

*Editorial*

## **Modelling and predicting invertebrate abundance along environmental gradients**

Raphael K. Didham

*School of Biological Sciences, University of Canterbury, Private Bag 4800, Christchurch (raphael.didham@canterbury.ac.nz)*

One of the central goals of ecologists and entomologists is to understand the distribution and abundance of organisms. Because we are interested not only in *how* individuals of a given species are distributed, but also *why* they are distributed the way they are, we often test (or model) the relationship between invertebrate abundance and environmental variables. In addition to increasing our understanding of the life history and ecology of species, this type of modelling has the advantage of allowing predictions to be made about the likely location and size of un-sampled populations in areas with similar environmental parameters. What is more, repeated quantitative samples from the same site can allow invertebrate populations to be monitored in relation to changing environmental gradients (such as land use change or climate change). In all of these cases, measuring, monitoring and predicting species distributions traditionally relies on estimating the central tendency of abundance as a function of one or more dominant environmental gradients.

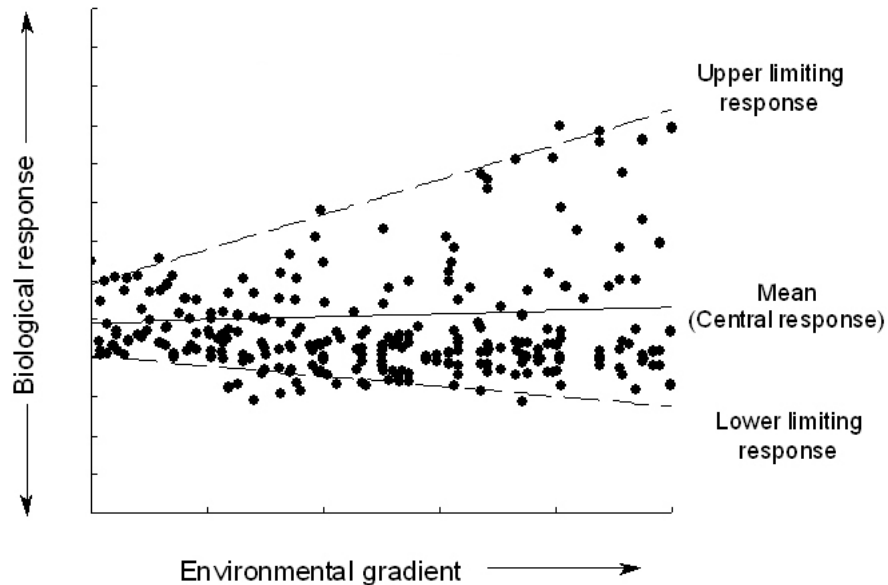
The problem with this approach is that conventional correlation and regression analyses fundamentally conflict with the basic ecological tenet of the 'law of limiting factors' (Liebig's law of the minimum). Liebig's law postulates that there is typically one factor which is least available, out of the total set of factors affecting growth, reproduction and survival of an organism, and that this factor controls population abundance even if conditions appear optimal for a range of other variables. In a landmark paper 10 years ago, Thomson *et al.* (1996) argued that if multiple factors limit the abundance of organisms, and if the law of limiting factors holds true, then the last thing we would expect from a correlation or regression analysis would be a tight statistical relationship between abundance and an environmental variable. More typically, we might expect an 'upper ceiling' imposed on abundance by the dominant limiting factor (Gaines & Denny 1993), below which abundance at individual sites might vary dramatically due to the interaction with other environmental variables (Thomson *et al.* 1996). Thomson *et al.* (1996) called these 'factor-ceiling' distributions and showed how they could produce statistically non-significant abundance-environment correlations, when in fact there were biologically important relationships with the measured environmental variables.

Thomson *et al.* (1996) probed *ad hoc* ways of addressing the problem (statistically), but it was Koenker *et al.* (1994) and Terrell *et al.* (1996) who first highlighted how advances in quantile regression theory made 20 years earlier in the fields of economics and statistics (e.g. Koenker & Bassett 1978), could be applied to ecological data. Since 1996 there has been a slow increase in the number of papers utilising these techniques in ecology (see Cade *et al.* 1999, Cade & Noon 2003, Cade *et al.* 2005; although there are a range of other techniques available, Gaines & Denny 1993, Garvey *et al.* 1998, Scharf *et al.* 1998). However, the significance of factor-ceiling distributions in abundance-environment relationships is still not widely appreciated.

It is timely, then, that a recent paper by Lancaster & Belyea (2006) renews the challenge (more specifically for entomologists) to carefully consider how to test and interpret the relationship between invertebrate abundance and environmental variation. As in the recent papers mentioned above, Lancaster & Belyea (2006) question the utility of simply fitting the average statistical response of abundance to environmental variation, because this ignores biologically important information contained in the inherent variability that is frequently observed around mean abundance values (Benedetti-Cecchi 2003). Such variability around the mean is often considered to be sampling error or unwanted ‘noise’ in the data, but it might also stem from important interactions among multiple limiting variables along the dominant environmental gradient. For example, in the hypothetical data plotted in Figure 1, there is a high degree of scatter in the data, resulting in the lack of a statistically significant relationship between mean biological response (y axis) and the environmental variable (x axis). However, the data clearly do not fill the entire parameter-space (Fig. 1), but instead have quite tightly defined maximum and minimum values in relation to the environmental gradient (a factor-ceiling distribution). This might occur because the environmental variable places tolerance limits on maximum and/or minimum abundance, but it is these limiting responses that are often blithely ignored by standard statistical techniques (Gaines & Denny 1993, Thomson *et al.* 1996).

### **Defining a hypothesis**

Lancaster & Belyea (2006) consider that abundance-environment relationships for invertebrates would be better modelled with limiting response (LR) functions, rather than with central response (CR) functions (see also Thomson *et al.* 1996, Cade *et al.* 1999). They strongly advocate the use of quantile regression techniques for fitting LR models (Scharf *et al.* 1998, Cade *et al.* 1999), rather than ordinary least squares regression (see below). However, the distinction is much more fundamental than simply a choice of different curve-fitting methods. The use of LR models goes right to the heart of how we might want to express our hypothesis about variation in abundance along an environmental gradient. Does the environmental predictor determine a *central tendency* response, around which the abundance at a particular site varies at random, or does it *limit* the maximum response, below which abundance varies because of the constraints



**Figure 1.** Example of an abundance-environment relationship showing a ‘factor-ceiling’ distribution. The relationship between average biological response and environmental variation (the central response function) is non-significant, despite the environmental variable placing strong constraints on maximum abundance (the upper limiting response).

imposed by other limiting variables? So, if we suppose that the hypothetical environmental gradient in Figure 1 represents a rainfall gradient, then would it be more appropriate to ask whether *average abundance* at a site is dependent on rainfall (in which case the conclusion would be that there is no direct relationship, although variability among sites becomes greater in areas with higher rainfall), or would it be more appropriate to ask whether rainfall imposes a limit on the *maximum abundance* that can be attained at a site (in which case the conclusion would be that there is a strong relationship, but additional factors must constrain abundance below the maximum at many sites)?

Depending on how the initial hypothesis is expressed, the conclusions drawn about the abundance-environment relationship can be quite different. Lancaster & Belyea (2006) argue that ecologists are (or should be) more interested in the limiting responses to a set of environmental variables, rather than the central tendency of response to the statistically dominant gradient, particularly as there is already widespread acceptance that multiple factors limit local abundance.

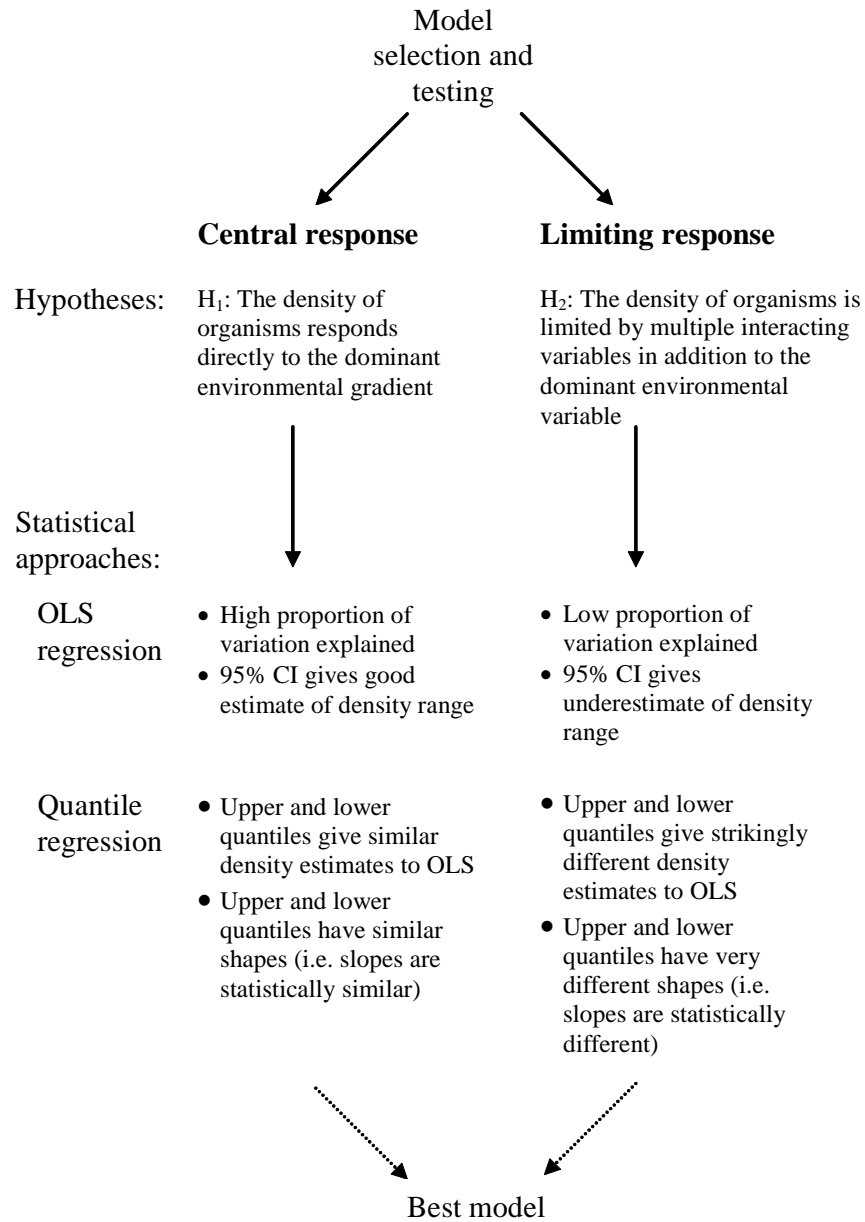
### **How to discriminate between CR and LR models**

In their article, Lancaster & Belyea (2006) test whether a CR or LR model best describes the relationship between stream invertebrate density and near-bed flow velocity in the Lammermuir Hills region of southeast Scotland. They make the point that tight relationships between flow velocity and invertebrate density are often observed in the laboratory, and yet there are surprisingly few well-defined density-flow relationships derived from field data (Lancaster & Belyea 2006; and for a New Zealand example see Collier 1993). This is almost certainly because so many other factors, in addition to near-bed flow, affect local variability in invertebrate abundance, including disturbance history of the site, resource availability, competition, predation and so on. This is a classic case in which multiple limiting factors may impose constraints on local abundance, and confound our ability to detect a statistically significant central response in the density-flow relationship (if one in fact exists). In many cases, however, it is not simply a matter of one particular model (e.g., LR) fitting the data and the other model (e.g., CR) not fitting (as in Fig. 1). Both models may also express qualitatively the same relationship, and in these cases the challenge is to discriminate which model gives the better quantitative description of the data.

Lancaster & Belyea (2006) present a table of ‘characteristic’ results from ordinary least squares (OLS) and quantile regression techniques that might allow discrimination between CR and LR models (depicted in Fig. 2). OLS is the standard regression technique used in the majority of studies testing normally-distributed response data across a continuous environmental gradient. Quantile regression, on the other hand, fits a ‘family’ of response functions to the data, including the median and all other quantiles (Cade & Noon 2003). For the 90<sup>th</sup> quantile, for example, the algorithm finds the predicted curve below which 90 % of the observed data points fall. The median (50<sup>th</sup> quantile) is analogous, but not necessarily identical, to the mean response curve in OLS. The key indicators of a limiting response (LR) model being more appropriate for a particular dataset are (1) if the proportion of variance explained by standard OLS regression is very low, (2) if the 95 % confidence interval for density provides an unrealistic underestimate of the true range of variation in the data, and (3) if the slopes of the upper and lower bounds of the quantile regression are markedly different in shape (Fig. 2).

### **Mayflies and stoneflies in a Scottish stream: an empirical example**

Lancaster & Belyea (2006) counted the abundance of mayflies (Ephemeroptera) in the families Baetidae and Heptageniidae, and stoneflies (Plecoptera) in the family Leuctridae, from 100 replicate 0.04 m<sup>2</sup> Surber samples collected from a single 27 m long riffle in a single 36 hr period, in order to determine the relationship between abundance and flow velocity (measured 6 cm above the streambed, immediately following the invertebrate sampling). The purpose of the large number of replicates was to allow robust quantile regression fitting up to the 90<sup>th</sup> quantile (Scharf *et al.* 1998), and the purpose of the



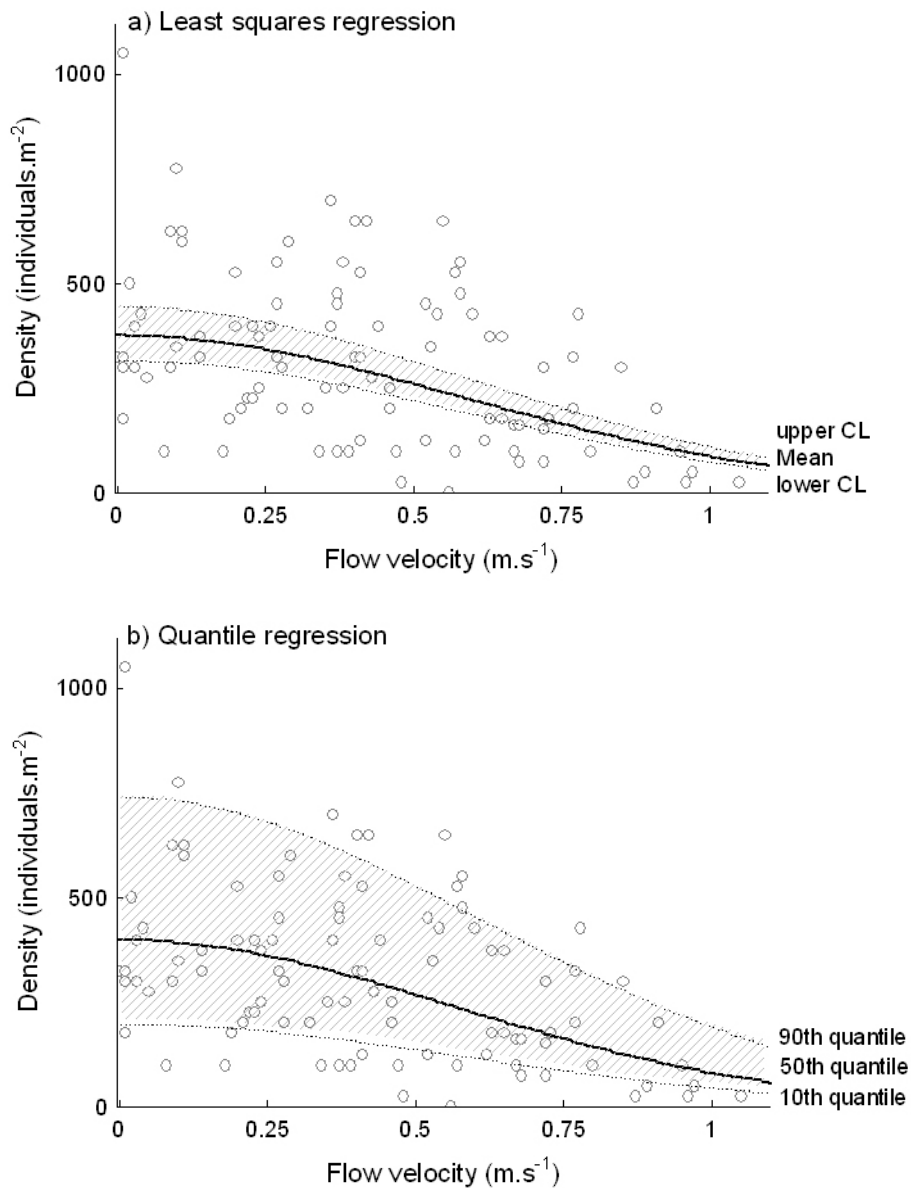
**Figure 2.** Criteria for discriminating between two alternative models, a central response (CR) model and a limiting response (LR) model, used to describe abundance-environment relationships. OLS, ordinary least squares. CI, confidence interval.

small spatial and temporal scale of sampling was to limit variation in extraneous environmental factors, other than flow velocity. What they found was that both abundance and flow varied massively, and despite restricting sampling to a single riffle at a single time, they still did not observe a well-defined density-flow relationship for any of the taxa studied. Taking the Heptageniidae as an example (Fig. 3), what they found instead was that the observed data conformed more closely to a factor-ceiling distribution, with a defined upper bound, and wide scatter of data points below. Standard OLS regression (Fig. 3a) explained little of the variance in the data (27 %), and the predicted 95 % confidence limits on the relationship were a poor reflection of the range of densities observed at each sampling location. Although quantile regression (Fig. 3b) gave qualitatively the same median slope of the relationship as OLS regression, the upper (90th) and lower (10th) quantiles were significantly different in shape, and provided a much more accurate reflection of the range of variation in observed densities. According to Lancaster & Belyea's (2006) criteria (Fig. 2), then, the density-flow relationship for Heptageniidae is best described by a limiting response (LR) model.

### **Why is this important?**

There are two major reasons why it is important to consider the density-flow relationship as an LR model, rather than as a CR model. The first is that we should always be thinking about our data in terms of the biological mechanisms driving the abundance-environment relationship, not just in terms of the statistical fit. It makes intuitive sense that heptageniids have a limited ability to withstand high flow velocities because they do not have exceptional morphological or behavioural adaptations to avoid dislodgement, such that maximum density would be expected to decline with increasing flow. Conversely, at low flow rates (to the left of the x axis in Fig. 3) where heptageniids have the ability to survive hypoxic conditions by gill-beating, it does not make intuitive sense that flow velocity would be able to explain the biological mechanisms causing abundance to be lower than the observed maximum at some sites and not others. What biological value, then, is there in stating that *average* abundance of heptageniids is higher at low flow velocities? Clearly, other (unmeasured) environmental variables must account for the variability below the upper quantile.

The second reason this is important is because of the increased predictive power gained by the LR model producing a more accurate description of the range of variation in the observed data. Particularly at high flow rates, the confidence limits from the OLS regression would be of almost no use in attempting to compare or extrapolate abundance between sites (or to monitor abundance in the same stream through time). In an applied context, Lancaster & Belyea (2006) give a good example of how the differing predictive power of CR versus LR estimates can dramatically affect our interpretation of whether target invertebrate densities have been successfully met in stream rehabilitation and restoration projects.



**Figure 3.** Comparison of (a) least squares regression, and (b) quantile regression functions for the relationship between heptageniid mayfly density versus flow velocity in Faseny Water, Lammermuir Hills, Scotland (redrawn from Lancaster & Belyea 2006). CL, 95 % confidence limit.

**What does this mean for entomologists?**

Rightly or wrongly, we tend to interpret greater abundance at a particular site as an indication of ‘habitat preference’, as if this somehow represents the physiologically ‘optimal’ habitat for the species. This may be a fair reflection of a species distribution in some instances, but there is widespread evidence that organisms may also be concentrated in physiologically stressful, suboptimal habitats because they are limited by competitors, predators or pathogens (amongst other factors) at more ‘optimal’ sites. This is directly analogous to the conceptual distinction between the ‘realized niche’ and the ‘fundamental niche’ (Hutchinson 1957). Of course, since G. Evelyn Hutchinson’s time we have also learned that the realized niche can *exceed* the fundamental niche if populations are able to persist in unfavourable sink habitats due to continuous immigration of dispersing individuals from source habitats (Hanski 1998). In attempting to understand the distribution and abundance of invertebrates, we should more clearly consider (and measure) the interaction among multiple limiting factors, with a mind to interpreting the biological mechanisms driving abundance-environment relationships, rather than simply estimating the central response to the statistically dominant environmental gradients.

We would be well advised not to use the average statistical response of abundance along an environmental gradient to predict the likely location or abundance of populations at un-sampled locations. The upper (90th) quantile is likely to give a better estimate of the maximum distributional extent of locations in which the organism may occur (in relation to the particular environmental gradient in question). Of course, the actual locations across which the organism will be distributed, and their abundances, will depend on the interaction with other limiting factors. These issues will have a major impact on any form of predictive mapping or modelling of pre-human, present and future distributions and abundances of organisms.

Finally, it will be important to determine whether a CR or LR model best describes the abundance-environment relationship before selecting the range of (target) density estimates to use when monitoring responses to changing environmental conditions (such as pollution, land-use change or climate change), or when testing the recovery of populations following ecosystem restoration.

**Acknowledgements**

Thanks to Rob Ewers for helpful comments on the manuscript. Note that Lancaster & Belyea (2006) carried out quantile regression using the *Quantreg* package ([cran.r-project.org/src/contrib/Descriptions/quantreg.html](http://cran.r-project.org/src/contrib/Descriptions/quantreg.html)) in the R-project statistical software (available free at [www.R-project.org](http://www.R-project.org)). Software to conduct a variety of quantile regression analyses is also available for various software packages, including S-Plus, Matlab, Stata, Shazam, SAS (see comprehensive details at Roger Koenker’s webpage [www.econ.uiuc.edu/~roger/research/home.html](http://www.econ.uiuc.edu/~roger/research/home.html)) and Blossom ([www.fort.usgs.gov/](http://www.fort.usgs.gov/)

products/software/blossom/blossom.asp).

## **References**

- Benedetti-Cecchi L. 2003. The importance of the variance around the mean effect size of ecological processes. *Ecology* 84: 2335-2346.
- Cade BS, Noon BR. 2003. A gentle introduction to quantile regression for ecologists. *Frontiers in Ecology and Evolution* 1: 412-420.
- Cade BS, Terrell JW, Schroeder TL. 1999. Estimating effects of limiting factors with regression quantiles. *Ecology* 80: 311-323.
- Cade BS, Noon BR, Flather CH. 2005. Quantile regression reveals hidden bias and uncertainty in habitat models. *Ecology* 86: 786-800.
- Collier KJ. 1993. Flow preferences of larval Chironomidae (Diptera) in Tongariro River, New Zealand. *New Zealand Journal of Marine and Freshwater Research* 27: 219-226.
- Gaines SD, Denny MW. 1993. The largest, smallest, highest, lowest, longest, and shortest: extremes in ecology. *Ecology* 74: 1677-1692.
- Garvey JE, Marschall EA, Wright RA. 1998. From star charts to stoneflies: detecting relationships in continuous bivariate data. *Ecology* 79: 442-447.
- Hanski I. 1998. Metapopulation dynamics. *Nature* 396: 41-49.
- Hutchinson GE. 1957. Concluding remarks. *Cold Spring Harbour Symposia on Quantitative Biology* 22: 415-427.
- Koenker R, Bassett G. 1978. Regression quantiles. *Econometrica* 50: 43-61.
- Koenker R, Ng P, Portnoy S. 1994. Quantile smoothing splines. *Biometrika* 81: 673-680.
- Lancaster J, Belyea LR. 2006. Defining the limits to local density: alternative views of abundance-environment relationships. *Freshwater Biology* 51: 783-796.
- Scharf FS, Juanes F, Sutherland M. 1998. Inferring ecological relationships from the edges of scatter diagrams: comparison of regression techniques. *Ecology* 79: 448-460.

Terrell JW, Cade BS, Carpenter J, Thompson JM. 1996. Modeling stream fish habitat limitations from wedge-shaped patterns of variation in standing stock. *Transactions of the American Fisheries Society* 125: 104-117.

Thomson JD, Weiblen G, Thomson BA, Alfaro S, Legendre P. 1996. Untangling multiple factors in spatial distributions: lilies, gophers, and rocks. *Ecology* 77: 1698-1715.