

Accounting for Multilevel Data Structures in Fisheries Data using Mixed Models

ABSTRACT: Multilevel data structures are those that have a hierarchical structure, in which response variables are measured at the lowest level of the hierarchy and modeled as a function of predictor variables measured at that level and higher levels of the hierarchy. For example, a multilevel data structure may consist of measurements taken on individual fish (lower level) that are nested within lakes or streams (higher level). Multilevel data structures are a common feature in fisheries research. We provide simulated fisheries data examples, similar in structure to other published studies, to illustrate the application of multilevel models and discuss how hypothesis testing and inferences can be incorrect if multilevel data structures are ignored. Ignoring multilevel data structures has implications for the use of commonly-used ordinary least squares (OLS) approaches to test hypotheses and to make inferences. Multilevel models are an alternate approach that circumvents problems associated with traditional OLS methods and allows valid inferences to be made.

Tyler Wagner
Daniel B. Hayes
Mary T. Bremigan

All authors are at the Department of Fisheries and Wildlife, Michigan State University, East Lansing. Wagner is a doctoral student and can be reached at wagnerty@msu.edu. Hayes and Bremigan are associate professors.

INTRODUCTION

Aquatic systems are often viewed as being hierarchically organized, with lower levels of organization nested within higher levels (e.g., Frissell et al. 1986; Imhof et al. 1996). For example, a hierarchy may consist of headwater streams nested within sub-watersheds that are nested within larger watersheds, or lakes nested within ecoregions. This hierarchical organization provides a conceptual basis for testing hypotheses, often leading to sampling designs that are also hierarchically organized. A common feature of such sampling designs is that the response variable is measured at the lowest level (i.e., finest scale) of the hierarchy and is modeled as a function of predictors measured at that level as well as one or more higher levels. This hierarchical organization leads to multilevel data structures for which traditional methods of statistical inference are often inappropriate (Raudenbush and Bryk 2002).

The fisheries literature is replete with examples of studies that have collected data with a multilevel structure. Despite this prevalence, the hierarchical structure of the data is often ignored during statistical analysis. A fundamental problem with ignoring the multilevel structure during analyses is that observations measured within a higher level (e.g., measurements made within the same stream) are likely to be more similar to each other compared to observations between levels (e.g., measurements made in different streams). Therefore, analyses that ignore the multilevel structure of the data violate a critical assumption to commonly-used analyses, namely the assumption of independence. Although the importance of accounting for the correlation structure of repeated measurements on individuals has received attention in the fisheries literature, especially with respect to laboratory studies and analyzing size-at-age data

obtained from scales and otoliths of fishes (e.g., Jones 2000; Schaalje et al. 2002; Pedersen and Malte 2004), multilevel data structures in field settings have been largely ignored.

The goals of this article are (1) to explain how to use multilevel models that account for multilevel data structures in fisheries data to test hypotheses and (2) to discuss how the analytical approach affects hypothesis testing and inferences. To accomplish these goals we provide two examples using simulated data similar in structure to published studies. We illustrate the analysis of these data with a commonly used statistical package, SAS®. The first example uses data with a two-level data structure to emphasize how hypothesis testing and inferences are affected depending on the statistical approach used, while the second example provides a detailed example of how to model data with a three-level data structure.

Example 1: Examining the effects of in-stream barriers on fish density

To introduce a simple multilevel data structure, we present the following example. Data were generated to emulate a commonly-used field study design to examine the effects of instream barriers on fish density. The simulated dataset contains sample sites (level 1) nested within streams (level 2; Figure 1). We used these data to test the null hypothesis that fish densities do not differ between sites above and below barriers or between streams with or without barriers. The dataset contains measurements for eight streams: four “treatment” streams that contain instream barriers and four “control streams” that lack instream barriers. For each treatment stream, fish density measurements were generated for three sites below the barrier and three sites above the barrier (or “reference line” for control streams; Figure 1; Table

1). The data were then analyzed two ways—using a general linear model (GLM) and a multilevel model (MIXED) in Statistical Analysis System (SAS), and using the GLM and MIXED procedures, respectively (Littell et al. 1996; SAS Institute Inc. 2000; Littell et al. 2002). The data were generated to represent a case where mean fish densities were reduced in treatment streams as a whole relative to control streams, and with a greater reduction in sites above the barriers in the treatment streams. In the control streams, mean fish density was similar in sites above and below the reference line.

Figure 1. An illustration of the study-design used for the stream barrier example. Each stream type was replicated four times (treatment streams A–D and control streams E–H).

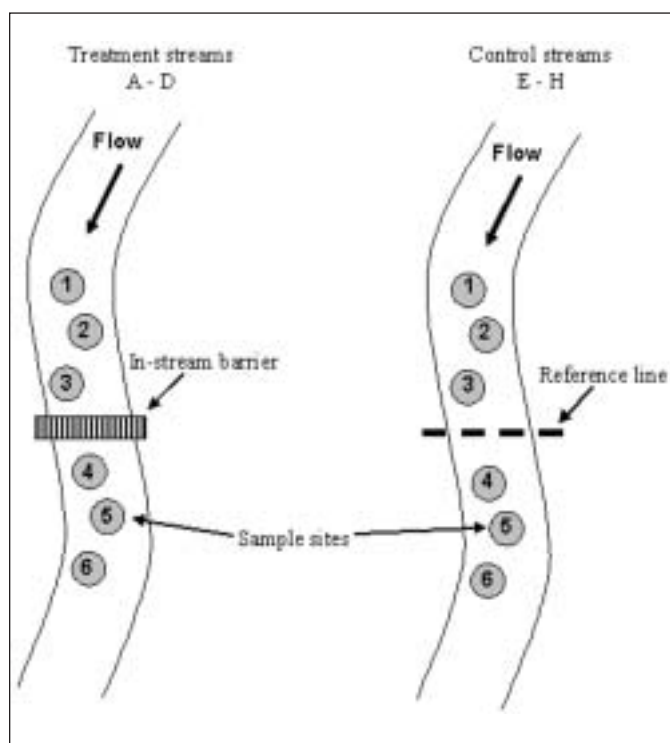


Table 1. Simulated dataset used for example 1. Numbers represent simulated fish densities (fish/m²) for sites nested within treatment (barrier) and control (no barrier) streams and for sample sites above or below the barrier or reference line. Values used to calculate within group means are shown outlined in a dashed line.

Stream		Barrier				Control			
Site	Position	A	B	C	D	E	F	G	H
1	Above	15	49	14	12	46	67	45	74
2	Above	14	47	15	9	41	70	45	67
3	Above	7	39	25	18	42	82	57	72
4	Below	38	45	21	20	49	76	39	66
5	Below	13	50	31	24	50	76	37	64
6	Below	24	50	33	30	55	74	35	72

CONTRASTING TRADITIONAL AND MULTILEVEL MODELS

Traditional approaches—Two ordinary least squares (OLS) approaches are commonly used to analyze multilevel data: an aggregating and a disaggregating approach. For the aggregated approach, observations within each higher level group are combined (analysis is performed at the higher level). For our stream barrier example, aggregation would occur if mean densities were calculated for each stream based on the six sample sites within each stream (Figure 1). When this approach is used, within-group variation is ignored (e.g., within stream), which may be a large proportion of the total variance, resulting in a loss of information and statistical power. For the disaggregated approach, all observations are used, but the higher level grouping factor (e.g., stream) is not factored into the analysis. For our stream barrier example, disaggregation would occur if each measurement of fish density was treated as an independent replicate sample from each stream. This approach is inappropriate however, because the experimental unit is actually the stream, not individual sites, and because the between-group variation is ignored (analysis is performed at the lower level). When this occurs “replicate” samples from a higher level grouping factor are assumed to be independent, which is often an invalid assumption and results in pseudoreplication (Hurlbert 1984). This approach can underestimate standard errors and thus increase the probability of type I errors, i.e., finding a significant difference when one does not actually exist. Furthermore, groups with the largest sample size may dominate the coefficient estimates.

Multilevel models—Multilevel models have received much attention in the past several years, especially in the social and behavioral sciences. Their increase in popularity is partly due to methodological advances and advances in statistical computing over the past several decades. As a result, several excellent references on the theory and application of multilevel models in the social and behavioral sciences are available; we refer readers to these references for more detailed information (e.g., Hox 2002; Raudenbush and Bryk 2002; Duan and Reise 2003).

Multilevel models are represented in the literature under a variety of names including mixed-effects models, hierarchical linear models, random-effects models, and random coefficient regression models. Multilevel models circumvent the problems described above associated with using OLS approaches. For example, multilevel models estimate standard errors correctly and result in improved estimation of fixed effects when multilevel data structures exist. Furthermore, both continuous and categorical variables can be specified to have fixed or random effects. A factor is fixed if it represents all possible levels of a factor for which inferences are to be made. For the instream barrier example, if the streams used in the analysis were the only streams for which inferences were to be made (e.g., if the researchers did not want to generalize their results to other streams) then “stream” would be specified as a fixed effect. A factor is random if it represents a random sample of a larger set of potential factors. For the instream barrier example, if the study streams represented a sample of streams from a larger population of streams with and without barriers, “stream” would be specified as a random effect.

Another way to illustrate the difference between fixed and random effects was presented by Bennington and Thayne

(1994). Their definition is presented in terms of the null hypothesis being tested for each effect.

Consider two effects, *A* and *B*, where *A* is fixed, *B* is random, and there is an interaction (*A* × *B*) possible between them. For a given dependent variable, the null hypothesis concerning *A* is that there is no difference in means among the levels of *A* in the experiment. For *B*, the null hypothesis is that there is no variability among all possible levels of *B* (including those not sampled), not that there are no differences among levels of that effect included in the experiment. For the interaction term (*A* × *B*), the null hypothesis is that variability among levels of *B* is the same for all levels of *A*. This differs from the case for fixed effects in that the null hypothesis for an interaction between two fixed effects (*A* and *C*) is that the response of the dependent variable is not different among specific levels of *A* depending upon the particular level of *C*.

Analysis of Example 1

Ordinary least squares—We first analyzed the simulated stream data assuming all factors are fixed effects (we assumed streams were not randomly selected from a larger population of streams) while ignoring the fact that sites are nested within streams. This is equivalent to using a disaggregated OLS approach, and was performed using the GLM procedure in SAS. An aggregated approach could also be implemented with these data, but for illustration purposes we restrict our analysis and discussion to the disaggregated approach, which is a common approach used in the analysis of fisheries data. Fish density was the response variable and site position (above or below a barrier) and stream type (barrier or no barrier) were fixed effects. The general form of this model is as follows:

$$Y_{ijk} = u + \text{Position}_i + \text{Stream_Type}_j + (\text{Position} \times \text{Stream_Type})_{ij} + e_{ijk} \quad (1)$$

where Y_{ijk} is the k^{th} measurement on the i^{th} position, in the j^{th} stream, u is the overall mean, $(\text{Position} \times \text{Stream_Type})_{ij}$ is the interaction effect, and errors (e_{ijk}) are assumed independent and $e_{ijk} \sim N(0, \sigma^2)$. An example of the data structure needed for analyzing the dataset in SAS is given in Appendix I. The SAS code for performing the GLM analysis is as follows.

```
PROC GLM DATA = barrier_data;
CLASS stream stream_type position;
MODEL density = stream_type position stream_type*position
/SOLUTION;
LSMEANS stream_type position stream_type*position /
STDERR PDIF ADJUST = tukey;
RUN;
```

For a detailed description of the SAS syntax, see Littell et al. (2002). Briefly, the CLASS statement contains the classification variables (categorical independent variables), the MODEL statement defines the model to be fit, and the SOLUTION option requests the parameter estimates. The LSMEANS statement requests that the least-squares (LS) means be calculated for each classification variable listed in the statement. Least-squares means are within-group means adjusted for other effects in the model and are also known as the population marginal means (Searle 1987). The PDIF option reports the results of the hypothesis test of the differences between LS means (H_0 : LS

mean_{*i*} = LS mean_{*j*}). The ADJUST = tukey statement requests a multiple comparison test with adjusted *P*-values and confidence limits for the LS means using the Tukey-Kramer method. This adjustment controls for the overall experiment-wise error rate (e.g., controls for type I error rate). Note that PROC GLM allows for random terms; however, the standard errors from the LSMEANS statement are usually not computed correctly (Littell et al. 1998). The GLM procedure in this example was run with only fixed effects.

The type III sums of squares tests for the significance of the fixed effects, which account for the other effects in the model, are as follows, stream type $F = 58.31$, $P = < 0.0001$, position $F = 1.01$, $P = 0.319$, stream type × position interaction $F = 1.71$, $P = 0.197$. The analysis indicates there is a significant difference in mean fish density between stream types, with barrier streams having significantly lower mean density levels compared to control streams (barrier stream $\bar{X} = 26.8$ fish/m², standard error (SE) = 2.92; control stream $\bar{X} = 58.4$ fish/m², SE = 2.92). Table 2 contains the LS means and standard error estimates for each stream type and site position. The analysis did not detect any interaction between site position and stream type, although we had simulated an interaction effect in the dataset.

Multilevel model—Because multilevel models have received more attention in the social and behavioral sciences and thus references are not available specifically for the natural sciences, we use symbols consistent with Raudenbush and Bryk (2002) in our description of multilevel models. For this analysis, the dependent variable was the same as in the GLM analysis; however, we took into account the nested structure of the data (sites nested within streams) and analyzed the data using the MIXED procedure in SAS with random effects. Position, above or below the barrier, was the site-level (level 1) predictor and stream type (with or without a barrier) was the stream-level predictor (level 2). Stream was regarded as a random effect. Typically, it is likely that measurements of fish density from the same stream are correlated (i.e., lack statistical independence); one way to model this correlation is by treating each stream as having a random effect. Furthermore, we assumed that streams used in the study represented a random sample of a larger population of streams; therefore, we can generalize our results to other similar systems.

Table 2. Least-squares means (LSM) and standard error (SE) estimates for example 1 analyzed in SAS using a general linear model (PROC GLM) and a multilevel model (PROC MIXED). For each analysis (column) least-squares means with different superscripted letters are significantly different ($P < 0.05$ using Tukey-Kramer multiple comparison test). Least-squares means correspond to the arithmetic within group means which can be calculated using the values in Table 1.

Stream type	Position	General linear model		Multilevel model	
		LSM	SE	LSM	SE
Barrier	Above	22.0 ^a	4.14	22.0 ^a	7.38
Barrier	Below	31.6 ^a	4.14	31.6 ^b	7.38
Control	Above	59.0 ^b	4.14	59.0 ^b	7.38
Control	Below	57.8 ^b	4.14	57.8 ^b	7.38

The model can also be viewed in two levels and in a combined form as follows:

$$\text{Level 1 model: } Y_{ij} = \beta_{0j} + \beta_{1j} (\text{Position}) + r_{ij} \quad (2)$$

where Y_{ij} is the density of fish in site i in stream j , β_{0j} is the mean outcome for stream j , β_{1j} is the coefficient for the fixed effect of site position on fish density, r_{ij} is the level-1 error, where $r_{ij} \sim N(0, \sigma^2)$, and σ^2 is the variance at level 1 after controlling for the effects of position.

$$\text{Level 2 model: } \beta_{0j} = \gamma_{00} + \gamma_{01} (\text{Stream_Type}) + u_{0j}, \text{ and}$$

$$\beta_{1j} = \gamma_{10} \quad (3)$$

where γ_{00} is the grand mean density, γ_{01} is the estimated coefficient for the fixed effect of stream type (i.e., barrier or control) on stream mean fish density, γ_{10} is a fixed effect representing the coefficient for the effect of position on fish density, and u_{0j} is the residual, where $u_{0j} \sim N(0, \tau_{00})$ and τ_{00} is the conditional variance (the stream-level variance after controlling for stream type). The combined model can then be written to contain the site position \times stream type interaction as follows:

$$Y_{ij} = \gamma_{00} + \gamma_{10} (\text{Position}) + \gamma_{01} (\text{Stream_Type}) + \gamma_{11} (\text{Position} \times \text{Stream_Type}) + u_{0j} + r_{ij} \quad (4)$$

where γ_{11} is the estimated coefficient for the interaction term and all other variables are defined as above. The model can be implemented using the following code:

```
PROC MIXED COVTEST DATA= barrier_data;
CLASS stream stream_type position;
MODEL density = stream_type position stream_type*position /
SOLUTION;
RANDOM intercept / SUBJECT = stream;
LSMEANS stream_type position stream_type*position / PDIFF
ADJUST = tukey;
RUN;
```

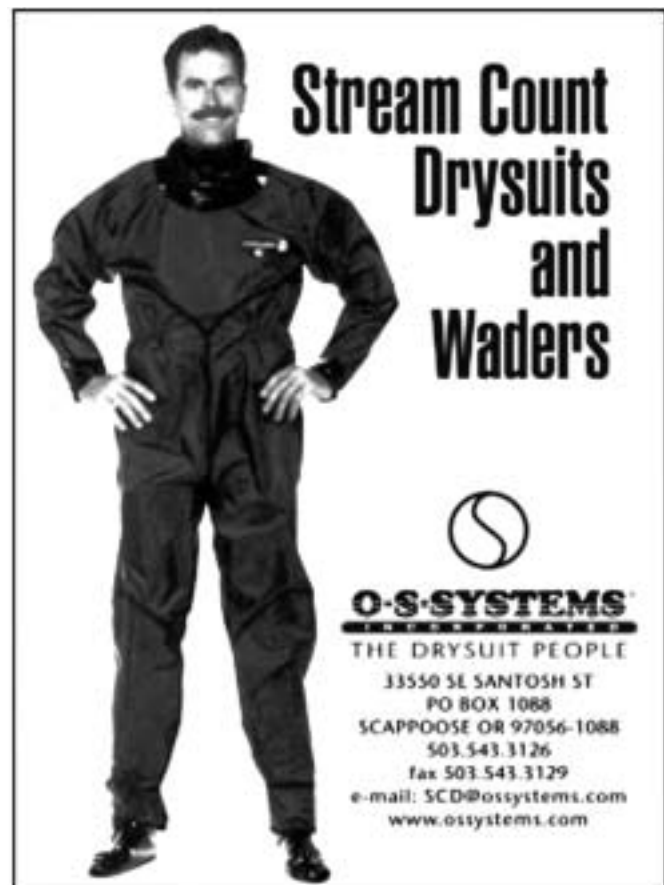
For an extensive explanation of the PROC MIXED syntax, see Littell et al. (1996). The syntax is similar to that used in the GLM procedure; however, important differences exist. The COVTEST statement produces asymptotic standard errors and Wald Z-tests for the covariance parameter estimates, $\hat{\sigma}^2$ and $\hat{\tau}_{00}$. The CLASS statement is the same as described for the GLM procedure, while the MODEL statement lists the dependent variable and only the fixed effects. The SOLUTION option after the MODEL statement requests the parameter estimates and their standard errors for the fixed effects. The RANDOM statement specifies the random effects in the model. The intercept is designated as random in this model because it is assumed that the stream-level intercepts are from a larger population of stream-level intercepts. The SUBJECT option identifies the subject(s) in the multilevel model. Specifying a subject is equivalent to nesting all effects in the RANDOM statement within the subject effect (Littell et al. 1996). Therefore, the above syntax is modeling fish density while accounting for the data being clustered (grouped) by streams. As in the GLM procedure, the LSMEANS statement requests the LS means estimates for the specified fixed effects and the PDIFF option for the LSMEANS statement requests the differences between the LS means.

The analysis indicated that there was a significant position \times stream type interaction. The overall type III tests for the fixed


effects are as follows: stream type, $F = 9.44$, $P = 0.0039$; position, $F = 5.56$, $P = 0.024$; and position \times stream type interaction, $F = 9.39$, $P = 0.004$. Note that because PROC MIXED uses a likelihood-based approach to estimation, it does not directly compute or display the sums of squares; however, the type III tests are equivalent to those produced by PROC GLM. Table 2 contains the LS means estimates and standard errors for each stream type and site position. In treatment streams, sites located above the barrier had significantly lower mean fish density estimates compared to sites below the barrier; whereas, in control streams mean fish density did not differ between sites located above or below the barrier reference line. Treatment streams had lower fish density estimates compared to control streams, regardless of position, but considering only sites below the barriers, treatment and control streams did not differ significantly in mean fish density estimates.

Both the traditional (OLS) and multilevel analyses resulted in the same LS means point estimates and they were equal to the arithmetic means of the values outlined in Table 1. Least-squares means will be equivalent to arithmetic means for cases with balanced designs, as in this example. However, for unbalanced designs, which are common in ecological studies, the LS means estimates will typically not equal the arithmetic means.

There are two major differences between the traditional and multilevel analyses that have implications for hypothesis testing and inferences: (1) the difference in standard error estimates and (2) the specification of random versus fixed effects. In the multilevel analysis, the standard error estimates of the means are about two times larger compared to the OLS estimates. This difference is due to the fact that the standard error in the traditional analysis is calculated using only the residual variance (the residual variance is the only



**Stream Count
Drysuits
and
Waders**



O-S-SYSTEMS[®]
INCORPORATED
THE DRYSUIT PEOPLE
33550 SE SANTOSH ST
PO BOX 1088
SCAPPOOSE OR 97056-1088
503.543.3126
fax 503.543.3129
e-mail: SCD@ossystems.com
www.ossystems.com

variance component in fixed effects models); whereas, the standard error in the multilevel analysis is calculated using two variance components: the residual variance and a between-stream variance. The smaller standard errors estimated using the traditional approach can lead to increased probabilities of type I error rates, i.e., finding a significant difference when one does not actually exist.

The second major difference between our two analyses is the specification of stream as a random effect in the multilevel analysis. The specification of random effects has implications for what inferences can be made based on the results of the analyses. For the traditional approach, with stream as a fixed effect, results can not be generalized to streams that were not used in the analysis and must be restricted to the eight streams used in the study. For the multilevel analysis, where stream was specified as a random effect, inferences can be generalized to a larger population of barrier and no-barrier streams. Often a goal of a study is the ability to generalize results found from a subset of study streams or lakes to a larger population of streams or lakes of interest. The use of random effects allows for such inference, whereas purely fixed effects models do not.

Example 2: Examining the relationship between aquatic macrophyte percent cover and stomach fullness of yellow perch (*Perca flavescens*) in inland lakes

For this example, we focus our analysis on the multilevel modeling approach to demonstrate how to analyze and interpret datasets with three levels. As a result, we do not compare this multilevel analysis with a traditional OLS analysis. The limitations of using

traditional OLS approaches as discussed previously, however, do exist for analyzing these data. We realize that some of the following model details are fairly dense; however, it is our goal that these details will aid in the understanding and interpretation of the model.

Data were generated to emulate a field study designed to examine the effects of macrophyte cover on percent stomach fullness of yellow perch in inland lakes. The dataset contains individual fish (level 1), nested within sampling sites (level 2), nested within lakes (level 3; Figure 2). In this example, sample sites were assumed to be randomly chosen within lakes and lakes were randomly selected from a larger population of lakes. The data were used to test the null hypothesis that percent stomach fullness is not related to percent macrophyte cover while controlling for the effect of individual fish weight on stomach fullness. The dataset contains measurements of percent stomach fullness and weight (g) for individual fish sampled from eight sample sites within each of four lakes. Sample sizes of individual fish varied among sites and lakes, ranging from 0–46 fish per site and from 160–245 fish per lake for a total of 751 observations. Weight and percent stomach fullness of individual fish ranged between 110–293 g and 14.4–42.2%, respectively. Percent macrophyte cover was generated for each sample site and ranged from 2–97%.

We introduced complexity to this dataset by not only introducing a third level to the hierarchy, but by also including predictors at multiple levels: the predictor at level 1 (the individual fish level) is fish weight and the predictor at level 2 (the site level) is percent macrophyte cover. Because of this complexity, we are going to view this analysis as a two-stage process.

Stage 1—The first stage involves obtaining initial estimates of the total variance, how the variation is partitioned (i.e., obtain variance estimates that describe how much variation in stomach fullness there is due to individual differences of fish within sites nested within lakes, among sites nested within lakes, and among lakes). These variance estimates are also used, along with variance estimates obtained in stage 2, to determine the percent variation explained at each level of the model by the predictor variables. The model that produces these estimates is a one-way ANOVA with random effects and is also referred to as an unconditional model because it does not contain any predictor variables. This one-way ANOVA with random effects can be viewed as a three-level model as follows:

$$\text{Level 1 model: } Y_{ijk} = \pi_{0jk} + e_{ijk} \quad (5)$$

where Y_{ijk} is the percent stomach fullness of fish i in site j and lake k , π_{0jk} is the mean stomach fullness of site j in lake k , and e_{ijk} is the random “fish effect,” and $e_{ijk} \sim N(0, \sigma^2)$ where σ^2 is the residual variance component due to individual differences of fish within sites nested within lakes.

Figure 2. An illustration of the study-design used for the yellow perch stomach fullness example. Yellow perch (level 1 of the hierarchy) were sampled from eight sample sites (level 2 of the hierarchy) within four lakes (level 3 of the hierarchy; lakes A–D). Stomach fullness (dependent variable) and weight (level 1 predictor variable) were determined for each fish and percent macrophyte cover (level 2 predictor variable) was determined for each sample site.

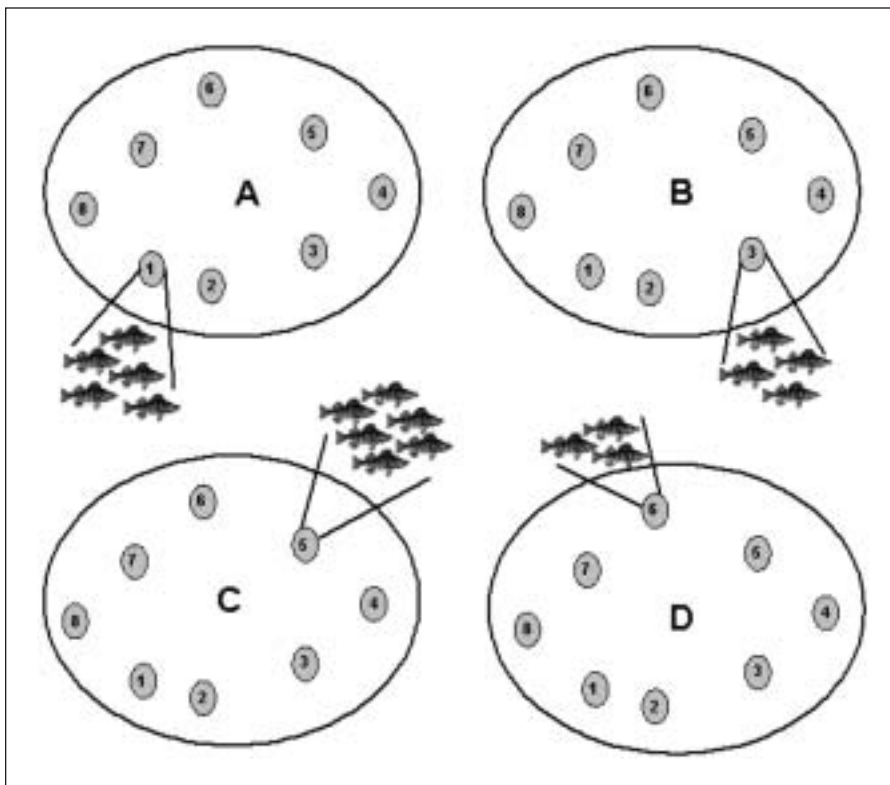


Table 3. Final parameter and variance estimates, standard errors (SE) and *P*—values for example 2.

Parameter	Estimate	SE	<i>P</i> -value
Intercept ($\hat{\gamma}_{000}$)	5.92	1.31	0.02
Weight ($\hat{\gamma}_{100}$)	0.10	0.003	<0.0001
Percent cover ($\hat{\gamma}_{010}$)	5.02	1.26	<0.0001
Weight × percent cover ($\hat{\gamma}_{110}$)	-0.005	0.004	0.24
Variance components			
Among fish within sites nested within lakes ($\hat{\sigma}^2$)	1.07	0.05	<0.0001
Among sites within lakes ($\hat{\tau}_{\pi}$)	5.15	1.48	0.0002
Among lakes ($\hat{\tau}_{\beta}$)	1.71	1.95	0.19

Level 2 model: $\pi_{0jk} = \beta_{00k} + r_{0jk}$ (6)

where β_{00k} is the mean fullness in lake *k*, r_{0jk} is the random “site effect,” and $r_{0jk} \sim N(0, \tau_{\pi})$, where τ_{π} is the variance between sites nested within lakes.

Level 3 model: $\beta_{00k} = \gamma_{000} + u_{00k}$ (7)

where γ_{000} is the overall grand mean fullness, u_{00k} is the random “lake effect,” and $u_{00k} \sim N(0, \tau_{\beta})$, where τ_{β} is the variance

between lakes. The combined unconditional model, therefore, has a fixed effect (γ_{000}) and three random effects, (u_{00k} , r_{0jk} , and e_{ijk}), and is as follows:

$$Y_{ijk} = \gamma_{000} + u_{00k} + r_{0jk} + e_{ijk} \quad (8)$$

Examining the initial variance estimates provides information regarding how much total variation there is at each level that can subsequently be modeled with predictor variables. The code required for performing the one-way ANOVA with random effects for this example is:

```
PROC MIXED COVTEST DATA = lake_data;
CLASS lake site;
MODEL fullness = / SOLUTION;
RANDOM intercept / SUBJECT = lake;
RANDOM intercept / SUBJECT = site(lake);
RUN;
```

The syntax for this unconditional model is similar to that described in example 1; however, there is an additional RANDOM statement which specifies that sites are nested within lakes.

Results from the one-way ANOVA with random effects show that the grand mean ($\hat{\gamma}_{000}$) stomach fullness over all lakes is 24% and the estimates of variance among-fish-within-sites-nested-within-lakes ($\hat{\sigma}^2$), among-sites-within-lakes ($\hat{\tau}_{\pi}$), and among-lakes ($\hat{\tau}_{\beta}$) are 8.67 (SE = 0.46, *P* < 0.0001), 7.44 (SE = 2.18, *P* = 0.0003), and 2.18 (SE = 2.6, *P* = 0.202), respectively. The percent variance among fish within sites nested within lakes, among sites within lakes, and among



fish@halltechaquatic.com

www.halltechaquatic.com

HT-2000 Battery Backpack Electrofisher

The HT2000 meets and exceeds all aspects of the Electrofishing Guidelines for Safety and Functionality.

Contact us today to find out why so many Federal, State and Local Authorities are choosing the HT2000 for their Fisheries Research Monitoring and Stream Assessments.



Toll Free: 1-866-425-5832 Fax: 1-519-766-0729
129 Watson Road South, Guelph, Ontario Canada N1L 1E4

Visit www.htex.com for Rugged Data Collection Systems, GPS Solutions and more Field Research Products.

lakes is 47%, 41%, and 12%, respectively. We now have information on how the variance is partitioned in our dataset and a conditional model can be specified in stage 2.

Stage 2—The predictor of interest in this study is percent macrophyte cover. However, we also need to account for the effect of individual fish weight on stomach fullness in the model. Therefore, our two predictors are fish weight, modeled at level 1, and percent macrophyte cover, modeled at level 2. Level 3 (the lake level) is left unconditional, with no predictor variables (covariates). Again, the model can be viewed as three levels and in a combined form as follows:

$$\text{Level 1 model: } Y_{ijk} = \pi_{0jk} + \pi_{1jk}(\text{Weight})_{ijk} + e_{ijk} \quad (9)$$

where Y_{ijk} is the stomach fullness for fish i in site j in lake k , π_{0jk} is the intercept for site j in lake k , π_{1jk} is the estimated coefficient for the fixed effect of fish weight on stomach fullness, and e_{ijk} is the level-1 random effect.

$$\text{Level 2 model: } \pi_{0jk} = \beta_{00k} + \beta_{01k}(\text{Percent_Cover})_{jk} + r_{0jk} \\ \pi_{1jk} = \beta_{10k} \quad (10)$$

where β_{00k} is the intercept for lake k , β_{01k} is the estimated coefficient for the fixed effect of percent macrophyte cover, β_{10k} is a fixed effect representing the coefficient for the effect of fish weight on stomach fullness, and r_{0jk} is the level-2 random effect.

$$\text{Level 3 model: } \beta_{00k} = \gamma_{000}, \beta_{10k} = \gamma_{100}, \beta_{01k} = \gamma_{010} \quad (11)$$

where γ_{000} represents the coefficient for the level 2 intercept, γ_{100} represents the coefficient for the fixed effect of fish weight on stomach fullness, and γ_{010} represents the coefficient for the fixed effect of percent macrophyte cover on stomach fullness. The combined model is as follows:

$$Y_{ijk} = \gamma_{000} + \gamma_{100}(\text{Weight}) + \gamma_{010}(\text{Percent_Cover}) + \gamma_{110}(\text{Weight} \times \text{Percent_Cover}) + u_{00k} + r_{0jk} + e_{ijk} \quad (12)$$

where γ_{110} is the estimated coefficient for the interaction term and all other variables are defined as above. The model can be specified using the following code:

```
PROC MIXED COVTEST DATA = lake_data;
CLASS lake site;
MODEL fullness = weight cover weight*cover / SOLUTION;
RANDOM intercept / SUBJECT = lake;
RANDOM intercept / SUBJECT = site(lake);
RUN;
```

The syntax is similar to that described in the unconditional model; however, we now have specified the full model in the MODEL statement.

The analysis indicated that both fish weight and percent macrophyte cover were significantly and positively associated with percent stomach fullness and that there was not a significant interaction effect (Table 3). Variance estimates obtained from the final model (full model) can be used along with the variance estimates from the unconditional model to determine how much variation was explained at each level as follows:

$$\text{Percent variance explained} = \frac{\text{Variance}_{\text{unconditional}} - \text{Variance}_{\text{full}}}{\text{Variance}_{\text{unconditional}}} \quad (13)$$


where $\text{Variance}_{\text{unconditional}}$ is an estimate of $\hat{\sigma}^2$, $\hat{\tau}_\pi$, or $\hat{\tau}_\beta$ from the unconditional model and $\text{Variance}_{\text{full}}$ is an estimate of $\hat{\sigma}^2$, $\hat{\tau}_\pi$, or $\hat{\tau}_\beta$ from the full model. For example, to determine how much of the variation in stomach fullness among fish within sites nested within lakes was explained by fish weight we perform the following calculation:

$$\frac{8.67(\hat{\sigma}^2_{\text{unconditional}}) - 1.07(\hat{\sigma}^2_{\text{full}})}{8.67(\hat{\sigma}^2_{\text{unconditional}})} = 0.88 \quad (14)$$

thus, fish weight explained 88% of the among-fish-within-site-nested-within-lake (level 1) variation in stomach fullness. Using equation 13, we can determine that percent macrophyte cover explained 31% of the variation in fish stomach fullness among sites within lakes (level 2). Because we did not have predictors at the lake-level (level 3), we do not need to calculate percent variation explained at this level. However, if predictors were included at level 3, the same calculation could be performed to determine percent variation explained.

In this example, the multilevel model accounted for the fact that fish were nested within sample sites within lakes and sample sites were nested within lakes. As in example 1, the specification of random effects allowed us to account for the lack of independence of observations within sites and lakes with similar implications for inferring (e.g., the ability to generalize to a larger population of lakes). The analysis also allowed for the partitioning of variance among the three levels. Variance partitioning provides valuable information on how much variation is contained at each level. Knowledge of how much variation exists at each level can also help guide future data collection efforts by allowing researchers to focus data collection at the level (e.g., spatial scale) that contains much of the variability that needs to be explained.

CONCLUSION

Multilevel models provide several advantages over the more commonly-used OLS approaches when analyzing data with a hierarchical structure. Because hierarchical structures are common to both experimental and field (observational) studies in fisheries research, we encourage the use of multilevel models where appropriate. Some other examples of where hierarchical data structures may arise in fisheries research, where multilevel models would be applicable, include investigations of fishing tournament-related mortality, where fish are nested within tournaments and tournaments are nested within lakes, and investigations of landscape features on lake or stream attributes, where waterbodies are nested within watersheds and watersheds are nested within ecoregions. Furthermore, due to statistical computing advances, multilevel models can be implemented in widely available statistical software packages. These approaches provide better estimation of fixed effects, allow for the partitioning of variance components across levels, and allow for generalizations beyond the particular groups (e.g., streams or lakes) used in the study. 

ACKNOWLEDGMENTS

We would like to thank Patricia A. Soranno, Todd Wills, Andrea Jaeger, and two anonymous reviewers for constructive comments that improved earlier drafts of this manuscript. Funding for this research was provided by the Michigan Department of Natural Resources, Fisheries Division.

APPENDIX

An example of the data structure required to analyze the stream barrier dataset (example 1) using SAS. The DATA statement specifies the name of the dataset that is generated. The INPUT statement specifies the variables that are read from the program editor window and the "\$" designates variables as character variables. The data corresponding to the variables listed in the INPUT statement are entered after the DATALINES statement.

```
DATA barrier_data;
INPUT density stream $ site stream_type $
position $;
DATALINES;
15      A      1      Barrier  Above
14      A      2      Barrier  Above
7       A      3      Barrier  Above
49      B      1      Barrier  Above
47      B      2      Barrier  Above
39      B      3      Barrier  Above
14      C      1      Barrier  Above
15      C      2      Barrier  Above
25      C      3      Barrier  Above
.
.
.
72      E      6      Control  Below
;
```

REFERENCES

- Bennington, C. C., and Thayne, W. V. 1994. Use and misuse of mixed model analysis of variance in ecological studies. *Ecology* 75:717-722.
- Duan, N., and S. P. Reise (editors). 2003. Multilevel modeling. Methodological advances, issues, and applications. Lawrence Erlbaum Associates, Mahwah, NJ.
- Frissell, C. A., W. J. Liss, C. E. Warren, and M. D. Hurley. 1986. A hierarchical framework for stream habitat classification: viewing streams in a watershed context. *Environmental Management* 10:199-214.
- Hox, J. J. 2002. Multilevel analysis: techniques and applications. Lawrence Erlbaum Associates, Mahwah, NJ.
- Hurlbert, S. H. 1984. Pseudoreplication and the design of ecological field experiments. *Ecological Applications* 54:187-211.
- Jones, C. M. 2000. Fitting growth curves to retrospective size-at-age data. *Fisheries Research* 46:123-129.
- Imhof, J. G., J. Fitzgibbon, and W. K. Annable. 1996. A hierarchical evaluation for characterizing watershed ecosystems for fish habitat. *Canadian Journal of Fisheries and Aquatic Sciences* 53:312-326.
- Littell, R. C., G. A. Milliken, W. W. Stroup, and R. S. Wolfinger. 1996. SAS System for Mixed Models. SAS Institute Inc., Cary, North Carolina.
- Littell, R. C., P. R. Henry, and C. B. Ammerman. 1998. Statistical analysis of repeated measures data using SAS procedures. *Journal of Animal Science* 76:1216-1231.
- Littell, R. C., W. W. Stroup, and R. J. Freund. 2002. SAS® for linear models. SAS Institute Inc., Cary, North Carolina.
- Pedersen, L. E., and H. Malte. 2004. Repetitive acceleration swimming performance of brown trout in fresh water and after acute seawater exposure. *Journal of Fish Biology* 64:273-278.
- Raudenbush, S. W., and A. S. Bryk. 2002. Hierarchical linear models (2nd ed.). Sage Publications, Thousand Oaks, CA.
- SAS Institute Inc. 2000. SAS/STAT user's guide. SAS Institute Inc., Cary, NC.
- Schaalje, G. B., J. L. Shaw, and M. C. Belk. 2002. Using nonlinear hierarchical models for analyzing annulus-based size-at-age data. *Canadian Journal of Fisheries and Aquatic Sciences* 59:1524-1532.
- Searle, S. R. 1987. Linear models for unbalanced data. John Wiley & Sons, Inc., New York.

Your Tags Your Way

FLLOY TAG

The World Leader & Innovator in Fish Tags - For Over 50 Years

- Shellfish, Lobster & Crustacean Tags
- T-Bar Anchor Tags, Spaghetti Tags, Dart Tags & More
- Net, Trap & Line Tags

- Laminated Disc and Oval Tags
- Dart, Fingerting, Streamer, Intramuscular Tags
- Guns and Tag Applicators, Extra Needles, etc.

...and almost any other kind of custom tagging solution you might need.

"Why Risk Your Research To The Copy-Cats"



...When You Can Have The Original?"

**View our latest catalog at www.floytag.com, or email us at: sales@floytag.com
or call to discuss your custom tagging needs: (800) 843-1172**