

Multinomial regression for analyzing macroinvertebrate assemblage composition data

Author(s): Song S. QianThomas F. CuffneyGerald McMahon

Source: Freshwater Science, 31(3):681-694. 2012.

Published By: The Society for Freshwater Science

DOI: <http://dx.doi.org/10.1899/11-026.1>

URL: <http://www.bioone.org/doi/full/10.1899/11-026.1>

BioOne (www.bioone.org) is a nonprofit, online aggregation of core research in the biological, ecological, and environmental sciences. BioOne provides a sustainable online platform for over 170 journals and books published by nonprofit societies, associations, museums, institutions, and presses.

Your use of this PDF, the BioOne Web site, and all posted and associated content indicates your acceptance of BioOne's Terms of Use, available at www.bioone.org/page/terms_of_use.

Usage of BioOne content is strictly limited to personal, educational, and non-commercial use. Commercial inquiries or rights and permissions requests should be directed to the individual publisher as copyright holder.

Multinomial regression for analyzing macroinvertebrate assemblage composition data

Song S. Qian¹

Nicholas School of the Environment, Duke University, Durham, North Carolina 27708 USA

Thomas F. Cuffney²

US Geological Survey, 3916 Sunset Ridge Road, Raleigh, North Carolina 27607 USA

Gerald McMahon³

Department of the Interior, Southeast Climate Science Center, Department of Biology, 240 David Clark Labs, North Carolina State University, Raleigh, North Carolina 27695-7617 USA

Abstract. Macroinvertebrate species composition data are often expressed as proportional abundances when assessing water-quality conditions or responses to disturbance. Proportional abundances represent the probability of belonging to one of many mutually exclusive and exhaustive groups (taxa). Proportional abundances have some unique properties that must be considered when analyzing these data: 1) the probabilities of group membership must sum to 1 and 2) a change in any 1 group affects all other groups. We used multinomial regressions to analyze changes in proportional abundances along gradients of urbanization in 9 metropolitan areas across the USA. Multinomial regression can be used to address multiple nonlinear responses simultaneously, whereas simple linear regressions must be used to analyze linear or polynomial responses of each group independently. We established that: 1) abundance ratios of tolerant and moderately tolerant groups responded consistently (3–5% increase in the ratios for every 1% increase in developed land cover in the watershed) across the urban gradient, 2) functional groups did not change significantly, and 3) ratios based on assemblage metrics were better indicators of environmental disturbance than ratios based on individual taxa. Multinomial regression, with its flexible model form, can capture patterns of species succession along a resource or stressor gradient. Our results also demonstrate that users of multinomial regression may encounter numerical problems with rare taxa, especially when these taxa have a complete separation along the gradient. Consequently, multinomial regressions are more suitable for analyzing aggregations of taxa or taxon traits.

Key words: generalized linear model, multinomial distribution, urbanization, species tolerance, functional groups, benthic macroinvertebrates, disturbance, multinomial regression.

Biological assemblages are commonly used to assess ecological conditions (Barbour et al. 1999, Davies and Jackson 2006) and to quantify the effects of anthropogenic disturbances in streams (Walsh et al. 2001, Kennen and Ayers 2002, Roy et al. 2003, Brown et al. 2009, Barbour and Paul 2010). Measures of richness, abundance, and composition (relative abundance) of

taxa or aggregations of taxon traits (i.e., assemblage metrics) are commonly used to describe responses or expected conditions. These measures often do not follow a normal distribution so they do not meet the normality assumption that underlies analytical methods, such as linear regression, analysis of variance, or parametric multivariate analyses. Consequently, analysis of these data involves transforming the data or using methods that can accommodate the appropriate distributions (e.g., generalized linear models [GLM]; McCullagh and Nelder 1989). Composition data (relative abundance as proportions) differ from richness and abundance data in that the proportions that represent the variables (taxa or

¹ Present address: Department of Environmental Sciences, Mail Stop 604, University of Toledo, 2801 West Bancroft St., Toledo, Ohio 43606-3390, USA. E-mail: mdqian@gmail.com

² E-mail addresses: tcuffney@usgs.gov

³ gcmahon@usgs.gov

metrics) must sum to 1 within a sample and a change in one group affects all other groups. These characteristics and the issue of normality must be addressed in the analyses. Because of these characteristics, the relative abundance of a specific taxon or group of taxa often responds to disturbance non-linearly, as represented by the species-packing model where the relative abundance of a taxon or group along a disturbance gradient is modeled by a bell-shaped curve (Whittaker 1967, Jongman et al. 1995). Unfortunately, responses based on composition data frequently are modeled with separate linear regressions for each variable (Cuffney et al. 2005, 2010), which can result in models in which the estimated relative abundances greatly deviate from the species-packing model and can lead to estimated relative abundances that do not sum to 1 (i.e., sum of proportions $>$ or $<$ 1).

We introduce the use of multinomial regression models as a quantitative method for analyzing composition data and as a tool for evaluating trends and changes in biological assemblages. We demonstrate multinomial regression models for taxon composition (relative abundance of Ephemeroptera taxa) and the composition of assemblage metrics (pollution-tolerance classes and functional groups). This method is an extension of the familiar GLMs that address issues related to the distribution of the composition data (Fahrmeir and Tutz 1994, Agresti 2002), and we illustrate it by reanalyzing macroinvertebrate assemblage data from the US Geological Survey (USGS) study on the effects of urbanization on stream ecosystems (Brown et al. 2009, Cuffney et al. 2010). The emphasis of our paper is on the discussion of the method, graphical presentation of the results, and model interpretation.

Methods

Study areas and data

Investigators for the USGS study on the effects of urbanization on stream ecosystems examined responses of aquatic biota (fish, invertebrates, and algae), water chemistry, and physical habitat in 9 metropolitan areas in different environmental settings: Atlanta, Georgia (ATL); Boston, Massachusetts (BOS); Birmingham, Alabama (BIR); Denver, Colorado (DEN); Dallas–Fort Worth, Texas (DFW); Milwaukee–Green Bay, Wisconsin (MGB); Portland, Oregon (POR); Raleigh, North Carolina (RAL); and Salt Lake City, Utah (SLC). In each metropolitan area, a multimetric urban intensity index (MA-NUII) was used to identify representative gradients of urbanization within relatively homogeneous environmental settings (McMahon and Cuffney 2000,

Cuffney and Falcone 2008). Data collected from these areas were used to examine the rate and form of biological responses, the physiochemical characteristics most strongly associated with biological responses, and responses among urban areas (Coles et al. 2004, Cuffney et al. 2005, 2010, Brown et al. 2009).

We quantified urban intensity as % developed land in each basin rather than with the MA-NUII to facilitate comparisons with other studies of urbanization. Percent developed land (National Land Cover Database Type I, class 20; http://www.mrlc.gov/nlcd_definitions.php) is 1 of 3 urban measures that were combined to form the MA-NUII (Cuffney and Falcone 2008) and is strongly correlated with MA-NUII. We characterized responses of benthic macroinvertebrate assemblages as the composition (relative abundance) of mayfly taxa, tolerance classes, and functional groups. We obtained pollution-tolerance values (TV) from Barbour et al. (1999) and NCDENR (2006) and used them to derive 4 tolerance groups: intolerant ($TV \leq 3$), moderately tolerant ($3 < TV < 7$), tolerant ($TV \geq 7$), and unknown (TV not available). We used TVs from NCDENR (2006) to supplement values from Barbour et al. (1999) for metropolitan areas in the South (ATL, BIR, DFW, and RAL). We derived TVs for other metropolitan areas from Mid-Atlantic (BOS), Upper Midwest (MGB), and Northwest (DEN, SLC and POR) TVs from Barbour et al. (1999). We derived 9 functional groups (shredder, piercer, collector-gatherer, collector-filterer, omnivore, predator, parasite, scraper, and unknown) from Barbour et al. (1999). Among the 9 metropolitan areas, watersheds in DEN, DFW, and MGB had high percentages ($>79\%$) of agricultural land cover prior to urbanization (antecedent agricultural land cover [AAG]), whereas watersheds in the other 6 areas had relatively low AAG ($<25\%$).

Probability distribution of species compositional data

We stored benthic macroinvertebrate data (counts for each taxon expressed as no./m^2) in matrix format with rows representing sampling sites and columns representing taxa. An additional column contained % developed land for each site. In analyzing the assemblage composition data, the response variables were the taxon count variables and % developed land was the predictor variable. The statistical basis for the method we used is the assumption that taxon count variables can be approximated by the multinomial distribution.

The multinomial distribution is a generalization of the binomial distribution, the probability distribution of the number of successes in n independent Bernoulli trials. In the context of taxon composition data, an independent Bernoulli trial is the process of

identifying an organism from a sample, and a trial is a success when the organism is identified as a group of interest (taxon or metric). For example, if the relative abundance of intolerant taxa in a sample were used as an indicator of water quality, then these data would be summarized by 2 counts: the total number of individuals (sample) and the number of individuals belonging to the intolerant group (successes). The statistical model to describe the randomness of the process is the binomial distribution of the number of successes (x) in n trials:

$$x \sim \text{binom}(\pi, n) \quad [1]$$

where π is the probability of success representing the mean of the distribution (or the relative abundance). The statistical model for response variables with binomial distribution is the logistic or probit regression, where the relative abundance π (after logit or probit transformation) is modeled as a linear function of ≥ 1 predictors. The logit transformation of π is $\text{logit}(\pi) = \log(\pi/[1 - \pi])$, the log odds that an organism in the sample belongs to the intolerant group. This is a bivariate response variable problem. That is, for each observation, the response is a vector of 2 elements—the number of intolerant taxa and the number of non-intolerant taxa (moderately tolerant, tolerant, and unknown). We can use π_1 as the probability of being intolerant and π_2 as the probability of being nonintolerant ($\pi_1 + \pi_2 = 1$). The logit transformation of π_1 can be expressed as $\log(\pi_1/\pi_2)$, the logarithm of the probability ratio of one over the other. A statistical assumption of the model is that both π_1 and π_2 are strictly positive, implying that there are always individuals belonging to the intolerant group in a given location. When π_1 is small, the chance of observing 0 intolerant taxa in a sample is high. In other words, an observed 0 in a sample does not imply that intolerant taxa do not exist, rather that the chance of seeing an individual in that group is low. The commonly used empirical relative abundance is an estimate of the true relative abundance, which is subject to estimation uncertainty. GLM should be used with a proper probabilistic assumption on the raw count data to account for this uncertainty properly.

When there are >2 taxon groups (e.g., intolerant, moderately tolerant, tolerant, and unknown), the response variable is a vector of >2 count variables with each representing the observed counts of 1 taxon group. For example, when the relative composition of the tolerant group is considered, the response variable is a vector of 4 count variables (y_{int} , y_{mod} , y_{tol} , y_{unk}), representing the counts of intolerant, moderately tolerant, tolerant, and unknown taxon groups, respectively. The statistical model describing the distribution of these

4 count variables is the multinomial distribution (Evans et al. 2000) with 4 parameters representing the (unobservable) true relative abundances (π_{int} , π_{mod} , π_{tol} , π_{unk}):

$$\{y_{int}, y_{mod}, y_{tol}, y_{unk}\} \sim \text{multinomial}(\pi_{int}, \pi_{mod}, \pi_{tol}, \pi_{unk}, N),$$

where $N = y_{int} + y_{mod} + y_{tol} + y_{unk}$ is the total count. The 4 relative abundances are constrained to sum to unity (i.e., $\pi_{int} + \pi_{mod} + \pi_{tol} + \pi_{unk} = 1$). In general, if there are r taxon groups, we need r relative abundances (π_1, \dots, π_r) to describe the composition, but only $r - 1$ free parameters (π_2, \dots, π_r and $\pi_1 = 1 - \pi_2 - \dots - \pi_r$). If we pick the 1st group as a reference group, the logit transformation under a multinomial distribution is a set of $r - 1$ log odds ratios:

$$\text{logit}(\pi_j) = \log \frac{\pi_j}{\pi_1} \quad [2]$$

for $j = 2, \dots, r$. Statistical inference about composition is done with the logit-transformed probabilities.

Statistical models

The multinomial response model (Venables and Ripley 2002) can be described in 2 steps, as in all applications of GLM. First, a distributional assumption is made on the response variable. In analyzing species compositional data, the response variable $Y = \{y_1, \dots, y_r\}$, the observed number of occurrences, is assumed to be from a multinomial distribution:

$$Y \sim \text{multinomial}(\pi, N)$$

where $N = \sum_{j=1}^r y_j$ is the observed total abundance and $\pi = \{\pi_1, \dots, \pi_r\}$, the variable of interest, is the vector of relative abundances (with constraint $\sum_{j=1}^r \pi_j = 1$).

Second, the mean of individual relative abundance is linked to a linear function of the predictors through a link function. In the multinomial case, the mean variable is the vector of relative abundances π . Because these probabilities sum to 1, only $r - 1$ sets of free parameters need to be estimated. Setting π_1 as the baseline, the probability of occurrence is linked to the predictors through the generalized logit transformation:

$$\log \frac{\pi_j}{\pi_1} = X\beta_j \quad [3]$$

for $j = 2, \dots, r$, where the notation $X\beta_j$ represents a linear function of predictor variables, or $X\beta_j = \beta_{j0} +$

$\beta_{j1}x_1 + \dots + \beta_{jp}x_p$. Because $\sum_{j=1}^r \pi_j = 1$, we have $\pi_1 = \frac{1}{1 + \sum_{k=2}^r e^{X\beta_k}}$, and $\pi_j = \frac{e^{X\beta_j}}{1 + \sum_{k=2}^r e^{X\beta_k}}$, for $j = 2, \dots, r$. Alternatively, let $\eta_j = X\beta_j$ for $j = 1, \dots, r$, and set $\beta_1 = 0$ (so that $\eta_1 = 0$ and $e^{\eta_1} = 1$), the probability formula can be simplified to:

$$\pi_i = \frac{e^{\eta_i}}{\sum_{k=1}^r e^{\eta_k}} \quad [4]$$

for $i = 1, \dots, r$.

Multinomial regression typically is used for analyzing interrelated categorical data (Agresti 2002), such as observed counts of taxa in an assemblage. Species compositional data can also be represented by counts of taxa, which are commonly analyzed with Poisson regression for individual taxa (e.g., Cuffney et al. 2011). These 2 methods (multinomial and Poisson regressions) are statistically the same when the analysis considers the interrelatedness of the taxon groups (Baker 1994). In our paper, all 0s are assumed to be sampling 0s (Agresti 2002). When structural 0s are present, different methods should be used (Zuur et al. 2009).

Computation

The maximum likelihood estimator of the multinomial regression model is implemented in the R function *multinom* from the R package *nnet* in the VR bundle (Venables and Ripley 2002). Other implementations of the multinomial regression are available in R, but we chose *multinom* because it is well documented and has many supporting examples.

As with most statistical software, the function *multinom* will return the estimated model coefficients (β s in Eq. 3). These coefficients describe changes in the log ratios of relative abundance of a given species group over the relative abundance of the reference group. The relationship between the relative abundance and the predictor (x) is nonlinear. Consequently, the statistical significance test associated with model coefficients is often related to specific conditions and should not be interpreted as in a linear model. The tobacco budworm example in Venables and Ripley (2002) illustrates this point. Ecological interpretation of model results should be based on the relative abundances estimated using Eq. 4, which also can be presented graphically. The choice of reference group is mathematically inconsequential, and different software may select different reference groups. The R function *multinom* uses the left-most column of the species group matrix as the reference group. In

these analyses, the reference groups were the intolerant taxa for tolerance groups, piercers (PI) for functional groups, and varied among metropolitan areas for the Ephemeroptera (*Caenis* in ATL and DFW, *Baetis tricaudatus* in DEN and SLC, *Plautidius* in RAL, *Stenonema vicarium* in BOS, *Pseudocloeon propinquum* in BIR, *Dipheter hageni* in MGB, and *Epeorus* in POR). A multinomial model cannot be identified when data are completely separated, that is, a taxon (or group) is present (absent) when the predictor x (urban gradient) is less than a threshold and absent (present) when the predictor is larger than the same threshold (Section 5.8 of Gelman and Hill 2007). Generally, a completely separated taxon should be removed from the analysis, or taxa should be combined to avoid this problem.

Results

Tolerance classes

The fitted model coefficients (Table 1) describe linear models between the log ratios (relative abundance of tolerant vs intolerant [Tol], moderately tolerant vs intolerant groups [ModTol], and unidentified tolerance group vs intolerant [Unknown]) and % developed land. The intercept is the mean log ratio when % developed land is 0, and the slope is the increase of the log ratio per 1% increase in developed land. The ratio is in a logarithmic scale, so it increases at a fixed multiplicative rate for every % increase in developed land (see Qian 2010, pp. 156–157). For example, in BOS, every 1% increase in developed land would result in ~4 and 5% increase in the ratios of moderately tolerant over intolerant and tolerant over intolerant, respectively. The approximate 95% confidence intervals of these slopes (mean ± 2 SE) can be used to check for statistical significance. If the confidence interval includes 0, the slope is not statistically different from 0 at a significance level of ~0.05. Slopes from the 3 metropolitan areas with high AAG (DEN, DFW, and MGB) were not statistically different from 0, and slopes from the other 6 areas were different from 0. Many macroinvertebrate metrics from the 3 metropolitan areas with high AAG also were unresponsive to urbanization (Cuffney et al. 2010, Kashuba et al. 2010, Qian et al. 2010). Interpretation of statistical significance of model coefficients is related specifically to the ratio of relative abundance of an individual group over the reference group, not the relative abundance of the group in question. See the Appendix (available online from: <http://dx.doi.org/10.1899/11-026.1.s1>) for model results using a different reference group.

We used the estimated model coefficients to calculate the relative abundance of each of the 4 groups along an urban gradient using Eq. 4. (Note that the choice of

TABLE 1. Estimated model coefficients (Coef) and their standard errors (SE) for responses of moderately tolerant (Modtol), tolerant (Tol), and unknown tolerance (Unknown) groups to urbanization relative to responses of intolerant (Intol) groups. See Fig. 2 for location codes.

City	Ratio to Intol	Intercept		Slope	
		Coef	SE	Coef	SE
ATL	Modtol	1.5372	0.1755	0.0425	0.0097
	Tol	-0.0233	0.2028	0.0519	0.0100
	Unknown	0.0120	0.2050	0.0441	0.0102
BIR	Modtol	0.6349	0.1539	0.0158	0.0040
	Tol	-0.6601	0.1917	0.0265	0.0045
	Unknown	-1.1563	0.2251	0.0239	0.0051
BOS	Modtol	0.2072	0.0948	0.0392	0.0053
	Tol	-1.3007	0.1383	0.0485	0.0061
	Unknown	-2.4987	0.2264	0.0443	0.0086
DEN	Modtol	1.8831	0.2417	0.0096	0.0055
	Tol	1.5007	0.2482	0.0085	0.0056
	Unknown	-1.1133	0.4181	0.0123	0.0084
DFW	Modtol	3.0901	0.3316	0.0013	0.0088
	Tol	3.1614	0.3312	0.0005	0.0088
	Unknown	2.5080	0.3375	-0.0015	0.0090
MGB	Modtol	1.5603	0.1669	0.0180	0.0057
	Tol	0.7849	0.1800	0.0192	0.0059
	Unknown	-0.1164	0.2109	0.0204	0.0063
POR	Modtol	0.4577	0.1147	0.0234	0.0039
	Tol	-0.6098	0.1456	0.0264	0.0043
	Unknown	-2.0101	0.2484	0.0211	0.0064
RAL	Modtol	0.9518	0.1700	0.0241	0.0047
	Tol	0.0381	0.1924	0.0289	0.0049
	Unknown	-0.3039	0.2091	0.0255	0.0051
SLC	Modtol	-0.0503	0.1875	0.0264	0.0037
	Tol	-1.1384	0.2342	0.0357	0.0042
	Unknown	-3.0216	0.4816	0.0342	0.0076

reference group will not change the outcome.) We estimated uncertainty about the estimated relative abundance with a Monte Carlo simulation that drew random samples of model coefficients assuming that the coefficients followed a multivariate normal distribution. In Fig. 1, model-predicted relative abundance (50 and 95% credible intervals) was compared to observed relative abundance from BOS. The rapid decrease of the intolerant group and the rapid increase of the tolerant group reflected the definition of these groups even though the TVs in Barbour et al. (1999) and NCDENR (2006) were not specific to urbanization. The other 5 metropolitan areas with low AAG had patterns similar to BOS, whereas the 3 metropolitan areas with high AAG (DEN, DFW, MGB; Cuffney et al. 2010) generally showed no response in the relative abundance of the 3 tolerance groups to urbanization (Fig. 2).

Functional groups

The multinomial regression models relating the relative abundance of the functional groups to urbanization did not respond strongly in any metropolitan areas (Fig. 3). Instead, responses were generally flat

across the gradient of urban intensity because the apparent increases or decreases in some groups were statistically insignificant when considering the estimation uncertainty. The flat responses indicate that the relative composition of the functional groups remained constant and, unlike relative composition of tolerance classes, was not affected by the amount of AAG in the metropolitan area. The difficulty of graphically presenting multinomial model results is clearly shown in Figs 2 and 3. For example, each panel of Fig. 3 can be expanded into 8 separate ones (similar to Fig. 1) to show the details of each group's response. Figs 2 and 3 should be used as tools to examine the general patterns. Regression coefficients for these models are in Table S2 in the Appendix.

Ephemeroptera taxa

The multinomial regression models relating mayfly taxa to urban intensity showed strong responses in most of the metropolitan areas. However, many of the models were adversely affected by the presence of rare taxa that created a complete separation between

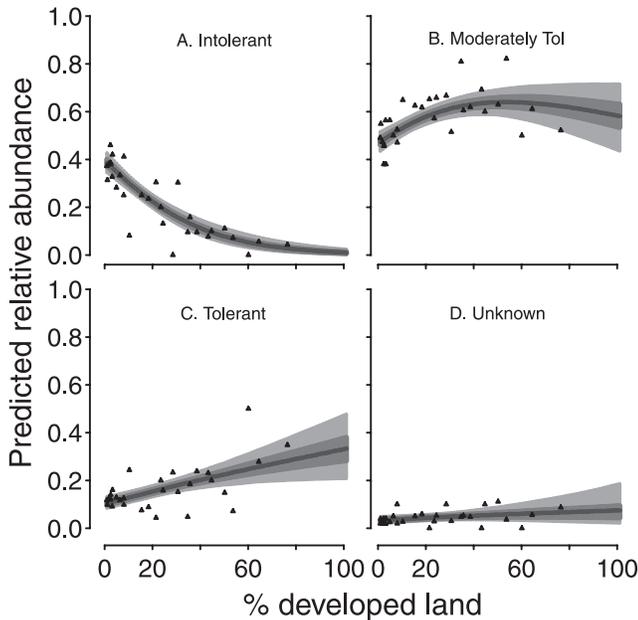


FIG. 1. Relative abundances of intolerant (A), moderately tolerant (B), tolerant (C), and unknown tolerance (D) groups estimated with a multinomial model (dark circles) compared to the observed relative abundances (triangles) along a gradient of urbanization (% developed land) for the Boston metropolitan area. The 50% and 95% confidence intervals of the estimations are represented by the dark and light shaded polygons, respectively. The prediction is extrapolated beyond the data range to show the nonlinear patterns of the fitted models.

presence and absence along the urban gradient. For example, the model for POR (Fig. 4A) was strongly affected by the perfect separation of Leptohyphidae and *Epeorus* sp., both with all 0 counts, except 1 positive count at the low end of the urban gradient. This positive count caused the estimated relative abundance for *Epeorus* sp. to be close to 1 and forced the estimates for all others to be close to 0. (See Appendix for further discussion of the influence of rare taxa.) Rare taxa represent a small fraction of the total and often exhibit little change in relative composition. When a rare taxon creates a complete separation, this single taxon causes a very high level of uncertainty in the estimated model coefficients and leads to distortions in the response patterns of the other taxa because the modeled probabilities must sum to 1. These distortions are most evident at the low end of the urbanization gradient because rare mayfly taxa are much more likely to be associated with the low end of the gradient (Fig. 4A). Removing taxa with complete separations can resolve the numerical problem (Fig. 4B) and clarify patterns of responses for the other taxa, particularly at the low end of the urban gradient.

Responses of the mayfly taxa differed substantially among metropolitan areas. The most complicated response was observed in ATL (Fig. 5A) where many taxa showed responses indicative of species replacement across the urban gradient. In contrast, in agriculturally affected MGB, only 2 taxa showed strong responses. These taxa displayed asymptotic increases (*Baetis flavistriga*) and decreases (*Leucrocuta* sp.) in relative abundance over only a portion of the urban gradient (0–20% developed land, Fig. 5B). No changes in assemblage composition were evident in MGB when % developed land was >20%.

We summarized responses in other metropolitan areas based on the types of responses observed in ATL and MGB, i.e., Gaussian responses (e.g., *Pseudocloeon* sp. in ATL), monotonically declining responses (e.g., *Leucrocuta* sp. in MGB), and monotonically increasing responses (e.g., *Baetis flavistriga* in MGB). Each of these patterns of response had an associated level of urbanization that indicated peak relative abundance (Gaussian) or the point at which monotonic increases began or decreases ended. These characteristics were used to compare responses among taxa and across metropolitan areas (Table 2). A more complete and comprehensive representation of mayfly responses is provided in the Appendix (coefficients: Table S3, graphs of composition changes: Figs S1, S2).

Mayfly responses were defined by just 2 or 3 taxa in most (6 of 9) metropolitan areas. BOS and ATL were exceptions in that responses were defined by 6 and 8 taxa, respectively. The composition of mayfly assemblages varied among metropolitan areas, and no taxon was present in all metropolitan areas. Of the mayfly taxa that showed strong responses to urbanization, the most widely distributed was *Baetis flavistriga* (6 metropolitan areas) followed by *Baetis intercalaris* (5), *Baetis tricaudatus* (4), *Acentrella turbida* (3), *Paraleptophlebia* (4), and *Isonychia* (4). The pattern of responses varied widely among taxa within a metropolitan area and for taxa across metropolitan areas (Table 2). The most consistent response was observed for *Baetis flavistriga*, for which relative abundance increased as urbanization increased in the 6 metropolitan areas in which it occurred. However, the increases began immediately in 3 metropolitan areas and at intermediate levels of urbanization (20–50% developed land) in 3 others. *Acentrella turbida* decreased at intermediate levels of urbanization (20–40% developed land) in 3 metropolitan areas, but did not respond in ATL. Responses of *Baetis tricaudatus* encompassed all possibilities (none, increasing, decreasing, and Gaussian) depending upon the metropolitan area. Similar differences in responses were observed for other taxa.

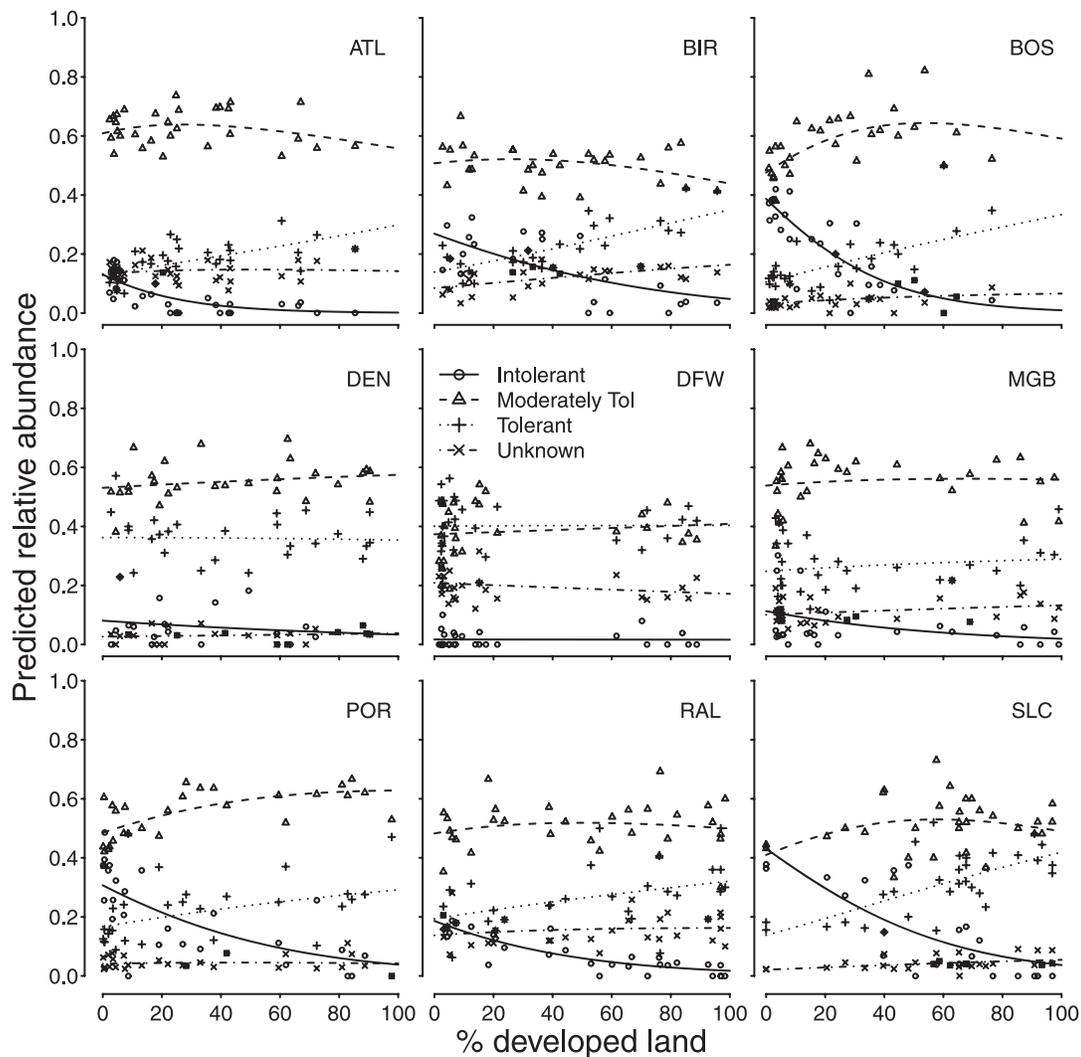


FIG. 2. Relative abundances of tolerance groups estimated with a multinomial model compared to the observed relative abundances along a gradient of urbanization (% developed land) for the 9 metropolitan areas (Atlanta, Georgia [ATL]; Birmingham, Alabama [BIR]; Boston, Massachusetts [BOS]; Denver, Colorado [DEN]; Dallas–Fort Worth, Texas [DFW]; Milwaukee–Green Bay, Wisconsin [MGB]; Portland, Oregon [POR]; Raleigh, North Carolina [RAL]; and Salt Lake City, Utah [SLC]). These plots compare responses among taxon groups and metropolitan areas. When the response of an individual group in a specific area is of interest, plots should include confidence intervals (Fig. 1A–D).

These results indicate that responses based on relative abundance differed among metropolitan areas and were not consistent with the pattern observed for assemblage metrics.

Discussion

Why multinomial regression?

Expressing community responses as a proportion of total abundance emphasizes the interconnections of the structural components (species or metrics) of the community as components respond to disturbance. Accurate depiction of these responses is important to

developing an integrated understanding of community rather than individual taxon responses. Multinomial regression is better suited to capturing these nonlinear interrelated responses than is simple linear regression. Thus, multinomial regressions capture patterns of response that are not obvious when other modeling techniques are used.

Mathematically, interrelated responses are modeled by the multinomial distribution, which connects observed counts to relative abundances. These relative abundances are not observed directly (although the calculated relative abundances are estimates), and they are linked to a disturbance variable through

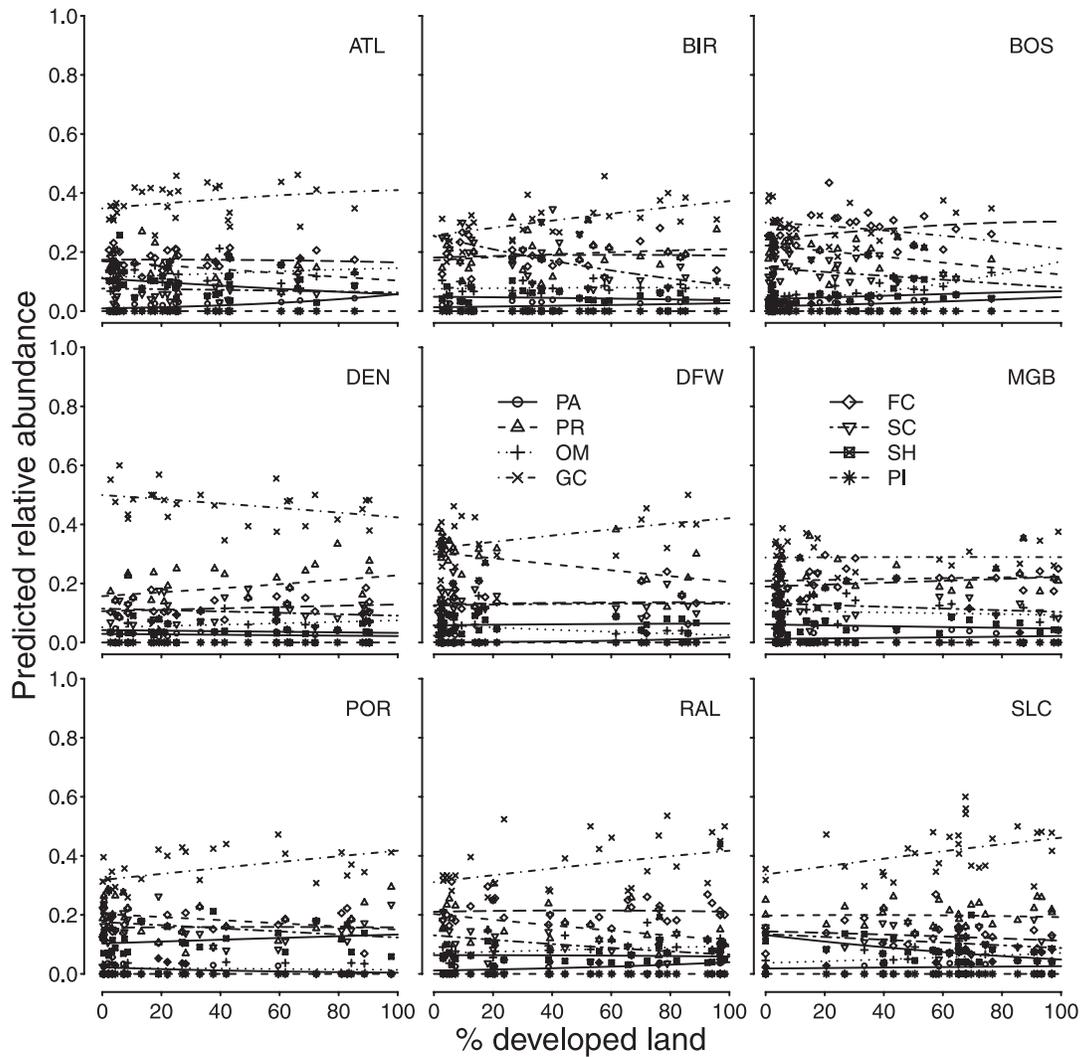


FIG. 3. Relative abundances of the 8 functional groups estimated with a multinomial model compared to the observed relative abundances along a gradient of urbanization (% developed land) in 9 metropolitan areas (see Fig. 2 for location codes). This plot compares responses among taxon groups and metropolitan areas. When the response of an individual group in a specific area is of interest, plots should include confidence intervals (Fig. 1A–D). PA = parasite, PR = predator, OM = omnivore, GC = collector-gatherer, FC = collector-filterer, SC = scraper, SH = shredder, PI = piercer.

a nonlinear function as in Eq. 4. The nonlinear multinomial regression model (Eq. 4) can encompass many types of responses, including monotonic increasing or decreasing responses and various shapes of unimodal curves. Simple regression models with appropriate data transformations also can be used to match the nonlinear pattern shown in our study (e.g., models in Fig. 1 can be approximated by a quadratic model). However, multinomial regression is far more efficient in that the individual models shown in Eq. 4 are fit without a tedious model-selection process. Furthermore, a multinomial model properly accounts for the interaction among multiple taxon groups through the constraint that relative abundances sum

to 1. When graphically presented, a multinomial regression model provides an effective means of displaying the change in the relative importance of various taxon groups or biological traits along a gradient.

Model interpretation

Multinomial regression is more complicated than simple regression, and model coefficients are interpreted differently. In our model, the coefficients (β s) of the multinomial regression model estimate the rate at which the log of the odds ratio of 2 categories changes as the predictor variable changes (% change

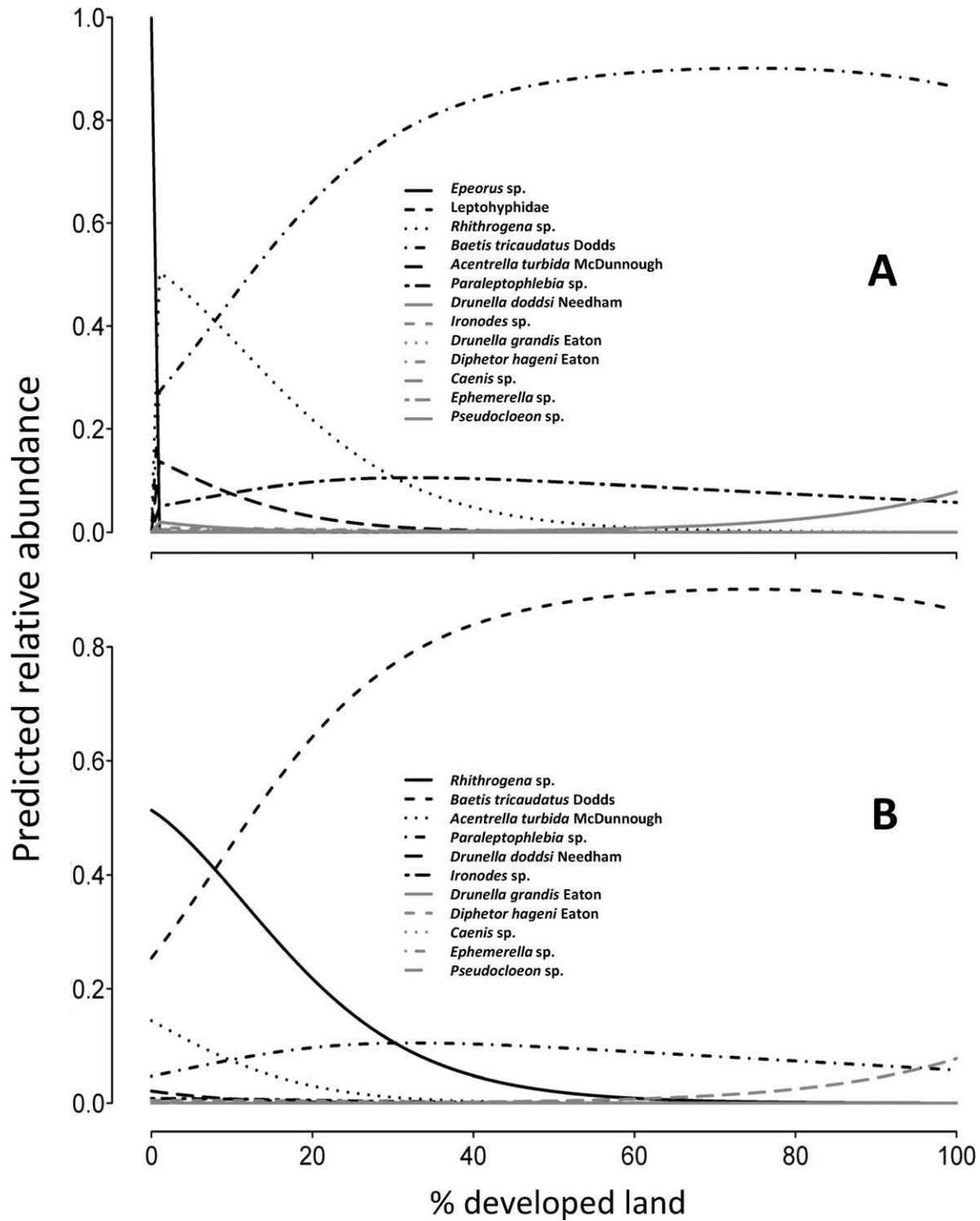


FIG. 4. Relative abundances of mayfly taxa estimated with a multinomial model for the Portland (POR) metropolitan area showing the effect of a perfect separation where taxa *Epeorus sp.* and Leptohiphidae were present at the low end of the urban gradient and absent elsewhere (A) and the effect of removing these taxa (B). In panel A, the uncertainty in the estimates of *Epeorus sp.* and Leptohiphidae obscure estimates for the other taxa, whereas in panel B, their removal reveals the responses of the other taxa.

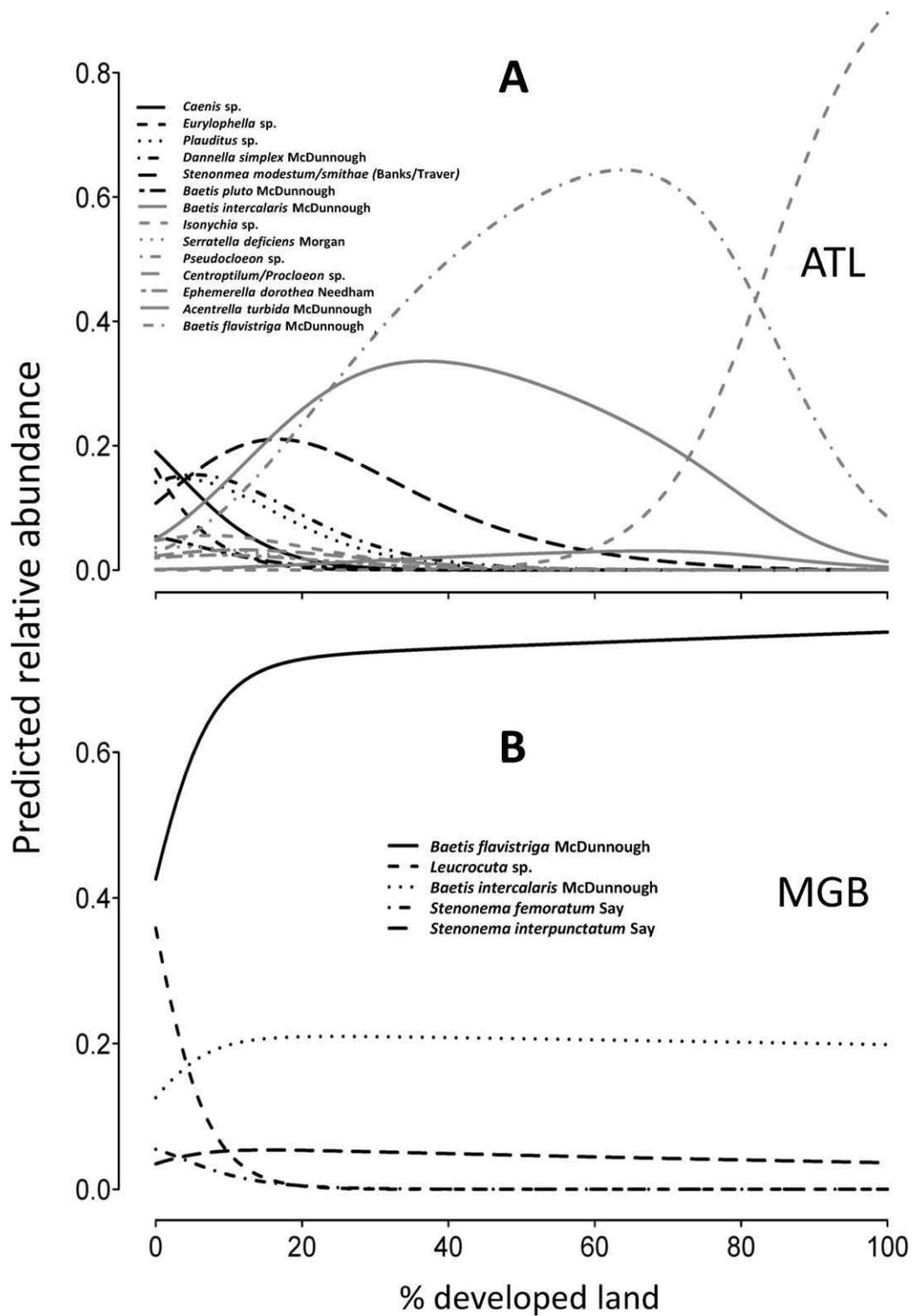


FIG. 5. Relative abundances of mayfly taxa estimated with a multinomial model for an urban area with low antecedent agricultural (AAG) land use (Atlanta [ATL]) (A) and one with high antecedent agriculture (Milwaukee–Green Bay [MGB]) (B). Percent developed land induced a succession of taxa along the urbanization gradient in a metropolitan area with low AAG, whereas little or no response was detected in a metropolitan area with high AAG.

TABLE 2. Taxa that displayed large changes (≥ 0.15) in modeled relative abundance over the urban gradient in ≥ 1 metropolitan area. Arrows indicate the general pattern of change in relative abundance over the urban intensity gradient (% developed land) and the numbers indicate the beginning, end, or peak location of the change (e.g., $\downarrow 25$ indicates a monotonic decrease over the range of 0–25%, $\uparrow 50$ indicates a monotonic increase over the range of 50–100%, and $\uparrow 50 \downarrow$ indicates a Gaussian response with maximum relative abundance at 50%), \leftrightarrow indicates little change over the gradient, – indicates that the taxon was not present in the metropolitan areas or the change was < 0.15 . See Fig. 2 for location codes.

Taxon	ATL	BIR	BOS	POR	RAL	SLC	DEN	DFW	MGB
Caenidae	–	–	–	–	–	–	–	–	–
<i>Caenis</i>	$\downarrow 25$	–	–	–	–	–	–	$\uparrow 18 \downarrow$	–
Ephemereididae	–	–	–	–	–	–	–	–	–
<i>Dannella simplex</i>	$\uparrow 7 \downarrow$	–	–	–	–	–	–	–	–
<i>Drunella flavilinea</i>	–	–	–	–	–	57 \uparrow	–	–	–
<i>Eurylophella</i>	$\downarrow 20$	–	–	–	–	–	–	–	–
<i>Serratella deficiens</i>	\leftrightarrow	\leftrightarrow	$\downarrow 30$	–	–	–	–	–	–
<i>Serratella serrata</i>	–	–	$\downarrow 16$	–	–	–	–	–	–
Leptohyphidae	–	–	–	–	–	–	–	–	–
<i>Tricorythodes</i>	–	\leftrightarrow	–	–	–	–	$\downarrow 100$	$\downarrow 100$	–
Leptophlebiidae	–	–	–	–	–	–	–	–	–
<i>Paraleptophlebia</i>	–	–	$\downarrow 20$	\leftrightarrow	\leftrightarrow	–	\leftrightarrow	–	–
Baetidae	–	–	–	–	–	–	–	–	–
<i>Acentrella turbida</i>	\leftrightarrow	$\downarrow 40$	–	$\downarrow 35$	–	–	–	–	–
<i>Baetis flavistriga</i>	50 \uparrow	0 \uparrow	0 \uparrow	–	20 \uparrow	–	20 \uparrow	–	0 \uparrow
<i>Baetis intercalaris</i>	$\uparrow 38 \downarrow$	\leftrightarrow	–	\leftrightarrow	$\uparrow 50 \downarrow$	–	\leftrightarrow	–	–
<i>Baetis tricaudatus</i>	–	–	\leftrightarrow	0 \uparrow	–	$\downarrow 100$	$\uparrow 70 \downarrow$	–	–
<i>Fallceon quilleri</i>	–	–	–	–	–	–	\leftrightarrow	0 \uparrow	–
<i>Plauditus</i>	$\uparrow 6 \downarrow$	–	\leftrightarrow	–	$\downarrow 80$	–	–	–	–
<i>Pseudocloeon</i>	$\uparrow 62 \downarrow$	–	–	–	–	–	–	\leftrightarrow	–
Heptageniidae	–	–	–	–	–	–	–	–	–
<i>Leucrocota</i>	–	\leftrightarrow	–	–	$\downarrow 80$	–	–	–	$\downarrow 20$
<i>Rhithrogena</i>	–	–	–	$\downarrow 70$	–	–	–	–	–
<i>Stenonema modestum</i>	$\uparrow 18 \downarrow$	–	$\uparrow 18 \downarrow$	–	\leftrightarrow	–	–	–	–
Isonychiidae	–	–	–	–	–	–	–	–	–
<i>Isonychia</i>	\leftrightarrow	\leftrightarrow	$\uparrow 18 \downarrow$	–	$\downarrow 80$	–	–	–	–

in ratio per 1 unit [1%] change in predictor). The odds ratios are expressed relative to a baseline condition (e.g., relative abundance of intolerant taxa). Consequently, the interpretation and ecological relevance of the coefficient depends on the category chosen as the baseline condition. For example, knowing that the ratio of the relative abundances of tolerant and intolerant taxa increases by 3 to 5% per 1 unit increase in urbanization (1% developed land) in metropolitan areas with low AAG is ecologically relevant, as is knowing that this rate is lower ($< 2\%$ per 1 unit change in urbanization) or not statistically significant in areas with high AAG. In contrast, if the unknown tolerance category were used as the baseline condition, the regression coefficients would represent the rate of change in the ratio of the occurrence of tolerant taxa to the occurrence of taxa whose functional group could not be identified. This ratio is not particularly relevant to understanding or managing the response of invertebrates to urbanization.

Selection of baseline condition becomes increasingly problematic as the data are divided into more categories resulting in a larger number of possible baseline conditions (e.g., modeling individual taxa rather than metrics). The mayfly examples presented in our paper had 5 to 14 possible baseline taxa per metropolitan area compared to 4 and 8 possible baseline conditions for tolerance and functional groups, respectively. An analysis of the entire macroinvertebrate assemblage would have involved 116 to 162 taxa making the process of determining the most ecologically relevant baseline a daunting task. Consequently, analyses that rely on interpretation of the coefficients of the multinomial regression must be done with careful consideration of the number of possible baseline conditions to be evaluated and of the ecological or management relevance of the relative abundance ratios. Interpretations of coefficients are most likely to be meaningful when the number of categories in the analysis is relatively small. Thus,

analyses of aggregations of taxa (assemblage metrics) that focus on a small number of ecologically relevant traits are more likely to be amenable to interpretation of the multinomial regression coefficients than analyses of large numbers of individual taxa.

Interpretations of the regression coefficients are affected by the choice of baseline condition, but the estimated relative abundances are not. Consequently, plotting the relative abundances against the predictor variable provides a simple and consistent means of visualizing responses across the gradient and comparing the model prediction with the observed data (Fig. 1A–D). Plots of the tolerance classes clearly showed that the relative abundance of the intolerant group decreased as urban intensity increased, whereas the relative abundance of the tolerant group increased for metropolitan areas with little AAG. These plots also revealed that metropolitan areas with high levels of AAG (DEN, DFW, and MGB) exhibited little response across the urban gradient. These patterns of response are consistent with our previous analyses and with results of other studies in which decreases in tolerance have been associated with increasing urbanization and agriculture. Functional groups showed little response to urbanization or agriculture. These results are consistent with our previous analyses, but they differ from those of others who have reported changes in functional structure of the invertebrate assemblages in response to urbanization and agriculture (Bacey and Spurlock 2007, Turner et al. 2008).

Species traits vs individual taxa and the problem of perfect separation

Plots of the estimated relative abundances for the mayfly taxa showed strong responses in most metropolitan areas. However, these plots revealed that multinomial regression, like other univariate and multivariate methods of analysis, are strongly affected by taxa that occur at only a few sites (i.e., rare taxa). Regression coefficients for these taxa cannot be determined with precision, often because of the problem of perfect separation. Errors in the estimation of the response of rare taxa may affect multinomial regressions even more strongly than simple linear regressions because the estimated relative abundances for the categories in a multinomial regression must sum to 1. Therefore, a large error in the estimation of the relative abundance of a rare taxon will affect the estimates of the relative abundances of other taxa. In contrast, simple regressions estimate the occurrence (i.e., relative abundance) of each category independently, and errors in one regression do not directly affect estimates in other regressions.

In our study, the problem of rare taxa was addressed by removing them from the analysis (especially those that created a complete separation). This approach is commonly used in univariate and multivariate analyses of taxon distributions (Gauch 1982, Clarke 1993, Jongman et al. 1995, McCune et al. 2002). The approach is based on the assumption that rare taxa do not carry useful information because their responses cannot be determined with precision (i.e., rare taxa constitute noise in the data set). However, in the case of the mayfly taxa, rare taxa did not represent random noise. Instead, most of the rare taxa (21 of 30 cases) occurred at sites with $\leq 10\%$ developed land (Fig. 2), a result indicating that they are far more characteristic of sites with low urbanization than of sites with high urbanization, which is why perfect separation is often associated with rare taxa. Eliminating rare taxa from the analysis removes this information and may result in a somewhat biased view of the distribution of mayfly taxa across the urban gradient. In contrast, assemblage metrics are not affected by rare taxa and are able to incorporate information that is lost in the analysis of taxon distributions.

Analysis of assemblage metrics (aggregations of taxon traits) offers some advantages compared to analyses of individual taxa. Metrics aggregate data into a smaller number of categories. This characteristic simplifies selection of an appropriate reference category that has ecological or management significance when compared to other categories. Metric categories often have direct ecological significance (e.g., tolerance or functional groups) that arises from traits associated with the taxon. Metrics also are able to incorporate information from rare taxa (e.g., tolerance values) that would be discarded in the analysis of individual taxa. This ability can be important, as demonstrated by our study, because 36 to 47% of the taxa in each metropolitan area occur at just 1 or 2 sites. The mathematical inability of multinomial regression to handle rare taxa is another consideration when deciding whether a functional or a structural approach is more appropriate for assessing the effects of land use on stream ecosystems (Dolédec et al. 2006, Townsend et al. 2008).

Concluding remarks

Multinomial regression is a statistically appropriate method for modeling changes in invertebrate assemblage composition (relative abundances) that involve >2 categories. The model can be used to analyze raw count data obtained from quantitative (e.g., Surber) or semiquantitative (e.g., kick net) sampling methods, as

long as the sampling effort is consistent across all sites. As a GLM method, multinomial regression is able to handle the distribution (multinomial vs normal) of the data in a manner that addresses the interrelatedness of the relative abundance data. This interrelatedness could not be addressed with the correlations and simple linear regressions used in our previous analyses of relative abundances (Cuffney et al. 2005, 2010, Brown et al. 2009). Application of the multinomial regressions to our data supported our original conclusions regarding the pattern of responses observed for tolerance, functional groups, and mayfly taxa. The relative abundance of intolerant taxa decreased as urbanization increased and the relative abundance of tolerant forms increased, a pattern well established in the literature (Paul and Meyer 2001, Roy et al. 2003, Walsh et al. 2005, Wenger et al. 2009). Functional groups were not observed to respond to urbanization or agricultural development in any of our analyses though other investigators have shown changes in functional composition associated with urbanization and agriculture (Bacey and Spurlock 2007, Turner et al. 2008). The multinomial regressions were also able to show patterns in the distribution of mayfly taxa across the gradient that were not obvious in our previous analyses (e.g., approximate Gaussian responses in ATL). We considered only 1 predictor (% developed land) in the examples presented here, but multiple predictors can be used in multinomial regression just as in a multiple regression problem.

The absence of change in the functional composition of the macroinvertebrate communities in response to perturbations caused by urbanization and degree of agricultural development (i.e., amount of AAG) differed from results of other studies in which functional groups changed with urbanization. For example, Turner et al. (2008) observed changes in functional groups and taxon richness in neotropical stream communities affected by agriculture and urbanization, and Bacey and Spurlock (2007) observed increases in scrapers at urban sites and filter-feeders at agricultural sites in California Central Valley streams. Our results suggest that macroinvertebrate communities retain their functional characteristics despite urban and agricultural development in the watershed, at least in terms of relative abundances.

Acknowledgements

We thank Roxolana Kashuba and Ibrahim Alameddine for their insightful discussions and comments. Comments and suggestions from Lester Yuan and John Van Sickle, who provided formal reviews of an early version of this manuscript, are greatly appreciated. The

research was completed while SSQ was partially supported by the USGS through an USGS-Duke University corporative agreement (08HQAG0121). Any use of trade, product, or firm names is for descriptive purposes only and does not imply endorsement by the US Government.

Literature Cited

- AGRESTI, A. 2002. *Categorical data analysis*. 2nd edition. Wiley, New York.
- BACEY, J., AND F. SPURLOCK. 2007. Biological assessment of urban and agricultural streams in the California Central Valley. *Environmental Monitoring and Assessment* 130: 483–493.
- BAKER, S. G. 1994. The multinomial-Poisson transformation. *Statistician* 43:495–504.
- BARBOUR, M. T., J. GERRITSEN, B. D. SNYDER, AND J. B. STRIBLING. 1999. *Rapid bioassessment protocols for use in streams and wadeable rivers: periphyton, benthic macroinvertebrates and fish*. 2nd edition. EPA 841-b-99002. Office of Water, US Environmental Protection Agency, Washington, DC.
- BARBOUR, M. T., AND M. J. PAUL. 2010. Add value to water resource management through biological assessment of rivers. *Hydrobiologia* 651:17–24.
- BROWN, L. R., T. F. CUFFNEY, J. F. COLES, F. FITZPATRICK, G. MCMAHON, J. STEUER, A. H. BELL, AND J. T. MAY. 2009. Urban streams across the USA: lessons learned from studies in 9 metropolitan areas. *Journal of the North American Benthological Society* 28:1051–1069.
- CLARKE, K. R. 1993. Non-parametric multivariate analyses of changes in community structure. *Australian Journal of Ecology* 18:117–143.
- COLES, J. F., T. F. CUFFNEY, AND G. MCMAHON. 2004. *The effects of urbanization on the biological, physical, and chemical characteristics of coastal New England streams*. U.S. Geological Survey Professional Paper 1695. US Geological Survey, Reston, Virginia.
- CUFFNEY, T. F., R. A. BRIGHTBILL, J. T. MAY, AND I. R. WAITE. 2010. Responses of benthic macroinvertebrates to environmental changes associated with urbanization in nine metropolitan areas of the conterminous United States. *Ecological Applications* 20:1384–1401.
- CUFFNEY, T. F., AND J. F. FALCONE. 2008. Derivation of nationally consistent indices representing urban intensity within and across nine metropolitan areas of the conterminous United States. U.S. Geological Survey Scientific Investigations Report 2008–5095. US Geological Survey, Reston, Virginia.
- CUFFNEY, T. F., R. KASHUBA, S. S. QIAN, I. ALAMEDDINE, Y. K. CHA, B. LEE, J. F. COLES, AND G. MCMAHON. 2011. Multilevel regression models describing regional patterns of invertebrate and algal responses to urbanization across the USA. *Journal of the North American Benthological Society* 30:797–819.
- CUFFNEY, T. F., H. ZAPPALÀ, E. M. P. GIDDINGS, AND J. F. COLES. 2005. Effects of urbanization on benthic macroinverte-

- brate assemblages in contrasting environmental settings: Boston, Massachusetts; Birmingham, Alabama; and Salt Lake City, Utah. Pages 361–407 in L. R. Brown, R. M. Hughes, R. Gray, and M. R. Meador (editors). *Effects of urbanization on stream ecosystems*. Symposium 47. American Fisheries Society, Bethesda, Maryland.
- DAVIES, S. P., AND S. K. JACKSON. 2006. The biological condition gradient: A descriptive model for interpreting change in aquatic ecosystems. *Ecological Applications* 16:1251–1266.
- DOLÉDEC, S., N. PHILIPS, M. SCARSBROOK, R. H. RILEY, AND C. R. TOWNSEND. 2006. Comparison of structural and functional approaches to determining landuse effects on grassland stream invertebrate communities. *Journal of the North American Benthological Society* 25:44–60.
- EVANS, M., N. HASTINGS, AND B. PEACOCK. 2000. *Statistical distributions*. 3rd edition. Wiley, New York.
- FAHRMEIR, L., AND G. TUTZ. 1994. *Multivariate statistical modelling based on generalized linear models*. Springer, New York.
- GAUCH, H. G. 1982. *Multivariate analysis in community ecology*. Cambridge University Press, New York.
- GELMAN, A., AND J. HILL. 2007. *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press, New York.
- JONGMAN, R. H. G., C. J. F. TER BRAAK, AND O. F. R. VAN TONGEREN (EDITORS). 1995. *Data analysis in community and landscape ecology*. Cambridge University Press, New York.
- KASHUBA, K., C. YOON KYUNG, I. ALAMEDDINE, B. LEE, AND T. F. CUFFNEY. 2010. Multilevel hierarchical modeling of benthic macroinvertebrate responses to urbanization in nine metropolitan regions across the conterminous United States. U.S. Geological Survey Scientific Investigations Report 2009-5243. US Geological Survey, Reston, Virginia.
- KENNEN, J. G., AND M. A. AYERS. 2002. Relation of environmental characteristics to the composition of aquatic assemblages along a gradient of urban land use in New Jersey, 1996–98. U.S. Geological Survey Water Resources Investigations Report 02-4069. US Geological Survey, West Trenton, New Jersey.
- MCCULLAGH, P., AND J. A. NELDER. 1989. *Generalized linear models*. 2nd edition. Chapman and Hall/CRC Press, Boca Raton, Florida.
- MCCUNE, B., J. B. GRACE, AND D. L. URBAN. 2002. *Analysis of ecological communities*. MjM Software Design, Glendeden Beach, California.
- MCMAHON, G., AND T. F. CUFFNEY. 2000. Quantifying urban intensity in drainage basins for assessing stream ecological conditions. *Journal of the American Water Resources Association* 36:1247–1261.
- NCDENR (NORTH CAROLINA DEPARTMENT OF ENVIRONMENT AND NATURAL RESOURCES). 2006. *Standard operating procedures for benthic macroinvertebrates*. Biological Assessment Unit, Division of Water Quality, North Carolina Department of Environment and Natural Resources, Raleigh, North Carolina. (Available from: <http://www.esb.enr.state.nc.us/BAUwww/benthossop.pdf>)
- PAUL, M. J., AND J. L. MEYER. 2001. Streams in the urban landscape. *Annual Review of Ecology and Systematics* 32:333–365.
- QIAN, S. S. 2010. *Environmental and ecological statistics with R*. Chapman and Hall/CRC Press, Boca Raton, Florida.
- QIAN, S. S., T. F. CUFFNEY, I. ALAMEDDINE, G. MCMAHON, AND K. H. RECKHOW. 2010. On the application of multilevel modeling in environmental and ecological studies. *Ecology* 91:355–361.
- ROY, A. H., A. D. ROSEMOND, M. J. PAUL, D. S. LEIGH, AND J. B. WALLACE. 2003. Stream macroinvertebrate response to catchment urbanization (Georgia, U.S.A.). *Freshwater Biology* 48:329–346.
- TOWNSEND, C. R., S. S. UHLMANN, AND C. D. MATTHAEI. 2008. Individual and combined responses of stream ecosystems to multiple stressors. *Journal of Applied Ecology* 45:1810–1819.
- TURNER, D., D. D. WILLIAMS, AND M. ATKINS-KOO. 2008. Longitudinal changes in benthic community composition in four neotropical streams. *Caribbean Journal of Science* 44:380–394.
- VENABLES, W. N., AND B. D. RIPLEY. 2002. *Modern applied statistics with S*. 4th edition. Springer, New York.
- WALSH, C. J., A. H. ROY, J. W. FEMINELLA, P. D. COTTINGHAM, AND P. M. GROFFMAN. 2005. The urban stream syndrome: current knowledge and search for a cure. *Journal of the North American Benthological Society* 24:706–723.
- WALSH, C. J., A. K. SHARPE, P. F. BREEN, AND J. A. SONNEMAN. 2001. Effects of urbanization on streams of the Melbourne region, Victoria, Australia – I. benthic macroinvertebrate communities. *Freshwater Biology* 46:535–551.
- WENGER, S. J., A. H. ROY, C. R. JACKSON, E. S. BERNHARDT, T. L. CARTER, S. FILOSO, C. A. GIBSON, W. C. HESSON, S. S. KAUSHAL, E. MARTÍ, J. L. MEYER, M. A. PALMER, M. J. PAUL, A. H. PURCELL, A. RAMÍREZ, A. D. ROSEMOND, K. A. SCHOFIELD, E. B. SUDDUTH, AND C. J. WALSH. 2009. Twenty-six key research questions in urban stream ecology: an assessment of the state of the science. *Journal of the North American Benthological Society* 28:1080–1098.
- WHITTAKER, R. H. 1967. Gradient analysis of vegetation. *Biological Reviews* 49:207–264.
- ZUUR, A. F., E. N. IENO, N. J. WALKER, A. A. SAVELIEV, AND G. M. SMITH. 2009. *Mixed effects models and extensions in ecology with R*. Springer, New York.

Received: 22 March 2011

Accepted: 5 March 2012