Regional Environmental Publications

189

Kristian Meissner

Experimental methods in the assessment and monitoring of rivers: benefits, limitations and integration with field surveys

VAASA 2000

WEST FINLAND REGIONAL ENVIRONMENT CENTRE

ISBN 952-11-0802-9
ISSN 1238-8610

Picture on cover: Kristian Meissner
Layout: Tuula Nyqvist
Place of printing: Multiprint, Vaasa
2000

Content

Introduction

Stream biomonitoring programs often start with a decision to allocate resources to investigate a specific situation over an extended period of time. The managerial decisions and aims are often beyond the realm of the investigator, are vaguely stated and, as such, cannot be readily explored. To succeed in the monitoring task, investigators will often have not only to rephrase the managerial aims into testable hypotheses but also to solve problems of limited funding and suboptimal data (Treweek 1996). Clearly, successful monitoring is related to the ability of making testable predictions about system responses and choosing the most adequate approach.

This requires a sound understanding of the system dynamics at several scales of observation. While there are many approaches to biomonitoring in streams, linking patterns to their causes through manipulative experiments is by far the most powerful and accurate method available (Cooper & Barmuta 1993). It is nevertheless important to realize that the temporal and spatial scale of a study will always affect the results and their extrapolation to other, usually wider, scales (for examples, see Thrush et al. 1997a, 1997b). Thus, iteration between experimental work, theory and field patterns is necessary to link observed patterns to their cause(s) (Werner 1998).

Working on large spatial scales (e.g. entire streams) is usually desirable, but due mainly to high costs of replication, it will often be problematic or unattainable (Carpenter 1989). Although field assessments provide a powerful tool for inferring human impacts, proper data preceding man-induced environmental accidents are often scarce or entirely missing. Similarly, in Environmental Impact Assessments (EIA), the time for sampling prior to an anticipated impact is usually limited, or sampling is conducted only after the onset of the impact. While these are often the realities the investigator must cope with, properly designed and executed field assessments are able to identify impacts of human interventions and should be used whenever possible. To be applicable, however, these techniques need both pre- and post-impact data. As will be shown in this report, there are ways to use field assessment designs even if pre-treatment data are missing but this requires extensive baseline data (Underwood 1996).

This report contributes to develop appropriate monitoring methodologies in rivers as a part of the 3-year EU RiverLife project completed in 2001 (Internet: http://www.vyh.fi/ympsuo/projekti/lifeppo/riverl/riverlife.htm).

2 Approaches to biomonitoring

Approaches to biomonitoring in streams can be roughly placed into four categories:

(i) Inventories and field surveys,
(ii) laboratory experiments,
(iii) field experiments and
(iv) field assessments.

There is no single, all-purpose approach to biomonitoring. Depending on the question the investigator has to trade off between three basic goals which cannot be easily incorporated into a single research program: (i) precision, (ii) realism and (iii) generality (Levins 1968). Different approaches to biomonitoring rank differently in relation to these concepts (Fig. 1).

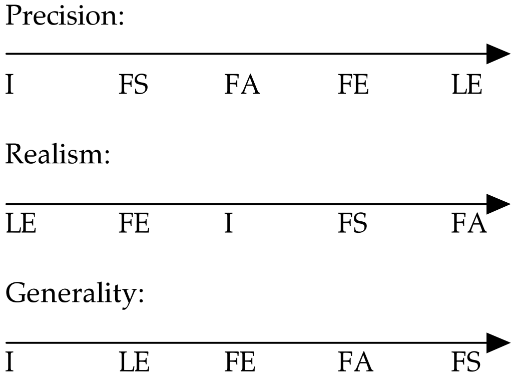

Fig. 1. Rank order of the different approaches used in monitoring in relation to different priorities set by the investigator. Abbreviations stand for inventory (I), field survey (FS), field assessment (FA), field experiment (FE) and laboratory experiment (LE), respectively.

It is beyond the scope of this report to explicitly deal with the many aspects of each approach (for such reviews, see Eberhardt & Thomas 1991, Cooper & Barmuta 1993). Similarly, ecotoxicological biomonitoring methods will, for brevity not be addressed in detail. Rather, I will describe the general nature and design of the various biomonitoring approaches and some of the most common problems associated with each of them.

2.1 Inventories, surveys and natural experiments

Although inventories have provided fundamental biological knowledge and clearly will continue to have an important status in biological monitoring (i.e. in providing valuable background information for more intricate analysis), they have often been misused as the sole approach to monitoring situations. There are severe problems related to the use of unreplicated surveys as the only monitoring tool. Similarly, monitoring using repeated tailor-made inventories may be inefficient. Inventories typically collect large sets of biotic and abiotic data from single or multiple times or sites with the purpose of finding important relationships between some of the measured factors. Differences in e.g. benthic densities between sites or times with different environmental conditions (e.g. pre- and post impact) are then related to differences in environmental conditions. However, data produced are purely correlative and can only indicate, not test, relationships. This is because perceived differences between sites or times could also be caused by other, possibly unmeasured factors.

The general problem is the interpretation of results, i.e. did the system monitored change after intervention and did this intervention cause the change (Carpenter et al. 1989). Obviously, in some cases the only plausible explanation to observed responses is the **intervention itself**. Note, however, that the observer has to explore any **plausible alternative mechanisms** (e.g. through small-scale studies) before defining the main cause of change. This may be easier for situations in which the observed change following an intervention is very large and where the system returns to its former state after the removal of the intervening factor.

Example: Mittelbach et al. (1995) provide an excellent example of an unreplicated, long-term large-scale study in which the system response was large and reversible. In their study, successive winterkills removed the top predator (largemouth bass) from a lake. The lake was then monitored for several years after the winterkill and a dramatic increase of planktivorous fish was recorded, along with a change in zooplankton community structure. Mittelbach et al. (1995) expanded their survey and conducted an experiment where they reintroduced the top predator to the affected lake. Changes in the community structure were monitored for seven years and it appeared that the system returned to its pre-winterkill state, with the predator population reaching its former level. In this study, any alternative hypotheses (e.g. broad-scale climatic changes or a change in nutrient status) could not explain the results, thus strengthening the conclusion that the removal of the top piscivore was the only plausible cause of impact. In this case, conclusions gained further credibility by long-term monitoring of the study system, reversibility of the effect, and by theoretical connections to trophic cascades.

In many surveys, however, patterns may be indiscernible due to a limited number of data points and high sampling variation. Whether such results can be extrapolated to sites or times not included in the survey, is often uncertain (e.g. Cooper & Barmuta 1993). The problem of extrapolation is, however, not limited to surveys or natural experiments but is common to any approach (see e.g. Frost et al. 1988, Carpenter & Kitchell 1988, Thrush et al. 1994, Englund & Olsson 1996). As sampling in survey monitoring is often done only once or at regularly spaced intervals it is possible to miss important events (e.g. due to a sudden drop of pH or an accidental spill of a chemical) that have long lasting effects on species populations. It may therefore be difficult, or even impossible, to relate observed responses to their true cause(s). Thus, survey methods should be mainly used to back-up other approaches in benthic monitoring (Cooper & Barmuta 1993).

2.2 Laboratory experiments

The focus of laboratory experiments is often to test possible mechanisms leading to observed patterns in nature. The advantage of laboratory experiments is that they allow: (i) standardization of conditions, (ii) use of sophisticated statistical designs, and (iii) adequate replication. This permits statistically valid conclusions at a relatively low cost. However, experimenters working with aquatic mesocosms often favor replication at the expense of the spatial extent and duration of the experiment (Petersen et al. 1999). As laboratory conditions already differ in many ways from those experienced in the field, reduction in the spatial and temporal extent of experiments increases the risk of artifactual results. The target organisms are often separated from natural conditions in relation to: a) weather, b) specific habitat features, c) resources and/or d) inter- and intraspecific interactions. These simplifications may help to reduce "noise" commonly associated with data from field studies. However, when designing laboratory experiments, the experimenter has to carefully balance the benefits of simplification against the threat of artificiality.

Widely used standard toxicity tests pose additional problems through their use of species that have unknown relative sensitivities. Further, genetically homogenous test populations (e.g. Sweeney et al. 1993) may lack the adaptive abilities of natural populations (Pontasch et al. 1989). Extrapolation of results from standard single-species tests to higher levels of organization without prior validation through extensive field studies has long been known to be problematic (Cairns 1983), and the common practice of using non-lotic species further complicates the issue.

Examples: Pontasch et al. (1989) investigated adverse effects of an effluent on test populations of laboratory daphnids and several lotic species in the laboratory and in the field. Single-species tests with daphnids predicted the effects of the effluent reasonably well for most lotic taxa, with the exception of mayflies. Pontasch et al. (1989) concluded that multispecies tests using indigenous lotic organisms are more laborious and costly but provide concise predictions of the effects of the effluent (e.g. inhibition or stimulation of insect emergence) in the receiving system.

In another laboratory experiment, Clements et al. (1989) exposed hydropsychids and stonefly predators to heavy metal stress. Their results showed that hydropsychids became more vulnerable to predation by stoneflies as heavy metal stress increased. Although the results were not verified through planned field experiments, and thus a laboratory artifact cannot be ruled out, it is important to notice that a single-species test would have been inadequate for detecting this pattern.

Thus the use of single-species laboratory tests as surrogates for lotic communities, or even for the estimation of the effect on the field populations of the same species, may not be unambiguous. Sensitivity to metal pollution has been shown to differ within families and across life-stages, being highest at early life-stages (Kiffney & Clements 1994, 1996) and during molting (Pontasch et al. 1989). As experimenters routinely use larger, more resistant life-stages in toxicity tests, underestimation of the toxicity effects may be common. The only safeguard against the threat of artifacts is to compare the laboratory results to the actual pattern observed in the field. Ideally, experiments should use multi-scale approaches (Petersen et al. 1999) and different life-stages (Kiffney & Cements 1996, Sweeney et al. 1993) in both field and laboratory studies.

2.3 Field experiments

Field experiments enable the experimenter to combine the realism of natural systems with proper experimental design. **Field experiments differ from natural experiments** in that the experimenter **actively controls** the target variable(s) and **treatments are truly replicated** (Cooper & Barmuta 1993). When biological variables are used as target variables the level of taxonomic resolution used is an important issue. The level of resolution is dependent on the **expected magnitude of the impact** (i.e. "effect-size"). Generally, minor impacts may require high taxonomic resolution while drastic ones will permit less resolution (Cooper & Barmuta 1993). Since natural environments are heterogeneous across multiple spatial scales (e.g. Dutilleul 1993, Legendre 1993), a major problem is to distinguish the true effect from variance or "noise" around the effect. There are several factors contributing to random error variance, which are all to some degree related to scaling issues. First, the experimenter has to consider the mismatch of the scale of the question and the analysis; since biological processes cannot be linearly extrapolated, it will be of little use to address questions operating at whole-river scales through the use of small-scale experiments, which cannot provide answers beyond their extent. Deciding on the size of individual units of observation (grain) and on the overall area of the study (extent) is thus an integral part of designing a field experiment. Patterns apparent at one scale of observation may be missed at other scales of sampling (e.g. Thrush et al. 1994), making observed variability conditional on the scale of description (Levin 1992). Changes in the scale of measurement affect the variance of the measured variable; for example, increasing the grain, while keeping the extent constant, decreases spatial variance. A wider grain will capture a greater proportion of spatial heterogeneity within each sample, and thus some resolution is lost, while between-grain heterogeneity (variance) is decreased. An increase in extent, while keeping the grain constant, will include more patches or landscape elements in the area studied. This will lead to an increase in the between-grain variance (Wiens 1989). Defining the appropriate spatial and temporal extent of an experiment requires good knowledge of the system characteristics. Working simultaneously on several temporal or spatial scales may be inevitable in order to discern the scale of impacts, whenever knowledge about the appropriate scale(s) is unavailable. The experimenter must focus on finding scales for which statistical behavior is more regular (Levin 1992) and which are best matched with the ecological question asked.

Another important concern in the design of experiments is statistical power, as it has important implications for the efficiency of the program. Testing the hypotheses that a certain intervention causes an impact can have two outcomes: either we conclude there is an impact or we reject this conclusion. However, our conclusion about the nature of the impact may be either right or wrong, resulting in the conventional error table (Table 1) for statistical hypotheses testing.

Table 1. Statistical conclusions of hypothesis testing in relation to detecting environmental impacts. Probabilities of error types are given in parentheses (modified from Peterman 1990, Fairweather 1991).

		Conclusion of study IMPACT	NO IMPACT
Actual state	IMPACT	correct (1-α)	Type II error (β)
of nature	NO IMPACT	Type I error (α)	correct (1-β)

Our ability to correctly specify whether there was an impact depends on the power of the test and the significance levels assigned to α and β. While the consequences of making a Type I error is not serious from an environmental point of view (i.e. "only" creating a false alarm), this risk is usually minimized to $\alpha = 0.05$ by convention (Peterman 1990, Fairweather 1991). Committing a Type II error in an EIA situation, however, can have drastic consequences. The risk of making an erroneous decision is controlled by the value of β. The statistical power of a test result is expressed as 1-β. Values close to 1 indicate low risk of a Type II error, and ideally the power of tests should be at least 0.8, or 1-α (Cohen 1988). The value for β is directly proportional to variation in the data and depends on the effect size to be detected (i.e. the maximum tolerable environmental impact), the sample size, and the α-value. Normally, only the sample size and α-level are within the control of the experimenter, thus complicating active control of β.

There are two important applications of power analysis. First, if the statistical behavior of the system is known (from surveys or pilot studies), power analysis can indicate the sample size needed to detect an impact of a specific size in the planning stage of an experiment. Second, power analysis can be used *a posteriori* to interpret any nonsignificant results and help decide whether there really was no effect, or that there simply was not enough power to detect an impact, even if one were present. Power analysis for common tests is available in many standard statistical packages, and a multitude of programs exist for its calculation (see review by Goldstein 1989). Power is generally increased through increased replication, which also improves the accuracy of other statistical estimates. Since increased replication is especially difficult in large-scale experiments (see below), another possibility to increase power is to relax the a-criterion (Fairweather 1991, Mapstone 1995). This, however, as Mapstone (1995) pointed out, has to be done *a priori* for any monitoring program and involves the difficult task of assigning costs of Type I and Type II errors, and a clear specification of the maximum acceptable impact.

Perhaps the most common statistical procedure for the analysis of experiments is analysis of variance (ANOVA) because of its statistical power and versatility to different experimental situations. For example, multifactorial ANOVA can deduce interactions between multiple experimental factors. Analysis of covariance (ANCOVA) allow the introduction of confounding factors as covariates, and repeated measures ANOVA permit the estimation of effects of factors on multiple measures of the same experimental units. More generally, ANOVA models allow the partitioning of variance into factor(s) controlled by the experimenter and "noise" or random error variation. Three main steps are typically used to minimize random error variation: randomization, control and replication (Hurlbert 1984). In any ANOVA-type analysis the use of randomized factorial designs, in which the levels of the treatment factor(s) are randomly assigned through space, reduce the unexplained variance quite efficiently. Completely randomized designs can, however, have their pitfalls (Hurlbert 1984, Dutilleul 1993), and in many cases randomized block designs are more useful by allowing additional partitioning of random deviation. Blocking may be an effective strategy in environments where the variability between the experimental units within a block is smaller when compared to the variation among blocks (due e.g. to environmental gradients). As the variation in blocked experiments is partitioned into three terms (treatment, block and error) instead of just two (treatment and error), blocking will reduce some degrees of freedom but will generally be more powerful (Potwin 1993). In cases where the scale of impact is unknown, the use of nested sampling and hierarchical ANOVA models (e.g. Underwood 1981) may identify the correct domain of scale. It should be pointed out that the use of completely randomized block designs is largely restricted to small and intermediate scales for logistical reasons. Assigning adequate control areas and the use of proper replication at larger scales can be especially problematic in field experiments (e.g. Carpenter 1990). Often upstream control areas are used, but since these tend to influence and differ from downstream conditions they are not truly independent controls *sensu stricto* (Hurlbert 1984). Some degree of pseudoreplication may be unavoidable because of prohibitive costs involved in true replication of large-scale experiments (Carpenter et al. 1989). Ideally, assigning an equal number of treatment and control streams would circumvent pseudoreplication in large-scale experiments and is a prerequisite for generalizations.

Let us consider the validity of several different approaches to the following example: three streams are to be restored (Fig. 2), and fish densities have been chosen as the target variable. In the first scenario, fish densities are estimated only once after restoration. Clearly, in this scenario, it is impossible to assess the effect of restoration (Fig. 2A), since data before restoration are missing, and there is no control. There are ways to salvage after-only data (Underwood 1994, 1996), and these will be discussed later. Note, that the tests used to analyze after-only data are generally less powerful than tests on properly designed assessments. In the second scenario, we now have data on fish densities once before and once after restoration (Fig. 2B). One might be tempted to judge that restoration increased fish densities. However, since control is missing, we are unable to assess whether the increased densities are due to restoration or whether there is a general increase in fish density in the area irrespective of restoration (e.g. due to an exceptionally benign winter). To be able to judge whether the observed increase from year 4 to year 5 was related to restoration, we would not only need longer time series for both the Before and After periods (Fig. 2C), but ideally also from an equal amount of control streams (Fig. 2D). Strictly speaking, only design D (Fig. 2D) allows us to draw conclusions about the general effects of restoration on fish density.

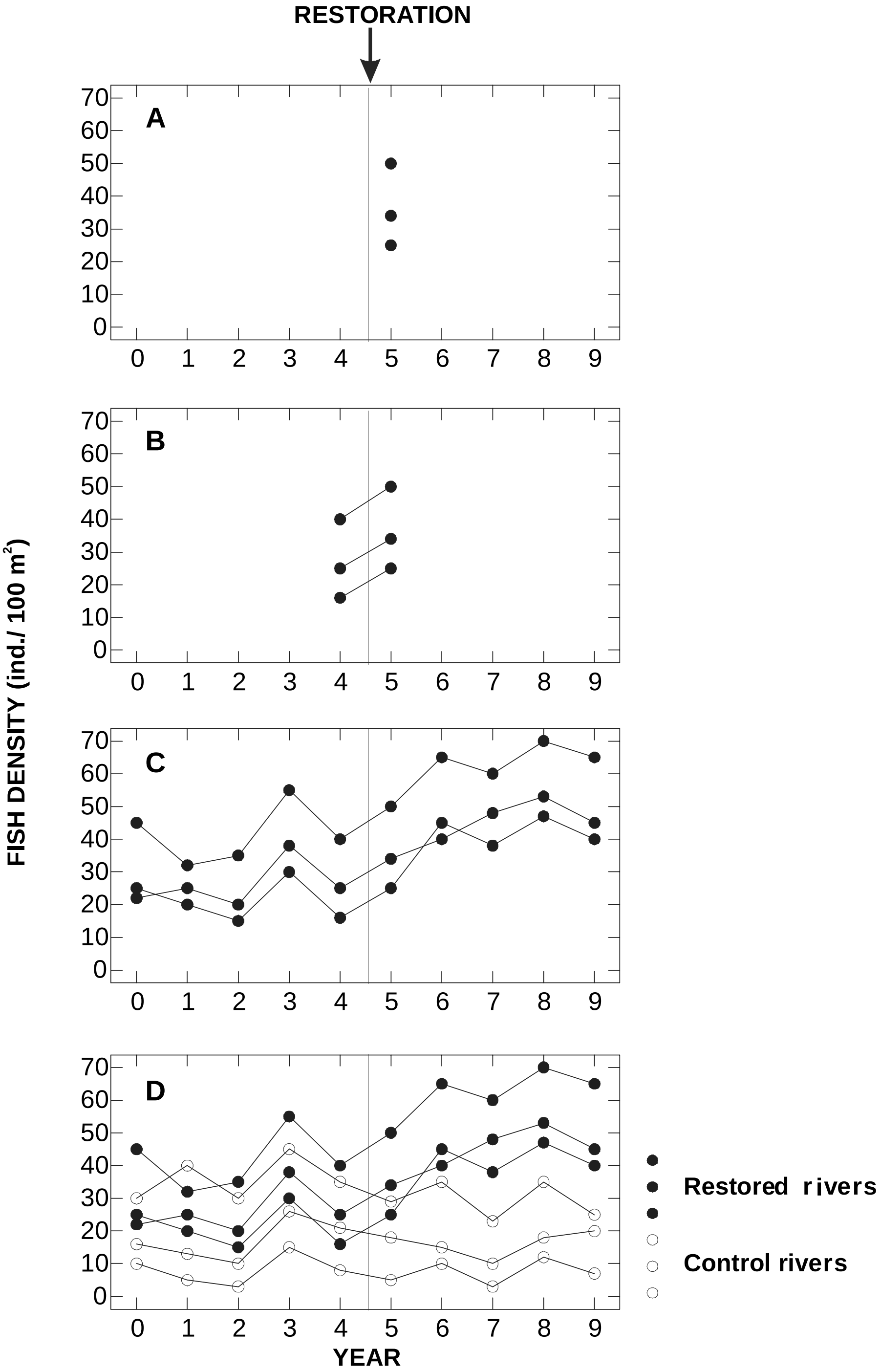

Fig 2. Four monitoring designs to evaluate effects of restoration on fish densities in streams. Fish densities are monitored: (A) one year after restoration in the impact streams, (B) one year both before and after restoration in the impact streams, (C) five years before and after restoration in the impact streams, and (D) five years before and after restoration in the impact and control streams (modified from Mäki-Petäys et al. 1999). See text for a more detailed discussion of the relative strengths and weaknesses associated with each approach.

2.4 Field assessments

In EIA, one is not necessarily interested in general impacts, but rather in a particular impact in a particular location (the impact of a new power plant, impacts of an oil-spill, etc.) (Stewart-Oaten et al. 1986). True replication of treatment (i.e. impact streams) in these cases is either unreasonable, undesirable or even unethical (e.g. in the case of environmental accidents) (Wiens & Parker 1995). Let us consider Green's (1979) example of a single impact location (i.e. an effluent is released into a river) with a single control location. He proposed the following design: samples are taken upstream (control) and downstream (treatment, or impact) of the discharge point. In his "optimal design" he suggested taking observations at *multiple sites* in the Control and the Impact area *once* before and *once* after the introduction of the effluent (similar to design B in Fig. 2, but with a control river and spatial pseudoreplication). An interaction of the *Area X Time* ANOVA-term would then be interpreted as a discharge effect. This is because differences between species abundances changed after the discharge began (Green 1979). Hurlbert (1984) rejected this design and pointed out that differences in abundances could only be demonstrated between locations, and such differences are not necessarily related to the impact. He reasoned that the *Area X Time* interaction term could only indicate an impact if the impact-unrelated difference between the control and impact location remained constant over time. Since the magnitude of this basic difference, however, inevitably changes over time (Hurlbert 1984), an impact will be indiscernible. In an important paper addressing Hulbert's criticism, Stewart-Oaten et al. (1986) laid the basis for a productive discussion about the sampling designs needed to monitor specific impacts. In the following, I will outline the most prominent field assessment approaches that evolved from this debate.

3 BACIPS (Before After Control Impact Paired Series) designs

Unlike in our example on stream restoration (Fig. 2), anthropogenic impacts are normally neither spatially replicated nor randomized. EIA of such single impact situations (e.g. effluents of a plant, building of nuclear power stations) call for special designs. In an extension of Green's "optimal BACI design", Stewart-Oaten et al. (1986) introduced the BACIPS (Before After Control Impact Paired Series) design. Instead of using the initial values of the Control and Impact in the analysis, the BACIPS design focuses on the *difference* in the parameter value between the Impact and the Control area ($\Delta_{Bi} = I_{Bi} - C_{Bi}$). The basic data are formed by the deltas of *multiple sampling occasions* before (Δ_{Bi}), as well as after (Δ_{Ai}), the impact. The mean of Δ_{Bi} is the basic difference between the Impact and Control site, and thus approximates the mean delta expected for the After period in the absence of an impact. The magnitude of the actual impact ("net effect") is calculated as the difference between the means of Before and After deltas (effect size = $\Delta_B - \Delta_A$). Variation in deltas among sampling dates in the Before and After periods (S_Δ) and the number of replicates (i.e. sampling dates; $n_B + n_A = n$) in each period provide confidence intervals for the effect size estimate (Osenberg et al. 1994). In the simplest BACIPS design the variability (S_Δ) and sample size are assumed to be equal in the Before and After periods, but more complex designs can also be analyzed (Stewart-Oaten et al. 1992).

Ultimately, BACIPS designs allow the test of whether the Before data differ from the After data, and commonly a t-test is used to conduct this comparison (Stewart-Oaten et al. 1992). There are several prerequisites for the use of two-sample t-tests in a BACIPS-type analysis. First, the effects of site and time are assumed to be additive. Thus, in the absence of an effect, the deltas (Impact-Control differences) should be equal among sampling occasions (Stewart-Oaten et al. 1986). There are different sources of non-additivity, however, the most prominent one being the lack of temporal coherence (*sensu* Magnusson et al. 1990) of target parameters (e.g. Korman & Higgins 1997, Smith et al. 1993). Second, deltas for different sampling occasions are assumed to be independent (i.e. there is no serial correlation). Serial correlation can occur either as a result of too closely spaced sampling occasions or as a consequence of local-scale events that influence only one site. Note that regional events influencing both areas equally do not lead to serial correlation since they cancel out when subtracting the After from the Before deltas (Stewart-Oaten et al. 1992). Third, the distribution of the deviation for the Impact-Control differences is assumed to be normal, same for each sampling within each period, and same between the Before and After periods (Stewart-Oaten et al. 1992). Careful consideration of these assumptions is needed, as at least some are likely to be violated. In some cases transformations or choice of an appropriate variate of t-test may alleviate the problems caused by violations of the additivity assumption (see Stewart-Oaten et al. (1992) for a detailed account). However, data cannot always be transformed to satisfy the additivity criteria (e.g. Smith et al. 1993, Korman & Higgins 1997), leading to inefficient tests or even artifactual results (Stewart-Oaten et al. 1986, 1992).

Example: Korman & Higgins (1997) simulated salmon population responses to habitat alterations. When control and impact site abundances exhibited poor tracking (i.e. non-additivity), BACIPS had only a 50% chance to detect a threefold change in abundances over a post-impact monitoring period of 10 years (n=20), at an α-error level of 0.20. When the additivity assumption was not violated, the power of detecting an impact increased by 15% (Korman & Higgins 1997). Bence et al. (1996) give some additional examples of non-additive data and provide a solution to the problem: the "predictive" BACIPS approach (see below).

A violation of the independence assumption (i.e. autocorrelation) is likely to occur whenever sampling intervals are too tightly spaced, because deltas close in time tend to be similar, leading to underestimation of the variance of Δ_B and Δ_A (Stewart-Oaten et al. 1992). First-order autocorrelation typically increases the risk of committing a Type I error (concluding there is an impact when there is none) and is easily detected through the Durbin-Watson test (Stewart-Oaten et al. 1986). The correct spacing between sampling occasions, in order to circumvent autocorrelation, depends on the organisms sampled (Stewart-Oaten et al. 1992) and on the target parameter (individual vs. population-level) (Osenberg et al. 1994).

Example: In a field assessment on the effects of wastewater produced during production of oil on a marine ecosystem, Osenberg et al. (1994) concluded that the sampling interval could be 60 days without yielding substantial autocorrelation. By contrast, in a BACIPS examining the effects of an acoustic deterrent on alewives (*Alosa pseudoharengus*) at the water intake of a nuclear power station, a spacing of two days between sampling occasions was deemed sufficient (Ross et al. 1996). More generally, Stewart-Oaten et al. (1992) concluded that autocorrelation of 0.30 was enough to invalidate t-tests.

The choice of the target parameter will also influence the detectability of an impact. To assess the power of different target parameter choices, Osenberg et al. (1994) calculated a standardized effect size ($|\Delta_B - \Delta_A|/(2\ S_\Delta)$). Large standardized effect size values indicate powerful parameters, as the absolute effect size will not be swamped by noise. Osenberg et al. (1994) observed that the variation in deltas (Impact-Control differences) was generally lowest for chemical-physical parameters, intermediate for individual-based (e.g. individual length, gonadal somatic index) and highest for population-based parameters (densities of various organisms). However, when calculating standardized effect sizes, the individual-based parameters scored higher than the chemical-physical or population-based ones. Their data showed that, to detect an impact of one standardized effect size ($\alpha = 0.05$, power of 80%), most individual-based parameters required less than 20 sampling intervals, while the chemical-physical and population-based parameters needed 90 sampling intervals (Osenberg et al. 1994). Low power of the chemical-physical parameters, despite low variability in deltas, was due to a comparably small effect size of the impact. In contrast, while population parameters were highly responsive and had large effect sizes, power tended to be swamped by high natural variability (Osenberg et al. 1994, see Korman & Higgins 1997 for similar results). High natural variation in target variables can be reduced by the collection of multiple samples from each area on each date (Stewart-Oaten et al. 1986). Nevertheless, spatial pseudoreplication should be avoided by taking only one value per area (i.e. the average of the multiple samples) when calculating the delta value for a single sampling occasion.

3.1 The "predictive" BACIPS approach

Bence et al. (1996) described an alternative method, a "predictive BACIPS approach", in which the Impact values are predicted through the data of the Control site. A major benefit of this approach is that it relaxes the additivity assumptions associated with the basic BACIPS model (Bence et al. 1996). In this approach, two functions are derived: one that models the dependency between the Impact and Control values in the Before data and another modeling the dependency between the Impact and Control values in the After period (Stewart-Oaten 1996). Note that (i) any type of function can be used to model the dependency between the Impact and Control data within a period, and that (ii) different functions for the After and Before period can be used, provided that the regression model fits the data well and that residuals are uncorrelated when plotted against the control values (Bence et al. 1996). Effect size estimates for any given Control value can be obtained by simple subtraction of the Before period estimate from the After period estimate. This yields a new function, which plots the estimated effect size (e.g. expected loss), with approximate confidence intervals against control values (Stewart-Oaten 1996, Bence et al. 1996). Thus the effect size is variable and only dependent on the magnitude of the control value. However, also an "average" effect size (similar to the net effect in the basic BACIPS approach), with confidence intervals, can be calculated through the use of all actual Control data points (i.e. in both Before and After periods). This average effect size may be important for predicting, for example, the mean percent loss in a species abundance. To calculate the confidence interval for this average effect size, a jackknife procedure is used (Bence et al. 1996). An underlying assumption of the predictive BACIPS approach is that differences between the expected Impact values and the particular Control values are only due to the perturbation (Stewart-Oaten 1996). In practice, this may not always be the case, and therefore the Before and After Control values should preferably exhibit similar ranges and variability (Bence et al. 1996). Although the predictive approach has the benefit of relaxing additivity assumptions, the approach is still vulnerable to violations of independence through serial correlation. Thus the same precautions as in the standard BACIPS approach are advisable, namely ample spacing of sampling intervals. As a general rule, whenever independence is violated, confidence intervals should be used, with great caution in the interpretation of data (Bence et al. 1996).

3.2 RIA (Randomized Intervention Analysis)

Another method for analyzing before-after monitoring data, Randomized Intervention Analysis (RIA), was presented by Carpenter et al. (1989). Because RIA is based on random permutations, it is unaffected by non-normality and is also robust to heterogeneity of variances (Carpenter et al. 1989). As temporal trends and lagged responses often cause non-normal error distributions, these features of RIA might be beneficial. Additionally, although RIA is affected by autocorrelation, Carpenter et al. (1989) showed that lack of independence in sequential data points does not necessarily cause ambiguous results. In a careful evaluation of RIA and BACIPS assumptions, Stewart-Oaten et al. (1992) state that the user of RIA has to adhere to all other assumptions of BACIPS, or to show that RIA is valid when they fail, but normality. The benefit of robustness against non-normality, however, is rather small, since t-tests are robust, unless sample size is very small (Stewart-Oaten et al. 1992). However, in cases with small sample sizes, RIA may be unable to discern an effect (Carpenter et al. 1989). Nevertheless, the procedure underlying RIA will be shortly presented here, as it is a viable alternative to BACIPS, and has been used successfully in several stream studies (e.g. Wallace et al. 1996, 1997, 1999).

RIA, like BACIPS, is based on paired observations of impact and control locations for both the before and the after period (Carpenter et al. 1989). Intersystem differences are calculated by subtraction of the Impact from the Control values in both before ($D_{Bi} = I_{Bi} - C_{Bi}$) and after ($D_{Ai} = I_{Ai} - C_{Ai}$) periods. Next, the average of the intersystem differences is calculated ($D_B = \sum D_{Bi}/n_{Bi}$; $D_A = \sum D_{Ai}/n_{Ai}$), and the test statistic is formed as the absolute value of the difference between D_B and D_A (i.e. "actual" $|D_B - D_A|$) (Carpenter et al. 1989). Finally, the frequency distribution of intersystem differences is estimated by random permutations. The intersystem differences are randomly assigned to Before and After periods, and all possible permutations have the same chance of being observed (Carpenter et al. 1989). The generated frequency distribution of $|D_B - D_A|$ values is compared against the actual $|D_B - D_A|$, and the proportion of values more extreme than the one observed ($|D_B - D_A|$) form the approximate *P* value (Carpenter et al. 1989). Non-random change in inter-ecosystem differences (i.e. an impact) is indicated by a low *P*-value. Carpenter et al. (1989) showed in simulations that 10 samples were unable to detect changes twice the size of the standard deviation, whereas simulations with 40 samples detected changes the size of one standard deviation. Note that RIA, like BACIPS, can demonstrate only that a change has occurred, not that the change was caused by the intervention (Hurlbert 1984). A sound knowledge of the system and exploration of alternative mechanisms are needed to be able to attribute the change to the intervention.

3.3 Beyond-BACIPS techniques

Although BACIPS techniques solve the problem of temporal replication (Stewart-Oaten 1986), the design is still spatially confounded because it uses a single control site (Underwood 1991, 1992). Since many natural populations fluctuate differently from one place to another (e.g. Korman & Higgins 1997), using a single control site leaves one unable to explicitly demonstrate that the perceived impact was caused by a human intervention, not by natural fluctuations (Hurlbert 1984). BACIPS designs can show that there are differences in temporal patterns between the two sites, but will not be able to exclusively attribute the difference to a potential impact (Underwood 1991). This problem could be overcome through proper spatial replication but, as stated earlier, there seldom is opportunity, or desire, to replicate the impacted sites. Nevertheless, as Underwood (1991, 1992) points out, wherever possible, there is no reason not to replicate the control. Replicate control sites do not have to be identical but they must stem from a population of apparently similar sites, and they should be independently arranged to avoid autocorrelation (Underwood 1992). If there is no true temporal replication (i.e. only one sampling occasion before and after the impact), the asymmetrical ANOVA design suggested by Underwood (1992) is a repeated measures design with two groups: a random factor formed by a group of control locations, the impact location being the sole member of the other group (Underwood 1992). An impact should cause the mean at the impacted site to differ more than expected from the mean at the control sites. The interaction terms of the asymmetrical design allow us to differentiate a human impact from any general changes. In this temporally non-replicated design, an impact is detected by a significant *Before vs. After X Impact vs. Control* interaction (Underwood 1992). Differences between the impact and the control sites are analyzed through orthogonal a priori contrast (Underwood 1992). However, fluctuations in abundances unrelated to the impact can also cause the interaction to be significant. The use of an asymmetrical design allows an easy detection of changes unrelated to the impact through a significant interaction between time of sampling and differences between the control locations (i.e. *Before vs. After X among Controls*) (Underwood 1992). A true impact will affect the interaction from before to after the disturbance between the impacted and the average of the control locations (Underwood 1992). Thus the proper test to account for changes unrelated to impact is the *F*-ratio of the mean square for *Before vs. After X Impact vs. Control* divided by the mean square for *Before vs. After X among Controls* (see Underwood 1992 for details). Note that while a *Before vs. After X Impact vs. Control* interaction clearly indicates an impact, lack of such a relationship does *not* mean there is *no* impact, but may simply mean that the effects of the impact were too small to cause a noticeable change, provided there was ample power to detect changes in the first place (Underwood 1992).

The use of multiple sampling occasions and multiple control sites enables a clearer identification of an impact of an intervention than in the simple asymmetrical design described above. In a temporally replicated case, an impact is detected as an interaction between differences among locations before vs. after the intervention (Underwood 1992). If there is no temporal interaction among controls, an impact is detected by the *F*–ratio of the *among After times X Impact vs. Control*-interaction term divided by the residual error term. Properly replicated, this asymmetrical ANOVA-design bears several advantages to BACIPS. First, it is able to discern an impact through the examination of specific interactions terms, even under conditions of poor temporal coherence, random fluctuations and a coinciding general change. Second, this design can identify not only sustained, "press" disturbances, but also more short-lived, "pulse" disturbances (*sensu* Bender et al. 1984). Third, whenever the focus of biomonitoring is, for example, on detecting changes in mean abundances of a species, Underwood (1991) proposed that detecting changes in the *temporal variance* of a species abundance may be equally important. This is especially true for endangered species, because large fluctuations around a low mean value may increase the risk of extinction (Simberloff 1986). Increased variation around the mean is undetectable through the standard BACIPS and RIA techniques and their detection calls for sampling at different time scales (Underwood 1991). Asymmetrical ANOVA may be extended to incorporate nested sampling (Underwood 1991, 1992). While Underwood (1991) suggests that this actually may not involve additional costs or workload, it is more likely that in order to maintain adequate power, the effort invested in such an approach will be several times that of a standard design. Consider the differences between a temporally replicated (e.g. 5 periods before and after impact) standard BACIPS approach, an equally replicated asymmetrical ANOVA with three controls, and a similar, but nested asymmetrical ANOVA. Let us assume that this degree of replication provides sufficient power to detect a biologically important impact. In each sampling period, six benthic samples are taken in each location to control for random error variation. Thus, the number of samples needed in a BACIPS design (2 x 2 x 5 x 6 = 120, Fig. 3A) will be half of a standard asymmetrical design with equal temporal replication but three control areas (4 x 2 x 5 x 6 = 240, Fig. 3B). Subsampling within periods in a nested approach would multiply the number of samples by the times subsampled in each location (e.g. Fig. 3C, three times nested within a period: 3 x 4 x 2 x 5 x 6 = 720!). Because of the manifold increase in sampling effort, it is highly probable that the nested approach, despite its potentially superior ability to discern patterns in temporal variation, will be rarely used in an EIA. Note, however, that in some cases it may be the only viable option in an EIA, especially when populations of endangered species are being monitored.

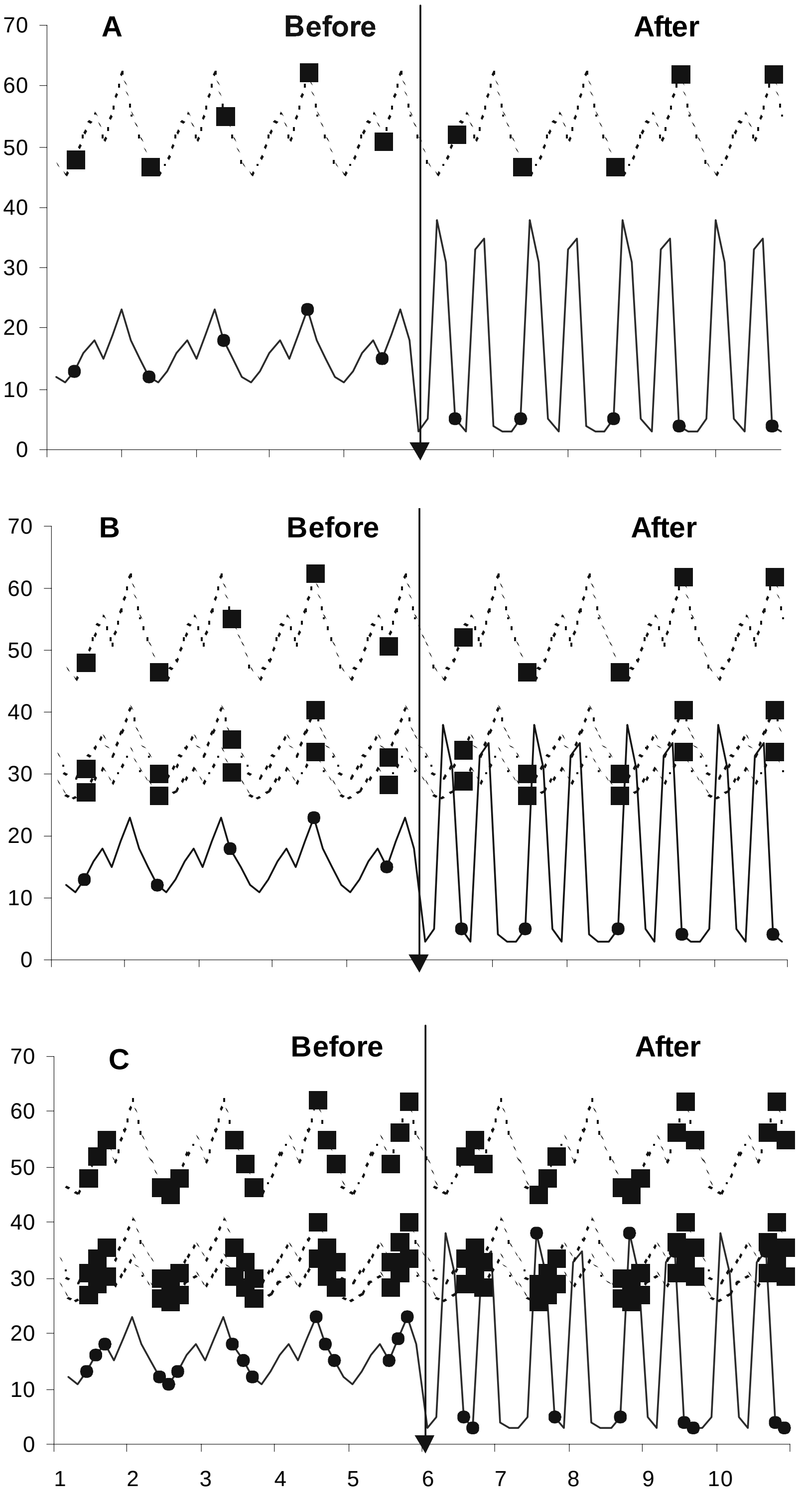

Fig 3. Example of an impact that increases variability around the mean abundance. Sampling intensity for BACIPS and asymmetrical variance designs are shown. Design (A) represents a typical BACIPS design with one impacted area (solid line) and a single control (broken line). Circles and squares represent sampling dates. Multiple control locations in an asymmetrical ANOVA (B) double the amount of sampling, while temporal nesting in design (C) necessitates a 6-fold sampling effort compared to (A). Note that in this case, design (C) is the only one capable of identifying the true nature of the impact (see text for details).

The asymmetrical design is in many ways more efficient in assigning environmental impacts to their cause than the BACIPS approach, but as with any field assessment, there are some problems associated with its use. In order to be efficient, the asymmetrical design must have sufficient power to conclude that an impact actually did occur. In an asymmetrical ANOVA design, power will always be smaller when compared to a balanced design, and whenever possible, balanced designs should be preferred. Calculation of power involves prior consideration of what level of change of abundance is considered to be biologically important (Peterman 1990). As power is related to variation, knowledge of natural variation in the abundances of target species is especially important (e.g. Osenberg et al. 1994), since this will affect the numbers of replicate samples needed to detect an impact and thus also the choice of species to be monitored (Underwood 1992). While Underwood (1992), using simulated data sets, achieved adequate power to detect major impacts with 4-5 non-nested temporal replicates, it is not clear whether this degree of replication is generally sufficient. In some cases, post-hoc pooling of variance components can be used to improve power of tests, but this procedure increases the risk of committing a type II error (see Underwood 1981).

While analysis of variance is quite robust to violations against normality and heterogeneity of variances, violations against independence can cause serious problems (Underwood 1992). The same precautionary measures as with the BACIPS approach (see above) apply. Compared to other approaches (e.g. BACIPS), the asymmetrical design will generally be more costly and labor-intensive. The additional cost may be outweighed by the gained benefit of being able to exclusively attribute an impact to a specific cause (Underwood 1991). This may be especially important in cases of environmental accidents where responsibility is often disputed and ambiguous, and inconclusive results may lead to expensive lawsuits (Underwood 1992).

3.4 The use of random time series to substitute for missing pre-treatment data

Investigators involved in EIA are often faced with the problem of lack of appropriate Before data (e.g. in the case of an environmental accident). Obviously, an impact cannot be unambiguously discerned based solely on post-impact data, even if asymmetrical designs are used (e.g. Glasby 1997). Assigning strict causality between a perceived impact and a probable cause may, however, sometimes be necessary to establish legal responsibility. In the absence of proper Before data, time series from randomly sampled, **undisturbed key habitats** (i.e. habitats similar to the one being subject to an impact) can serve as Before data in an asymmetrical design (Underwood 1994). Because human developments very likely disturb certain stream habitats more than others, and because time and resources for proper monitoring before an impact may be limited, there clearly is a need for such a time series from target habitats in multiple undisturbed streams (Underwood 1996). Basically, the set of streams to be sampled for the purpose of such a series should be randomly selected (i.e. chosen as a "random factor") from the entire population of similar streams (see Underwood 1992). As a consequence of the theory of unbiased representative sampling (e.g. Feller 1968), variation in the mean of a target parameter among the randomly selected streams provides an estimate of the variance in the entire population of similar streams. Furthermore, as differences among locations are seldom constant in time, sampling should be repeated at several time scales to estimate temporal variance (see Underwood 1996 for details). These estimates are representative of any set of streams chosen from the entire population of similar streams, and thus could be used as a substitute for the missing Before data. To assure reliability of fit between the general and the observed data from impacted sites, some control locations should be sampled in the after period and compared to the general data (Underwood 1996). If sampled properly, data from the randomly chosen set can then be used to substitute missing Before data from the Impact location, and may also serve as data from the Control site(s).

4 Case studies of biomonitoring in streams

This chapter describes some examples of approaches to monitoring situations in stream environments. Experimental designs used in these examples do not conform to the more efficient designs described in the previous section; unfortunately, no such data seem to be available for any lotic environment. Nonetheless, in the absence of ideal data, a combination of multiple approaches may still enable the investigator to explain the observed patterns. This chapter is by no means an exhaustive summary of stream biomonitoring attempts. The examples were mainly chosen to demonstrate certain strengths and inadequacies of published monitoring studies. First, section 4.1 demonstrates the need of basic research into system dynamics at multiple spatial scales to be able to identify the underlying mechanisms and predict system responses to catastrophic events. Second, to identify the effects of stream habitat modification, relatively long time series are needed (sections 4.2 & 4.3). The last example (section 4.3) stresses the importance of understanding inter-ecosystem linkages in biomonitoring studies.

4.1 Large-scale pathogen outbreaks test predictions of small-scale experiments

The herbivorous caddisfly *Glossosoma nigrior* is known to be a strong interactor in benthic communities of stony-bottomed trout streams in Michigan (e.g. McAuliffe 1984). Kohler (1992) assessed the strength of direct competitive interactions between *Glossosoma* and *Baetis* mayfly nymphs in well replicated (n=6), short-term (1 d), small-scale (115 cm^2) laboratory experiments. These experiments showed that the growth of another grazer (*Baetis*) was negatively affected by the presence of *Glossosoma,* and vice versa (Fig. 4).

Both grazers significantly depressed the amount of periphyton in the laboratory experiments. To confirm the results from laboratory studies, Kohler (1992) conducted a replicated small-scale field experiment. As *Glossosoma* is a case bearing caddisfly, it is rarely found in the water column and was thus readily excluded from substrate patches by lifting them from the stream bottom (treatment) with reinforcing bars. Substrate patches allowing free colonization by *Glossosoma* served as a control (Kohler 1992). A substrate patch consisted of six tiles, which had been left in the stream for one year. Control and treatment patches were paired to form a block, and three such blocks were placed into a riffle section of a brook. The field experiment lasted for ten months and samples were collected at about monthly intervals. The duration of the experiment spanned two generations of both *Glossosoma* and *Baetis*. Repeated measures ANOVA was used to analyze the data. In addition, several smaller field experiments were conducted to examine colonization by chironomids, and the effects of interference competition from *Glossosoma*.

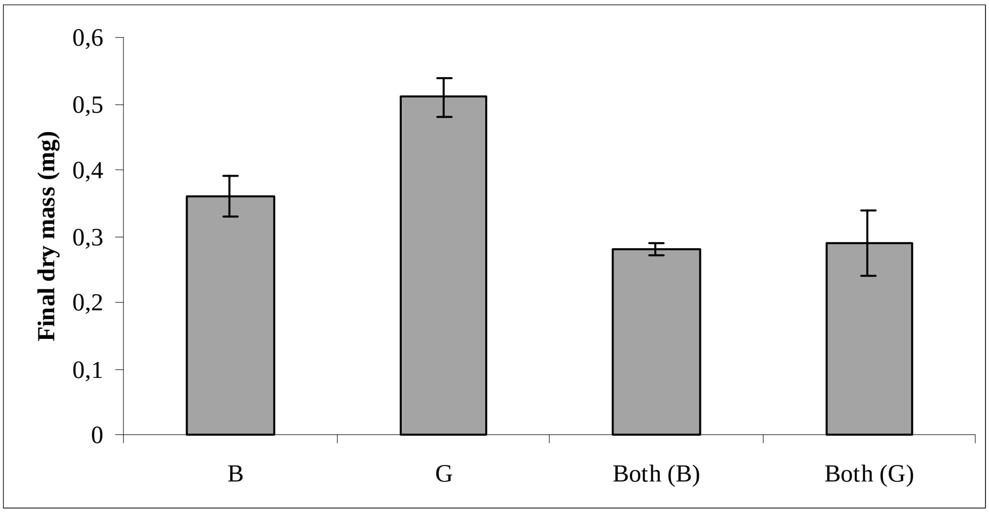

Fig. 4. Dry mass of Baetis and Glossosoma larvae at the end of the laboratory experiment. Capital letters represent treatments (mean, ± 1 SE) containing only Baetis (B), only Glossosoma (G) or both species (modified from Kohler1992).

The main result was that periphyton biomass was two times higher in the *Glossosoma* exclusion treatment than in the control (Fig. 5). Further, grazer/collector-gatherer species were found to be more abundant in the exclusions. In addition, the colonization experiment showed that chironomids grew faster in the absence of *Glossosoma*.

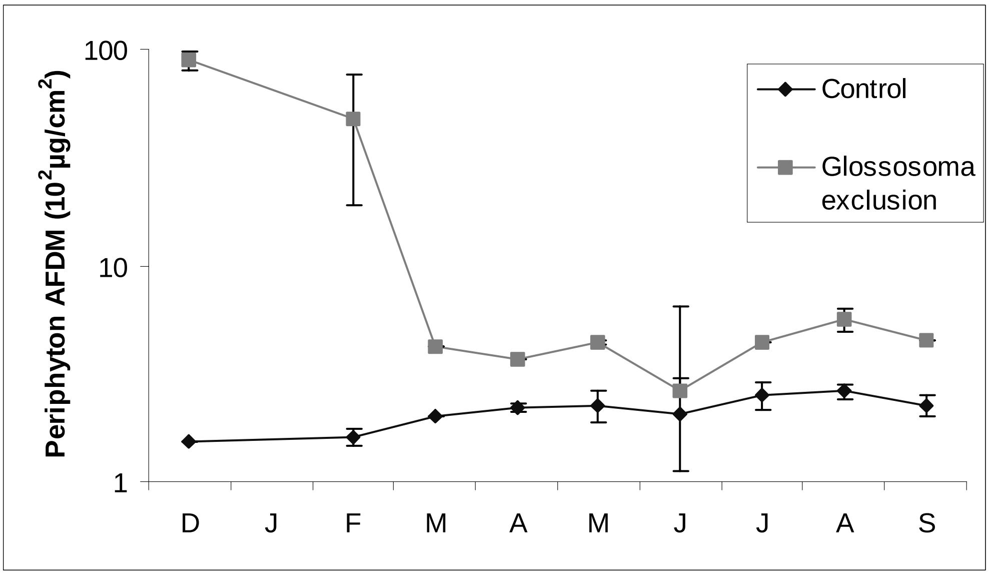

Fig. 5. Ash free dry mass (AFDM) (mean, ± 1 SE) of periphyton in a field experiment in response to Glossosoma exclusion (modified from Kohler 1992).

The results of the laboratory study were thus generally corroborated by the small-scale field experiments. However, in the field experiment, *Baetis* was unable to suppress periphyton levels in the absence of *Glossosoma* (Fig. 5). This discrepancy probably resulted from artifactual laboratory conditions, where *Baetis* was forced to spend more time on food patches than they would have in the field. Kohler (1992) concluded that the competitive nature of the relation between *Glossosoma* and *Baetis* tended to be overestimated by laboratory studies. Small-scale field experiments suggested that many taxa would increase in abundance if the absence of *Glossosoma* were maintained over larger spatial and longer temporal scales. Note that Kohler (1992) was only able to identify an artifactual laboratory result (depression of periphyton levels by *Baetis*) through simultaneous laboratory and field experiments.

Although a fine example of the strength of experimental work in ecology, the field results might still be representative only of the scale of observation used. Thus, it was not clear from these experiments whether the effects of *Glossosoma* exclusion would translate into larger scales of observation (i.e. whole stream reaches).

One year after these experiments, Kohler & Wiley (1992) noted outbreaks of a microsporidian parasite (*Cougourdella*) of *Glossosoma* in many streams. Formerly high population densities of *Glossosoma* (>2500 individuals/m^2) suddenly crashed to abundances of less than ten individuals/m^2, with a concurrent increase in the proportion of *Glossosoma* infected with *Cougourdella*. In the light of similar results in at least 20 streams in northern and Southwest Michigan, infections caused by *Cougourdella* seemed the only plausible cause for the dramatic reductions of *Glossosoma* abundances. The situation was further monitored in six streams for which before-outbreak data on both periphyton and invertebrate abundances were available. Differences in periphyton and invertebrate densities before and after outbreak were tested with one-tailed paired t-tests with streams being the units of replication. One-tailed tests were used because the results of the small-scale experiments indicated that periphyton abundances and invertebrate densities should *increase* following *Glossosoma* collapse (for details, see Kohler & Wiley 1997). On average, *Cougourdella* outbreaks coincided with a 25-fold reduction in *Glossosoma* densities. *Glossosoma* collapse was reflected as a significant increase in periphyton biomass (Fig. 6) and a 2-5-fold increase in grazer and filter-feeder densities. Some previously rare species showed dramatic increases in population size, typically with a lag of at least one year following the *Glossosoma* collapse.

To compare the predictions of the small-scale experiments conducted in Spring Brook to patterns observed during the large-scale survey, two approaches were used. First, Kohler & Wiley (1997) assessed whether the small-scale experiment correctly predicted the responses of invertebrate populations and periphyton standing crop in Spring Brook. Effect sizes from the small-scale experiments and the whole stream perturbation were calculated for each taxon using a procedure described by Sarnelle (1997). If the small-scale experiments were to correctly predict large-scale effects of the *Glossosoma* collapse, effect sizes for both scales should be positively correlated with a slope of 1. Kohler and Wiley (1997) tested this prediction using linear regression. Second, to estimate whether the results from the small-scale experiments could be extrapolated beyond Spring Brook, effect size correlations were also calculated for data on all six streams monitored. The small-scale experiment predicted the effects on the large scale rather poorly for both Spring Brook and for the six stream survey data. The reason for the poor fit between the small- and the large-scale results was that, rather unexpectedly, small-scale experiments tended to *underestimate* the effects (i.e. magnitude) of *Glossosoma* population collapse (Kohler & Wiley 1997).

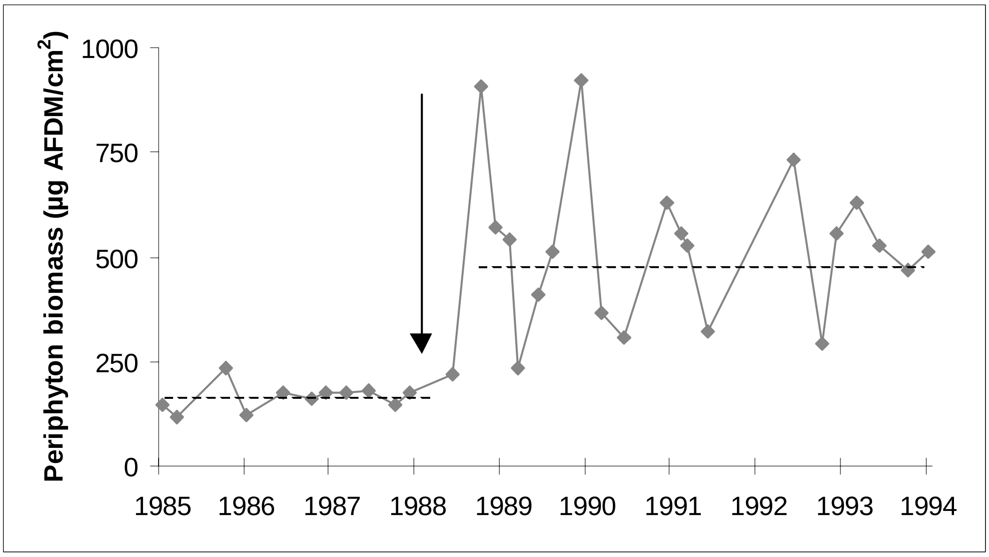

Fig. 6. Periphyton biomass in Spring brook, Michigan. The arrow indicates time of Glossosoma collapse. Horizontal dashed lines are the overall mean periphyton biomass for the pre- and post-collapse periods (redrawn and modified from Kohler & Wiley 1997).

In conclusion, this example shows the value of the combination of laboratory and field experiments in identifying the mechanisms underlying field patterns. Using this combination, Kohler (1992) was able to identify a laboratory artifact (*Baetis* is not as efficient a grazer as laboratory results suggested). Although the direction of the results and predictions of small-scale field experiments were generally mirrored by the large-scale survey data (i.e. increases in grazer and filter-feeder densities, increases in periphyton biomass), the magnitude of these effects was not correctly predicted. Thus, while small-scale studies served well to identify specific mechanisms underlying ecosystem responses, studies on larger temporal and spatial scales were needed to accurately quantify population- and community-level consequences of species interactions (Kohler & Wiley 1997). Note, however, that the large-scale survey lacked proper control since all the streams studied were affected by *Cougourdella* outbreaks. Thus, strictly speaking, Kohler & Wiley (1997) were unable to exclusively attribute the system level responses to the collapse of *Glossosoma*.

4.2 Effects of habitat enhancement on trout populations

Although stream habitats are frequently modified to increase the number and size of game fish, the effects of such alterations have rarely been rigorously tested. In a manipulative long-term study, Gowan & Fausch (1996) assessed the effect of log drop structures on stream habitat features, and on abundance, biomass, growth, survival and movement of fish in six Colorado trout streams. The experiment spanned eight years, divided in a two-year pre-treatment period and six years of post-impact monitoring. In each stream, a 500 m reach lacking large woody debris and pools was chosen as the experimental section. Each section was divided into an upper and a lower part, and treatments were randomly assigned among the subsections. Treatments consisted of installing ten log drop structures (see Riley & Fausch (1995) for details), while controls were left unmodified. To assess changes in habitat features, Gowan and Fausch (1996) measured several habitat variables along permanent transects. Differences between the treatment and the control section in overhead cover, wetted area and pool area were tested using repeated measures split-plot analysis of variance. Streams were used as replicates, stream sections as the whole-plot factor (control vs. treatment) and time as the subplot factor (pre-treatment vs. post-treatment). No significant differences in habitat structure existed between the control and the treatment sections in the pre-manipulation period. However, following the installation of enhancement structures, pool volume, total cover, percentage of fine substrate and mean depth were significantly greater in the treatment sections.

To assess fish responses to habitat manipulations, trout were electrofished and marked to indicate the section they were originally caught in. Fish were divided into two age classes, juvenile (1-year-old) and adult (2-year-old or older fish) fish, and biomass for each species and age-class was estimated for each section. Differences in salmonid abundance between the sections in the before period were tested using the same repeated measures ANOVA design as for the physical variables. A repeated measures multivariate analysis of variance (MANOVA) was used to test for differences between the control and the treatment sections in the six-year after-period, because population samples taken in successive years were likely to be autocorrelated. A significant stream section MANOVA term would indicate that log drop structures induced a change in a target variable.

Trout abundance and biomass did not differ between the sections in the pre-treatment data. Habitat manipulation increased adult but not juvenile trout abundance (Fig. 7) and biomass in the post-treatment period across all streams. Gowan and Fausch (1996) named four mechanisms that could have caused the changes observed in adult trout populations: (i) increased recruitment of juveniles, (ii) increased adult survival, (iii) enhanced growth, or (iv) net immigration. Juvenile recruitment, adult survival and adult growth did not differ between the treatment and the control sections. Analysis of movements of marked fish indicated that immigration into the treatment and the control sections was high: however, immigrants into the treatment sections seldom originated from the control sections or vice versa. During the initial post-treatment period, immigration into the treatment sections was higher than into the controls. Gowan & Fausch (1996) concluded that the increased biomass and abundance of adult trout in the treatment sections was mainly related to fish immigrating from outside the study reach.

Although the cause for the increase in adult trout biomass and abundance in the treatment sections was apparently correctly identified, this experiment bears three potentially serious problems: (i) non-independent experimental units, (ii) potentially autocorrelated data and (iii) unbalanced temporal design. First, the use of adjacent upstream and downstream areas may be problematic because of non-independence. This stems from the fact that changes in the treatment area could affect the control area, leading to non-independence of data. Changes in the control section may not be independent of those in the treatment section and tests of treatment effects are therefore invalid. Further, changes in the physical nature of an upstream treatment section could have created spillover effects on the downstream control. Gowan & Fausch (1996) reported that movement between the sections was low and probably did not cause any serious bias to the results, but this does not fully remove the problem of interdependence among the data. Rather than using adjacent controls, treatments should have been assigned to separate streams or, if this were not possible, control and treatment sections should have been widely spaced within a single stream (see e.g. Underwood 1996). A further problem is created when (as in this study) treatment and control sections are randomly assigned to upstream vs. downstream locations, since this causes individual replicates to be heterogeneous, due to a potential position effect. Second, although Gowan & Fausch (1996) noted that their fish samples were potentially autocorrelated, this was not tested (e.g. by using a Durbin-Watson test). The use of autocorrelated data violates assumptions of most ANOVA-designs and typically increases the risk of committing a type I error. The preemptive use of MANOVA on possibly autocorrelated data without prior testing seems unwarranted. Third, although Gowan and Fausch (1996) had both Before and After data for many fish parameters, standard BACIPS was not readily applicable due to the unbalanced temporal nature of the design. However, if assumptions for BACIPS in Gowan & Fausch's (1996) trout data were met (which is unlikely to be the case), one way to use individual Before data points would have been to calculate deltas (see chapter 3) for individual streams. Deltas for the Before and After period could have then been analyzed using unbalanced one-way ANOVA (Shaw & Mitchell-Olds 1993) with stream (n=6) as a blocking factor.

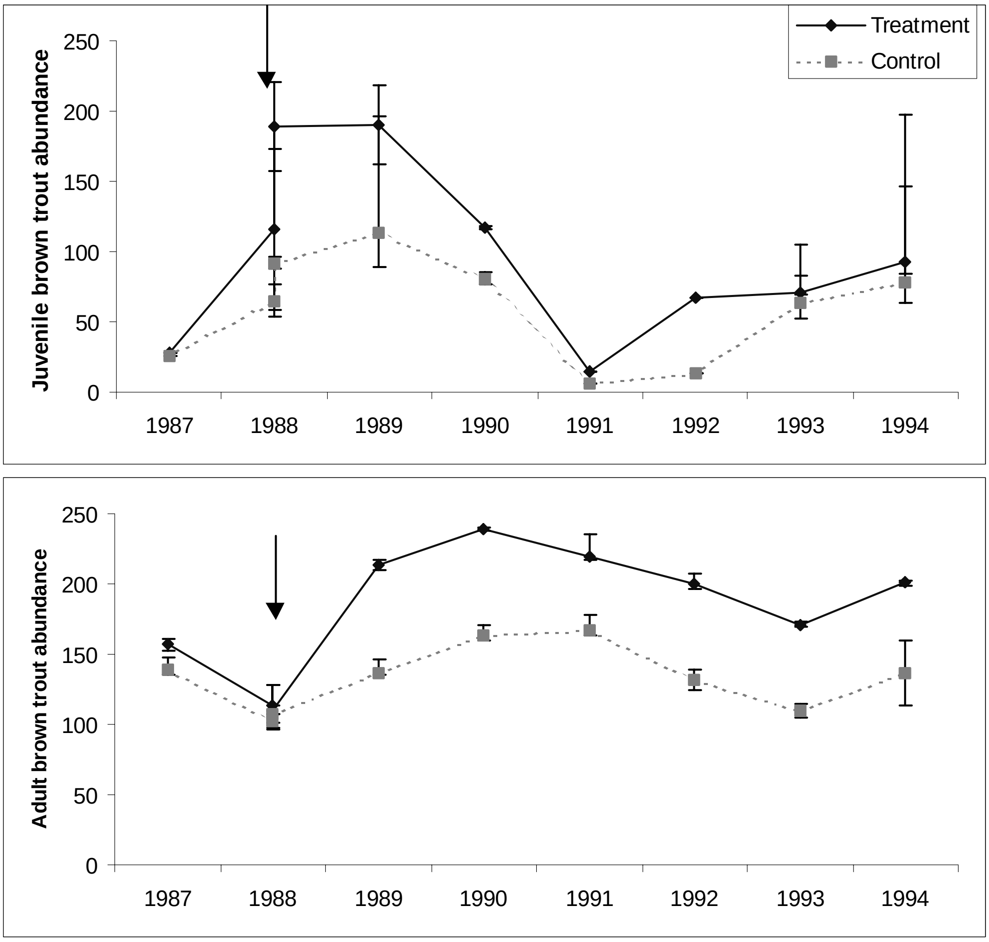

Fig. 7. Juvenile and adult brown trout abundance in treatment and control sections (no./250 m section). Vertical bars are 95% profile likelihood confidence intervals. Arrows depict the time of the habitat manipulation (redrawn and modified from Gowan & Fausch 1996).

Despite obvious design flaws, this experimental assessment of the value of habitat enhancement is one of the very few long-term studies published on this issue. This makes the rather inefficient use of the Before data especially regrettable. While the inadequacies of the design are partly outweighed by strong treatment responses in this particular study, in many other cases this might have resulted in ambiguous results and serious problems of interpretation.

In conclusion, increases in adult trout biomass and abundance in the treatment sections was not linked to juvenile abundance, improved survival or increased growth rates. Data on marked fish revealed that the increase in adult trout abundance resulted from between habitat movement. Curiously, fish that moved into the treatment sections generally did not originate from the adjacent control areas, indicating that a considerable proportion of trout may move over relatively long distances. Although no direct responses in survival were found at the scale of the experiment, it is likely that survival of adult trout on larger scales was positively affected by the habitat manipulation. As long distance movement of trout is probably related to search of suitable habitat, the absence of such habitat could have resulted in death of moving individuals.

4.3 The importance of terrestrial litter input to stream ecosystem functioning

In a study on the effects of leaf litter exclusion to a headwater stream, Wallace et al. (1997, 1999) provided an excellent example of the importance of understanding the functional linkages between different ecosystems. Wallace et al. (1997) monitored a control and a future impact stream one year before and four years after a leaf litter exclusion experiment. Leaf litter was excluded by using an exclusion canopy made of gill netting and lateral fences. The canopy was constructed across the bankful channel width and spanned a 180 m long stream section in the treatment stream, while the control stream remained unmanipulated. The objectives of the study were to assess the impact of the treatment (leaf litter exclusion) on (i) stream organic matter inputs and standing crops; (ii) benthic animal abundances and biomasses; and (iii) secondary production. Macroinvertebrate abundances and biomasses were estimated from replicate benthic samples collected at monthly intervals from two distinct habitats in each stream: moss-covered bedrock vs. mixed (cobbles, gravel and sand) substrates. Randomized intervention analysis (RIA) was used to test for changes caused by the manipulation. RIA indicated a significant difference between the treatment and the control stream in leaf input: most (>94 %) of the litter input into the treatment stream was excluded through the use of the exclusion canopy and lateral fences. Similarly, RIA showed that abundances and biomass of total invertebrates, as well as those of shredders and predators, decreased significantly in the mixed substrate habitat in response to litter exclusion. Thus, the majority of taxa (58 %), responsible for 93 to 97 % of production in this habitat type, showed significant decrease in response to the treatment. Secondary production was also negatively affected by the treatment in the mixed substrate habitat. Predators in particular seemed to be affected by a decline in their resources, leading to a 76 % reduction in their secondary production (Wallace et al. 1999). Wallace et al. (1997) concluded that forest streams are subsidized, strongly donor-controlled systems (i.e. dependent on allochthonous inputs). In the bedrock substrate, by contrast, the abundance and biomass of invertebrates and secondary productivity were unaffected by the treatment. This seems curious since the two habitat types were in close proximity. The lack of response in the moss-covered bedrock habitat was caused by a different structure of the benthic community in this habitat type. The bedrock community was dominated by filterers, gatherers and predators and was thus more dependent on fine particulate organic matter (FPOM) trapped within mosses than on coarse particulate organic matter (CPOM). Levels of FPOM were unaffected during the first two years of the experiment and only slowly decreased during the last two years. Wallace et al. (1999) anticipated a steady decrease of accumulated FPOM if the experiment were to be continued, eventually resulting in similar changes of abundance and biomass in the bedrock habitat as already observed for the mixed habitat.

In conclusion, these results demonstrate the importance of inter-ecosystem subsidies in detrital-based streams. This study shows the potential of strong bottom-up regulation in forested headwater streams. Note, however, that RIA cannot exclusively link the observed changes to the experimental manipulation. In light of the magnitude of the observed change and the absence of plausible alternative explanations, it is likely that the litter exclusion caused the observed patterns. Thus the implication of this study is that severe disruptions in the flow of CPOM from the riparian zone to the stream may indeed result in profound changes in benthic community structure. While there is some evidence suggesting that similar changes might result from other factors causing degradation of the riparian zones of streams (e.g. Stone & Wallace 1998), this prediction has not been rigorously tested yet.

5 The choice of appropriate target variables

While many factors influence the choice of target variable(s) (spatial and temporal scale, costs, etc.), the choice will ultimately depend on the managerial question behind monitoring. Funding for biomonitoring is limited and thus investigators often have to select certain indicator variables to represent a greater set of variables of interest. Depending on the aim of monitoring, this will pose varying demands on the "ideal" properties that such indicators should possess (Jones & Kaly 1996). Although many authors have listed desirable qualities for indicator species (longevity, sedentary life style, stable populations, etc.), traits that are beneficial in one study may prove disadvantageous in another assessment situation. Let us assume that we are interested in the overall long-term effects of a pollutant on a stream benthic community. If we chose to monitor only the most sensitive species of the community, effects of the pollutant will likely be overestimated. That is, while the short-term extinction of a sensitive species may be caused by the release of the pollutant, this does not necessarily mean that the rest of the community will be affected on a long-term basis (Jones & Kaly 1996). Rather than concentrating on the most sensitive species, focus on certain abundant key taxa will likely provide the most accurate answer to the specific questions addressed in a study.

Abundant species are commonly used as target variables in monitoring, mainly because fewer samples are needed to adequately estimate their sample means (Morin 1985) and less replicates are needed to demonstrate an effect in quantitative studies (Cooper & Barmuta 1993). However, there are ways to overcome the difficulties of monitoring rare species.

Example: In a long-term study on the effects of global warming on arctic plant populations, Lesica & Steele (1996) used a study design based on temporal resampling of permanent plots, thus deviating from a fully random sampling. Subsequently a modified repeated-measures ANOVA model was used that accommodated for the effects of high frequency variation and allowed an assessment of the significance of long-term trends (see Lesica & Steele 1996 for details). Thus, although some general rules may apply, finding appropriate indicators and being able to accurately predict the effects of future impacts will ultimately be related to clearly stated monitoring aims, sound prior knowledge of the system dynamics and of the reliability of the indicators. In the absence of such knowledge, literature reviews and theoretical considerations should guide the *a priori* choice of indicator taxa.

6 Final remarks and recommendations

Although investigators ultimately will be bound by the managerial aims and by the funding available, there is no reason to address monitoring and environmental questions with less precision and rigor than those of "academic" experimental ecology (Underwood 1996). While the focus in this report has been on hypotheses testing, this should not be used as a substitute for defining the biological importance and effect size of impacts. The key to high-standard biomonitoring lies in converting the managerial aim into testable hypotheses, choosing the appropriate scale and study design, using the best-suited organisms as the target species, and defining the magnitude and biological importance of an impact. This involves, as with any ecological research, working on several spatial and temporal scales, and constantly realigning results with predictions and general ecological theory (Werner 1998).

While the fact that one cannot draw conclusions about the state of nature and predict human impacts in the absence of proper data has generally been recognized, the lack of such data is an often-faced reality in biomonitoring (Christensen et al. 1996, Treweek 1996). In particular, the scarcity of pre-impact data constitutes a major problem to the applicability of the field assessment designs outlined above. There are usually no legal requirements to monitor streams subject to future human developments prior to the onset of the development. Yet, surveys lacking pre-impact data cannot demonstrate *any* change in stream communities and thus are a complete waste of money and effort. As has been shown, this problem can be overcome if random time series from a set of similar streams are available (see section 3.3, Underwood 1994). In Finland, there is a great and urgent need for such time series from headwater streams, as these are most prone to adverse impact from human developments. Clearly, such long-term series are to be created on a nationwide basis, with emphasis on ecoregions where streams are most likely to be influenced by human developments in the future. Within each region, a minimum of 15-20 streams should be *randomly* chosen for monitoring. This sample size is required to ascertain that a sufficient range of habitats will be included. In total, about 80 streams should be included in such a nationwide benthic biomonitoring program. In order to provide adequate precision for the estimates, individual streams must be randomly sampled three times/year (seasonally restricted), with a minimum of five replicate samples per sampling date. Note that in order to provide information about the variation of estimates within each stream, the replicate samples taken on each sampling date *must not* be pooled. This series should be carried out for at least 6-8 years to provide baseline information about interannual variation in benthic abundances. The establishment of this series thus necessitates taking ca. 300 samples/region/year. While this may seem like an enormous investment of resources, such a long-term data set could substitute for the lack of Before and Control data in future EIA situations. Clearly, this would reduce tremendously any future investment of resources in field assessments, providing investigators with a tool to link impacts to their cause(s) without the need for specific-data in each individual case.

References

Bence, J. R., Stewart-Oaten, A. and Schroeter, S. C. (1996). Estimating the size of an effect from a before-after-control-impact paired series design: The predictive approach applied to a power plant study. In Detecting ecological impacts: Concepts and applications in coastal habitats (ed. R. J. Schmitt and C. W. Osenberg), pp. 133-148. San Diego: Academic Press.

Bender, E. A., Case, T. J. and Gilpin, M. E. (1984). Perturbation experiments in community ecology: Theory and practice. Ecology 65: 1-13.

Cairns, J. J. (1983). Are single species toxicity tests alone adequate for estimating environmental hazard? Hydrobiologia 100: 47-57.

Carpenter, S. R. (1989). Replication and treatment strength in whole-lake experiments. Ecology 70: 453-463.

Carpenter, S. R. (1990). Large-scale perturbations: Opportunities for innovation. Ecology 71: 2038-2043.

Carpenter, S. R. and Kitchell, J. F. (1988). Consumer control of lake productivity. Bioscience 38: 764-769.

Carpenter, S. R., Frost, T. M., Heisey, D. and Kratz, T. K. (1989). Randomized intervention analysis and the interpretation of whole-ecosystem experiments. Ecology 70: 1142-1152.

Christensen, N. L., Bartuska, A. M., Brown, J. H., Carpenter, S., D'Antonio, C., Francis, R., Franklin, J. F., MacMahon, J. A., Noss, R. F., Parsons, D. J., Peterson, C. H., Turner, M. G. and Woodmansee, R. G. (1996). The report of the Ecological Society of America Committee on the Scientific Basis for Ecosystem Management. Ecological Applications 6: 665-691.

Clements, W. H., Farris, J. L., Cherry, D. S. and Cairns, J., Jr. (1989). The influence of water quality on macroinvertebrate community responses to copper in outdoor experimental streams. Aquatic Toxicology 14: 249-262.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences. Hillsdale, NJ: L. Erlbaum Associates.

Cooper, S. D. and Barmuta, L. A. (1993). Field experiments in biomonitoring. In Freshwater Biomonitoring and benthic macroinvertebrates (ed. D. M. Rosenberg and V. H. Resh).

Dutilleul, P. (1993). Spatial heterogeneity and the design of ecological field experiments. Ecology 74: 1646-1658.

Eberhardt, L. L. and Thomas, J. M. (1991). Designing environmental field studies. Ecological Monographs 61: 53-73.

Englund, G. and Olsson, T. (1996). Treatment effects in a stream fish enclosure experiment: Influence of predation rate and prey movements. Oikos 77: 519-528.

Fairweather, P. G. (1991). Statistical power and design requirements for environmental monitoring. Aust. J. Mar. Freshwat. Res. 42: 555-567.

Feller, W. (1968). An introduction to probability theory and its applications. New York: Wiley.

Frost, T. M., DeAngelis, D. L., Bartell, S. M., Hall, D. J. and Hurlbert, S. H. (1988). Scale in the design and interpretation of aquatic community research. In Complex interactions in Lake Communities (ed. S. R. Carpenter), pp. 229-258. New York: Springer-Verlag.

Glasby, T. M. (1997). Analysing data from post-impact studies using asymmetrical analyses of variance: A case study of epibiota on marinas. Australian journal of ecology 22: 448-459.

Goldstein, R. (1989). Power and sample size via MS/PC-DOS computers. The American Statistician 43: 253-260.

Gowan, C. and Fausch, K. D. (1996). Long-term demographic responses of trout populations to habitat manipulation in six Colorado streams. Ecological Applications 6: 931-946.

Green, R. H. (1979). Sampling design and statistical methods for environmental biologists. New York: John Wiley.

Hurlbert, S. H. (1984). Pseudoreplication and the design of ecological field experiments. Ecological Monographs 54: 187-211.

Jones, G. P. and Kaly, U. L. (1996). Criteria for selecting marine organisms in biomonitoring studies. In Detecting ecological impacts: Concepts and applications in coastal habitats (ed. R. J. Schmitt and C. W. Osenberg), pp. 29-45. San Diego: Academic Press.

Kiffney, P. M. and Clements, W. H. (1994). Effects of heavy metals on a macroinvertebrate assemblage from a Rocky Mountain stream in experimental microcosms. Journal of the North American Benthological Society 13: 511-523.

Kiffney, P. M. and Clements, W. H. (1996a). Effects of metals on stream macroinvertebrate assemblages from different altitudes. Ecological Applications 6: 472-481.

Kiffney, P. M. and Clements, W. H. (1996b). Size-dependent response of macroinvertebrates to metals in experimental streams. Environmental Toxicology and Chemistry 15: 1352-1356.

Kohler, S. L. (1992). Competition and the structure of a benthic stream community. Ecological Monographs 62: 165-168.

Kohler, S. L. and Wiley, M. J. (1997). Pathogen outbreaks reveal large-scale effects of competition in stream communities. Ecology 78: 2164-2176.

Korman, J. and Higgins, P. S. (1997). Utility of escapement time series data for monitoring the response of salmon populations to habitat alteration. Canadian Journal of Fisheries and Aquatic Sciences 54: 2058-2067.

Legendre, P. (1993). Spatial autocorrelation: Trouble or new paradigm? Ecology 74: 1659-1673.

Levins, R. (1968). Evolution in changing environments. Monographs in Population Biology 2. Princeton, Princeton University Press.

Lesica, P. and Steele, B. M. (1996). A method for monitoring long-term population trends: An example using rare arctic-alpine plants. Ecological Applications 6: 879-887.

Levin, S. A. (1992). The problem of pattern and scale in ecology. Ecology 73: 1943-1967.

Magnuson, J. J., Benson, B. J. and Kratz, T. K. (1990). Temporal coherence in the limnology of a suite of lakes in Wisconsin, U.S.A. Freshwater biology 23: 145-159.

Mapstone, B. D. (1995). Scalable decision rules for environmental impact studies: Effect size, Type I, and Type II errors. Ecological Applications 5: 401-410.

McAuliffe, J. R. (1984). Resource depression by a stream herbivore: Effects on distributions and abundances of other grazers. Oikos 42: 327-333.

Mittelbach, G. G., Turner, A. M., Hall, D. J., Rettig, J. E. and Osenberg, C. W. (1995). Perturbation and resilience: A long-term, whole-lake study of predator extinction and reintroduction. Ecology 76: 2347-2360.

Morin, A. (1985). Variability of density estimates and the optimization of sampling programs for stream benthos. Canadian Journal of Fisheries and Aquatic Sciences 42: 1530-1534.

Mäki-Petäys, A., Vehanen, T., Huusko, A. and Muotka, T. (1999). Virtavesien kunnostuksen arviointi ja seuranta. Suomen kalastuslehti 7: 8-11.

Osenberg, C. W., Schmitt, R. J., Holbrook, S. J., Abu-Saba, K. E. and Flegal, A. R. (1994). Detection of environmental impacts: Natural variability, effect size, and power analysis. Ecological Applications 4: 16-30.

Peterman, R. M. (1990). Statistical power analysis can improve fisheries research and management. Canadian Journal of Fisheries and Aquatic Sciences 47: 2-15.

Petersen, J. E., Cornwell, J. C. and Kemp, W. M. (1999). Implicit scaling in the design of experimental aquatic ecosystems. Oikos 85: 3-18.

Pontasch, K. W., Niederlehner, B. R. and Cairns, J., Jr. (1989). Comparisons of single-species, microcosm and field responses to a complex effluent. Environmental Toxicology and Chemistry 8: 521-532.

Potwin, C. (1993). ANOVA: Experiments in controlled environments. In Design and analysis of ecological experiments (ed. S. M. Scheiner and J. Gurevitch), pp. 46-68. New York: Chapman & Hall.

Riley, S. C. and Fausch, K. D. (1995). Trout population response to habitat enhancement in six northern Colorado streams. Canadian Journal of Fisheries and Aquatic Sciences 52: 34-53.

Ross, Q. E., Dunning, D. J., Menezes, J. K., Kenna, M. J., Jr. and Tiller, G. (1996). Reducing impingement of alewives with high-frequency sound at a power plant intake on Lake Ontario. North American Journal of Fisheries Management 16: 548-559.

Sarnelle, O. (1997). Daphnia effects on microzooplankton: Comparisons of enclosure and whole-lake responses. Ecology 78: 913-928.

Shaw, R. G. and Mitchell-Olds, T. (1993). ANOVA for unbalanced data: An overview. Ecology 74: 1638-1645.

Simberloff, D. (1986). The proximate causes of extinction. In Patterns and processes in the history of life (ed. D. M. Raup and D. Jablonski), pp. 259-276. Berlin: Springer-Verlag.

Smith, E. P., Orvos, D. R. and Cairns, J., Jr. (1993). Impact assessment using the before-after-control-impact (BACI) model: Concerns and comments. Canadian Journal of Fisheries and Aquatic Sciences 50: 627-637.

Stewart-Oaten, A. (1996). Problems in the analysis of environmental monitoring data. In Detecting ecological impacts: Concepts and applications in coastal habitats (ed. R. J. Schmitt and C. W. Osenberg), pp. 109-130. San Diego: Academic Press.

Stewart-Oaten, A., Murdoch, W. W. and Parker, K. R. (1986). Environmental impact assessment: "Pseudoreplication" in time? Ecology 67: 929-940.

Stewart-Oaten, A., Bence, J. R. and Osenberg, C. W. (1992). Assessing effects of unreplicated perturbations: No simple solutions. Ecology 73: 1396-1404.

Stone, M. K. and Wallace, J. B. (1998). Long-term recovery of a mountain stream from clear-cut logging: The effects of forest succession on benthic invertebrate community structure. Freshwater Biology 39: 151-169.

Sweeney, B. W., Funk, D. H. and Standley, L. J. (1993). Use of the stream mayfly Cloeon triangulifer as a bioassay organism: Life history response and body burden following exposure to technical chlordane. Environmental Toxicology and Chemistry 12: 115-125.

Thrush, S. F., Pridmore, R. D. and Hewitt, J. E. (1994). Impacts on soft-sediment macrofauna: The effects of spatial variation on temporal trends. Ecological Applications 4: 31-41.

Thrush, S. F., Cummings, V. J., Dayton, P. K., Ford, R., Grant, J., Hewitt, J. E., Hines, A. H., Lawrie, S. M., Pridmore, R. D., Legendre, P., McArdle, B. H., Schneider, D. C., Turner, S. J., Whitlatch, R. B., Wilkinson, M. R. and et al. (1997). Matching the outcome of small-scale density manipulation experiments with larger scale patterns an example of bivalve adult/juvenile interactions. Journal of Experimental Marine Biology and Ecology 216: 153-169.

Thrush, S. F., Schneider, D. C., Legendre, P., Whitlatch, R. B., Dayton, P. K., Hewitt, J. E., Hines, A. H., Cummings, V. J., Lawrie, S. M., Grant, J., Pridmore, R. D., Turner, S. J. and McArdle, B. H. (1997b). Scaling-up from experiments to complex ecological systems: Where to next? Journal of Experimental Marine Biology and Ecology 216: 243-254.

Treweek, J. (1996). Ecology and environmental impact assessment. Journal of Applied Ecology 33: 191-199.

Underwood, A. J. (1981). Techniques of analysis of variance in experimental marine biology and ecology. Oceanography and Marine Biology: An Annual Review 19: 513-605.

Underwood, A. J. (1991). Beyond BACI: Experimental designs for detecting human environmental impacts on temporal variations in natural populations. Aust. J. Mar. Freshwat. Res. 42: 569-587.

Underwood, A. J. (1992). Beyond BACI: The detection of environmental impacts on populations in the real, but variable, world. Journal of Experimental Marine Biology and Ecology 161: 145-178.

Underwood, A. J. (1994). On beyond BACI: Sampling designs that might reliably detect environmental disturbances. Ecological Applications 4: 3-15.

Underwood, A. J. (1996). Spatial and temporal problems with monitoring. In River restoration (ed. G. Petts and P. Calow), pp. 182-204: Blackwell Science.

Wallace, J. B., Grubaugh, J. W. and Whiles, M. R. (1996). Biotic indices and stream ecosystem processes: Results from an experimental study. Ecological Applications 6: 140-151.

Wallace, J. B., Eggert, S. L., Meyer, J. L. and Webster, J. R. (1997). Multiple trophic levels of a forest stream linked to terrestrial litter inputs. Science 277: 102-104.

Wallace, J. B., Eggert, S. L., Meyer, J. L. and Webster, J. R. (1999). Effects of resource limitation on a detrital-based ecosystem.

Werner, E. E. (1998). Ecological experiments and a research program in community ecology. In Experimental Ecology: Issues and perspectives (ed. W. J. Resetarits and J. Bernando), pp. 3-26. New York: Oxford University Press.

Wiens, J. A. (1989). Spatial scaling in ecology. Functional Ecology 3: 385-397.

Wiens, J. A. and Parker, K. R. (1995). Analyzing the effects of accidental environmental impacts: Approaches and assumptions. Ecological Applications 5: 1069-1083.

Suomenkielinen lyhennelmä

Virtavesien seuranta on usein keskittynyt kuvaamaan, todentamaan ja seuraamaan ihmisen toiminnasta aiheutuneita vaikutuksia jokiluontoon. Seurannasta vastaavan tutkijan on pystyttävä muodostamaan testattavia hypoteeseja usein väljästi muotoilluista ympäristöjohtamistavoitteista kyetäkseen vastaamaan kysymyksiin jokisysteemien tilasta. Hypoteesien kehittäminen vaatii tutkijalta sekä hyvää tutkittavan systeemin tuntemusta että usein myös työskentelyä useilla lähestymistavoilla ja mittakaavoilla. Seurannassa käytettävät lähestymistavat voidaan karkeasti luokitella neljään luokkaan: kenttäkartoitukset (survey), laboratoriokokeet, kenttäkokeet ja kenttäarvioinnit. Kartoituksien käyttöä ainoana menetelmänä ihmistoiminnan vaikutuksien arvioinnissa olisi vältettävä, sillä niiden suuri heikkous on replikoimattomuus ja kyvyttömyys kiistatta todentaa syy-seuraus suhteita toiminnan ja havaitun vaikutuksen välillä (Cooper & Barmuta 1993). Vaikka toisaalta toistetuilla laboratoriokokeilla kyetään syy-seuraus suhteita kiistatta osoittamaan, niiden käyttö ainoana lähestymistapana seurannoissa ei myöskään ole ongelmaton, koska saatuja tuloksia ei voida suoraan yleistää suuremmille mittakaavoille. Tämä johtuu siitä, että pienen mittakaavan ilmiöt eivät yleensä ole lineaarisessa suhteessa isoimmilla mittakaavoilla tapahtuvien prosessien kanssa (esim. Thrush et al. 1997). Laboratoriokokeiden käyttöä seurannoissa saattaa vaikeuttaa lisäksi koeolojen liiankin suuri poikkeaminen luonnonoloista, mikä voi aiheuttaa keinotekoisia koetuloksia (Kohler 1992). Ympäristöön kohdistuvien vaikutuksien arvioimiseksi kentällä toteutettu kokeellinen lähestymistapa on edellä mainittuja menetelmiä voimakkaampi ja tarkempi. Kuten laboratoriokokeita, myös kenttäkokeita voidaan usein toistaa riittävästi, ja näin ollen syy-seuraus suhteita pystytään kiistatta osoittamaan. Kentällä tehtyjen kokeiden olosuhteet ovat todellisempia kuin laboratoriokokeissa, mutta käytetty mittakaava on useimmiten rajallinen. Vaikka kenttäkokeiden toteuttaminen suurilla mittakaavoilla on suositeltavaa, käytännössä riittävien toistojen saavuttaminen voi aiheuttaa suuria kustannuksia (Carpenter 1989). Lisäksi tietyissä tapauksissa voi olla eettisesti arveluttavaa toistaa tiettyä vaikutusta (esim. öljyonnettomuuden sattuessa) (Wiens & Parker 1995). Kenttäarvioinnissa (field assessment) käytetyt koeasetelmat pystyvät myös osoittamaan syy-seuraus suhteita. Kenttäarviointi vaatii toimiakseen sekä tietoa vaikutusta edeltäneestä että sen jälkeisestä tilasta (Stewart-Oaten et al. 1986). Vaikka ihmistoiminnan vaikutuksien arvioinnissa (environmental impact assessment) tieto kohde- ja kontrollisysteemien tilasta ennen vaikutusta on usein puutteellinen, oikein toteutettu kenttäarviointi on voimakkain ja suositeltavin työkalu ihmisvaikutuksien todentamiseksi (Underwood 1994).

Seurannoissa yleinen ongelma on erottaa kohdemuuttujissa tapahtunutta varsinaista muutosta ns. ”hälystä” eli havaitun vaikutukseen liittyvästä epävarmuudesta. Epävarmuus lisääntyy, jos kysymyksen mittakaava ja analyysin mittakaava eivät kohtaa toisiaan (Levin 1992). Käytännössä tämä tarkoittaa, ettei koko jokea koskeviin kysymyksiin voi saada vastausta pienellä mittakaavalla tehtyjen kokeiden kautta. Oikean mittakaavan valinta on sidoksissa tutkijan systeemituntemuksen kanssa, ja jos oikea mittakaava on tuntematon, työskentely useilla mittakaavoilla voi olla välttämätöntä. Toinen epävarmuuteen liittyvä tekijä on tilastollinen voimakkuus (power). Tutkijan kyky tulkita tilastollisen testin tulos oikein riippuu ratkaisevasti testin voimakkuudesta. Tilastollisen testauksen seurauksena voidaan todeta joko että tilastollisesti merkitsevää vaikutusta oli tai ei ollut. Voimakkuuden ollessa pieni päätelmä testin tuloksesta voi kuitenkin olla virheellinen ja saattaa johtaa tai II-tyypin virhepäätelmään (päätelmä, ettei vaikutusta ollut, vaikka todellisuudessa sitä esiintyi). Ihmistoiminnan vaikutukseen suunnatuissa seurannoissa I-tyypin virhe (päätelmä että vaikutusta oli, vaikkei se todellisuudessa esiintynyt) ei yleensä ole vakava, sillä se aiheuttaa ainoastaan ″väärän hälytyksen″. Tyypillisesti I –tyypin virheen riskiä minimoidaan asettamalla se H_0:n hylkäämistaso (α) 0,05:een (Peterman 1990, Fairweather 1991). Sen sijaan II-tyypin virhe voi aiheuttaa seurannassa olevalle jokisysteemille huomattavasti vakavampia seurauksia. Myös tätä virhettä vastaan on mahdollista suojautua esimerkiksi lisäämällä toistojen määrää ja kasvattamalla H_0:n hylkäämistasoa (esim. 0,05 sijasta 0,2) (Mapstone 1995). Vaihtoehdoista ensimmäinen on usein vaikeaa toteuttaa, sillä toistojen määrän lisäys suurilla mittakaavoilla lisää huomattavasti kustannuksia. H_0:n hylkäämistason muuttaminen voimakkuuden kasvattamiseksi etukäteen vaatii tietoa I-tyypin ja II-tyypin virheistä aiheutuvista kustannuksista ja suurimman sallitun vaikutustason asettamista (Mapstone 1995).

Usein seurantaa suorittava tutkija ei ole kiinnostunut vaikutuksien yleisestä todentamisesta (esim. syanidipäästöjen vahingollisuus yhteisötasolla), mikä vaatisi tasapainoisen, replikoidun ja satunnaistetun koeasetelman, vaan hän on kiinnostunut jonkun tietyn ihmistoiminnan vaikutuksien (esim. Tizsajoen joutuneiden syanidipäästöjen vaikutus eliöstöön) arvioimisesta. Tällainen suuren mittakaavan replikoimattomien ja ei-satunnaistettujen ihmistoiminnan vaikutuksien arviointi vaatii erikoisia kenttäarviointiasetelmia. Tyypillistä näille asetelmille on, että ne hyödyntävät tietoa kohdealueelta ja kontrollialueelta useilta ajankohdilta sekä ennen siihen kohdistunutta vaikutusta että sen jälkeen (BACIPS, ennustava BACIPS ja RIA). Tavallisessa BACIPS (Before After Control Impact Paired Series) asetelmassa saman ajankohdan kohdealueen (I) ja kontrollialueen (C) kohdeparametrin arvosta muodostetaan erotus (esim. $\Delta_{Bi} = I_{Bi}$- C_{Bi}) (Stewart-Oaten et al. 1986). Erotukset lasketaan erikseen jokaiselle ajankohdalle ennen (before, Δ_{Bi}) ihmistoiminnan aloittamista ja jälkeen (after, Δ_{Ai}) sen aloittamisen. Keskimääräinen Δ_{Bi} on kohdealueen ja kontrollialueen erotus joka estimoi keskimääräistä eroa tapauksessa, jossa systeemiin ei ole kohdistunut mielenkiinnon kohteena olevaa ihmisvaikutusta. Ihmistoiminnan vaikutuksien suuruutta pystytään arvioimaan laskemalla keskimääräisten ennen ja jälkeen erojen pohjalta ns. vaste-ero (Δ_B-Δ_A); vaste-erolle on myös mahdollista laskea luottamusvälit (Stewart-Oaten et al. 1986). Lisäksi BACIPS asetelma mahdollistaa testauksen, joka kertoo, poikkeavatko kohdemuuttujan arvot ennen toiminnan aloittamista toiminnan aloittamisen jälkeisistä arvoista. Ennen ja jälkeen ajankohtien erojen testauksessa käytettävä t-testi asettaa aineistolle tiettyjä rajoitteita, joista tärkeimmät ovat riippumattomuus, additiivisuus ja normaalijakautuneisuus (Stewart-Oaten et al. 1992). BACIPS asetelman tiukat rajoitteet ovat johtaneet uusien menetelmien kehittelyyn jotka poistavat tiettyjä vaatimuksia. Näistä läheisintä sukua BACIPS lähestymistavalle on ns. ennustava BACIPS, joka ennustaa kohdealueen arvoja kontrollialueen arvojen perusteella (Bence et al. 1996). Ennustavassa BACIPS-asetelmassa mallinnetaan kaksi funktiota. Ensimmäinen esittää kohde- ja kontrolliarvojen välistä suhdetta ennen toiminnan aloittamista, ja toinen suhdetta kohde ja kontrolliarvojen välillä toiminnan aloittamisen jälkeen. Laskemalla erotus näiden kahden funktion välille mahdollistetaan vaste-eron suora estimointi tiettyä kontrolliarvoa kohden. Näiden funktioiden erotuksista syntyy kolmas funktio, joka kuvaa estimoitua vaste-eroa ja sen luottamusvälejä kontrolliarvoja vastaan. Ennustavan BACIPS-asetelman etuna on, että toisin kuin perinteinen BACIPS-asetelma, se ei oleta additiivisuutta (Bence et al. 1996).

Toisen vaihtoehdon perinteiselle BACIPS asetelmalle tarjoaa RIA (Random Intervention Analysis) (Carpenter 1989). Kuten BACIPS, RIA käyttää analyysin perustana kohde- ja kontrollialueen erotuksia ennen ja jälkeen ihmistoiminnan aloittamista (esim. $D_{Bi} = I_{Bi}-C_{Bi}$). Yksittäisistä erotuksista lasketaan ajankohtien keskimääräiset erotukset (esim. $D_B = \sum D_{Bi}/n_{Bi}$) ja testauksessa käytetään näiden keskimääräisten erotuksien itseisarvoa ($|D_B-D_A|$). Erotuksien frekvenssijakaumaa estimoidaan satunnaisella uudelleenotannalla, jossa alueiden väliset erot satunnaisesti arvotaan joko ennen tai jälkeen ajankohtiin kuuluviksi, riippumatta niiden alkuperäisestä ajankohdasta (Carpenter 1989). Tuotettua frekvenssijakaumaa verrataan alkuperäistä $|D_B-D_A|$ vastaan ja arvioidaan se osuus arvoista, joka on alkuperäistä itseisarvoerotusta äärevämpi. Äärevämpien arvojen osuus vastaa P-arvoa; pieni P-arvo ilmentää ei-satunnaista muutosta systeemissä (Carpenter 1989). RIA, kuten BACIPS, ei suoraan pysty osoittamaan, että tietty toiminta on aiheuttanut vaikutuksen (Stewart-Oaten et al. 1992). Merkitsevät testitulokset voisivat johtua myös muista, samanaikaisesti sattuneista tapahtumista (esim. myrskyjen, tulvien tms. johdosta). Siten tutkijan kyky arvioida vaihtoehtoisten selitysmekanismien osuutta tuloksissa on keskeinen osa näiden kenttäarviointimenetelmien soveltamista. On kuitenkin olemassa menetelmä, jonka avulla pystytään kiistatta osoittamaan ja yhdistämään aiheuttaja vaikutuksiinsa. Tätä menetelmää kutsutaan asymmetriseksi asetelmaksi tai Beyond-BACI menetelmäksi (Underwood 1991, 1992). Asymmetrisen asetelman taustana on epätasapainoinen varianssianalyysi. Vaikka kohdealueita ei useimmiten voida replikoida, on kontrollialueiden replikointi sen sijaan usein mahdollista (Underwood 1994). Kontrollialueiden replikoinnin ansiosta tilastollinen merkitsevyys ja vaikutuksien aiheuttaja ovat selvästi pääteltävissä varianssianalyysin yhdysvaikutuslausekkeista. Vaikka asymmetrinen asetelma on kontrollialueiden toiston kautta usein tavallista BACIPS asetelmaa työläämpi, on sen etuna kyky kiistattomasti todentaa syy-seuraus suhteita. Jos näytteenotto lisäksi suoritetaan hierarkkisesti, asetelma kykenee havaitsemaan myös vaikutuksia, jotka eivät ilmene keskiarvon muutoksina, vaan varianssin kasvuna (Underwood 1991). Käytännössä tällä ominaisuudella on suuri merkitystä esimerkiksi silloin, kun seurannan kohteena on uhanalainen laji, koska pienen populaation keskiarvoon littyvä suurempi vaihtelu lisää sukupuuttoon kuolemisen riskiä (Simberloff 1986). Vaikka hierarkkinen, asymmetrinen asetelma on ainoa kenttäarviointimenetelmä, joka kykenee havaitsemaan muutoksia varianssissa, on menetelmää valitessa muistettava, että tämän asetelman käyttö moninkertaistaa työmäärää.

Kenttäarviointiasetelmien vaatimaa tietoa ennen vaikutuksien alkamista vallinneesta tilanteesta ei aina ole saatavilla (esim. vaikutuksien äkillisen ilmenemisen yhteydessä). Asymmetrisessa asetelmassa puuttuvaa ennen tieto voidaan kompensoida, jos on saatavilla aikasarjoja samankaltaisista jokisysteemeistä kuin kohdesysteemi (Underwood 1994). Näiden aikasarjojen tulee muodostua satunnaisesti valituista jokisysteemeistä, joita on seurattu useilla ajallisella mittakaavoilla. Otosteorian oletuksena on, että vaihtelu, joka ilmenee satunnaisesti valittujen populaatioyksikköjen keskiarvossa, tulisi olla sama kuin vaihtelu, joka esiintyy koko populaatiossa (esim. Feller 1968). Tämän seurauksena voi satunnaisesti valittujen populaatioyksikköjen keskiarvoa käyttää kuvaamaan puuttuvaa ennen-tietoa. Otoksen antaman estimaatin tarkkuutta voidaan arvioida suorittamalla näytteenottoa vaikutuksen jälkeen useilla kontrollialueilla ja vertailemalla näiden antamia arvoja satunnaissarjojen antaman estimaatin kanssa (Underwood 1996). Jos estimaatti ja satunnaisotannalla valittujen jokisysteemien muodostama keskiarvo ovat samanlaisia, voidaan aikasarjan antamilla estimaateilla korvata puuttuvaa ennen-tietoa.

Lähestymistavan, kuten kohdemuuttujankin, valinta on pitkälti sidoksissa seurantaohjelman kysymyksenasetteluun. Useimmiten rahoitus seurantaohjelmia varten on rajallinen ja tämä luo tarpeita ilmentäjälajien käytölle (Jones & Kaly 1996). Perinteisesti kohdelajeina jokiseurannassa käytetään runsaasti esiintyviä pohjaeläin- tai kalalajeja, koska näiden keskiarvojen luotettavaan arvioimiseen tarvitaan vähiten näytteitä (Morin 1995). Toisaalta harvinaisten lajien seurantakin voi tietyissä tapauksissa olla mielekästä, ja tähän on olemassa omat menetelmänsä (esim. Lessica & Steele 1996).

Kysymyksenasettelu seurantaohjelman taustalla vaikuttaa niin menetelmän kuin myös kohdelajin valintaan. Vaikka tutkijoilla usein on hyvin rajallinen rahoitus seurantaohjelmien toteuttamiseksi, on selvää, ettei ilman oikeanlaista asetelmaa ja aineistoa pystytä toteamaan ihmisen aiheuttamia muutoksia jokiekosysteemissä (Christensen 1996, Treweek 1996). Koska ihmistoiminnan aloittamista edeltävä tieto puuttuu usein kokonaan, jäävät tehokkaat kenttäarviointimenetelmät usein käyttämättä jokiseurannoissa. Tämän puuttuvan tiedon kompensointi luo tarpeen kansallisille aikasarjoille ihmistoiminnalle alttiina olevista jokiekosysteemeistä. Suomessa tällaisia kohteita ovat lähinnä metsäpurot ja pienet joet. Aikasarjoja tulisi luoda ekoregioittain siten, että kussakin ekoregiossa valittaisiin satunnaisesti 15-20 jokea/puroa, joita seurattaisiin kahtena tai kolmena ajankohtana vuodessa. Kokonaisuudessa tämä tarkoittaisi 70-80 puron/joen seuraamista kansallisella seurantaohjelmalla. Jokaisella näytteenotolla otettaisiin vähintään viisi rinnakkaisnäytettä: jotta jokaisen joen sisäistä vaihtelua pystyttäisiin arvioimaan, rinnakkaisnäytteitä ei tulisi yhdistää. Pohjaeläintiheyksissä esiintyvän vuosienvälisen vaihtelun estimoinnin helpottamiseksi aikasarjaa tulisi jatkaa kuudesta kahdeksaan vuotta. Vaikka tällaisen aikasarjan perustaminen näyttäisi vaativan suuria investointeija, käytännössä aikasarjasta saatu tieto toimisi jatkossa kompensoivana tiedonlähteenä tilanteissa, joissa etukäteistietoa ihmistoiminnan kohteeksi joutuvasta jokiekosysteemistä ei ole. Näin ollen aikasarja säästäisi tulevaisuudessa tehtävien ihmistoiminnan arviointiin liittyvien seurantojen vaatimaa resurssointia, ja siten mahdollistaisi tehokkaimpien mahdollisten seurantamenetelmien käytön.

Documentation page

Publisher	West Finland Regional Environment Centre	Date	November 2000
Author(s)	Kristian Meissner		
Title of publication	Experimental methods in the assessment and monitoring of rivers: benefits, limitations and integration with field surveys		
Parts of publication/ other project publications			

Abstract

Investigators conducting biomonitoring in running waters are faced with the difficult task of choosing the most suitable approach for a specific problem. Individual approaches rank differently with respect to three basic goals: (i) precision, (ii) realism and (iii) generality. As there is no all-purpose approach maximizing all three goals simultaneously, the priorities set by the investigator, issues of scale, statistical power and target variables will influence the final choice. This report outlines the strengths and weaknesses associated with the most prominent approaches used in biomonitoring of rivers. Many situations (e.g. toxic spills) will require the use of field assessment designs. Various BACI- designs (Before After Control Impact) are described in detail in this report. To be applicable, however, these designs require pre- and post-impact monitoring of the target and the control systems. In practice, pre-impact data are often missing, which in most cases makes the use of powerful BACI designs impossible. Methods to substitute missing pre-impact data through random time series are discussed. Finally, this report presents three examples of biomonitoring case studies and examines the solutions used in these studies to solve biomonitoring problems.

Keywords	biomonitoring, experiments, field assessment, scale, power, BACI (Before After Control Impact)
Publication series and number	Regional Environmental Publications 189
Theme of publication	
Project name and number, if any	
Financier/ commissioner	West Finland Regional Environment Centre
Project organization	

ISSN	1238-8610	ISBN	952-11-0802-9
No. of pages	42	Language	English
Restrictions	Public	Price	40 FIM

For sale at/ distributor	West Finland Regional Environment Centre, P.O.Box 262, FIN-65101 VAASA tel. +358-(0)6-367 5211, fax. +358-(0)6-367 5251
Financier of publication	West Finland Regional Environment Centre
Printing place and year	Multiprint, Vaasa 2000

Kuvailulehti

Julkaisija	Länsi-Suomen ympäristökeskus		Julkaisuaika	Marraskuu 2000
Tekijä(t)	Kristian Meissner			
Julkaisun nimi	Kokeellisten menetelmien käyttö jokivesien tila-arvioinnissa ja seurannassa: mahdollisuudet, rajoitukset ja yhdistäminen kenttämenetelmien kanssa			
Julkaisun osat/ muut saman projektin tuottamat julkaisut				
Tiivistelmä	Virtavesien seurannasta vastaava tutkija voi kohdata ongelmia yrittäessään valita sopivinta lähestymistapaa seurantaongelmaansa. Käytettävissä olevat lähestymistavat poikkeavat toisistaan kolmen päätavoitteen, (i) tarkkuuden, (ii) todenmukaisuuden ja (iii) yleistettävyyden suhteen. Mikään lähestymistapa ei pysty täysin täyttämään kaikkia päätavoitteita yhtaikaa, ja näin ollen tutkijan ennakkotavoitteet, tutkimuksen mittakaava, tilastollinen voimakkuus ja kohdemuuttuja vaikuttavat lopulliseen koeasetelman valintaan. Raportissa käsitellään jokien seurannassa käytettyjen lähestymistapojen vahvuuksia ja heikkouksia. Monet seurantatilanteet vaativat kenttäarviointimenetelmien käyttöä (esim. äkillinen myrkkypäästö). Tähän tarkoitukseen sopivia, voimakkaita BACI- asetelmia (Before After Control Impact) käsitellään perusteellisesti tässä työssä. Toimiakseen nämä BACI- asetelmat vaativat tietoa kohdealueelta ja kontrollialueelta useilta ajankohdilta ennen ja jälkeen vaikutusta. Käytännössä tietoa joen tilasta ennen vaikutusta on harvoin olemassa, mikä tekee BACI- menetelmien käytön monesti mahdottomaksi. Puuttuvaa tietoa joen tilasta ennen siihen kohdistunutta vaikutusta voidaan kuitenkin korvata kattavien aikasarjojen avulla. Tähän tarkoitukseen kehitetty menetelmä esitellään tarkemmin raportissa. Työssä pohditaan lisäksi kattavien korvaavien aikasarjojen perustamistarpeita Suomen virtavesissä ja tarkastellaan kolmen esimerkkitapauksen avulla mahdollisia ratkaisumalleja monimutkaisiin seurantatilanteisiin.			
Asiasanat	seuranta, kokeellinen lähestymistapa, kenttäarviointi, tutkimuksen mittakaava, tilastollinen voimakkuus, BACI (Before After Control Impact)			
Julkaisusarjan nimi ja numero	Alueelliset ympäristöjulkaisut 189			
Julkaisun teema				
Projektihankkeen nimi ja projektinumero				
Rahoittaja/ toimeksiantaja	Länsi-Suomen ympäristökeskus			
Projektiryhmään kuuluvat organisaatiot				
	ISSN	1238-8610	ISBN	952-11-0802-9
	Sivuja	42	Kieli	Englanti
	Luottamuksellisuus	Julkinen	Hinta	40 mk
Julkaisun myynti/ jakaja	Länsi-Suomen ympäristökeskus, PL 262, 65101 VAASA, puh. (06) 367 5211, fax. (06) 367 5251			
Julkaisun kustantaja	Länsi-Suomen ympäristökeskus			
Painopaikka ja -aika	Multiprint Vaasa, 2000			

Presentationsblad

Utgivare	Västra Finlands miljöcentral		Datum	November 2000
Författare	Kristian Meissner			
Publikationens titel	Användningen av experimentella metoder i bedömningen och monitoringen av tillståndet i älvar och åar: möjligheter, begränsningar och kombinering med fältmetoderna			
Publikationens delar/ andra publikationer inom samma projekt				
Sammandrag	Forskare som ansvarar för monitoringen av rinnande vatten kan stöta på problem i valet av det lämpligaste sättet att närma sig de problem som uppkommer i monitoringen. De tillvägagångssätt som finns att tillgå avviker från varandra när det gäller de tre huvudmålsättningarna, (i) noggrannhet, (ii) realism och (ii) generaliserbarhet. Inget av dessa tillvägagångssätt kan helt uppfylla alla huvudmålsättningar samtidigt och därför påverkar forskarens på förhand uppställda mål, forskningens omfattning, statistiska intensitet och syftesvariabel det slutliga valet av metod. I rapporten beskrivs styrkan och svagheterna hos de metoder som används i monitoringen av älvar och åar. Många situationer kräver användning av fältbedömningsmetoder (t.ex. plötsliga giftutsläpp). De BACI-metoder (Before After Control Impact) som är lämpliga och effektiva för ändamålet behandlas grundligt i detta arbete. För att dessa BACI-metoder skall fungera krävs information om forsknings- och kontrollområdet från flera tidpunkter före och efter påverkan. I praktiken finns det sällan uppgifter om älvens tillstånd före påverkan, vilket innebär att användningen av BACI-metoder ofta är omöjligt. Den information som saknas om älvens tillstånd innan den påverkades kan dock ersättas med hjälp av täckande tidsserier. Metoden som har utvecklats för detta ändamål presenteras mera ingående i rapporten. Dessutom begrundas behovet av att utarbeta ersättande tidsserier för rinnande vatten i Finland. Med hjälp av tre exempel granskas eventuella modellösningar för komplicerad monitoring.			
Nyckelord	monitoring, experimentellt tillvägagångssätt, fältbedömning, forskningens omfattning, statistisk intensitet, BACI (Before After Control Impact)			
Publikationsserie och nummer	Regionala miljöpublikationer 189			
Publikationens tema				
Projektets namn och nummer				
Finansiär/ uppdragsgivare	Västra Finlands miljöcentral			
Organisationer i projektgruppen				
	ISSN	1238-8610	ISBN	952-11-0802-9
	Sidantal	42	Språk	Engelska
	Offentlighet	Offentlig	Pris	40 mk
Beställningar/ distribution	Västra Finlands miljöcentral, PB 262, 65101 VASA, tel. (06) 367 5211, fax. (06) 367 5251			
Förläggare	Västra Finlands miljöcentral			
Tryckeri/ tryckningsort och -år	Multiprint, Vasa 2000			