ indicates where the split is done. Hence we have $(n-1)$ splits for each of the p predictors giving at most $(n-1)p$ splits.

2. For each partition we take the mean of the response in that partition. We then compute

$$RSS(partition) = RSS(part_1) + RSS(part_2)$$

   and choose the partition that minimizes the residual sum of squares. Total residual sum of squares is defined as

$$RSS = \sum(y_i - \bar{y})^2.$$

3. Subpartitioning of the partitions is done in a recursive manner, choosing the subpartitions that lead to the greatest reduction in the RSS. Partitions are only allowed within existing partitions and not across them. This makes it natural to represent the partitioning with a tree.

We look at the steps described above in a little more detail. Imagine that we have m regions and in each region we model the response as a constant c_j. Then for every splitting variable, we seek to minimize $\sum(y_i - c_j)^2$. It follows from mathematical statistics that the minimizer $\hat{c}_j$ is the mean of the observations in region R_j. That is,

$$\hat{c}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} y_i$$

where n_j is the number of observations in the R_j region. Generally, Hastie et al. (2009) propose that starting with all the data, consider splitting on variable j and split point s. We define half-planes

$$R_1(j,s) = \{X|X_j \leq s\} \text{ and } R_2(j,s) = \{X|X_j \geq s\}.$$

Then what is required are the splitting variable j and split point s along that variable that solves

$$\min_{j,s}\left[\min_{c_1} \sum_{x_i \in R_1(j,s)} (y_i - c_1)^2 + \min_{c_2} \sum_{x_i \in R_2(j,s)} (y_i - c_2)^2\right].$$

One of the key advantages of tree methodologies over the linear model is their ability to use complex interactions among the predictors to explain variability in the response and at the same time making interpretation fairly simple. Splitting on a variable and making further splits on other variables is analogous to looking at higher order interactions among the predictors. This is a luxury we often cannot reasonably have with linear models, partially because the number of interactions can be quite large with a large set of explanatory variables.

We can get predictions out of the regression tree as well. First, if we have say k partitions, we can make a prediction at each of the k terminal nodes. The mean of the observations at each node is used as the prediction,

$$\hat{f}(X) = \sum_{i=1}^{k} \hat{c}_i I\left\{(X_1, ..., X_p) \in R_i\right\}.$$

The indicator function evaluates to 0 or 1 depending on whether or not the particular values of the predictors belong in region R_i.

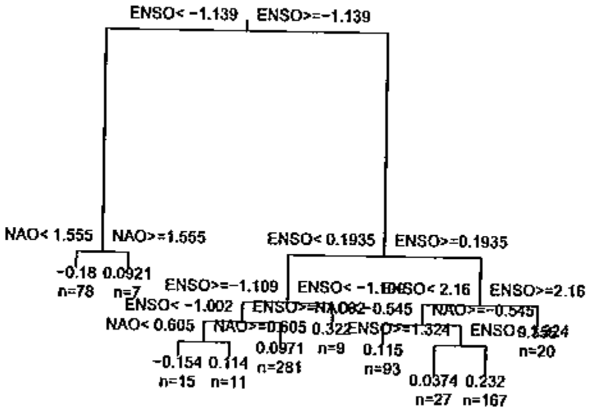
To illustrate the method described above, we will use the temperature anomaly time series. The estimated tree can be thought of as

$$\widehat{Temp}_t = \hat{f}(ENSO_t, NAO_t)$$

and it is displayed in Figure 3. The tree shows 10 terminal nodes, providing predictions of the NH monthly temperature anomaly in terms of ENSO and NAO. Interpreting the tree in terms of ENSO and NAO values is straightforward. By looking at the tree, we know that in predicting temperature, the most important variable is perhaps ENSO. This is because the variable ENSO and split point -1.139 provide the largest reduction in residual sum of squares. We can also see that ENSO values ≥ -1.139 are associated with above average temperatures or positive anomalies, in general. ENSO values < -1.139 are associated with below average temperatures. The splits are recursively performed according to the steps outlined above, with each split providing the maximum reduction in the relevant residual sum of squares.

The tree above has 9 splits and hence 10 terminal nodes or leaves. Is this the optimal tree? That is, can we find another simpler tree that closely approximates this tree? Cross-validation methods discussed in the next section provide a way of assessing the performance of trees and choosing a tree that should be an optimal predictive tree.

Figure 3: A regression tree for the NH anomaly series.



2.1 Pruning

Trees are grown to explain as much variability as possible, with no concern about whether the tree is overfit and may be explaining noise in the observations instead of "real" structure in the population. We next consider the problem of how large a tree we should grow. The size of a tree is directly related to the complexity of the tree. Tree size is determined by how many terminal nodes the tree has. Just as complex linear models may potentially overfit the data, so will very large trees. In the same vein, smaller trees may not capture the inherent structure in the data. There is a need to balance the trade-off between the size of the tree and how much information it captures. Tree pruning, a method of reducing the over-fit tree, is analogous to doing backward model selection using AIC or other model selection criteria, making it a very important part of regression tree modeling.

A number of authors have prescribed approaches to doing this, among them are Hastie et al. (2009) and Venables and Ripley (2000). They both suggest an approach called cost-complexity pruning. In this approach, the largest tree, T_0, is grown, terminating the splitting process only when some minimum node size has been reached, n_0. A subtree defined as $T \subset T_0$ which is any tree that can be obtained by pruning T_0. The pruning is done by collapsing any number of internal nodes. Further, index terminal nodes by m with node m representing region R_m, and let $|T|$ be the number of terminal nodes or leaves in T. Then

$$N_m = \# \{x_i \in R_m\},$$

$$\hat{c}_m = \frac{1}{N_m} \sum_{x_i \in R_m} y_i,$$

$$Q_m(T) = \frac{1}{N_m} \sum_{x_i \in R_m} (y_i - \hat{c}_m)^2,$$

and the cost-complexity criterion is defined as

$$C_\alpha(T) = \sum_{m=1}^{|T|} N_m Q_m(T) + \alpha \, |T| \, .$$

The goal then is to find the subtree $T_\alpha \subset T_0$ to minimize the cost-complexity criterion, $C_\alpha(T)$, for each α. It is intuitive from how $C_\alpha(T)$ has been defined that for large values of α, smaller trees will minimize $C_\alpha(T)$ while larger trees will minimize $C_\alpha(T)$ for smaller values of α. We can think of alpha as some sort of penalty for the size of the tree. That is, if we can imagine that large trees will get a bigger penalty, then we can see that smaller trees will minimize $C_\alpha(T)$ and vice versa. It can be shown for each α that there exists a unique subtree T_α that minimizes $C_\alpha(T)$. Hastie et al. (2009) suggest weakest-link pruning to find T_α.

This is done by successively collapsing the internal node that gives the smallest per-node increase in $\sum_m N_m Q_m(T)$ and this is continued until the root node tree is produced. That is, at each pair of leaf nodes with a common parent, the error on testing data can be evaluated and we see whether the sum of squares would be smaller by removing those two leaves and making their parent a leaf. This is performed using a cross-validation (CV) technique. Breiman et al. (1984), Ripley (1996), and Venables and Ripley (1999) all provide further details. Estimation of α is achieved by k-fold cross-validation: $\hat{\alpha}$ is chosen to minimize the cross-validated sum of squares. An example of how this is done in practice is shown below.

The complexity parameter, *cp*, is the parameter α divided by the relative error, $R(T_\phi)$ for the root tree. That is, it is proportional to α. As was mentioned earlier, larger trees are associated

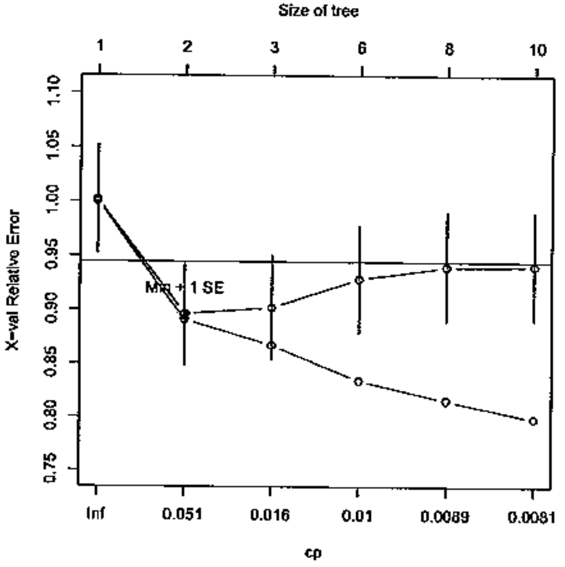
Table 1: Complexity parameter and other measures for trees.

cp	Number of Splits	Relative Error	CV Error	CV SD
0.1092	0	1.0000	1.002	0.050
0.02358	1	0.8908	0.897	0.049
0.0110	2	0.8672	0.902	0.048
0.0091	5	0.8342	0.929	0.050
0.0087	7	0.8160	0.940	0.051
0.0075	9	0.7987	0.941	0.051

with large values of α or cp and vice versa. The cross-validated error and cross-validated standard error are computed in the R (R Development Core Team, 2009) package `rpart` . In Table 2, we see the CV error for $T_{\hat{\alpha}}$ for each value of cp. For example, for a cp of 0.011 the subtree that minimizes $C_\alpha(T)$ has 2 nodes.

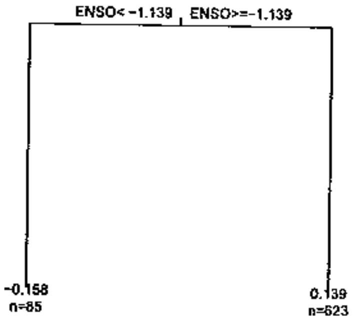
The best tree can be chosen based on either the minimum CV error or the 1-SE rule. One criterion for choosing an optimal tree is using the minimum CV error rule where the tree with the smallest CV error is chosen. The 1-SE rule is to choose the smallest tree with a cross-validated error within 1-SE of the minimum cross-validated error. In our example, 1-SE gives 0.89647 + 0.048510 = 0.94498, so the optimal tree is still the 1-split tree. In some cases, like we observed above, the two rules may select the same tree. The minimum CV error rule will often select larger trees than the 1-SE rule. Figure 4 contains a plot which summarizes the information in Table 2.

Figure 4: Plot of Cross-validated error versus Complexity parameter.



The pruned tree, based on $cp = 0.0236$, is shown in Figure 5. We can see that the predictor NAO is not used in the split in the pruned tree. The tree shows that above-average temperatures are associated with ENSO values greater than -1.139 whereas below-average temperatures are associated with ENSO values less than -1.139. The tree potentially suggests that there is one main threshold in ENSO in terms of explaining variability in temperature anomaly, not two as have been suggested elsewhere.

Figure 5: Pruned Tree.



3 GAM and GAMM

There are some situations where either because of *a priori* assumptions or information from plots, linearity assumptions are suspected to be false. One approach in modeling where linearity assumptions are not met is to transform the predictors and fit a linear model. This may not be the best thing to do in some cases. When it comes to nonlinear regression effects, Generalized Additive Models, GAMs, can also be used. The difference between a linear model and a GAM can be seen by taking a look at their functional forms:
for a GAM,

$$Y_i = \alpha + s(X_i) + \epsilon_i;$$

and for a linear model,

$$Y_i = \alpha + X_i\beta + \epsilon_i.$$

The only difference in the two equations is that in the GAM we incorporate a smooth function of the predictor(s), denoted as $s(X_i)$, whereas we only incorporate the raw form of the predictor(s) in the linear model. Wood (2006) discusses the construction of GAMs using spline basis functions and a penalized regression spline approach where the amount of smoothing for each predictor is controlled by a smoothing parameter. This is similar to how *cp*'s are used to select an optimal tree size. The smoothing parameter, λ_j, for each variable is selected using Generalized Cross-Validation (GCV). GCV is therefore used to estimate the wiggliness of the spline which determines the estimated degrees of freedom (*edf*). We can have more than one smooth function, $s(x)$, in our model in which case each would have its own *edf*.

Further GAMM (Generalized Additive Mixed Models) are an extension of GAM that allow correlations, spatial or temporal or both, to be incorporated into GAMs (Wood, 2006; Zuur et. al., 2009). GAMMs estimate the amount of smoothness of function differently using mixed model techniques. The amount of smoothing is chosen by estimation of random effects variances associated with the spline coefficients. We can state a GAMM as follows:

$$Y_i = \alpha + s(X_i) + Z_i b_i + \epsilon_i;$$

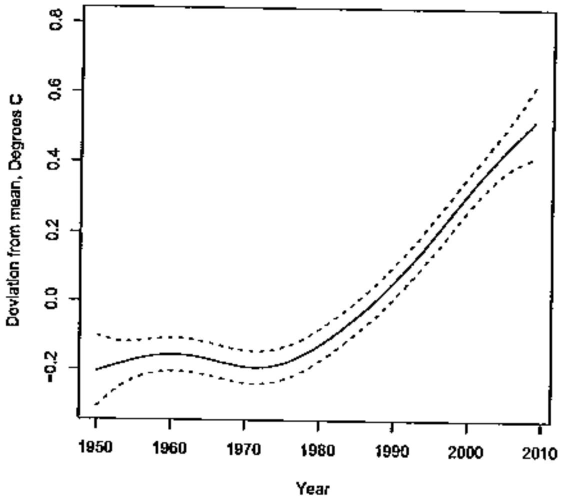
where $b_i \sim N(0, {\sigma_b}^2)$ is a random effect. GAMMs are more computer-intensive than GAMs which can result in numerical issues in some applications. We use the function `gamm` in the package `mgcv` (Wood, 2006) in R to fit the additive model. A potential additive model in our application is

$$NH_t = \alpha + s(Time) + \epsilon_t$$

where $\epsilon_t = \Phi\epsilon_{t-1} + w_t$, modeling correlation over time using a first-order autoregressive error structure.

The *edf* for $s(Time)$ in the GAMM are about 7.6, and are strong evidence of a non-linear trend in this model (p-value < 0.0001) as seen in Figure 6. The average temperature seems to increase

Figure 6: Smoothing function of Time in the Optimal GAMM.



slightly around 1960 and then increase more linearly after 1975. The estimated AR coefficient is $\hat{\phi} = 0.577$, showing a pretty strong month-to-month autocorrelation in temperature.

3.1 Trees for correlated data

Trees can also be used for correlated data. Abdollel et al. (2002), Lee (2005), and Segal (1992) demonstrate how trees can be used to handle correlated data in the context of repeated measures. Segal (1992) explicitly models only the the set of time periods that are in the training data, which means that the resulting tree cannot be used to forecast values from future time periods for individuals in the training data. Also, the approach adopted by Galimberti and Montanari (2002) require the covariances of the errors and random effects must be estimated prior to fitting the regression tree. In addition, the random effect values are not estimated which means they cannot be used for prediction.

In medical experiments, among others, where we have a random sample of subjects, one of the goals of statistical analyses is to estimate the variability among the subjects. The general idea is that there is a random effect due to each individual and variability of these random effects is what we refer to as subject-to-subject variability. Similar approaches have been developed by A. Hajjem et al. (2008) and Sela and Simonoff (2009) to modify trees to handle these types of random effects.

For the rest of this paper, we will adapt the latter's approach to modeling. They propose Random Effects Expectation-Maximization (RE-EM) trees. The tree and random effects portions of the model are estimated through a sort of expectation-maximization process. The underlying model is the general mixed effects model with additive errors. This is given by:

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

where $b_i \sim N(0, {\sigma_b}^2)$ is a random effect, f is a tree function and ϵ_{it} is some autoregressive (AR) error within subject. The only difference here is that we do not make any assumptions about the parametric form of f. A tree structure is used to estimate f in the expectation-maximization process. Sela and Simonoff use the following algorithm to do this:

1. Initialize the estimated random effects, $\hat{\mathrm{b}}_\mathrm{i}$, to zero.

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

   2.1. Estimate a regression tree approximating f, based on the target variable, $y_{it} - \mathrm{Z}_{\mathrm{it}}\hat{\mathrm{b}}_\mathrm{i}$, and predictors, $x_{it.} = (x_{it1}, ..., x_{itK})$, for $i = 1, ..., I$ and $t = 1, ..., T_i$.

   2.2. Use this regression tree to create a set of indicator variables, $I(x_{it.} \in g_p)$, where g_p ranges over all of the terminal nodes in the tree.

   2.3. Fit the linear mixed effects model, $y_{it} = Z_{it}\hat{\mathrm{b}}_\mathrm{i} + I(x_{it.} \in g_p)\mu_p + \epsilon_{it}$. Extract $\hat{\mathrm{b}}_\mathrm{i}$ from the estimated model.

Going back to the temperature anomaly data, we saw from the plot of the series that it contained a smooth non-linear trend. This makes it necessary to modify, somewhat slightly, RE-EM trees for our purposes. We call the modified RE-EM trees GAMM-trees. The idea is still the same except that this time we are estimating some smooth function of one or more of our predictors in addition to estimating a tree using the same EM process. The method could use GAMs or GAMMs, but is illustrated using the more general GAMMs. We only need to slightly modify the algorithm above to achieve this. The underlying model is only slightly different from what we saw above:

$$y_t = f(x_{t1}, ..., x_{tK}) + s(x_1) + ... + s(x_k) + \epsilon_t$$

where $s(x_i), ..., s(x_k)$ are the one or more smooth functions. The algorithm is detailed below:

- Initialize the estimated smooth functions, $\hat{s}(x_i)$, to zero.

- Iterate through the following steps until GAMM likelihood converges:

  1. Estimate a regression tree approximating f, based on the target variable, $y_t - \sum \hat{s}(x_i)$, and predictors, $\mathbf{x_t} = (\mathbf{x_{t1}}, ..., \mathbf{x_{tK}})$, and $t = 1, ..., T$.

  2. Use this regression tree to create a set of indicator variables, $I(\mathbf{x_t} \in \mathbf{g_p})$, where g_p ranges over all of the terminal nodes in the tree.

  3. Fit the GAMM, $y_t = \sum s(x_i) + I(\mathbf{x_t} \in \mathbf{g_p})\mu_\mathbf{p} + \epsilon_\mathbf{t}$. Extract $\sum \hat{s}(x_i)$ from the estimated model.

As can be seen, the underlying model has two parts: a tree function f and a set of smooth functions, $\sum s(x_i)$. GAMM-trees would be preferred to RE-EM trees when smooth functions of one or more of our predictors rather than random effects estimation is desired. The predictors that go into f are those that have thresholds in them, are thought to interact or it is unknown whether they might be important. Predictors that are thought to relate to the response through a smooth function enter the model through the spline effects.

4 Results

We demonstrate the methods discussed above in the analysis of the Northern Hemisphere temperature anomaly series. The model we propose is

$$NH_t = \alpha + f(ENSO_t, NAO_t) + s(Time) + \epsilon_t$$

where $\epsilon_t = \phi\epsilon_{t-1} + w_t$.

As we discussed above, how the predictors are handled in the GAMM-tree is based on prior information about how the predictors may impact the response. The tree portion of the model is based on ENSO and NAO since the literature suggests the possibility of interesting thresholds in these variables and interactions which would make a tree suitable. The smooth function in time is chosen because we saw a smooth non-linear trend in the temperature anomaly series. A first-order autoregressive (AR(1)) error structure is used. We shed more light on the reason for this below. We show the trend over time in the temperature anomaly series with a 20-split tree GAMM-tree model in Figure 5. To obtain the optimal tree, we have to prune the resulting tree. One way of doing this is described below.

Figure 7: Trend over time in Anomaly (With a 20-split tree).

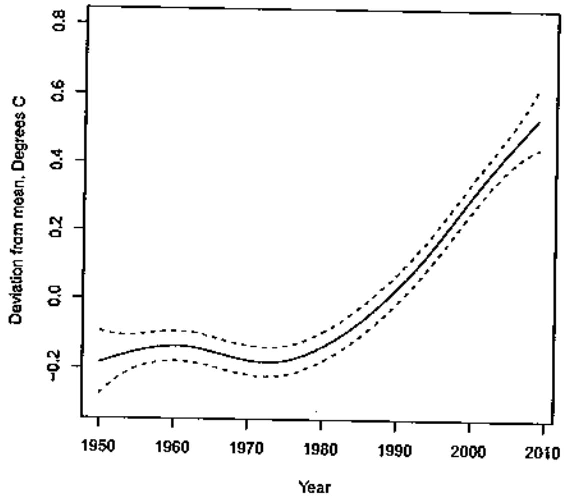
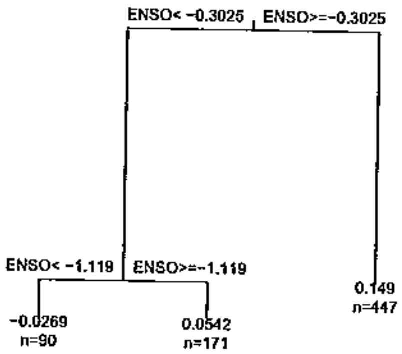


Figure 8: Pruned NH Temperature Anomaly Tree from GAMM-tree.



4.1 Iterative Tree Pruning

Having estimated an initial tree in a GAMM-tree, it is desirable to attempt obtain an optimal tree by pruning this tree. To identify an optimal GAMM-tree, we propose the iterative pruning algorithm.

1. Fit initial GAMM-tree model with a large tree, T, and select preliminary tree size, M_0, based on minimum CV of $Y_i - \hat{s}_T(x)$
2. Refit GAMM-tree with tree of size M_0 and estimate $\hat{s}_M(x)$.
3. Fit a tree to $Y_i^* = Y_i - \hat{s}_M(x)$.
4. Find next minimum CV tree size M^*.
5. Repeat steps 2 to 4 until no change in selected tree size.
6. Do steps 1 to 5 repeatedly and select tree with highest relative frequency.

There is another approach to selecting the optimal GAMM-tree where the tree of size M is cross-validated based on a trend for a size M tree, requiring estimation of the GAMM-tree for each potential tree size in the cross-validation (CV) table. Using this method, pruning is done conditional on the estimated trend. Although this method may be superior to the one described above, it is more computationally-intensive than the iterative approach discussed above.

Using the iterative tree pruning algorithm, the tree in Figure 8 is obtained 23 out of 30 times. The tree shows that negative ENSO events (El Nino) may be associated with above average temperatures while positive ENSO events (La Nina) may be associated with below average temperatures. The trend is shown again in Figure 9 but this time with the tree pruned. This does not look very different from the the trend with the tree unpruned (also has similar *edf* and p-value).

Having obtained the optimal tree, we have to refit our model and update the node estimates. Below are the estimates of NH temperature anomaly at the 3 terminal nodes, after adjusting for trend. The AR coefficient ϕ is also estimated as $\hat{\phi} = 0.52$ with a 95% confidence interval (0.45, 0.59). The tree estimation algorithm chooses the best split when growing a tree but there are other splits, called competitor splits, that may lead to the same or similar amount of reduction in residual

Figure 9: Trend over time with Pruned Tree (2-split tree).

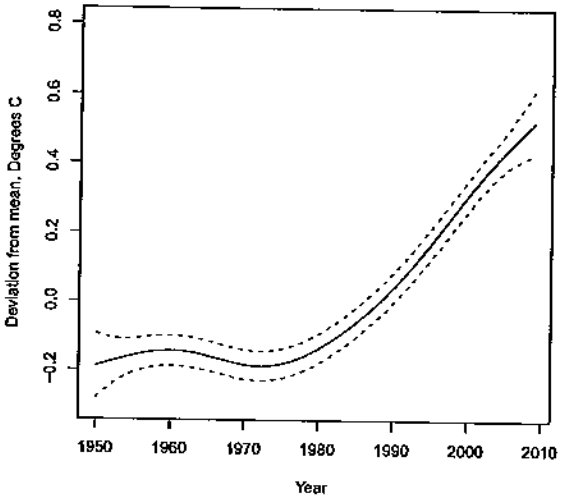


Table 2: Node Estimates and Confidence Intervals.

Node	Estimate	SE	CI
ENSO $\in (-\infty, -1.12)$	0.03037	0.02610	(-0.022, 0.083)
ENSO $\in [-1.12, -0.30)$	0.06566	0.01757	(0.031, 0.100)
ENSO $\in [-0.30, \infty)$	0.13144	0.01370	(0.104, 0.159)

sum of squares. In the pruned tree in Figure 8, the competitor split for each of the splits was NAO. However, those splits were considered as 'next best' since they provided a comparatively smaller reduction in residual sum of squares.

The results we obtain from the final tree (pruned tree) adjusted for trend are only slightly consistent with what the literature says about ENSO and NAO in terms of the thresholds that may exist in them. The tree suggests to us that ENSO ≥ -0.3025 and ENSO < -0.3025 may be important in predicting temperature anomalies given a trend adjustment. Literature suggests that the important thresholds in ENSO are ENSO > 0.5 and ENSO < -0.5 but we did not identify any splits on ENSO > 0.5 as being important. NAO is not found to be as important as ENSO.

4.2 Discussion

We show the sample Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots in Figure 10. These are based on just the initial GAMM unadjusted for the tree. The ACF seems to be dampening out for the most part which is somewhat consistent with a first-order autoregressive error structure or at least some sort of autoregressive error. The PACF has large lag 1 and 2 values and possibly at lag 4 as well, but generally cuts off. This suggests some sort of autoregressive error would be considered. AR(1) errors can serve as a simple approximation to other correlation structures, so we use this until methods to select GAMM-trees across correlation structures are developed.

Diagnostic plots and ACF and PACF plots are shown for the pruned GAMM-tree in Figure 11. There may be unconvincing hints of non-constant variance in the residuals versus fitted values plot but it does not look like we have any serious problem. The normal probability suggests slightly heavier tails than expected but no distinct outliers. The ACF and PACF plots are not very different from what we saw before. It may be possible to incorporate a more a complex error structure in the future.

Figure 12 plots the trend and the fitted values from the optimal tree (GAMM-tree) and the observed temperature anomalies. The GAMM-tree, super-imposed on the trend, shows how the tree adjusts the trend over time to explain variability in temperature. We can see from the plot

Figure 10: ACF(left) and PACF(right) Plots.

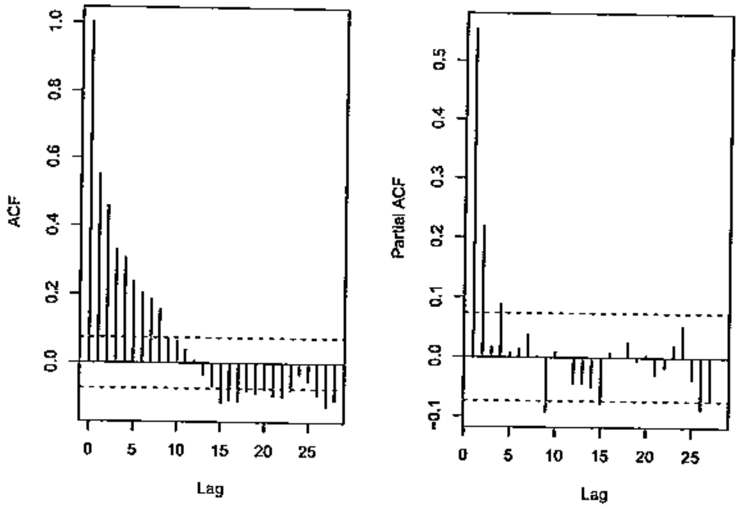
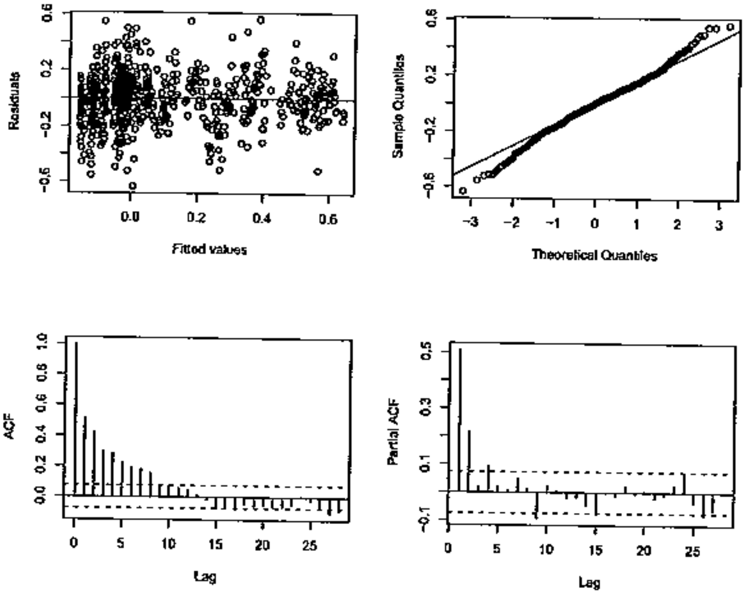
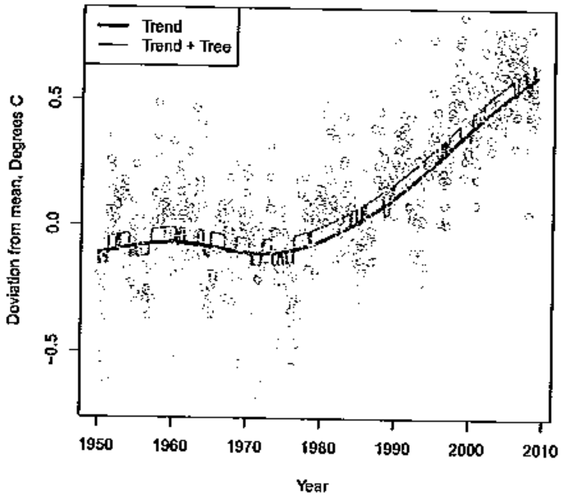


Figure 11: Residuals versus fitted (top left), normal probability plot (top right), ACF (bottom left) and PACF (bottom right) for the GAMM-tree.



that the tree only explains a small amount of variation relative to the trend. This suggests that either ENSO has a minimal effect on Northern Hemisphere temperatures or that our methods are not detecting its impacts.

Figure 12: Trend, Combined Tree and Trend, and Observations.



We have looked at trees in general and their applicability to statistical modeling. We have shown how GAMM-trees, a modification of RE-EM trees, can be applied to the Northern Hemisphere temperature anomaly series. In doing this, we improved on Sela and Simonoff's RE-EM trees approach by introducing an algorithm (iterative pruning algorithm) to identify the optimal tree for the GAMM-tree model. This approach to identifying the optimal tree can be applied to RE-EM trees also with minimal modification. There is also potential for spatial or spatio-temporal applications of GAMM trees.

Our results suggest a generally increasing trend in temperature over the last couple of decades. This is evident from the plot in Figure 12. However, we still have quite a bit of variability that is yet to be explained. The GAMM-tree does a good job of adjusting the trend over time, but it still needs some improvement. A bigger tree than we obtained through the iterative pruning process would explain more variability in temperatures, perhaps the method is conservative. The lack-of-fit could also be attributable to some covariate(s) that we did not include in the model.

GAMM-trees are a new approach for dealing with complex datasets that may contain both smooth and threshold or interacting non-linear effects. Finding a method to efficiently select

GAMM-trees across correlation structures is of key importance so far as future work is concerned. The concept of trees, fairly new to statistics, and the GAMM-tree approach to modeling would be enhanced greatly by further research in the area of regression trees.

References

[1] Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees.* Wadsworth International.

[2] Faraway, J. (2006). *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models with S (4 ed.).*Chapman & Hall/CRC.

[3] 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.

[4] Hastie, T., Tibshirani, R., and Friedman J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* New York: Springer.

[5] Jones, P. D. and Moberg A. (2003). *Hemispheric and Large-Scale Surface Air Temperature Variations: An Extensive Revision and Update to 2001.* Journal of Climate.

[6] R Development Core Team (2009). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing.

[7] Sela, R., and Simonoff, J. (2009). *RE-EM Trees: A New Data Mining Approach to Longitudinal Data.* Statistics Working Papers Series.

[8] Venables, W., and Ripley, B. (2002). *Modern Applied Statistics with S (4 ed.).*,New York: Springer.

[9] Wigley, T. M. L. (2000). *ENSO, volcanoes and record-breaking temperatures.* Geophys. Res. Lett., 27, 4101-4101.

[10] Wolter, K. and Timlin, M.S. (1993), Monitoring ENSO in COADS with a seasonally adjusted principal component index. *Proc. of the 17th Climate Diagnostics Workshop*, Norman, OK, NOAA/N MC/CAC, NSSL, Oklahoma Clim. Survey, CIMMS and the School of Meteor., Univ. of Oklahoma, 52-57.

[11] Wood, S (2000). *Modeling and Smoothing Parameter Estimation with Multiple Quadratic Penalties.* Journal of the Royal Statistical Society, Series B 62, 413-428.

[12] Wood, S (2006). *An Introduction to Generalized Additive Models with R.* Boca Raton, FL: CRC Press.

[13] Zuur, A. F., Ieno E. N., Walker, N. J., Saveliev, A. A., Smith, G. M. (2009). *Mixed Effects Models and Extensions in Ecology with R.* New York: Springer.