Freshwater Biology (2010), **55** (Suppl. 1), 60–79 **doi:10.1111/j.1365-2427.2009.02377.x**

Back to the future: little-used tools and principles of scientific inference can help disentangle effects of multiple stressors on freshwater ecosystems

BARBARA J. DOWNES
Department of Resource Management and Geography, University of Melbourne, Melbourne, Vic., Australia

SUMMARY

1. There are multiple tools for scientific inference that seem rarely used in research examining the effects of stressors on rivers caused by human impacts. Very few of these scientific tools are 'new'. While foundational to scientific methods, they seem to have been overlooked or forgotten. The thesis of this paper is that, by looking back to what used to be considered basic knowledge about scientific methods and the discipline of ecology, we may re-learn some useful ways of improving survey designs and re-framing scientific questions.
2. Two common barriers to strong inference are examined in detail in this paper: disentangling the effects of different stressors, so that we can confidently infer which ones are the causes of unacceptable environmental changes; and dealing with high variability among replicate observations. Poor information about causality means managers cannot know what rehabilitation or amelioration should be attempted. Poor fits of models to data lower confidence in inference. Commonly proffered solutions, which include large sample sizes; choosing 'representative reaches'; or using complex multivariate statistics, do not solve these problems.
3. The solutions lie within the basic components of good experimental design, which apply as much to surveys as they do to experiments. Several pieces of practical advice are offered and explained, which include (i) the necessity to specify a precise mechanism of cause and effect in hypotheses, and what changes to common approaches this entails; (ii) some difficulties caused by scale-ups that are implicit in the selection and measurement of variables, which necessitate changes to some standard protocols; (iii) the value of planned comparisons in surveys as ways of strengthening inference and employing approaches, like control species, where other forms of controls cannot be gained; (iv) the necessity to view random sampling as essential to the selection of sites, which means we should abandon the notion of 'representative' reaches; (v) to use sample compositing and sub-sampling to optimise sampling effort at those replicates that provide degrees of freedom for hypothesis tests while cutting costs (vi) to be open to new forms of analysis, like quantile regression, which tests non-traditional hypotheses about constraints, rather than mean or central responses, and which deals much better with sorting between the effects of multiple stressors.
4. *Thematic implications*: sorting between the effects of multiple stressors caused by human impacts needs the best possible scientific inference we can apply. Common forms of studies in the modern stream literature suggest we collectively know less now than we did

Correspondence: Barbara Downes, Department of Resource Management and Geography, The University of Melbourne, 221 Bouverie Street, Vic. 3010, Australia.
E-mail: barbarad@unimelb.edu.au

40–50 years ago because some fundamental aspects of strong inference and basic knowledge in ecology seem to have been forgotten or lost. This raises questions about the quality of ecological training provided at universities. Although some aspects of good design are seen as 'too expensive', cost *per se* is relative. A well-designed programme that has been optimised for the funds available is far cheaper than the costs of poorly designed surveys that provide inaccurate information and predictions, which are more likely to lead to poor management decisions.

Keywords: human impacts, rivers, sampling theory, survey design

The difference between the average scientist's informal methods and the methods of the strong-inference users is somewhat like the difference between a gasoline engine that fires occasionally and one that fires in steady sequence. If our motorboat engines were as erratic as our deliberate intellectual efforts, most of us would not get home for supper. (*Platt, 1964, 'Strong inference' p. 348*)

Introduction

Our capacity to predict correctly the effects of almost any human activity that can create unacceptable impacts on freshwater ecosystems is poor. Take, for example, the relations between river flows and the diversities, densities or identities of species present. This has been a research area for decades, yet we have virtually no capacity to predict the effects of human-caused changes to river flows, with the research instead producing vague and even conflicting expectations (e.g. Dewson, James & Death, 2007). Part of the reason is that almost all human impacts involve many different variables – multiple stressors – and it is difficult to disentangle their effects in ways allowing correct prediction for the future. The same statement about poor predictive capacity could be made for almost any human impact on freshwater ecosystems (as papers in the rest of this *Freshwater Biology* issue attest).

Why is our capacity to predict still so poor? The quotation given above comes from John Platt's forthright, plain language paper about the then rapid increases in correct predictions that were being made in some scientific disciplines while others languished. It has a message that remains relevant over 40 years later. If science contains problems that have resisted solution, it might be time to look critically at the methods being employed, particularly the logic underpinning decisions about the veracity of hypotheses.

Much of the logic allowing inferences about hypotheses is captured in the rubric of 'design', which Kirk (1995 p. 1) defines as: '…a plan for assigning subjects to experimental conditions and the statistical analysis associated with the plan'. The underlying principles are just as important when using observational (i.e., survey) data to test hypotheses, and they are independent of which branch of statistics or analysis (e.g. Bayesian or frequentist approaches) is used. The components of design are: (i) formulation of a clear hypothesis (ii) identification of relevant variables, (iii) specification of the subjects of the hypothesis and the statistical population from which they will be sampled in a representative way (iv) specification of a procedure for assigning subjects to different treatments and (v) determination of analysis. In this paper, I propose that the seemingly intractable difficulties of distinguishing between multiple stressors associated with human impacts might be a function of methods commonly used in studies, which compromise these design components.

Purpose and outline of this paper

This paper is aimed at an audience of people interested in reading about alternatives to standard approaches to assessing human impacts on freshwater ecosystems. In line with the author's background, the context will be human impacts on rivers and the perspective mainly that from frequentist hypothesis-testing, but problems in logic cut across types of ecosystems and analyses. The purpose of this paper is to highlight some effective tools and principles of scientific inference that seem little-used and little

understood. Presenting these tools and principles also highlights substantial problems with some methods that are commonly found in the freshwater literature, particularly in papers on river 'health' and 'bioassessment'. Where pertinent, these problems are discussed, but this paper is – most emphatically – not a review of the human impacts literature. Indeed, this paper does not focus much upon the 'river health' or 'bioassessment' literature at all, even though such papers tend to dominate because that literature does not focus on causality (as explained further below). Some reference citations are necessary of course but are simply examples from a very wide pool of possible papers – no criticism of any individual work is intended, and of course exceptions to common approaches can always be found (and some are cited).

There are three things to emphasise. First, this is an opinion paper but I do offer explanations and rationale for those opinions. Second, there is no pretence that I am suggesting anything 'new' *per se*. This paper discusses ecological knowledge and specific methods because they seem to have been forgotten or overlooked, not because they are 'new'. What I *am* suggesting is that we could greatly improve the predictive capacity of our science if this forgotten knowledge was reclaimed. Third, issues about methods are not just relevant to 'ivory tower' scientists tackling basic problems in ecology. Like many others (e.g. Underwood, 1995) I do not see that distinctions between 'basic' and 'applied' science are either useful or desirable, only distinctions between sound, evidence-based science and weak, evidence-poor science. In that regard, I think it is important that managers understand how to tell the difference as well, not only because we need management problems to be framed as questions that can be tackled using the best possible scientific methods, but also because managers are often involved in decisions about research funding.

I begin by describing two specific and common design problems that occur with observational data collected to disentangle the effects of multiple stressors. I then go on to discuss the basic tools that can be used to deal with these problems, some prospective sources of confusion and error in their implementation that are apparent in the literature, and offer some solutions to problems that are little explored or used. In each case, I relate both problems and solutions to the elements of design identified above. I finish by making some suggestions for reducing the cost of surveys without compromising the quality of information that is gathered and suggest that quantile regression, a tool that is relatively new to ecologists (e.g. Cade & Noon, 2003), offers an approach with great promise for making rapid progress in sorting the effects of different stressors.

Two problems common to survey designs: identifying causality and dealing with variability

Most human activities that create impacts on the environment do so via multiple, prospective pathways of cause-and-effect. To continue the example above, lowered discharges alter flood (i.e., disturbance) frequency, discharge variability at base flows, sediment loads, bed movement, population connectivity via dispersal through the water column, recruitment, flooding of wetlands, and so forth (review: Dewson *et al.*, 2007). If we observe unacceptable environmental degradation, like a loss of species diversity, in rivers downstream of dams or water abstraction, we are immediately confronted with the problem that the specific *causes* of that loss could be any one of the environmental changes associated with lowered flows or a combination of them. An environmental change that causes some response by a population of interest is called a stressor, a term that encompasses effects on individuals like lowered growth rate, lower fecundity or increased mortality (Underwood, 1989). Stressors often interact with each other, synergistically or antagonistically, producing unexpected and often unpredictable effects (Underwood, 1989; Folt *et al.*, 1999). When we add in time lags and also that stressors operate over different spatial scales, sorting out which stressors are the direct causes of an unacceptable change (like loss of species diversity) and which are not, i.e., are just correlates, is very difficult.

Nevertheless, identifying the specific causes of unacceptable changes to the environment is *pivotal* to managing human activities and attempting environmental repair. In other words, simply demonstrating a significant association between water abstraction (for example) and environmental changes is only the first step, because it does not provide any guidance to managers about what they can do to ameliorate or reverse the damage (given that banning all water abstraction is usually impossible). For that, we must identify the specific causes of change. To give a

specific example, two common ways in which weirs change flows are (i) mid-sized floods are retained behind weirs and they and their effects on the downstream channel disappear from the discharge regime and (ii) discharge below weirs is often fixed at a constant rate so that, apart from floods that overtop weirs, channels experience unnaturally constant flows. It has been observed that bryophyte diversity and abundances can be reduced downstream of weirs. Two commensurate hypotheses to explain losses downstream of weirs are (i) changes to frequencies of disturbance of substrata caused by changes to the frequencies and/or magnitude of floods and (ii) reduction in favourable habitat for bryophytes, which is the 'splash zone' on large, emergent rocks and which is shrunk by invariant discharge (Downes, Entwisle & Reich, 2003). If we wish to improve the density and diversity of bryophytes downstream of weirs, managers need to know whether to change disturbance frequencies (which relates to the frequencies and sizes of bed-moving floods downstream of weirs) or re-introduce variability into daily or weekly discharge rates. These are different management actions and possibly only one of them will work.

Experiments are arguably the best way to distinguish between different causal mechanisms (Kirk, 1995), but manipulative experiments to test hypotheses about human impacts are often impossible, and this means we are usually restricted to using observational (i.e. survey) data (Downes *et al*., 2002). Such surveys almost all rely on associations, such as correlations, for evidence of causality, even though it is well known that associations cannot *ipso facto* demonstrate causality (Toulmin, Rieke & Janik, 1979; Sokal & Rohlf, 1995). For example, a common survey design involves collecting samples of invertebrate density or diversity and simultaneously measuring prospective, causal variables, particularly physical variables. The latter variables are plotted against biological variables to search for significant associations, which are then used as evidence of causality (e.g. Dolédec *et al*., 2007) The problem is that many prospective, causal variables are related to each other, and identifying which of them are causal agents of biotic changes and which are just correlated with the causal agents is very difficult when all we have are correlations (Fig. 1).

Gaining strong inference about causality is the first problem discussed in this paper. A second and related

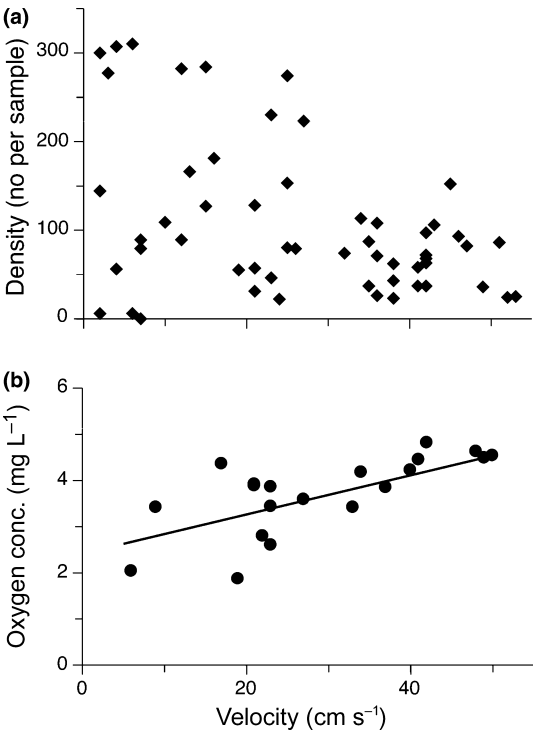

Fig. 1 (a) Numbers of a caenid mayfly, *Tasmanocoenis tillyardi* in 0.04 m^2 Surber samples collected from a single site in Hughes Creek, south-eastern Australia, plotted against water velocity taken simultaneously at the same spots. While there is a clear association between density and velocity, other factors also affect the density of *T. tillyardi*, leading to high amounts of scatter in the data and a triangular array of points. (b) Dissolved oxygen (DO) at the sediment surface in relation to water velocities taken at the same site in Hughes Creek. The line is a least squares regression line with a statistically significant fit. Physical variables like DO and velocity are often related, producing complex interactions across multiple environmental gradients for organisms and creating doubt concerning which variable, if either, is causally related to the densities of organisms. (Figures redrawn from Lancaster *et al*., 2009.)

problem is that replicate observations within surveys often exhibit a great deal of variability (Fig. 1), adding to uncertainty about both the existence and strength of causal relations. As will be discussed further below, sampling variability can produce high variances (i.e., a lot of scatter) when we sample at inappropriate spatial or temporal scales or use survey designs in which insufficient attention was given to identifying and isolating prospective sources of 'nuisance' variation ahead of time.

Both of the above problems – identifying true, causal relations among a set of prospective variables and dealing with high variability among replicate

observations – have been known about for a long time, so it is worth commenting briefly on common strategies used for dealing with them. Three strategies are (i) to have large sample sizes to 'strengthen' patterns (ii) to sample in 'representative' places to reduce 'nuisance' variation, and (iii) to use multivariate statistics to identify causal variables. I will deal with these in detail below; suffice to say here that none of these approaches solves the problems identified above. Increasing sample size does not strengthen patterns because sample variance is not reduced by increasing the sample size, and when we boost the number of observations, we increase the prospect of confusing a statistically significant result with an ecologically significant one. Taking observations in one place chosen to represent a larger spatial unit (like an entire catchment) means the estimates we derive from our data are highly likely to be biased and hence unreliable for prediction. Finally, multivariate statistics (like ordination, cluster analysis, MANOVA) cannot *ipso facto* provide causality because that springs from design; it is not a feature of our choice of statistics (Tabachnick & Fidell, 1996).

Improving evidence of causality in survey design

In this section, I consider first the basic requirements for the first two components of design (formulation of clear hypotheses, identification of relevant variables), then go on to consider potential problems with these components in common forms of study designs in the literature. These considerations suggest some of the standard approaches are not advisable for studies where we hope to examine causality and to tease apart the effects of different stressors. I then suggest ways in which survey designs can implement principles of experiments to improve inference about cause-and-effect (the fourth component of design).

Requirements for hypotheses and variables

As indicated above, the first step in design is to formulate hypotheses, and perhaps the most critical aspect of hypotheses is that each should contain predictions that are unique to a particular model (or 'explanation') for patterns (Underwood, 1990). If predictions are not unique to models, then there is no way of determining which explanation is correct if the predictions are found to be true. With multiple stressors, we have multiple competing hypotheses to explain a biological response to an environmental gradient. Ergo, we need to specify predictions unique to each prospective stressor, which will allow us to identify which are causal agents while others are simply correlates. Producing unique and explicit causal predictions is challenging, and only likely to come from thinking clearly about the biology of the organisms involved and how environmental gradients can cause a loss of fitness. That is, we need to identify gradients that produce increases in mortality, decreases in fecundity, decreases in growth rates, etc. (Lancaster & Downes, in press).

To illustrate, let us continue to consider the effects that changes to discharge can have on biota through changes to hydraulic variables, like spot water velocities, water depths, turbulence, etc. Many freshwater researchers presume there is a strong association between measures of flow and measures of density or occurrence of species (e.g. Jowett, 2003). In fact, the prevailing view is that flow is the prime driver of distribution in rivers (Lancaster, Downes & Glaister, 2009), and this would suggest that, of the human impacts we might consider, changes to flows ought to be the simplest to deal with. In actuality, associations between measures of flow and biological measures like densities, diversities or presence/absence could be driven by literally *dozens* of different direct and indirect causes. For example, densities of a species could respond to changes in water velocities because (i) concentrations of dissolved oxygen can be low at low velocities (e.g. Walling & Webb, 1992); (ii) individuals are unable to settle on the benthos at high velocities (e.g. Fonseca, 1999); (iii) species consume food, like fine detritus, which deposits in high concentrations in slow velocity areas (e.g. Lancaster *et al.*, 2009); (iv) species consume food, like suspended particles, that are most abundant at highest velocities (e.g. Hansen, Hart & Merz, 1991); (v) species avoid a generalist predator that lives primarily in slower velocities (e.g. Lancaster *et al.*, 2009) or faster velocities (e.g. Hansen *et al.*, 1991); (vi) species are excluded by a competitor for an essential resource that is most abundant at higher velocities (e.g. Hemphill, 1988), etc. Any and all of these explanations may apply to the assemblage of species inhabiting individual riffles, resulting in significant positive correlations between densities and water velocities for some species, and negative correlations for others. Moreover, some

correlations may extend over a wide range of velocity values, whereas others occur only over a limited range of values (as can occur in threshold relations). Many species may show no association with velocity at all. Thus, even when a measure of flow is 'important', we should expect to find mixed responses by different taxa contingent upon their respective biologies and accounting for indirect effects (i.e., where velocity is not the direct causal variable but might be correlated with the one that is). Such mixed responses are *not* trivial just because they operate over small spatial scales; they cannot be combined sensibly to generate a single 'community' response (as occurs in some multivariate approaches) without producing flow-density relations of dubious ecological sense (Lancaster & Downes, 2009), and without jettisoning sensible causal mechanisms that provide capacity to make useful predictions for the future.

Specifying causal mechanisms means also that the scales of space and time of effects of stressors on biota must be explicitly considered and stated. Suppose we detect a strong association between water velocity and density of an insect species, which is caused by low oxygen concentrations at slow velocities. We have detected a gradient operating over scales of perhaps metres at most, but most studies of stressors hope to make predictions at much larger spatial (and temporal) scales than this. To use a density–velocity relation measured over metres to predict effects at a catchment scale, we would have to assume no other processes affect oxygen concentrations or organism densities as we scale up over several orders of magnitude (Schneider, 1994). We already know such assertions are false, except in cases where human impacts are so extensive and so intense that stressors range over only extreme values, and variation over all small spatial scales has been nullified. Speaking more broadly, we know very little about how stressors change over different spatial and temporal scales in ecology, and that lack of knowledge represents an enormous barrier to generalisation (Schneider, 1994).

Hypotheses and variables in commonly used designs

The above discussion produces several conclusions we can draw about common forms of study designs used in research examining multiple stressors on rivers.

First, it is important to recognise that a large proportion of studies of human impacts on rivers – usually described as 'river health' or 'bioassessment' studies – cannot offer any insights into causality, especially in regards to disentangling the effects of multiple stressors. This is not a criticism. These studies do not set out to test *whether* there are cause-and effect relations; their purpose is to develop indices that can be used for environmental assessment. Consequently, the response of biota to human impacts is largely assumed, and comparisons between near pristine ('reference') locations and those affected by human activities are used to develop sensitivity scores for different taxa (e.g. Metcalfe-Smith, 1994) or complex multivariate models that can predict changes in number or identity of taxa expected to be present (Wright, 1995). The purpose is thus to produce tools that can be used to determine to what degree a site is degraded – not to explain the specific causes of any damage (Wright, 1995). Because this is such a huge literature (530 papers in the last 10 years alone, according to a search using the topics of 'river health' or 'bioassessment' in Web of Science), it is worth explaining why these studies do not (and cannot) make definitive statements about cause-and-effect. It is because they use reference sites, not control sites. The purpose of controls is to isolate the effect of a particular 'treatment' (Quinn & Keough, 2002). In the context of looking at human impacts, this logic means controls must be matched closely to the impact sites except for the instance of the human impact of interest. If impact sites have a suite of other, human impacts present, then so should the control sites. The logical basis for selecting and using control sites is thus completely different from reference sites, which, with their characteristics of being 'near pristine', offer some indication of what sites might look like in the virtual absence of human impact but provide little to no definitive information about the causes of any specific impact. Consequently, reference sites do not provide any information about causality except in the (unusual) situation that impact sites have a singular, human impact on them, in which case control sites would be 'near pristine'. This material is not new but the logic is infrequently presented, and it may otherwise be puzzling why so little reference is made in this paper to the many 'river health' or 'bioassessment' schemes.

A second issue is that many studies contain profound mismatches between the scales of sampling and the scales over which stressors are thought to act

and at which predictions are required. For example, macroinvertebrates are typically sampled over small spatial scales (fractions of a m^2 or 1–2 m^2 at most) and also temporal scales (samples collected over minutes to a few days). Stressors often act over entire catchments (100s–1000s of km^2) and they may be present for years. When we draw inferences about catchment-wide stressors using typical macroinvertebrate data (e.g. Dolédec *et al.*, 2007) we have done an implicit scale up. Ecologists have known for decades that scale ups do not produce useful predictions (Schneider, 1994). I suggest that the next wave of research on human impacts on rivers needs to tackle this mismatch between sampling scales and the scales needed for useful predictions as a priority, as is being done in ecology more broadly (e.g. Melbourne & Chesson, 2005), and some suggestions are given below.

A third implication concerns a common approach in the freshwater literature of using only coarse taxonomic resolutions, such as family-level data. Coarse taxonomy is problematic for studies focusing on causality because taxonomy is a poor guide to ecology. Even species within the same genus can respond to environmental stressors in markedly different ways (e.g. Macan, 1963). Variation between species means that it is virtually impossible to construct an ecologically sensible hypothesis – i.e. one containing a specific statement about cause-and-effect – if it refers to members of an entire phylum, family or genus. This is well-understood in other areas of the literature, where species-level identifications are the norm. Birds, for example, suffer when humans cut down trees, but examinations of the effects are pitched at the species level (e.g. Vale *et al.*, 2008), precisely because our knowledge about birds demonstrates that agglomerating different species produces nonsensensical results. I suspect the habit of doing this in freshwater research is mostly caused by a lack of familiarity with the specific resource and environmental requirements of species, coupled with their small size and seeming 'sameness'. A further motivation is that family-level identifications reduce laboratory processing time and hence save money, but losing predictive capacity by abandoning causality can be poor value for money and there are other ways to reduce costs without compromising the quality of information (explained below). The drawbacks of using taxonomic groups like families in freshwater studies were pointed out long ago (e.g. Resh, 1979), and it is clear that, if we are to group species, then groups need to be based around ecological and functional similarity (i.e., species that share a common response to environmental or resource gradients) rather than taxonomy. Indeed, I would suggest that identifying which species can be grouped, and why, continues to be an area in desperate need of targeted research.

It is sobering to realise that some of the earliest works on the responses of river biota to human impacts (usually considered to be those of R. Kolkwitz and R. Marsson, who developed the Saprobien classification system for identifying zones of organic pollution: Hynes, 1960) are now over 100 years old. Research carried out in the 1920s–1960s on species-specific tolerances to various environmental gradients produced vigorous debates about the value and utility of such classification schemes when species did not have fixed responses to pollutants and were difficult to group definitively. The theme was emphasised in Macan's (1963) text book on freshwater ecology, which saw understanding *why* some species are superior to others at tolerating environmental extremes as foundational to the study of freshwater ecology. In a landmark text on human impacts on rivers published in 1960, Hynes, (1960, p. 161) opined '*The effect of heavy pollution of* (sewage fungus) *is therefore very different from that of sewage. This is inevitable, because different creatures react differently to the various aspects of organic pollution*'. In other words, we already knew in 1960 that we could not even classify rivers exposed to sewage pollution with any confidence, let alone other forms of pollution or human impact. Why then, in the 21st century, does the literature continue to produce so many studies in which species are agglomerated into families (or even higher groups) and in which there are continuing expectations that we can use them to 'classify' rivers? The 2000 European Water Framework Directive, for example, is premised on the notion that river classification is possible, desirable and will lead to improved management outcomes (see for example Davy-Bowker *et al.*, 2006).

My guess (hypothesis) is that the determination to agglomerate species and to classify rivers might be a consequence of three factors. First, perhaps people working in this field lack training in ecology *per se* as

opposed to freshwater science. Without training in ecology there is no exposure to the 100+ years of ecological research trouncing the hypothesis that ecosystems or communities can be usefully 'classified' if one tries hard enough. By useful, I mean classification schemes that result in predictions of sufficient accuracy to test causal models in ecology, arguably the same level we need for effective management. Indeed, modern ecology began to abandon classification systems in the 1950s when plant ecologists produced data showing that such schemes did not work precisely because species responded to environmental gradients individually, i.e., they did not have fixed and easily categorised responses that allowed them to be grouped or plant communities to be usefully pigeon-holed (Kingsland, 1991). This debate between plant ecologists was analogous to the one being had among freshwater ecologists, mentioned above. Second, I suspect that the advent of multivariate statistics, followed by desktop computers with the power to do the analyses, opened the door to research where potentially illogical reasoning about hypotheses disappeared inside a figurative forest of statistical jargon. Indeed, illogic created by inappropriate interpretation of multivariate analyses is a common problem in many disciplines, leading Tabachnick & Fidell (1996) to point out bluntly that multivariate output '…can make garbage look like roses' unless we know how to decide when '…the output more closely resembles the fertilizer than the flowers' (Tabachnick & Fidell, 1996, p. 6). In other words, large, multivariate datasets rarely contain *no* patterns at all. The challenge is to identify whether patterns, like 'classes' of rivers appearing in ordination plots or cluster analyses, are real or are simply created by mathematical artefacts or by chance. To do that requires a high level of scepticism, severe tests of hypotheses (see below), and high expectations for predictive capacity, things that are all somewhat less likely if most researchers do not question the existence of such 'classes' in the first place and are determined to find them. Third, there is an understandable imperative to develop tools that can help with managing human impacts on rivers. Perhaps this imperative has drowned out those voices urging caution and questioning whether the level of predictive accuracy was ever going to be anywhere near adequate to protect freshwater ecosystems or guide their restoration (e.g. Lake, 2005).

Means for improving inference using planned comparisons within surveys

As mentioned above, experiments provide the best means for inferring causality when they are logistically possible so it is worth exploring what characteristics produce that capacity and how they can be incorporated into surveys. In short, experiments test hypotheses using planned comparisons between treatments and controls, where a control provides information about outcomes we expect to see in the absence of treatment (Quinn & Keough, 2002). An important principle is that the subjects of the experiment are randomly allocated to either a treatment or a control group. It is this randomisation that ensures that any characteristics of the subjects that might affect test outcomes are not systematically associated with any experimental group, and this is what gives us the confidence that any subsequent difference between treatments and controls *is* caused by the treatment (Kirk, 1995; Quinn & Keough, 2002).

We can apply this principle, to a degree, in surveys. Most researchers are familiar with a common type of design used to look at human impacts [called Before-After-Control-Impact (BACI) designs] that uses a group of control and impact locations each sampled both before and after human activities commence. A change over time at the impact location(s), relative to controls, that is coincident with the onset of human activity is used as evidence of human-caused change (for more explanation, see Underwood, 1993; Downes *et al*., 2002). Of course, a BACI-type survey design is not the same as an experiment because we do not get to randomise subjects like entire rivers to be in impact treatments while others are assigned to be controls, but in well-planned and executed BACI-designs, the most likely explanation if we reject the null hypothesis is that human beings caused a change (Underwood, 1993; Downes *et al*., 2002). Designs like these in which there are planned comparisons (i.e. unique, *a priori* predictions) produce severe tests of hypotheses (see Mayo, 1996), in which we are much less likely to mistakenly accept an erroneous explanation (i.e., an incorrect, alternative hypothesis) for our results. It follows that survey designs having such attributes produce greater confidence in conclusions about cause-and-effect than those where we have no *a priori* planned comparisons to test hypotheses.

The above material is obviously not new, but I have highlighted the benefits of planned comparisons to emphasise that they are highly pertinent to non-experimental research because they reflect a controlled implementation of sound, scientific reasoning (i.e. Platt's 'strong inference'). As most researchers know though, implementing these principles in surveys is not easy. A common difficulty is finding control locations, and this problem becomes especially acute where impacts occur over large spatial scales (Downes *et al*., 2002). It is important though not to abandon the whole principle of using planned comparisons, because that is tantamount to throwing the baby out with the bath water. If we cannot apply commonly used survey designs, like BACI-type designs, then the challenge is to find alternative ways of generating strong evidence of cause-and-effect, not to abandon the principles of inference altogether.

An example of a little-used, alternative means is that of control species. Control species are taxa that are predicted not to respond to the stressor of interest because of specific aspects of their biology or ecology. Species differ in their sensitivity to environmental gradients; this can be used to our advantage. If a group of control species are matched with a group of 'treatment' species, which *are* predicted to respond to a specific gradient, and there are otherwise no systematic differences between the 'treatment' and the 'control' species that might affect their responses to the stressor (e.g. in taxonomy, habitat, trophic levels, etc.), then we have improved our capacity to draw conclusions about causality. Effectively, we use species to predict, *a priori*, both the presence *and* absence of responses expected [although I note that predicting absences of responses generates problems with creating logical null hypotheses, and this requires careful thought – see Underwood (1990) for an explanation]. The likelihood that an alternative hypothesis can explain our results becomes remote because its action and effects would have to be closely correlated with sensitivity to the stressor of interest at the species level. The more species we use, the less likely any alternative explanation becomes, and the stronger is our case for inferring that responses are due to the stressor of interest. Improving inference using control species is not a new idea *per se*. The basis for this approach is rooted in discussions about scientific method, viz; '…a statement describing many tests (especially if they are independent of one another) will be less probable than a statement describing only some of these tests' (Popper, 1983, p. 247). In other words, from the frequentist's viewpoint, we seek evidence that is improbable to gain unless our hypothesis is correct, and the more severe tests that we have to bear on the question, the stronger is our case.

Control species have rarely been used in studies of stressors on freshwater ecosystems, although elements of the underlying reasoning can be found in some studies of pollution [see Rundle, Weatherley & Ormerod (1995) and Downes *et al*. (2002), page 234, for a worked example], and see Keough, Quinn & King (1993) for a particularly convincing example where control species were used to draw inferences about the effects of human collection of shellfish. Using control species can look superficially like tolerance indices (e.g. SIGNAL index: Chessman, 2003) but there are several critical differences. Using sets of control and treatment species produces multiple, *a priori*, causally based predictions for individual test, and the species are matched to remove taxonomic and other forms of systematic variation that might otherwise explain the collective results. None of those aspects are present in most indices, which, by calculating single numbers jettison much of the underlying inferential framework laid out above. (Of course, this is reasonable because most of these indices are designed to measure the degree of degradation, not test causal hypotheses.) The approach of gaining controls by using species, rather than searching for *spatial* controls (as is done in typical BACI designs) can theoretically be applied to any stressor and so this represents a hugely under-used means of inference. It can be used to strengthen inference in all designs, but is particularly valuable where few or no control locations exist or other design elements (like temporal data) are missing that create gaps in the logic needed to disprove different hypotheses. The method does require detailed knowledge of the biology of species, but as discussed above, this is essential anyway if we are to ensure causality takes centre stage in future.

Identifying and controlling unhelpful sources of variability in data

The second major problem I will discuss is that of high variation among replicate observations within data sets. Such variability is sometimes dismissed as

'noise' created by poor sampling methods or as 'nuisance' variation that obscures what is really happening (e.g. Dakou *et al*., 2007). However, much of this scatter is not 'noise' caused by measurement errors, but reflects sampling variability (or 'sampling error'), i.e., true variation among replicate observations (Quinn & Keough, 2002). Sampling variability is real variation unaccounted for in survey designs – the idea that variation in data is simply a 'nuisance' that obscures patterns was repudiated some time ago (e.g. Chesson, 1986). The presence of that view in the freshwater literature needs to be examined because some current methods used to reduce variation among replicates are actually likely to be adding noise, obscuring patterns and introducing bias. I discuss two such sources of concern: inappropriate standardisations and sampling protocols that do not abide by the logic of sampling theory. These aspects relate to the second and third components of design (variables; identifying statistical populations and sampling from them). Attention to these issues could improve the 'signal-to-noise ratio' in typical freshwater studies, and I also discuss some little used approaches that can keep the costs of sampling down without compromising on the quality of information.

'Noise' caused by inappropriate standardisations

One of the basic rules of taking samples is to keep the spatial and temporal units of sampling the same between replicates. Observations need to be collected over consistent units of space and time because varying the scales of sampling (like using different quadrat sizes) can change estimates of means and variances (Schneider, 1994) and often produce markedly different sampling precision (Krebs, 1989). Most sampling equipment and methods are designed to keep the scales of sampling consistent among replicates, but there are a variety of procedures and pieces of equipment that break this rule, many used in studies of human impacts on rivers. Two sampling methods that I will use for illustration are passive samplers, like drift nets, that rely on stream current to deliver organisms (and are used to assess stressors, e.g. effects of forestry: Hoover, Shannon & Ackerman, 2007) and the use of geomorphological units like riffles to form replicate study sites. Both examples have the problem identified above: the spatial sizes of samples varies between replicates. In drift samples, the volume of water varies between nets because flow varies between sampling spots. In the second example, the areas of riffles vary naturally between locations.

In both examples, the standard solution is to divide counts of organisms by the values of the spatial unit to correct between (or standardise) replicates, and herein is where the problem lies. When we conduct such corrections, we are assuming that there is a strong and largely linear relation between measures of space and biological measures like abundances. For such a linear relation to occur, individual organisms have to be distributed randomly through space, so that changes in the sizes of spaces sampled (i.e., areas or volumes) produce proportionate (i.e., linear) changes in numbers of organisms collected (Schneider, 1994). Of course, individuals are almost never randomly distributed. They are very often aggregated, and this means there will be unpredictable changes in counts of organisms with changes in area or volume between replicates. To illustrate using the drift example, the standard approach of dividing drift counts by the volumes of water sampled means we are assuming a linear relation between water volume and drift numbers (Fig. 2a,c). If that linear relation is not present and we ignore this assumption and standardise our replicates by water volume anyway, we simply add a lot of unhelpful noise that is likely to obscure patterns in the data completely (Fig. 2b,d). The situation is complicated even more if discharge rises or falls during sampling, which then dilutes or concentrates drifters in the water column but variably and unpredictably between replicates or over time. Curiously, although it is well known that drifters are not randomly distributed through the water column (e.g. Brittain & Eikeland, 1988), examinations of the relations between drift numbers and water volumes, and the effects they have on calculations of drift densities, are uncommon even though drift data are notoriously 'noisy' (Fig. 3).

The solution to this problem is critically dependent on the question being asked about drift. If the hypothesis is about quantifying the numbers of drifters entering and moving through study sites, the numbers of drifters per unit cross-sectional area arriving at the upstream end of study sites may be more relevant than numbers of drifters per unit volume of water, which means we avoid the need to standardise across varying sample volumes altogether

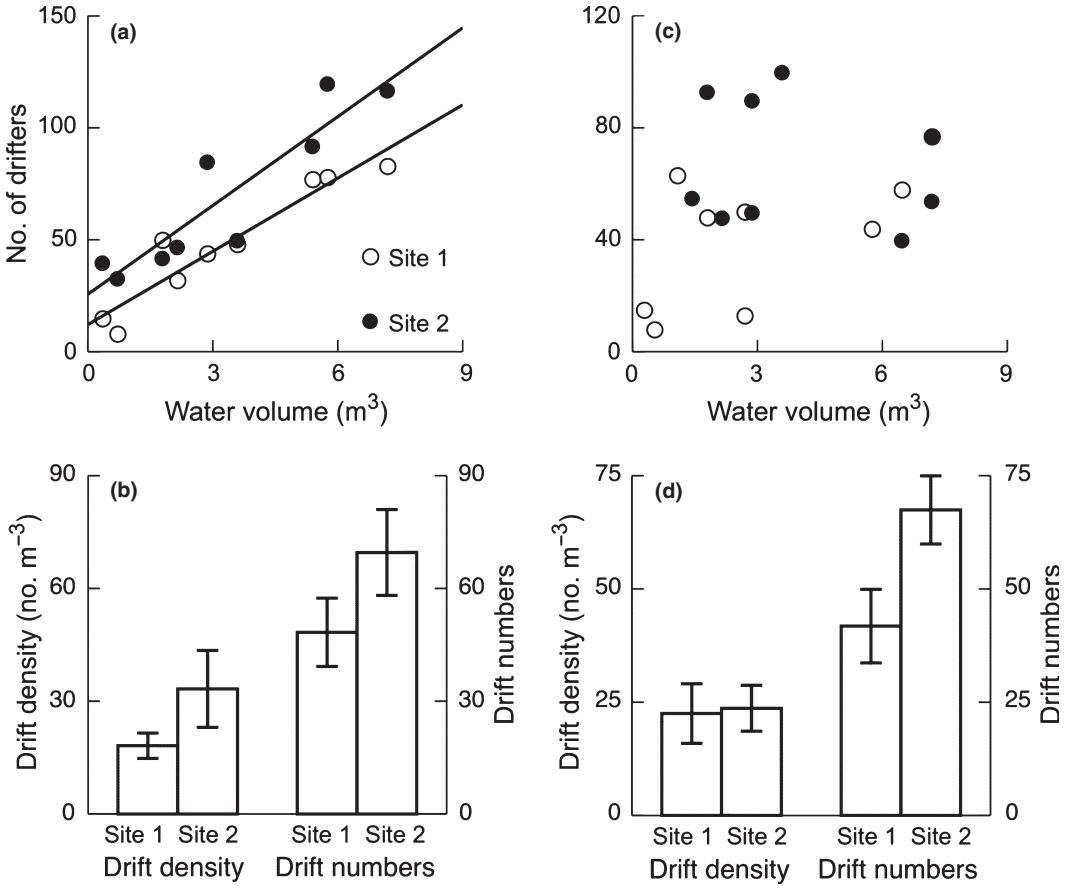

Fig. 2 Two sets of imaginary data that illustrate an important and common assumption behind the standard calculation of drift densities, in which drift numbers are divided by the volume of water sampled by nets. The assumption is that there is a strong, positive and largely linear relation between water volume and drift numbers, which thus allows us to standardise for the differing volumes of water sampled by replicate nets. If this relation does not exist, the calculation of drift densities is effectively based on a faulty assumption that can obliterate patterns. Shown are data that might hypothetically arise from drift being sampled with nine drift nets at each of two sites, 1 and 2, plotted as drift numbers versus the volume of water passing through each drift net (estimated by measuring water velocity at the mouth of each net, multiplied by the area of the net mouth). In both cases, overall numbers of drifters are higher at site 2 than at site 1. (a) There is a strong, positive and largely linear relation between numbers of drifters and the volumes of water sampled by nets. This can be demonstrated using least squares regression; the lines are highly significant fits to the data (note that slopes estimate density differences). When drift densities are calculated and expressed as means with standard errors (b, left-hand bars), the differences between sites in average drift numbers are preserved (b, right-hand bars). (c) An alternative outcome where there is no strong relation between drift numbers and volume of water sampled by nets, but overall numbers of drifters are still higher in site 2 than site 1. Dividing drift numbers by volumes to produce drift densities and then calculating the average drift density at each site is therefore based on a faulty assumption (that volume of water can be used to correct between samples), which adds so much noise to estimates that the real difference between sites in numbers of drifters evident in (d) (right-hand bars) disappears (d, left-hand bars). The procedure is equivalent to using a regression equation (i.e., $y = a+bx$) when no relation between x and y exists. Drift numbers are only likely to be closely related to water volumes if drifters are distributed completely randomly through the water column in space and time (Schneider, 1994), a highly unlikely circumstance (Brittain & Eikeland, 1988). (Reproduced from Downes & Lancaster, 2010).

(Downes & Lancaster, 2010). In other studies, densities of drifters in the water column is of direct concern (for example, if we want to quantify food available to drift-feeding fish), but we can consider using uncorrected drift numbers in some cases. As heretical as that may sound, uncorrected numbers are used with passive samplers in other environments (e.g. passive marine planktonic larval traps: Todd *et al.*, 2006) precisely because it is well known that there is no relation between numbers of organisms and volumes of water sampled.

In the second example, habitat units like riffles vary naturally in area, and when we need to use whole riffles as replicates (which is typical for larger

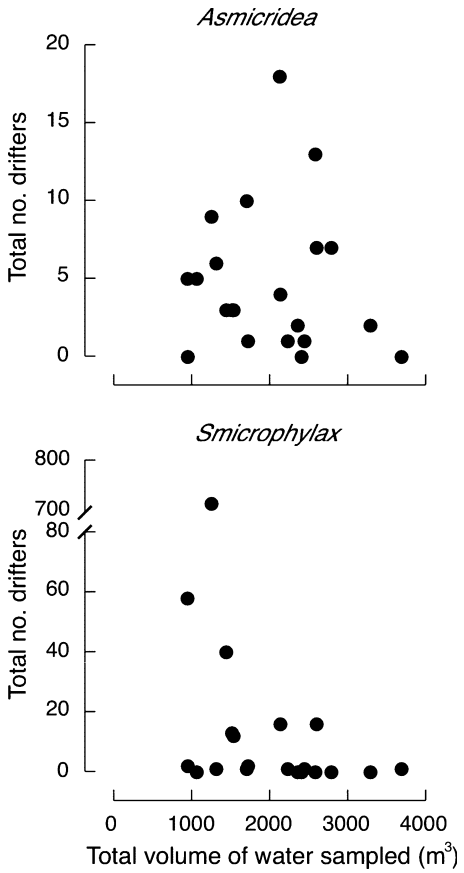

Fig. 3 The total number of drifting hydropsychids of two species, *Asmicridea* sp. AV1 and *Smicrophylax* sp. AV2, captured during five sampling occasions over a week for 21 site/times plotted against the total volume of water sampled by drift nets during that time. There is neither a strong nor a positive relation between numbers of drifters and sampled water volumes and the scatter plots show a lot of variation. (Reproduced from Downes & Lancaster, 2010).

organisms like fish), we have the same problem again of organisms being counted over different spatial areas. This presents the same issue described above: converting numbers to densities and correcting between observations involves assuming a linear relation between site areas and numbers that may not exist. If we have a large range in riffle area, we may add a great deal of noise to density estimates by standardising across replicates. While researchers can try to keep stream length the same between locations, it is usually impossible to keep total site area the same because channels vary in width. Upland channels are narrower than lowland channels, meaning that even if we keep stream lengths the same, upland sites will be smaller in area than lowland sites. We could vary the length of channel to keep area constant, but whether this approach is pertinent depends, again, on the question we are asking and the biology of the organism. For example, if we are sampling fish that swim along river margins and do not cross the channel (e.g. juvenile *Galaxia maculatus*: Hale, Downes & Swearer, 2008), then keeping stream length consistent between sites is critical, not the overall area.

Problems caused by varying the sizes of sample units are pervasive in ecology, especially when we consider that analogous problems pertain to temporal units as well. That is, if replicates vary greatly in the period of time over which they were collected and we divide through by a common unit of time to standardise them, we are making the assumption that organismal abundances are random over time: a patently false assertion. There are no standard solutions to temporal or spatial standardisations because much more research is needed on scaling rules, and this is an active area of research in cutting edge ecology (Schneider, 1994; Melbourne & Chesson, 2005). In the meantime, I suggest that much more attention needs to be given to critiquing 'standard' approaches if we are to reduce noise in the data collected in studies of human impacts. Again, all of the material above is not new, but an absence of discussion about these issues may reflect knowledge that has simply been forgotten. The problems of correcting across different scales of measurement and the short-comings of sampling equipment have been known about for decades (e.g. Macan, 1963; Krebs, 1978, 1989) and at least used to be part of basic training in ecology. There are many papers debating the best method for expressing drift, which incorporates some of the material above (e.g. Brittain & Eikeland, 1988 and references cited therein). It is nevertheless rare to see the assumptions underlying standardisations examined and tested or alternatives offered and rationalised in much of the freshwater literature.

Issues around statistical populations and random sampling

In almost all ecological research, we have to take samples because we cannot measure every possible individual case (or 'subject') of interest, whether our cases are invertebrate densities, water quality or quantity measures, or something else. As well, in

almost all projects we wish to draw inferences beyond just the values in our sample to all the cases that were not, in the end, measured. The beauty of sampling theory is that we do not need to count or measure every case to make reliable (i.e., unbiased) estimates like means or to make inferences in hypothesis tests (Sokal & Rohlf, 1995). Nevertheless, like any theory, sampling theory is based on assumptions that must be met if the estimates we derive from samples are to be unbiased and therefore reliable (Sokal & Rohlf, 1995; Quinn & Keough, 2002). Two of the most critical of these assumptions are that (i) the cases to be sampled come from a statistical population that has been clearly defined, and (ii) samples are representative of populations, something which is typically gained by ensuring random sampling (Sokal & Rohlf, 1995). A failure either to define the statistical population correctly or to sample from it in a representative way can produce biased data and hence misleading results, so it is thus worth discussing both assumptions further in the context of freshwater research.

A statistical population is defined as the entire collection of cases to which we want our estimates or inferences to apply (Sokal & Rohlf, 1995). 'Cases' can be invertebrates in quadrats, spot measures of water velocity, fish densities in riffles, etc. – any variable of interest. For each variable, we define the statistical population by outlining the criteria that determine which cases will be included and which will not. As a minimum, we need to specify the spatial boundaries within which we will sample, and which will determine the cases that will be included in our statistical population (the same argument can be made for temporal boundaries as well). The criteria will also reflect knowledge we already have. For example, it is well-established that filter-feeders occur in relatively fast flows; it would not make sense to sample pools or backwaters for them and the criteria for the statistical population can specify that samples will be collected only from fast-flowing places like riffles. Criteria for statistical populations should also acknowledge any places we cannot sample. For example, a Surber sampler cannot collect invertebrates from spots too close to the banks or in very deep water, in the middle of debris dams, and so forth. Specifying the criteria for the statistical population means we identify all the cases for which we can draw valid inferences, especially in space and in time, and those where we cannot. All statistical populations have such limits, which should of course relate to the hypothesis we are testing.

I have described a process for defining statistical populations in detail because it is rare for freshwater researchers to describe their populations explicitly. For example, it is common for researchers to sample invertebrates only in riffles, which means that this geomorphological unit or habitat forms part of the criteria for the statistical population (Downes & Reich, 2008). The same issue applies where replicate sites are used to test hypotheses about factors operating over entire catchments of 100s–1000s of km^2, as is common in the human impacts literature. In such circumstances, sites are a level of replication; they are simply 'big' quadrats. Clear criteria that specify which sites within catchments are considered part of the statistical population are required. For example, we could identify all channels within each catchment that have discharges (or stream widths) within a certain range as a defining characteristic.

When researchers do not specify statistical populations explicitly, there are two potentially unfortunate consequences. First, vagueness about the criteria for the statistical population increases the chances that conclusions are inadvertently drawn about cases that were never part of the statistical population. For example, we might sample only sites that meet a discharge criterion but draw conclusions about entire catchments. This is illogical and likely to produce poor predictive capacity. Second, great scatter within data can be generated when insufficient care has been given to specifying the types of locations to be sampled, and why. Alternatively, a carefully defined population of sites can reflect knowledge about environmental gradients that has already been demonstrated to be important and which can then be included in the sampling design to isolate the variation associated with it. As I will describe further below, a common method of choosing sites in much of the stream literature, especially that concerned with human impacts, does not comply with any of this logic or this theory.

The second important aspect of sampling theory is to ensure representativeness of observations, which is achieved by using random sampling. Random sampling has a definition: it means that each identified case within the statistical population has an equal chance of being sampled (Sokal & Rohlf, 1995). Random sampling is not the same as 'haphazard

sampling', where we try to avoid bias (e.g., by throwing a quadrat behind us so that we cannot see where it will land) but do not guarantee that every subject has an equal chance of being sampled. A common way of achieving random sampling within a site is to imagine the site on a set of x, y axes (x across the stream, y along the stream). It is simple to generate random x, y coordinates, which are then used to determine the position of samples. Spots within sites that are not part of the sampling population (e.g. too close to banks, within debris dams, within deep pools, etc.) guide which random coordinates can be legitimately discarded as sampling points. This approach can also be used where we are sampling species that are only found within particular kinds of habitats. It is still important to ensure that such habitats (or 'strata'), however they are defined, are sampled randomly (Manly, 1992). In other words, random sampling does not mean 'sampling everywhere', which in my experience is a common misunderstanding. It means sampling randomly from all cases defined by the statistical population.

Random sampling for small scale samples is generally well understood (although often not described in most papers), but, curiously, random sampling is rarely ever applied to the selection of replicates at large spatial scales, like sites or reaches (or even catchments). In studies of human impacts on rivers, it is common for sites to be chosen deliberately, even when the stated goal is to gain representation of a larger spatial unit like an entire catchment. In this approach, a thusly called 'representative reach', claimed to be typical of channels in the catchment, is deliberately chosen and sampled intensively to represent that catchment. The method is usually premised on the assumption that channel morphology controls the distribution of biota, hence the reach is chosen because it contains representative morphology, with the presumption that this will represent the biota (e.g. Maddock, 1999; Rogers & Biggs, 1999). Even in sampling programmes that do not use this named 'representative reach' method, individual lengths of stream are still chosen to 'represent' catchments with the expressed claim that these individual lengths of stream will capture the main hydromorphological and habitat variation (e.g. Furse *et al.*, 2006).

There are at least two, fatal problems with these approaches as sampling methods. First, humans cannot choose without introducing bias, which is why methods ensuring observer choice is removed from the *process* of data collection are considered a pivotal aspect of good, defensible science (Beveridge, 1961). Removing observer bias is behind the use of 'double-blind' experiments that are standard in psychology and medicine, for example (Kirk, 1995). When a length of stream is deliberately chosen, it will reflect beliefs already held by that researcher about the drivers of biotic distribution, but such beliefs are, of course, intrinsic to individuals and difficult to quantify. As such, there is no way to demonstrate *ipso facto* that any of the estimates derived from the data are unbiased and therefore reliable. Indeed, given human beings are inherently biased, the default assumption has to be that the data are unreliable. Bias is not avoided in situations where a careful protocol for making the choice has been set out (e.g. Furse *et al.*, 2006) if the final selections of places to collect data are still not random selections from a prospective group of possible places (i.e. a population). The choice still remains under the control of the individual (biased) researcher in the field.

Second, choosing one length of channel to 'represent' an entire catchment is *not a sampling process that is identifiably based on the logic underpinning sampling theory*. The latter presumes that sample units vary, which is why we need more than one. When we have a sampling protocol that uses a single site to allegedly 'represent' a much bigger area, like a catchment, we are actually imposing a particular model, which specifically assumes that variation between prospective sites within a catchment is close to zero or is at least trivial compared to catchment to catchment variation. Unfortunately, there are adequate numbers of studies demonstrating that this model is false. Sites of similar channel morphology and hydrology do not necessarily have comparable diversities or densities of species – sometimes they have quite the reverse; there are studies demonstrating that variation between sites separated by a few hundred metres of river length (or less) can be greater than that seen between whole catchments (see references and discussion in Townsend *et al.*, 2004; Heino, Louhi & Muotka, 2004). Such variability means that sampling protocols using 'representative' sites fatally mix site-to-site variation with catchment-to-catchment variation in unpredictable and uncontrolled ways, and that inevitably leads to false conclusions (Townsend *et al.*, 2004).

Using 'representative' reaches is a method that is entrenched in many parts of the freshwater literature, particularly in papers about management and assessment of rivers (although stellar exceptions can be found: Jeffers, 1998). Hundreds of millions of dollars and Euros have already been invested in river assessment protocols using this as a sampling method, and it continues to be seen as the standard approach (see for example *Hydrobiologia* (2006) vol. 566, issue 1). Given this investment, it might be tempting to brush aside the above arguments as mere 'pedantry', but describing the logical bases of sampling as an esoteric debate of no interest to those who are busy trying to solve the problems caused by human impacts on rivers is a serious mistake. It is rather analogous to investing millions of dollars building skyscrapers, struggling to get them as high as possible but allowing the builders to construct the foundations out of flour and water. Moreover, this problem cannot be solved after the fact statistically. We can apply statistical analysis to any data we like, but if sample observations do not abide by sampling theory then there is no legitimate, logical pathway from sample estimates to inferences about populations (a situation succinctly, if rather offensively, described by the old adage 'garbage in, garbage out': Tabachnick & Fidell, 1996), and it is the latter that tell us about whole catchments and regions.

I have no doubt that the suggestion that we should stop using 'representative' reaches or sites as a sampling method will be highly unpalatable and confronting to many freshwater researchers, but using sampling protocols grounded in sampling theory does have immediate, practical benefits. That is, this argument is not just about the logical niceties of abiding by sampling theory so that we can all feel smug about our data. The mathematics of sampling theory (which rest on probability theory: Sokal & Rohlf, 1995) show that estimates derived from truly representative samples (typically, random samples) *will* be more accurate than estimates gained from sampling processes that cannot, and do not, ensure representation. If estimates from samples are more accurate, then this means the level of predictive success we can generate from statistical analyses applied to those data will be improved, i.e. large scale regional sampling programmes could be improved simply by changing the sampling protocols to make them abide by the logic of sampling theory.

Doubtless it is the perceived expense of fully randomised sampling designs that have obviated their use in many quarters. Thus, in most studies of human impacts, the spatial scales of impacts are large – whole river channels or catchments – and our hypotheses contain predictions we expect to see at these spatial scales. Typically, we need replicate sites within catchments (or rivers) and replicate samples within sites, but, when hypotheses are about whole rivers or whole catchments, replicate sites and replicate samples within sites do not provide degrees of freedom (i.e. replicates) for tests of the relevant hypotheses. The role of replicate sites and samples within sites is simply to ensure catchment- or river-level estimates are accurate (Quinn & Keough, 2002). The need to replicate at large spatial scales while maintaining adequate, representative sampling at small scales usually produces a sobering amount of work though. For example, we might wish to sample invertebrates at ten sites in each of three catchments that vary in degree of human impact, but know that 12 Surber samples are needed to gain a sufficiently precise estimate of mean densities at each site, resulting in a prohibitive 360 samples. This is the sort of outcome that has probably deterred researchers from pursuing rigorous sampling across multiple spatial scales and encouraged them instead to pursue other ways of keeping down costs, such as family-level identification or 'representative' sites.

However, random sampling is not actually more costly to do. We can optimise designs to target sampling effort at those levels of replication that provide degrees of freedom for hypothesis tests. The solution lies in compositing samples across levels where the variance is of no interest or value, and then sub-sampling these pooled samples to derive estimates. In the example above, Surber samples could be pooled into a single sample for each site from which we then produce our estimate of average density for that site. Given that sites are the relevant replicates for the hypothesis (here, it is about differences between the catchments), each contributes only a single value and degree of freedom for the analysis, so little is gained, and a lot of time and money consumed, by enumerating each Surber separately. (Of course, counting all the animals in a pooled sample and dividing by the total area sampled is algebraically equivalent to enumerating each sample separately and then calculating the average density.) The total

number of animals may be in the tens of thousands in composited samples, but the latter can be sub-sampled in the laboratory (through use of splitters: Longhurst & Seibert, 1967 or sub-samplers: Waters, 1969). The analysis required to discover how many sub-samples need to be counted in order to estimate well the total within a composited sample is relatively straightforward (Krebs, 1989), even when dealing with rare species. The composited sample is mixed well in the laboratory, sub-samples taken, then we calculate how many need to be counted to achieve a minimum level of accuracy (usually expressed as the width of the confidence interval as a percentage of the mean: Krebs, 1989). This is a very effective way of reducing the time required to process samples. In the above example, instead of coming back from the field with 360 samples, we come back with 30. Although the time required to process each of these 30 samples is greater individually than it would be for each of the Surber samples (because they need to be split into sub-samples in the laboratory) the total time is far less, and costs in time and money can be reduced by as much as two thirds without compromising the quality of information.

Discussions about sub-sampling are present in the stream literature on human impacts but they largely focus on ways of cutting down the laboratory processing time needed to ensure representation of species in community-level data, particularly rare species, (see exchange of views in the *Journal of North American Benthological Society*, 15(3), for example). Very little of the discussion is about using pooling and sub-sampling in survey designs to target the level of replication appropriate to tests of hypotheses while cutting down the cost, an approach that seems uncommon in the freshwater literature looking at human impacts on rivers, but is not new by any means.

A short word about choice of analytical model

The fifth component of design is to specify the analysis before data collection has commenced. It is not possible to optimise sampling properly, or do power analysis to consider how many replicates need to be collected, without an analytical model. Different analytical models obviously pose different requirements for replication, and an important point is that there must be an explicit connection between the model used and the hypothesis posed. An interesting observation is that the great majority of commonly used statistics work with hypotheses about measures of central tendency like means (e.g. analysis of variance, MANOVA, centroids in various types of ordination) or central responses, such as when we look at the responses of organisms to environmental gradients (using, for example, simple least squares regression, multiple regression). Arguably though, it could be more informative to test hypotheses about limits or constraints on the distribution of organisms than those that are about 'average' responses. For example, a common management need is to determine the minimum discharges that need to be left downstream of dams for rivers to meet requirements for 'river health'. This is a question about the minimum of a distribution that relates a measurement of river health to a measure of discharge; it is not about the centre of that distribution.

To explain this, we can use a form of statistical analysis called quantile regression, which tests hypotheses about constraints but is not yet commonly used in ecology (Cade & Noon, 2003; Lancaster & Belyea, 2006). Quantile regression allows us to test whether factors set maximum or minimum limits to values – for example, we can test an hypothesis that water velocities set limits to maximum densities (Fig. 4). Asking questions about limiting responses involve fitting lines to the margins of sets of points (not the centre of them). Scatter below the line for maxima or above the line for minima does not affect the fit *per se* (Fig. 4). The advantage of this approach is that hypotheses about constraints specifically allow for factors other than the one of interest to affect the dependent variable without compromising the fit of the analytical model and hence the test of the hypothesis. When we fit central relations between an independent and dependent variable using survey data, other independent variables create variation that usually results in poorly fitting lines with a lot of scatter (even when we use multivariate approaches). Quantile regression is thus ideally suited to survey designs where we have measured biological variables along multiple prospective causal gradients and obtained scatter plots with triangular distributions of points (e.g. Fig. 1a). In many cases, such distributions are caused by two or more variables all affecting the dependent variable simultaneously. Quantile regression allows us to explore which of these variables is

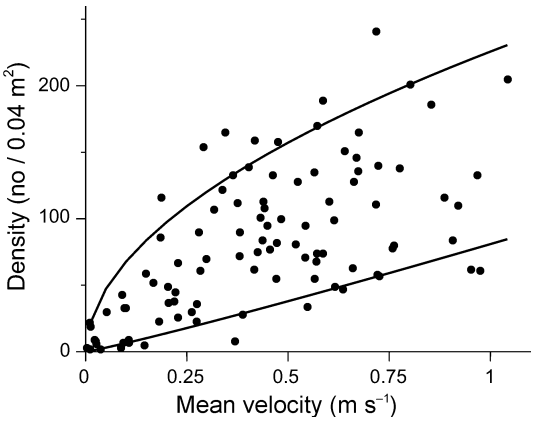

Fig. 4 Example of the application of quantile regression to survey data. Numbers of baetid mayflies, predominantly *Baetis rhodani*, in 0.04 m^2 Surber samples collected from a single riffle of Faseny Water in south-eastern Scotland, plotted against water velocity taken simultaneously, and analysis by quantile regression showing 10th and 90th regression quantile functions (solid lines). The lines are statistically significant fits to the margins of the points and provide evidence that nearbed velocities constrain both the upper and lower limits of densities of baetid mayflies. (Redrawn from Lancaster & Belyea, 2006.)

dominant, which means it has great potential to disentangle the effects of multiple stressors in ways that traditional analytical approaches in ecology cannot. Additionally, as suggested above, questions and hypotheses about what factors set limits to numbers (rather than those focusing on the average response) seem more sensible both ecologically and for management decisions.

Conclusions

The arguments in this paper show there are a raft of improvements that can be made to some standard survey designs and sampling methods in some areas of freshwater research, particularly in studies looking at human impacts on rivers. As has been acknowledged throughout this paper though, these improvements do not spring from new methodological advances in the last few years. They come from foundational aspects of sampling and scientific inference that at least used to be part-and-parcel of basic training in ecology. They come also from familiarity with the science of ecology and the progress that has been made over the last 100 or so years in tackling basic questions about distribution and abundance, the answers to which underpin effective management, conservation and repair of ecosystems.

The fact that these aspects seem less commonly understood and practised now than 40–50 years ago raises some serious questions about the education we are providing at universities and whether we are giving students sufficient or the right training to be effective, ecological scientists. However, I think there is another aspect to this as well. Understanding human impacts on rivers is an inter-disciplinary field that involves scientists other than ecologists, such as hydrologists, geomorphologists, engineers, (and of course people from other disciplines like economics, policy and planning). It seems likely that education in these latter disciplines does not include formal training in the science of ecology and its methods, even though these are specifically needed to tackle ecological questions. Thus, perhaps part of the problem is that research into human impacts on rivers does not typically involve true, and equal, collaboration between ecologists and other researchers even when the questions posed are inherently ecological (not hydrological, etc.). That suggests the nature of inter-disciplinary teams may need some careful thought in future.

Another reason why some foundational aspects of science may have been lost is a perception that it is simply too expensive and/or takes too long, and we need answers fast. As I hope I have demonstrated, there are ways to cut both costs and time that do not compromise the quality of information or our capacity to make strong inferences. Moreover, if research is unlikely to produce definitive tests of hypotheses, which are needed to provide definitive guidance for managers, then that is actually very poor value for money, and it might be better to spend the funds doing something else. For research targeted specifically at solving problems for management, we should recognise that a cost of great concern to managers is the cost of *making mistakes*. For example, water in Australia is a highly contested resource that costs a lot of money. Managers can argue for environmental flow releases, but these can be worth millions of dollars and must be warranted by arguments about the environmental benefits that will be produced. If environmental flows are based on science that is not the best possible, they are unlikely to achieve those benefits because the predictions about environmental changes are very likely to be wrong, producing not only a waste of money but also a heavy cost in public opinion given a likely backlash against 'squandering'

water on the environment. Against this, the cost of conducting research well in the first instance can be put into a much clearer perspective. These are the sorts of arguments about costs that should and can be martialled, if necessary, to have ecological research on human impacts funded properly.

Acknowledgements

I am grateful to the Freshwater Biological Association for the invitation to speak at the inaugural conference and be given an opportunity to publish these thoughts. These ideas have developed in discussion with many, many colleagues over a long period of time, and I also acknowledge relevant financial support from Land & Water Australia (Environmental Water Allocation Programme, Project UME71) and the Australia Research Council (Discovery Grant, DP0772854). Many thanks to Jill Lancaster for acting as a patient sounding board, for drawing Figs 1 and 4, and for comments on an early draft of the manuscript.

Conflicts of interest

The author has declared no conflicts of interest.

References

Beveridge W.I.B. (1961) *The Art of Scientific Investigation*, 3rd edn, Heinemann, London.

Brittain J.E. & Eikeland T.J. (1988) Invertebrate drift – a review. *Hydrobiologia*, **166**, 77–93.

Cade B.S. & Noon B.R. (2003) A gentle introduction to quantile regression for ecologists. *Frontiers in Ecology and the Environment*, **1**, 412–420.

Chessman B.C. (2003) New sensitivity grades for Australian macroinvertebrates. *Marine and Freshwater Research*, **54**, 95–103.

Chesson P. (1986) Environmental variation and the coexistence of species. In: *Community Ecology* (Eds J. Diamond & T. Case), pp. 240–256. Harper and Row, New York.

Dakou E., D'heygere T., Dedecker A.P., Goethals P.L.M., Lazaridou-Dimitriadou M. & De Pauw N. (2007) Decision tree models for prediction of macroinvertebrate taxa in the river Axios (Northern Greece). *Aquatic Ecology*, **41**, 399–411.

Davy-Bowker J., Clarke R.T., Johnson R.K., Kokes J., Murphy J.F. & Zahra'dkova S. (2006) A comparison of the European Water Framework Directive physical typology and RIVPACS-type models as alternative methods of establishing reference conditions for benthic macroinvertebrates. *Hydrobiologia*, **566**, 91–105.

Dewson Z.S., James A.B.W. & Death R.G. (2007) A review of the consequences of decreased flow for instream habitat and macroinvertebrates. *Journal of the North American Benthological Society*, **26**, 401–415.

Dolédec S., Lamouroux N., Fuchs U. & Mérigoux S. (2007) Modelling the hydraulic preferences of benthic macroinvertebrates in small European streams. *Freshwater Biology*, **52**, 145–164.

Downes B.J. & Lancaster J. (2010) Does dispersal control population densities in advection-dominated systems? A fresh look at critical assumptions and a direct test. *Journal of Animal Ecology*, **79**, 235–248.

Downes B.J. & Reich P. (2008) What is the spatial structure of stream insect populations? Dispersal behaviour of different life-history stages. In: *Aquatic Insects: Challenges for Populations* (Eds J. Lancaster & R. Briers), pp. 184–203. CAB International, Wallingford, UK.

Downes B.J., Barmuta L.A., Fairweather P.G., Faith D.P., Keough M.J., Lake P.S., Mapstone B.D. & Quinn G.P. (2002) *Monitoring Ecological Impacts: Concepts and Practice in Flowing Waters*. Cambridge University Press, Cambridge.

Downes B.J., Entwisle T.J. & Reich P. (2003) Effects of flow regulation on disturbance frequencies and in-channel bryophytes and macroalgae in some upland streams. *River Research and Applications*, **19**, 27–42.

Folt C.L., Chen C.Y., Moore M.V. & Burnaford J. (1999) Synergism and antagonism among multiple stressors. *Limnology and Oceanography*, **44**, 864–877.

Fonseca D.M. (1999) Fluid-mediated dispersal in streams: models of settlement from the drift. *Oecologia*, **121**, 212–223.

Furse M., Hering D., Moog O. *et al.* (2006) The STAR project: context, objectives and approaches. *Hydrobiologia*, **566**, 3–29.

Hale R., Downes B.J. & Swearer S.E. (2008) Habitat selection as a source of inter-specific differences in populations of two diadromous fish species. *Freshwater Biology*, **53**, 2145–2157.

Hansen R.A., Hart D.D. & Merz R.A. (1991) Flow mediates predator-prey interactions between triclad flatworms and larval blackflies. *Oikos*, **60**, 187–196.

Heino J., Louhi P. & Muotka T. (2004) Identifying the scales of variability in stream macroinvertebrate abundance, functional composition and assemblage structure. *Freshwater Biology*, **49**, 1230–1239.

Hemphill N. (1988) Competition between two stream dwelling filter-feeders, *Hydropsyche oslari* and *Simulium virgatum*. *Oecologia*, **77**, 73–80.

Hoover S.E.R., Shannon L.G.W. & Ackerman J.D. (2007) The effect of riparian condition on invertebrate drift in mountain streams. *Aquatic Sciences*, **69**, 544–553.

Hynes H.B.N. (1960) *The Biology of Polluted Waters*. Liverpool University Press, Liverpool, UK.

Jeffers J.N.R. (1998) The statistical basis of sampling strategies for rivers: an example using River Habitat Survey. *Aquatic Conservation: Marine and Freshwater Ecosystems*, **8**, 447–454.

Jowett I.G. (2003) Hydraulic constraints on habitat suitability for benthic invertebrates in gravel-bed rivers. *River Research and Applications*, **19**, 495–507.

Keough M.J., Quinn G.P. & King A. (1993) Correlations between human collecting and intertidal mollusc populations on rocky shores. *Conservation Biology*, **7**, 378–390.

Kingsland S.E. (1991) Defining ecology as a science. In: *Foundations of Ecology: Classic Papers with Commentaries* (Eds L.A. Real & J.H. Brown), pp. 1–13. University of Chicago Press, Chicago, USA.

Kirk R.E. (1995) *Experimental Design: Procedures for the Behavioral Sciences*, 3rd edn. Brooks/Cole Publishing Co, Pacific Grove, USA.

Krebs C.J. (1978) *Ecology: The Experimental Analysis of Distribution and Abundance*, 2nd edn. Harper & Row Publishers, New York, USA.

Krebs C.J. (1989) *Ecological Methodology*. Harper & Row Publishers, New York, USA.

Lake P.S. (2005) Perturbation, restoration and seeking ecological sustainability in Australian flowing waters. *Hydrobiologia*, **552**, 109–120.

Lancaster J. & Belyea L.R. (2006) Defining the limits to local density: alternative views of abundance-environment relationships. *Freshwater Biology*, **51**, 783–796.

Lancaster J. & Downes B.J. (2009) Linking the hydraulic world of individual organisms to ecological processes: putting ecology into ecohydraulics. *River Research and Applications*, in press.

Lancaster J., Downes B.J. & Glaister A. (2009) Interacting environmental gradients, trade-offs and reversals in the abundance–environment relationships of stream insects: when flow is unimportant. *Marine and Freshwater Research*, **60**, 259–270.

Longhurst A.R. & Seibert D.L.R. (1967) Skill in the use of Folsom's plankton sample splitter. *Limnology & Oceanography*, **12**, 334–335.

Macan T.T. (1963) *Freshwater Ecology*. Longmans, Green & Co, London.

Maddock I. (1999) The importance of physical habitat assessment for evaluating river health. *Freshwater Biology*, **41**, 373–391.

Manly B.F.J. (1992) *The Design and Analysis of Research Studies*. University of Cambridge Press, Cambridge.

Mayo D.G. (1996) *Error and the Growth of Experimental Knowledge*. University of Chicago Press, Chicago, USA.

Melbourne B.A. & Chesson P. (2005) Scaling up population dynamics: integrating theory and data. *Oecologia*, **145**, 178–186.

Metcalfe-Smith J.L. (1994) Biological water-quality assessment of rivers: use of macroinvertebrate communities. In: *The Rivers Handbook*, Vol 2 (Eds P. Calow & G.E. Petts), pp. 144–170. Blackwell Scientific Publications, London, UK.

Platt J.R. (1964) Strong inference. *Science*, **146**, 347–353.

Popper K.R. (1983) *Realism and the Aim of Science. Postscript to The Logic of Scientific Discovery*, vol. 3, series ed. Hutchinson, W.W. Bartley, London.

Quinn G.P. & Keough M.J. (2002) *Experimental Design and Data Analysis for Biologists*. Cambridge University Press, Cambridge, UK.

Resh V.H. (1979) Biomonitoring, species diversity indices, and taxonomy. In: *Ecological Diversity in Theory and Practice*. (Eds J.F. Grassle, G.P. Patil, W. Smith & C. Taillie), pp. 241–253. International Co-operative Publishing House, Fairland, Maryland.

Rogers K. & Biggs H. (1999) Integrating indicators, endpoints and value systems in strategic management of the rivers of the Kruger National Park. *Freshwater Biology*, **41**, 439–451.

Rundle S.D., Weatherley N.S. & Ormerod S.J. (1995) The effects of catchment liming on the chemistry and biology of upland Welsh streams: testing model predictions. *Freshwater Biology*, **34**, 165–175.

Schneider D.C. (1994) *Quantitative Ecology: Spatial and Temporal Scaling*. Academic Press, San Diego, USA.

Sokal R.R. & Rohlf F.J. (1995) *Biometry: The Principles and Practice of Statistics in Biological Research*, 3rd edn. W.H. Freeman and Co., New York.

Tabachnick B.G. & Fidell L.S. (1996) *Using Multivariate Statistics*, 3rd edn. HarperCollins College Publishers, New York, USA.

Todd C.D., Phelan P.J.C., Weinmann B.E., Gude A.R., Andrews C., Paterson D.M., Lonergan M.E. & Miron G. (2006) Improvements to a passive trap for quantifying barnacle larval supply to semi-exposed rocky shores. *Journal of Experimental Marine Biology and Ecology*, **332**, 135–150.

Toulmin S., Rieke R. & Janik A. (1979) *An Introduction to Reasoning*. Macmillan Publishing Co., New York, USA.

Townsend C.R., Downes B.J., Peacock K. & Arbuckle C.J. (2004) Scale and the detection of land-use effects on morphology, vegetation and macroinvertebrate communities of grassland streams. *Freshwater Biology*, **49**, 448–462.

Underwood A.J. (1989) The analysis of stress in natural populations. *Biological Journal of the Linnean Society*, **37**, 51–78.

Underwood A.J. (1990) Experiments in ecology and management: their logics, functions and interpretations. *Australian Journal of Ecology*, **14**, 365–389.

Underwood A.J. (1993) The mechanics of spatially replicated sampling programmes to detect enviromental impacts in a variable world. *Australian Journal of Ecology*, **18**, 99–116.

Underwood A.J. (1995) Ecological research and (and research into) environmental management. *Ecological Applications*, **5**, 232–247.

Vale M.M., Cohn-Haft M., Bergen S. & Pimm S.L. (2008) Effects of future infrastructure development on threat status and occurrence of Amazonian birds. *Conservation Biology*, **22**, 1006–1015.

Walling D.E. & Webb B.W. (1992) Water quality I. Physical characteristics. In: *The Rivers Handbook*, Vol 1 (Eds P. Calow & G.E. Petts). pp. 48–72. Blackwell Scientific Publications, London, UK.

Waters T.F. (1969) Subsampler for dividing large samples of stream invertebrate drift. *Limnology & Oceanography*, **14**, 813–815.

Wright J.F. (1995) Development and use of a system for predicting the macroinvertebrate fauna in flowing waters. *Australian Journal of Ecology*, **20**, 181–197.

(Manuscript accepted 16 November 2009)