Designing Mixedwood Experiments

Statistical Tools for Mixedwood Studies

VERA SIT

ABSTRACT

This paper discusses two statistical tools used for analyzing mixedwood data. The first involves regression when the data come from several populations. The second involves repeated measures analysis. In regression analysis where the data come from several populations, we must first check that a common regression model is suitable for each population before fitting a general model to all data. This can be accomplished by including the population effect in the regression model. This technique is demonstrated with an example. When data are collected repeatedly on the same subjects, we must take into account the dependence structure of the data in the analysis. Five approaches commonly used for analyzing repeated measures data are: 1) treating repeated measures as replicates, which is not recommended because the method ignores the correlation in the data; 2) analyzing the data at each time, which is legitimate but quite restrictive because it does not allow the examination of the response trend over time; 3) split-plot in time analysis, which is unsuitable for mixedwood studies due to the required sphericity assumption; 4) multivariate analysis, which does not require the sphericity assumption, but tends to have low power when the number of repeated measures levels is large relative to the number of experimental units; and 5) response curve analysis, which directly addresses the shape of the response trend over time and is more powerful than method 4).

INTRODUCTION

Proper data analysis is one of the key elements in the success of mixedwood studies. This paper examines two statistical tools for analyzing mixedwood data.

One of the objectives in a mixedwood study is to investigate the relationship between variables. Often, studies are carried out at several sites so as to cover the full range of variables of interest. Before developing a general model for all the sites, we must check that the assumption that using

a common model across sites is valid. The regression technique for analyzing these types of data is presented in the first part of this paper, demonstrated with an example.

Another objective of a mixedwood study is to look at response trend with respect to time. In this case, subjects are measured repeatedly over time. The second part of this paper describes the strengths and weaknesses of five common approaches used to analyze repeated measures data.

REGRESSION WITH DATA OBTAINED FROM SEVERAL POPULATIONS

We begin with a typical regression scenario. A study is designed to investigate the relationship between growth of a seedling and the amount of light reaching the seedling. One way to carry out the study is to randomly select a number of locations on a site, and, at each location, measure the height of the nearest seedling and the light at the top or mid-crown of that seedling.

Regression analysis can be used to estimate the relationship between light and seedling height growth. Like all statistical techniques, regression analysis has a number of assumptions:

1. The dependent variable (X) is non-random, observed with negligible error.
2. The experimental errors are uncorrelated, with zero mean, and constant variance. (The experimental errors must be normally distributed for the estimators to attain the property of minimum variance of the class of unbiased estimators.)
3. The underlying regression model is correct.

Let's modify the scenario slightly. Suppose the experimenter is interested in estimating the relationship between light underneath a broadleaf canopy and basal area of the broadleaf species. In order to have data that cover a full range of basal area, the experimenter needs to do the study at a number of sites, each thinned to a different density. At each site, one would locate a number of sampling points; at each sampling point, light and basal area would be measured within a circular plot centred on the measurement point. Since each site has a different density, the radii of the circular plots vary with approximately an equal number of trees in each plot. For this modified study, how should the experimenter do the analysis? What technique could be used? How does this set-up differ from the first situation?

A common approach is to treat the data the same as in the first case and use regression on all the data. The assumptions of independence, random errors with zero mean, and equal variance would be reasonable if the data were generated from a random process. However, the last assumption that a single model would be adequate for all sites might not be met if the sites had different characteristics. Therefore, this approach is not advisable unless one can justify a single model for all sites.

A more appropriate approach is to fit a separate model to each site, check the model parameters, and combine all the site data only if their

parameters are similar. We will illustrate the two approaches with an example.

Data used in this example come from a study done by Phil Comeau in 1993 (Comeau, B.C. Ministry of Forests, unpublished). One of the objectives of the study was to develop a relationship between the fraction of diffuse light penetrating the canopy (i.e., diffuse non-interceptance, DIFN) and birch basal area. The study was carried out at six sites (Mica Lake, TT1, TT2, Burton Creek, Adams, and Barrier). DIFN measurements were collected from a total of 31 measurement points from the six sites. At each measurement point, the diameter of every tree within a circular plot centred on the measurement point was recorded. Radius of the measurement plot varied from 1.78 m to 6.0 m, depending on the density and size of the birch. There were approximately 50 trees in each circular plot. Some of the sites had received juvenile spacing. Measurement plots were established on both spaced and unspaced plots. The sites differed in age, tree size, and density. As a result, each site covered a part of the full range of basal area. This was one of the reasons to incorporate several sites in the study. It was also of interest to see if the relationship applied to several sites. In other words, the sampling was stratified by site.

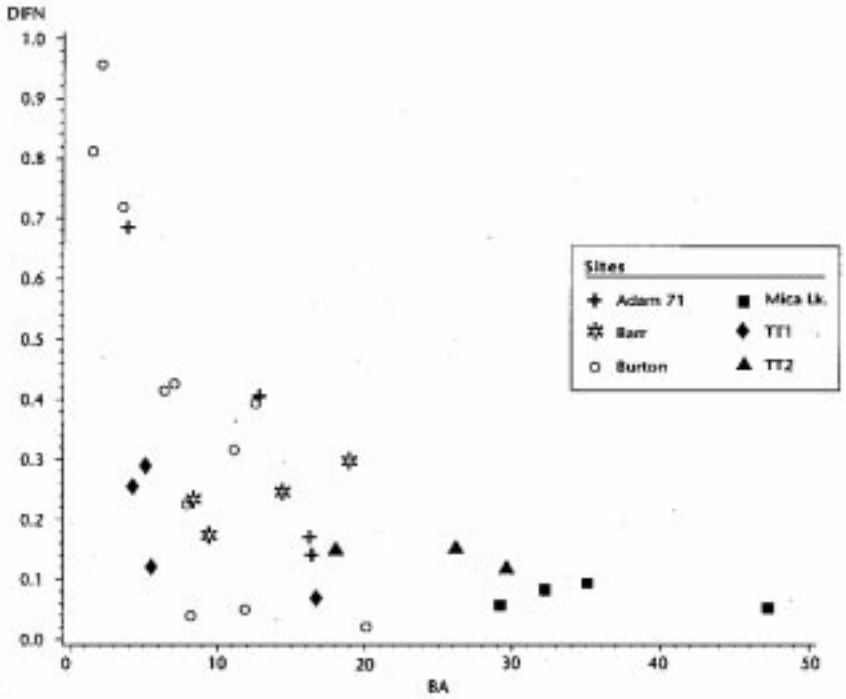

FIGURE 1 *DIFN versus basal area plot.*

Regression was used to develop the relationship between DIFN and birch basal area. The plot (Figure 1) of DIFN versus basal area (BA) showed that the relationship was non-linear. Natural log transformation on BA was performed to linearize the relationship. Figure 2 is a plot of DIFN versus LN_BA. As a first attempt, DIFN was regressed on LN_BA. The fitted model was:

$$\text{DIFN} = 0.814 - 0.226\ \text{LN\_BA}$$

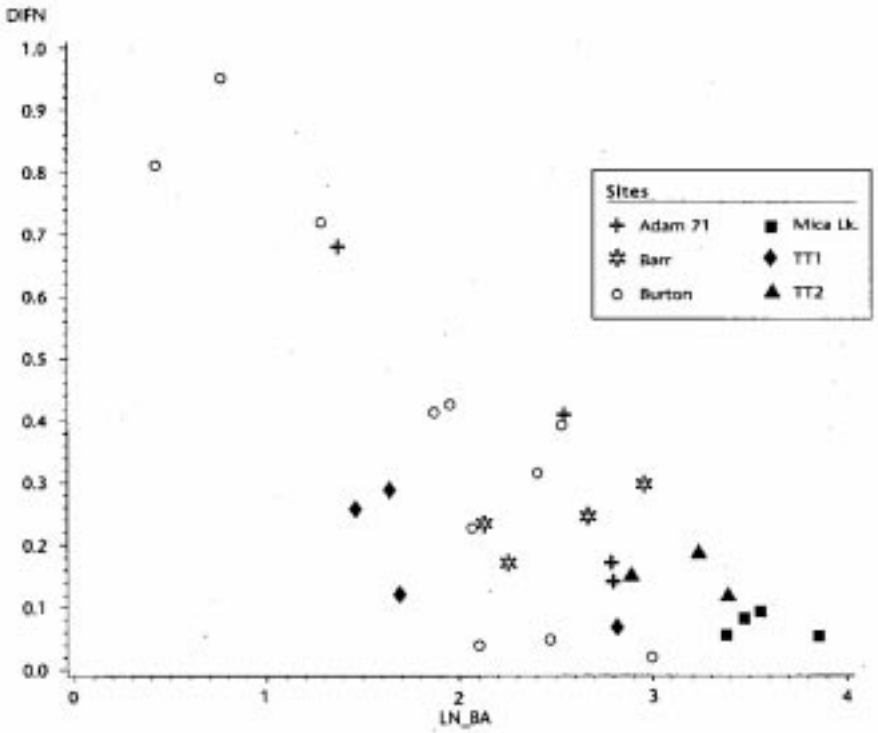

FIGURE 2 *DIFN versus LN (basal area) plot.*

R-square of this model was 0.61. This general model assumed that site-to-site variation was small and that a single model was adequate for all sites. Before accepting this model, we must check this assumption. This can be done by including SITE and LN_BA*SITE in the regression model to fit a separate line (using the same model structure) to data from each site. If the interaction term LN_BA*SITE was significant, we would conclude that the fitted lines for each site were not parallel and that an overall model for all sites was not possible. Otherwise, we would conclude that all the fitted lines had the same slope. We could then remove the LN_BA*SITE term from the model and proceed to check if all fitted lines had the same y-intercept. The analysis was performed using SAS, and the results are summarized in Table 1.

TABLE 1 *ANOVA table for testing if regression lines for the six sites are parallel*

Source	DF	Type III SS	Mean square	F value	Pr > F
LN_BA	1	0.03611734	0.03611734	2.53	0.1292
SITE	5	0.19107502	0.03821500	2.68	0.0560
LN_BA*SITE	5	0.14420193	0.02884039	2.02	0.1243

Because the p-value for LN_BA*SITE was large ($p = 0.1243$), we concluded that the fitted regression lines were parallel. The term LN_BS*SITE was then dropped from subsequent models.

To check if the lines had identical y-intercepts, DIFN was regressed on LN_BA and SITE. The fitted equations were:

Adams:	DIFN = 1.0635 − 0.3 LN_BA
Barrier:	DIFN = 0.9867 − 0.3 LN_BA
Burton:	DIFN = 0.9701 − 0.3 LN_BA
Mica:	DIFN = 1.1438 − 0.3 LN_BA
TT1:	DIFN = 0.7542 − 0.3 LN_BA
TT2:	DIFN = 1.0858 − 0.3 LN_BA

R-square for this model was 0.76. The analysis results are given in Table 2.

TABLE 2. *ANOVA table and parameter estimates for fitted parallel regression lines to the sites*

Source	DF	Type III SS	Mean square	F value	Pr > F
LN_BA	1	0.84487264	0.84487264	48.43	0.0001
SITE	5	0.25878496	0.05175699	2.97	0.0328
Parameter		Estimate	T for H0 Parameter = 0	Pr > |T|	Std Error of Estimate
INTERCEPT		1.085833784	6.91	0.0001	0.15709513
LN_BA		−0.300342353	−6.96	0.0001	0.04315700
SITE Adam71		−0.022273900	−0.21	0.8365	0.10671435
Barr		−0.099072958	−0.94	0.3558	0.10512233
Burton		−0.115780188	−1.13	0.2695	0.10232665
MicaLk		0.057969053	0.57	0.5761	0.10221419
TT1		−0.331672753	−2.88	0.0084	0.11499287
TT2		0.000000000			

The intercept for TT1 was significantly different from the others ($p=0.0084$). This implies that a general model for all sites was not appropriate. The fitted model for each site and the overall model are plotted in Figure 3.

Further examination of the sites revealed that more than 30% of the trees in site TT1 were cottonwood, with only a minor component of cottonwood in the other sites. Since birch basal area was the dependent variable in the analysis, we might have underestimated the "basal area" effect at site TT1. If we want to develop a general model for all sites, we may consider using total basal area (instead of just birch) as the independent variable, or incorporate species in the model.

Regression analysis is valid only if the underlying model is correct. When stratified sampling is used to obtain data for regression analysis, the stratification factor should be included in the model. An over-simplified model could produce incorrect results.

REPEATED MEASURES ANALYSIS

Let's consider a simple experiment. To assess the effects of three methods of site preparation (v-plow, hand screef, and a control), 30 rows of 25 seedlings were established at a single trial site. The three site preparation treatments were randomly assigned to each row of seedlings so that each treatment was applied to exactly 10 rows of seedlings. Seedling diameter and height were measured at the time of planting and 6 years later.

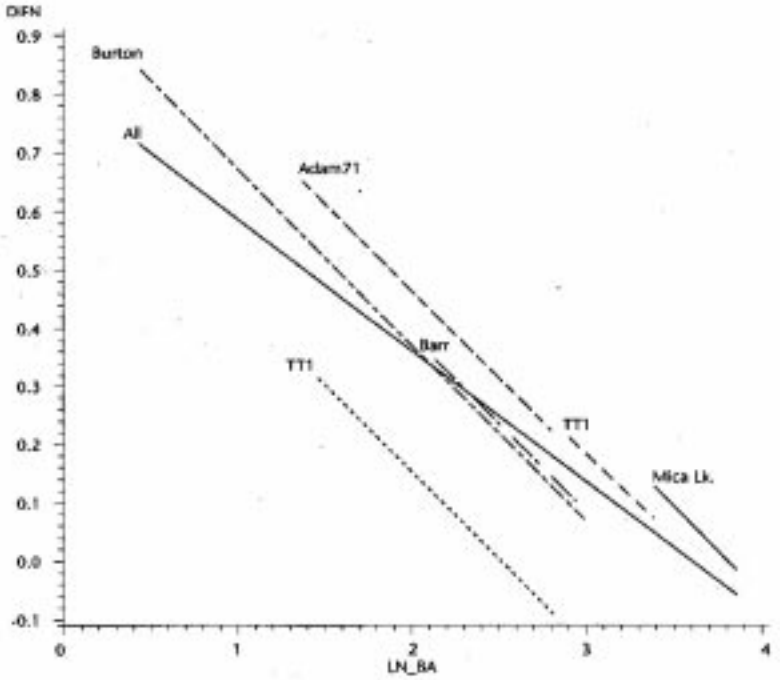

FIGURE 3 *Regression models: with and without site effect.*

Analysis of variance can be used to compare the effect of the three site-preparation treatments on seedling growth. Measurements taken at the time of planting can be used as covariables to increase the sensitivity, or power, of the test. Suppose measurements were taken at the time of planting and annually for the next six years. The experiment would result in repeated measures data. Repeated measures data are characterized by the fact that the same unit is measured repeatedly over time or space. Time or space is the repeated measures factor. The levels of the repeated measures factor cannot be assigned at random. For example, year 3 must come after year 2, or the top layer of soil must be above the bottom layer of soil. Consequently, the observations are serially correlated.

Serial correlation, also called autocorrelation, is the correlation between observations made at two, not necessarily consecutive, times. In the site-preparation treatment example, seedling heights are serially correlated because height of a seedling in any given year depends on its height in previous years. This dependence in the data violates the assumption of independence on which ANOVA and Regression methods are based.

A main objective of repeated measures analysis is to make comparisons between the site-preparation treatments and to examine the time effect for the site-preparation treatments. There are three main repeated measures hypotheses:

1. There is no interaction involving time. That is, if we plot seedling height versus time for each seedling, the lines corresponding to the three site preparation methods are parallel. This implies similar growth patterns for the different site preparation methods.

2. There is no change over time. The lines in the seedling height versus time graph are all flat. This implies no growth over time. This is usually not an interesting hypothesis because changes over time are expected in a repeated measures study.
3. There is no overall difference between groups. In this case, the lines corresponding to the three site preparation methods are identical. This implies no site preparation effect.

Usually, hypothesis 1) is tested first. If it is not rejected, then hypotheses 2) and 3) will be tested separately.

There are several ways to analyze repeated measures data. Five common approaches are described below:

1. ANOVA, treat repeated measures as replicates

   This approach ignores the dependence structure in the data and treats the yearly data as replicates. Since it violates the basic ANOVA assumption of independence, *its use is not recommended.* Also, because time effect is included in the experimental error, the experimental error is inflated, resulting in more conservative tests.

2. ANOVA at each time

   In this approach, one would carry out an ANOVA for data collected in each repeated measure factor level (e.g., yearly data). Hence, serial correlation between observations over time is ignored. Sometimes, a researcher may use this approach to locate the point in time at which treatment responses become significantly different. Although individual ANOVAs are valid, the separate tests are still not independent. Thus, inferences about the response through time are invalid. Also, the time at which significant differences are detected is a function of the number of experimental units. A change in the number of replications changes the point in time at which significance is detected. *This approach does not allow the estimation of the response trend over time.*

3. Split-plot in time analysis

   In this case, the duration of the study is viewed as the main plot which is divided into yearly "split-plots" for the time factor. In addition to the usual ANOVA assumptions of equal variances (over time and experimental units) and independence of measurements made on different experimental units, the correlation (covariance) between any pair of repeated measures is assumed to be the same for all times and all experimental units. For example, the correlation between row heights is the same for all pairs of years and all rows. The assumptions of homogeneous variances and covariances is known as the *sphericity assumption.* It is often not achievable in biological studies. Therefore, *this approach may not be suitable for most forestry studies.*

4. Multivariate analysis

   In the multivariate approach, an observation consists of a vector of measurements at the repeated measures factor levels. Comparisons of group means are replaced by comparisons of the group mean vectors;

error mean squares are replaced by error matrices; F-tests are replaced by multivariate test statistics (Wilks' Lambda, Pillai's trace, Hotelling-Lawley trace, and Roy's greatest root). Two vectors are declared different if at least one element in the vector is different for the two groups. For example, if the average seedling growth measured at year 2 for the three site preparation methods were different, we would conclude that the growth trends over time were different among seedlings treated by the three site preparation methods.

This method of analysis does not require the sphericity assumption. The power of the multivariate test is quite low due to the use of the error matrix. This becomes an important consideration when the number of repeated measures levels is large and the number of experimental units is small. Finally, *multivariate analysis does not exploit the underlying continuity of the response functions in time, thus may only indirectly answer the question of trend over time.*

5. Response curve analysis

A response curve is a plot of the data over time. In this approach, a curve is fitted to each set of seedling data (over time) to get a set of fitted parameters (a, b, . . .). Each set of parameters summarizes all the information in the response curve for each seedling. These fitted parameter values will be used as data in a multivariate analysis as described in method (4). This method of multivariate analysis is more powerful than that using raw data alone because the size of the vector is greatly reduced due to fewer time points and parameters in the response curve. If the multivariate analysis of variance of no difference among the parameter vectors is rejected, then univariate analysis of variance on each parameter may be carried out to identify which parameter is different among the treatment levels. This approach allows one to directly compare special features (e.g., maximum, minimum, end point, point of inflection) in the response curve.

The success of this approach depends on the suitability of the response curve. The selected curve must be adequate for all seedlings. It should have a minimum number of parameters and the parameters should correspond to interesting features on the curve. Sometimes, the curve may need to be re-parameterized to identify the features of interest. Many data may be needed to adequately define the response curve. If the response curve equation is unknown, or if there are not enough data, polynomials or any orthorgonal combinations of data may be used instead. See Sit and Poulin-Costello (1994) for a discussion on curve fitting techniques.

The advantages of the response curve analysis are:

1. The dimension of the response vectors are reduced from the number of time points to the number of parameters in the response curve, therefore simplifying the multivariate analysis.
2. The analysis directly addresses the shape of the response function by examining easily interpreted coefficients.
3. The analysis can be performed on only those experimental units with necessary data to estimate the response function of interest. Thus, small amounts of incomplete records can be used.

Some interesting papers on repeated measures analysis are: Moser et al. (1990), Potvin and Lechowicz (1990), Meriedith and Stehman (1991), Gumpertz and Brownie (1993), and Nemec (1996). An extensive list of references can be found in Koch et al. (1980).

CONCLUSION

Regression and repeated measures analysis are important tools for analyzing mixedwood data. We must take care to ensure that these methods are carried out properly. In regression analysis, one important but often ignored assumption is that the underlying model is correct. In this paper, we examined the consequences when this assumption was violated.

Five common approaches for analyzing repeated measures data are presented. The first three methods described are either incorrect or inadequate for repeated measures data. Response curve analysis is the best method for assessing trends.

LITERATURE CITED

Gumpertz, M.L. and C. Brownie. 1993. Repeated measures in randomized block and split-plot experiments. Can. J. For. Res. 23:625–639.

Koch, G.G., I.A. Amara, M.E. Stokes, and D.B. Gillings. 1980. Some views on parametric and non-parametric analyses for repeated measurements and selected bibliography. Int. Statist. Rev. 48:249–285.

Meriedith, M.P. and S.V. Stehman. 1991. Repeated measures experiments in forestry: focus on analysis of response curves. Can. J. For. Res. 21:957–965.

Moser, E.B., A.M. Saxton, and S.R. Pezeshki. 1990. Repeated measures analysis of variance: application to tree research. Can. J. For. Res. 20:524–535.

Nemec, A.F.N. 1996. Analysis of repeated measures and time series: an introduction with forestry examples. B.C. Min. For., Res. Br., Victoria, BC. Biometrics Handbook No. 6. Work. Pap 10/1996.

Potvin, C. and M.J. Lechowicz. 1990. The statistical analysis of ecophysiological response curves obtained from experiments involving repeated measures. Ecology 71(4):1389–1400.

Sit, V. and M. Poulin-Costello. 1994. Catalogue of curves for curve fitting. B.C. Min. For., Res. Br., Victoria, BC. Biometrics Handbook No. 4. 110 pp.