internal report **423**

Estimating the power of the ARR monitoring protocols: The MBACI(P) design

K McGuinness

June 2003

supervising scientist

FINAL REPORT
October 29, 2002

ESTIMATING THE POWER OF THE ARR MONITORING PROTOCOLS: THE MBACI(P) DESIGN

prepared by
Keith McGuinness
Centre for Tropical Wetlands Management
Faculty of Education, Health and Science
Northern Territory University

for eriss
Department of the Environment and Heritage
Commonwealth of Australia

Table of Contents

Preamble

The overall aim of this document is to assist in estimating the power of certain types of monitoring programs to detect changes which might result from artificial disturbances (*i.e.* "environmental impacts"). The issues involved, both statistical and ecological/environmental, are complex. An appreciation of these is important but an extended discussion of them here would be out of place. Further discussion of these issues is in the book *Monitoring Ecological Impacts: Concepts and Practice in Flowing Waters* (Downes et al. 2002) and I recommend that this be consulted. The analytical approach adopted here is the analysis of variance: for information on this I recommend the recent books by Quinn & Keough (2002) and Underwood (1997).

Power and statistical tests

Any statistical test of a null hypothesis will result in either the retention (or acceptance) or rejection of that null (Table 1). A *correct* decision is made if the *null is true and is retained*, or is *false and is rejected*. An incorrect decision is made if the null is *retained when false*, or *rejected when true*. The latter mistake is referred to as a Type I error.

Table 1. Decisions in statistical tests.

	Null is accepted	**Null is rejected**
Null is actually true	Correct decision	Type I error (α)
Null is actually false	Type II error (β)	Correct decision

The Type I error rate—that is, the probability of (incorrectly) rejecting the null when it is actually true—is determined by the significance level (α) used when interpreting the results of the statistical test. This rate is conventionally set at 5% ($\alpha = 0.05$), although several authors have concluded that this decision is arbitrary and warrants greater consideration than it usually gets. This is undoubtedly true, but a discussion of these issues would be out of place here (see Chapter 12 in Downes et al. 2002), so the conventional level will be adopted for illustrative purposes (unless stated otherwise).

In environmental monitoring situations, the Type II error rate is frequently of particular concern. This is because, in general terms, in such situations the null hypothesis, broadly stated, often takes the following form:

H_0: There is no difference between Control and Impact samples

Retaining this null when it is actually false (*i.e.* making a Type II error), may well result in an environmental impact going undetected, an outcome which is generally regarded as undesirable. Thus, much recent effort has been devoted to developing experimental designs which limit both the Type I and Type II error rates to acceptable levels. In other words, studies have attempted to describe designs which provide *powerful* tests of the null hypothesis, while limiting the Type I error rate.

The *power* of a statistical test, is the *probability that the test will correctly reject the null hypothesis* when it is actually false:

Power = $1 - \beta$

Unfortunately, the Type II error rate for any particular test—and, therefore, the power of that test—is not determined in as simple a fashion as the Type I error rate. The latter is, as noted earlier, set by the analyst. The former depends, among other things, on the sample size (or number

of observations) *and* the magnitude of any effect. This is easy to see. For instance, if there is a large effect—that is, a large difference between the control and impact samples—then a few observations will suffice to demonstrate this (*i.e.* a few observations will be adequate to reliably reject the null hypothesis of no difference). In contrast, if an effect exists but it is only small, then many observations will probably be required to reliably detect this.

In general, several pieces of information are required if the power of a test, to reject a particular null, is to be calculated.

- The *null hypothesis* must be stated—this is self-evident. This *may* be of the form:

  H_0: There is no difference between control and impact samples

- The *alternate hypothesis* must be stated specifying the *effect size* of interest—the alternate hypothesis must specifically define what magnitude of effect (or difference) is of interest. For instance:

  H_A: The control and impact samples will differ by at least 10 species.

- The *sample size* must be stated—other things being equal, more observations will result in a more powerful test (although this may depend greatly on the design of the study).
- The *statistical test* to be used must be stated—tests vary in suitability and power depending upon how well the assumptions of the test are satisfied by the particular situation under study.
- In most cases, the magnitude of *sampling variation* must be stated—more samples will be required for adequate power in systems which exhibit great variability.

Estimating power

In general, the power of a particular test can be estimated by one of two methods: direct calculation, or simulation.

- *Direct calculation*—If suitable formulae are available, the power of a particular test may be calculated directly. Usually, two steps are involved. First, information of the type listed above is used to calculate a defined parameter (*e.g.* delta) of an appropriate sampling distribution (*e.g.* the non-central F-distribution). Second, the power of the test is estimated from that distribution, based on the estimated parameter and measures of sample size (*e.g.* degrees of freedom), using charts or some other means. This method provides precise results but can be tedious and prone to error, as the calculations can be complicated. Increasingly, specialised software is becoming written to ease this process but that which is currently available only covers a limited set of simple situations.
- *Simulation*—A computer model of the system can be built, which generates "dummy" data which can be structured so that either the null hypothesis is true, or an effect of a particular magnitude is present. Multiple sets of data can be generated by this model and the null hypothesis tested. Power can be estimated by recording the number of times, for any situation, that the null is rejected. The tedious and complicated part of this process is the coding, and testing, of the computer simulation. Once this is done, comparing different scenarios is (or should be) easy. This method provides less precise estimates of power than direct calculation. Provided, however, that 1,000 – 2,000 sets of data are generated for each different situation, these estimates usually differ by only 1 – 2% from the calculated results (pers. obs.); and variations of this magnitude are unimportant.

After exploring both approaches, direct calculation, based on Underwood (1993), seemed to be the simplest and most reliable approach. *It is, however, critically important to recognise that the*

calculations described here apply specifically to a particular version of the ARR design (see box and below) and might not be valid for other designs.

Workbook notes. Throughout this document, references are made to "the workbook". This workbook does most of the tedious calculations required to estimate power, however, the calculations done are tailored *specifically to the ARR design* (of which more below) *and* to a version of this design which incorporates some limitations. I do not consider the effects of these limitations to be severe but their nature, and the reasons for them, are discussed elsewhere in this document. The workbook incorporates Underwood's (1991, 1992, 1993, 1994) more specific MBACI tests (and is based, in particular, on Underwood 1993). The workbook is "ARRPowerCalc.xls" (written using Excel97 and VBA) and its use is illustrated in these notes.

The ARR design

Exemplary data for this report were provided from baseline, macroinvertebrate community structure data gathered from streams in the vicinity of the Jabiluka mine site (period 1999-2001). Several basic aspects of the ARR design are pertinent (see Figure 1).

- *Streams*—Four streams are being monitored: three are "control" streams; the fourth is the putative "impact" stream.
- *Reach*—Samples are collected on both "upstream" and "downstream" sections of each stream. The "downstream" section of the impact stream will be downstream of the impact, when that commences. Note that the upstream and downstream reaches (or sites) are "paired", as they (obviously) occur together on the same stream, and their behaviour is likely to be correlated[1] as they are connected by flow (and see Chapter 8 in Downes et al. 2002).

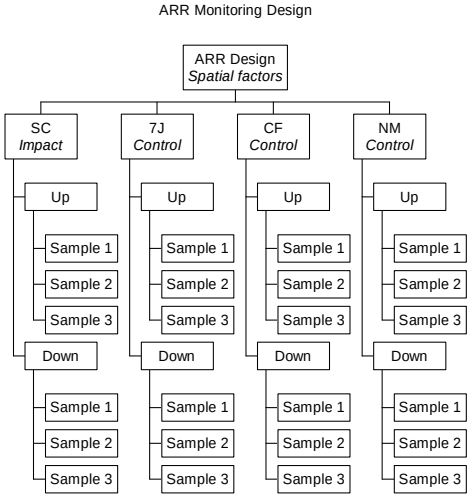

Figure 1. Spatial design of ARR monitoring program.

- *Year*—All streams will be sampled for some years prior to, and after, the start of the impact.
- *Time*—Samples will be taken several times each year (up to about 5), at approximately monthly intervals, from each of the streams.

[1] Sites on the same stream often exhibit similar changes through time—for instance, in response to a disturbance such as a flood event—because they are part of the same connected system. This "connectedness", or correlation, should not be confused with the problem of *non-independence of observations* (or replicates): the latter can cause problems with statistical analyses (*e.g.* ANOVA) if an appropriate model is not used (*e.g.* if replicates are treated as independent when they are not). In the designs analysed here, the replicates are based on the samples randomly, and independently, collected at each sampling time. Thus, the problem of non-independence does not arise (for more see Underwood 1997, Quinn & Keough 2002).

At present, three replicate samples are collected from each reach on each stream on each sampling occasion. The macro-invertebrates present are identified and counted. Tests for impacts may be made using several variables derived from these results.

- *Abundances* of any taxa, or group of taxa: see Appendix 1.
- *Diversity*, estimated simply from counts of the number of different taxa, or from indices incorporating information on abundance: see Appendix 1.
- *Differences in community structure* between upstream and downstream reaches, based on similarity (or dissimilarity) indices, such as the Bray-Curtis index (commonly used to derive dendrograms or ordinations). In order to have replicate readings for each stream at each time, the three upstream replicates are randomly paired with the downstream replicates. Each set of (randomly paired) samples, provides one replicate estimate of upstream–downstream similarity (or dissimilarity).

The latter, novel approach, deserves further discussion. Essentially, the Bray-Curtis index is used to measure the similarity (or dissimilarity) of assemblages on the upstream and downstream reaches of each stream. Any impact which occurs should predominantly affect the downstream reach of the impact stream, causing assemblages there to diverge from those in the upstream reach of that same stream. Thus, an impact should be detectable by observing that, after the impact starts, the upstream–downstream similarities on the impact stream decrease, relative to changes on the control streams.

One way—and arguably the most powerful way—to test for impacts in this system is using analysis of variance (ANOVA), utilising the "MBACI" types of designs described by Underwood (1991, 1992, 1993, 1994). These designs compare results from one impact site, to those from *several* control sites, all sampled several times before and after the start of the putative impact. The basic rationale is that a long-term impact occurs when the behaviour of the impact site *after the impact* differs from the average behaviour of the control sites (Figure 2). Underwood (1991, 1992, 1993, 1994) should be consulted for further details and justification (also see Keough & Mapstone 1995; Downes *et al.* 2002; for a contrasting view see Stewart-Oaten & Bence 2001).

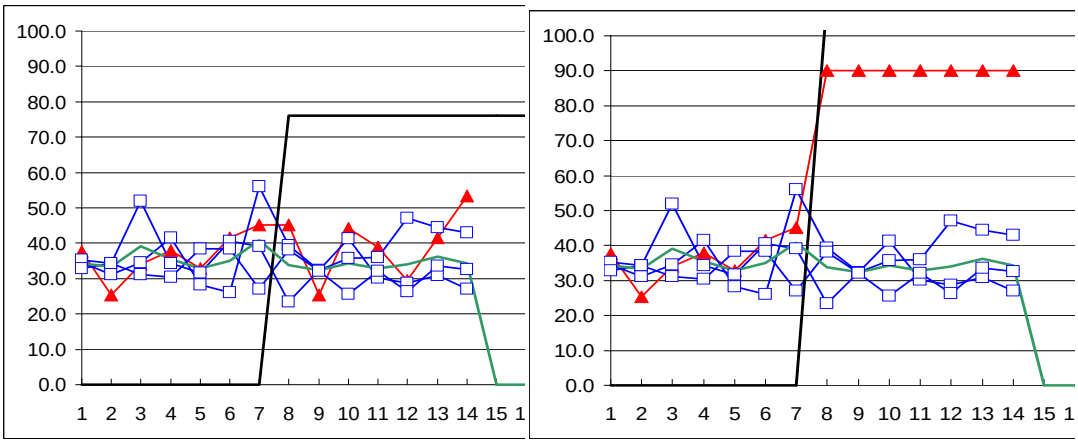

Figure 2. Examples of no impact (left) and impact (right) based on measuring dissimilarity (differences in community structure, y-axis) between upstream and downstream reaches. The lower (x) axis of the graph is sample number, through time. Open squares represent data from control streams, triangles, data from impact stream.
(Graphs copied from the power calculation workbook.)

A basic ANOVA model for this situation would incorporate two factors:

- *Stream*—the four streams sampled.
- *Time*—the multiple times sampled on each stream.

[Note that in this particular case, the variable analysed is the *difference in community structure between the upstream and downstream reaches* (on each stream). Thus, "reach" cannot be a factor in this design, although it might be in other analyses of other variables (*e.g.* number of taxa present); see Appendix 1.]

In such an analysis, an impact, if sufficiently large, would result in a significant *Stream* × *Time* interaction, indicating that the behaviour of the four streams, through time, was not the same. The existence of such an interaction would not, however, *uniquely* identify an impact. For instance, a *Time* × *Stream* interaction would arise if one of the control streams behaved in a different fashion to the other controls, before or after the impact.

Thus, more specific tests are required to uniquely identify the presence of a change indicative of an impact. Underwood (1991, 1992, 1993, 1994) derives such tests by subdividing the spatial and temporal factors in the design.

- *Stream*—At least two types of comparisons can be made here: (a) of the impact stream to the average control stream; and (b) of control streams to each other.
- *Time*—At least two types of comparisons can also be made here: (a) from before to after the impact; and (b) among times either before or after the impact.

In general, a change from before to after the impact, in how the impact stream compares to the control streams would indicate that an impact had occurred. (The situation may not be this simple as, depending upon the design of the sampling program, different sorts of effects may cause other patterns to occur: see Underwood (1991, 1992, 1993, 1994) for an extended discussion.)

The ARR design here incorporates one spatial scale[2]—*among streams*—but three temporal scales—*before versus after* the putative impact, *among years*, and *among times within years*.

On replication in this design

In the design as analysed here, there are (three) replicate "observations" on each reach on each stream at each sampling occasion. The replicate "observations" here are *not* the raw data themselves (*i.e.* the counts of numbers of individuals of each species) but the *dissimilarities between randomly-paired upstream and downstream samples*. These replicate dissimilarities are form the lowest level of replication in the design. As the samples themselves are initially randomly and independently collected, and are subsequently randomly paired, there is no problem of non-independence of observations (and see footnote 1). An alternate way of proceeding would be to pool the samples for each reach at each time and calculate one measure, or observation, of dissimilarity for each stream at each time. Doing this would, however, (among other things) eliminate tests for short-term differences among streams; information which is, potentially, of some significance. It should only be considered if costs become very prohibitive.

[2] Noting again that the spatial scale "reach" does not exist in this particular situation because of the nature of the observations: upstream–downstream differences.

Specifying the design

In an ANOVA model, factors may be either "fixed" or "random". In brief, in the case of a *fixed* factor, all relevant levels (*i.e.* conditions) are included in the study. In the current example, the temporal factor, *Before/After Impact*, is fixed because there are only two possible conditions of interest (*i.e.* before and after the impact commences) and both are included in the design.

In contrast, in the case of *random* factors, only a sub-set of the possible and available conditions are included in the study, and the particular instances included are selected, essentially, at random from the entire set available. For instance, the control sites included in an environmental study might be selected, essentially, at random from a larger set of suitable available sites.

The distinction between random and fixed factors, in ANOVA designs, is important because it determines both the way in which particular hypotheses are tested *and* the applicability of the conclusions drawn (see Keough & Mapstone 1995 for a discussion of these issues as they relate to environmental monitoring and Underwood 1981, Underwood 1997 and Quinn & Keough 2002 for a more general discussion). For instance, if control sites are drawn at random from a larger pool of such sites, then any conclusions made about "control sites" in the study should apply to the entire population of sites. In contrast, if particular sites are selected and used for specific reasons, then any conclusions from the study will obviously only apply to those particular sites.

It is also important to recognise that decisions as to which factors are fixed and which, if any, are random may affect the power of tests of particular null hypotheses (because such decisions determine which Mean Squares, and the degrees of freedom, are used to form the numerator and denominator when calculating the F-ratio used to test the null). For instance, if "Sites" and "Times" are both random, then the "Sites by Before/After" effect may be tested over the "Sites by Times" interaction; thus the degrees of freedom, and the power of this test, will be determined by the number of sites and times sampled, and not the number of replicates. This is discussed further later in this document.

Streams

The usual interpretation is that control sites are randomly selected from a "pool" of available sites (e.g. Underwood 1993; Keough & Mapstone 1995), although some authors dispute the validity of this (Stewart-Oaten & Bence 2001). That practice shall be followed here.

Times

Sampling times may, or may not, be randomly selected (Keough & Mapstone 1995). As noted above, sampling at three temporal scales is incorporated in the ARR design.

- *Before versus after the impact*—As noted earlier, this is a fixed factor as the only two possible conditions (before and after) are included in the study.
- *Years within B*/A—Years are sampled sequentially, and all years from the start of the sampling program until some (as yet undetermined) time after the impact. This factor is also clearly fixed.
- *Times within Years*—Within each year, samples are taken from each stream at approximately monthly intervals through the wet season. During the wet season, samples are usually taken from December through to April—5 approximately evenly spaced times—but in any given year there may be more, or fewer, times depending upon climatic and other conditions. Over the period December 1998 to April 2001, the interval between samples averaged 27 days but ranged from 19 to 44 days (-30% to +60%). Given also that the

temporal dynamics of flow are likely to differ among the four streams, it seems reasonable to consider these *Times within Years* to represent randomly selected samples from a range of possibilities. For essentially the same reasons, *Times* should also be considered to be nested within *Years*.

The factors listed above provide a complete analysis of the three temporal scales included in the design. The more factors included in the design, however, the more complex, obviously, is the analysis of the resulting data and the more complex the analysis of the power of any tests completed. The complexity of the analysis itself is a lesser consideration, as modern statistical packages can handle most designs with little difficulty. The estimation of the power of tests is a somewhat different issue, as this is still a cumbersome process. It is, therefore, worth examining the implications of not splitting the latter two temporal factors.

Combining (or not separating) "years" and "sampling times"

As discussed above, the ARR design incorporates sampling at three temporal scales: *before/after* the impact, sequential *years* before and after the impact, and *sampling times* (at approximately monthly intervals) within each year. Adding the spatial factor, *sites*, results in—*at its simplest*—a four factor ANOVA, with four main effects (*sites, before/after, years, sampling times*), six second order interactions (*sites by before/after*, etc.), four third order interactions (*sites by before/after by years*, etc.) and one fourth order interaction (*sites by before/after by years by times*). Admittedly, some of these terms would disappear if *times* is nested in *years* and, certainly, several are not of major interest when testing for impacts. Nonetheless, it is a somewhat cumbersome set of terms to deal with.

One further complication can arise: there may well not be the same number of times (*i.e.* months) sampled each year, because of varying conditions in the field (and possibly other factors). Having unequal numbers of sampling times each year may (in fact, will probably) complicate procedures for the eventual analysis of the results, but it certainly complicates the procedures for estimating power.

One solution is to "balance" the design by reducing the data to the smallest common number of sampling times, but this may omit much data. Another approach is to simply treat the various sampling times as just a series of observations collected before, and after, the start of the impact, ignoring the "years" factor. Doing this will "blur" together differences *among* years (before or after) with differences *within* years, but it is not obvious that this will *greatly* affect tests for impacts. There are likely to be some differences in the power of tests—between analyses including and excluding years—because of changes in the design and the degrees of freedom associated with some tests, but which analysis was more powerful would probably depend on the particular pattern of change observed.

This approach (of excluding the "years" factor) would be less tenable if pronounced annual differences were evident in the results, but this does not appear to be the case (Figure 3). There is little evidence, in the results to date, of marked differences among years; but there is certainly considerable fluctuation from one time to another (Figure 3). For these reasons, the factor "years" is not considered further in this discussion.

Workbook notes. One further simplification was made in developing the workbooks: the "impact" was assumed to start half way through the sampling period. In other words, if samples were collected on 14 occasions, it was assumed that the impact commenced between times 7 and 8. Doing this *greatly* simplifies the calculations required but, obviously, may well affect the results. This issue will be considered further below.

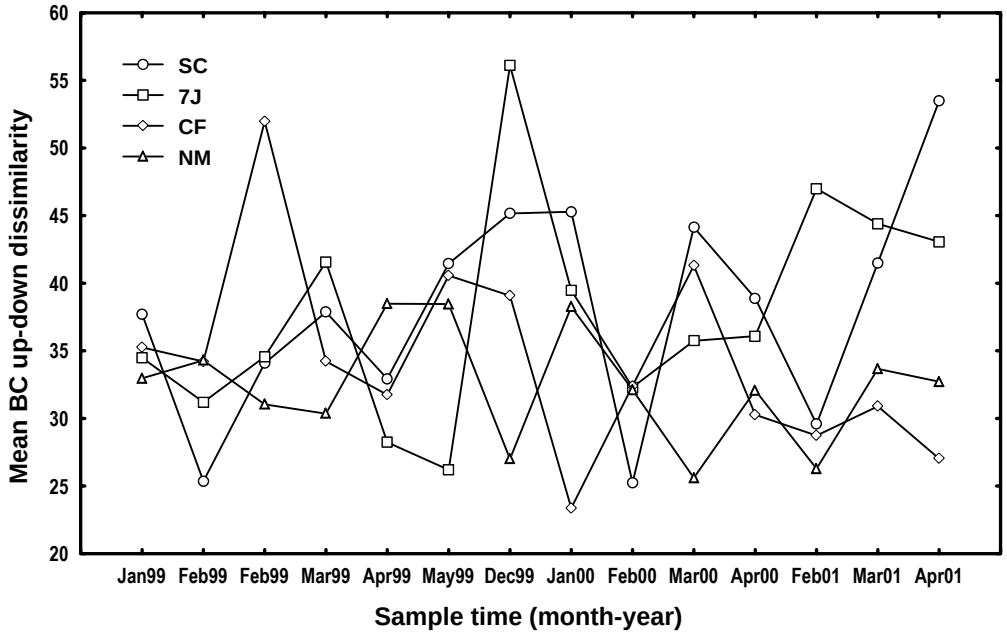

Figure 3. Mean dissimilarity between upstream and downstream replicates (randomly paired) at 14 times on each stream. The values shown are the means for the 4 streams for 14 sampling times (times 1, 13 and 14, in the original data, were omitted because missing data did not allow all upstream-downstream comparisons to be calculated). Note: results for the stream SC differ from those in Figure 4 because those in the latter figure reflect a particular impact "scenario" being examined.

Implications of the design used

As noted earlier, I have made specific decisions about some aspects of the design which will influence the estimates of power resulting. It is important to consider these further.

- *Specifying "Streams" (i.e. "Locations" or "Sites") and "Times within Before/After" (i.e. the separate sampling times before and after the impact) as random factors.* The most likely outcome of these decisions—*if they were subsequently determined to be inappropriate*—would be to make some of the estimates of power generated by my approach *conservative* (see Keough & Mapstone 1995). In other words, power might actually be higher than the estimates obtained stipulating both factors as random[3].

  Streams. Specifying "streams" as a random factor is consistent with the intent of the monitoring program—to compare the behaviour of the potentially impacted stream with that of other, similar streams in the region—and with the conclusions of most authors (Underwood 1991, 1992, 1993, 1994; Keough & Mapstone 1995; Downes et al. 2002; but see Stewart-Oaten & Bence 2001 for a contrary view). This decision is, however, not without complications and it does have important consequences.

  It can be argued that it is reasonable to regard the "control" streams as (more or less) randomly selected from a "pool" of suitable comparison sites (but see Stewart-Oaten & Bence 2001). It does not, however, seem equally reasonable to consider the "impact" stream as just one selection from a pool of alternative impact sites: after all,

[3] Note that it is *not* a simple matter to alter the spreadsheet to accommodate "streams" or "times" as fixed factors. When these factors are random, estimates of power are based on the "normal" central F-distribution (Underwood 1993; Keough & Mapstone 1995): if one, or both, factors were fixed, then some estimates would be based on the non-central F-distribution (Underwood 1993; Keough & Mapstone 1995).

both the specific type of impact, or its location, are selected for particular reasons. Thus, the "streams" factor may contain a mixture of fixed and random elements; in such cases, the factor appears to usually be regarded as random (this is the case with interactions between fixed and random factors in standard designs—see, e.g. Winer *et al.* 1991; Underwood 1997; Quinn *et al.* 2002)—and appears to be the case in Underwood's designs).

This decision—fixed or random—is not trivial as it affects both the way in which hypotheses about impacts are tested *and* how the power of such tests is estimated. For instance, if "streams" is considered to be a fixed factor, power is estimated using the non-central F distribution whereas if it is a random factor the central F-distribution is used (Keough *et al.* 1995). The calculations involved also differ. These issues is discussed further below (see "The special case of two streams: one impact and one control").

Times within Before/After. Opinion on the issue of sampling times is less settled (see references cited immediately above) and some may question my decision to regard this as a random factor. For this reason, I repeat and expand the arguments offered earlier. First, over the trial sampling period—December 1998 to April 2001—the interval between samples within a 12 month period averaged 27 days but ranged from 19 to 44 days. Second, sampling is only done during the wet season: the interval from the last sample of one wet season to the first sample of the next ranged from 218 to 280 days, although it was usually about 220–245 days. Thus, the interval between samples ranged from 19 to 280 days. Third, while the seasonal dynamics of rainfall will cause some similarities in the behaviour of the streams, small scale differences are likely to mean that temporal synchrony among the streams is limited. Indeed, this is evident in the results presented later (Figure 3; Table 2). Given these points, I consider it difficult to regard these "sampling times" as a fixed factor.

Consequences of these decisions. As noted above, stipulating "streams", and particularly "times within Before/After", as random factors has *considerable* implications for the power of statistical tests. Some tests done regarding "times(B/A)" as a fixed factor would usually[4] have greater power than tests with it as a random factor. For instance, the test for the *Before/After* × *Locations* interaction in Table 2 has 3 and 36 degrees of freedom if "times(B/A)" is random but 3 and 112 df if this factor is fixed: all other things being equal, the latter test would be more powerful. Thus, the estimates of power resulting from this approach (with "times(B/A)" random) are likely to be conservative, relative to estimates from alternate approaches[5].

- *Having the impact start half-way through the sampling program.* As noted, this decision was made solely to simplify the calculations required. Clearly, it is not feasible, when examining the power of alternative designs in complex situations, to explore *all* possible outcomes. The particular decision made here (*i.e.* impact starts half-way through the sampling program) does limit the range of outcomes which can be considered but still leaves many alternatives. Thus, the magnitude and duration of the impact, the length of the sampling program and the number of replicate samples can all be varied. Exploring changes in these should provide a good general idea of the potential power of different programs.

[4] As power is influenced by multiple factors, generalisations may not hold in particular circumstances.

[5] There is an alternate approach to ANOVA models in which this might not be the case (this alternate approach results in different tests for mixed-model designs like this one; see Quinn & Keough 2002). This approach is, however, not considered in standard analyses (*e.g.* Underwood 1991, 1992, 1993, 1994, 1997; Keough & Mapstone 1995) and is mentioned here only for the sake of completeness.

The special case of two streams: one impact and one control

If the design includes only two streams—one the impact stream and a second, "control" stream—then the assumption that "streams" is a random factor *may* be untenable. In many cases, the stream selected as the "control" might be chosen to match the "impact" stream as closely as possible. Indeed, much of the justification for comparisons between the two would seem to rely on the assumption that the two streams were very similar *except for the impact*. Thus, the comparison, in this case, would be of two similar streams, one of which had an "impact" applied to it. The factor "streams" is, then, being interpreted almost in an *experimental* fashion, with the "impact" being applied to one of two (replicate) streams. Under this rationale, both of the two possible levels of the experimental treatment—the "impact"—are included in the study and the factor would be regarded as fixed (the two levels being "impact present" and "impact absent"). As noted earlier, this would alter the way in which power was estimated.

This report includes only preliminary analyses of the data collected to date in the ARR monitoring program but even these analyses suggest that adopting the "two streams" design would be unwise. The analyses of dissimilarity (Table 2) and diversity-differences (Appendix 1: Table 5) both found evidence of short-term temporal variability in the behaviour of the streams (*i.e.* significant "T(BA) × L" interactions). These results indicate that the behaviour of the streams varies over the short-term, from one sampling time to the next, with perhaps some increasing, others stable and others decreasing (and see Figure 3; Figure 5). In these circumstances, it does not seem reasonable to select one of these "control" streams and consider it "similar to the impact stream, save for the impact".

The results of the analysis of diversity-difference are particularly pertinent here: this analysis found a *significant difference among streams in behaviour from "before" to "after" the "impact" even though the "impact" in this case merely represented a completely arbitrary division of the samples* (into two equal groups). In particular, stream "7J" showed a decrease, from "before" to "after", while the other streams were stable (or, perhaps, showed a slight increase). If stream 7J had been selected as the single, "control" stream, then an *entirely spurious "impact" might have been declared*[6]*!* This is the "fatal flaw" in the "single control" approach; and the reason why authors such as Underwood (1991, 1992, 1993, 1994) and Keough & Mapstone (1995) argue for multiple controls.

Analysis of differences in community structure

For the purposes of analysis, the 14 sampling times with full data were arbitrarily divided into two groups of 7 times: the first 7 representing samples taken "before" the impact and the second 7 representing those taken "after" the impact (although, of course, these samples are actually all "before" the impact).

The only significant term in this analysis is the "T(BA) × L" interaction (Table 2). This is consistent with the observations (Figure 3) and indicates significant short-term (time to time) temporal variation in each stream. In other words, mean upstream-downstream dissimilarity varies from time to time on each stream *but* all streams do not show the same pattern of change. For instance, three of the streams (CF, 7J, SC) on occasion exhibited high dissimilarities between upstream and downstream samples (>50) but did so at quite different times (CF in Feb99, 7J in Dec99, SC in Apr01; Figure 3). Upstream-downstream dissimilarity on each stream varied

[6] This result was in no way "contrived". To illustrate the types of analyses which might be done, and to estimate certain values required for the power analyses, I simply divided the 14 available, complete, sets of samples into two sets of 7. Further I did only two analyses; those reported here in Table 2 and Table 5.

through time, going up and down, but dissimilarity was increasing on some streams while it was decreasing, or stable, on others (Figure 3).

This is what Underwood (1993) called an "interactive system" in which the "control locations have variable short-term trends"[7]. This can, ultimately, effect the power of tests for impacts (and the sequence of tests done, if the procedure in Table 6 of Underwood (1993) is followed). For instance, if the "T(BA) × L" interaction is *not* significant, then the test for long-term changes due the impact will usually be more powerful (*i.e.* if *post hoc* pooling[8] is done, the "Before/After × Locations" term can be tested over the "Residual", which gives a more powerful test). Underwood (1993) discusses this and other issues related to testing for impacts in these designs.

Table 2. ANOVA on BC dissimilarities for 14 times at the four streams. The first 7 times were, arbitrarily, regarded as the "before impact" samples. The "den" column indicates which term was used as the denominator for that F-ratio. The only significant term is T(BA)xL; the interaction between locations and sampling times (within Before/After).

SOURCE	SS	df	MS	den	F	p-level
1-Before/After=BA	13.1	1	13.1	--	--	--
2-Times(BA)=T(BA)	1488.4	12	124.0	5	0.84	0.61
3-Locations=L	970.5	3	323.5	5	2.18	0.11
4-BAxL	880.1	3	293.4	5	1.98	0.13
5-T(BA)xL	*5342.9*	*36*	*148.4*	*6*	*2.67*	*0.00*
6-Residual	6214.0	112	55.5			
7-TOTAL	14909.1					

This analysis also provides estimates of one value critical for estimating power in this situation: Residual variance:

$$\text{Residual variance} = \sigma_e^2 = 55.5$$

Residual variance (σ_e^2) is simply estimated by the MS Residual.

Workbook notes. The value of the MS Residual must be derived from an analysis of relevant raw data, as an estimate of variation among the replicates is required to calculate power.

Stating the Alternate Hypothesis

As noted above, in order to calculate the power of a statistical test, the *alternate hypothesis* must be stated specifying the *effect size* of interest. In environmental situations there is always likely to be some dispute about the magnitude of effect which might be deemed "important".

Despite this, in simple situations actually specifying the alternate hypothesis, once a magnitude (or, perhaps, range of magnitudes) has been decided, is relatively simple. For instance, consider a situation in which the mean number of taxa at two sites—a control and an impact—is to be compared. The null hypothesis in this case would be:

[7] Strictly speaking, the test done in Table 2 is not the same as that in Table 6 of Underwood (1993) *but* the data analysed here were all collected *before* the disturbance so all locations are, effectively, "controls" and any distinction between "before" and "after" the disturbance is artificial.

[8] *Post hoc* pooling is a procedure (not supported by all statisticians) which may result in more powerful tests in some situations. It is described by Underwood (1981, 1997) and Quinn & Keough (2002). Underwood's (1993) paper describes it, and other procedures, with reference to MBACI designs.

- H_0: There is no difference between control and impact samples

If a difference of 10 species were deemed to be important, then the alternate hypothesis would simply be:

- H_A: The control and impact samples will differ by at least 10 species.

The situation with the standard "BACI" (control and impact sites sampled once before, and once after, the impact) design is a little more complex but still easily handled. Here it is a matter of specifying what magnitude of difference, between the control and impact sites, in the change from before to after is important.

The situation with more elaborate "MBACI" designs, incorporating several control sites and repeated sampling before and after the impact, is much more complex. For one thing, these designs may allow tests for several different types of impacts; for instance, short-term versus long-term impacts (see Underwood 1991, 1992, 1993, 1994 for details). For each of these different types of impacts there is a null hypothesis and, potentially, an alternate hypothesis which could be specified (for calculating power). Further, since the alternate hypotheses here do not constitute simple differences (*e.g.* between a control and impact) but *differences in patterns of temporal change*, specifying these hypotheses, and calculating power under different circumstances (*e.g.* different numbers of replicates), is not straightforward.

One approach which can be adopted in these situations, is to create "scenarios" which incorporate plausible outcomes *if an impact occurs* and explore the consequences—for power—of impacts of greater or less magnitude, greater and lesser numbers of replicates, and so on (see Keough & Mapstone 1995 for an example of this). Such an analysis, while not exhaustive, does indicate the likely consequences, for power, of designs with particular levels of replication and impacts of particular magnitudes. For instance, Figure 4 illustrates such a "scenario" using the ARR design and data (note that results for some sampling times were omitted from this graph because missing data made it impossible to calculate all the upstream–downstream comparisons).

Workbook notes. The graph shown in Figure 4 was copied directly from the "Model" page of the simulation workbook, after making suitable modifications to the values for the impact stream after Time 7. After this is done, power could be estimated directly.

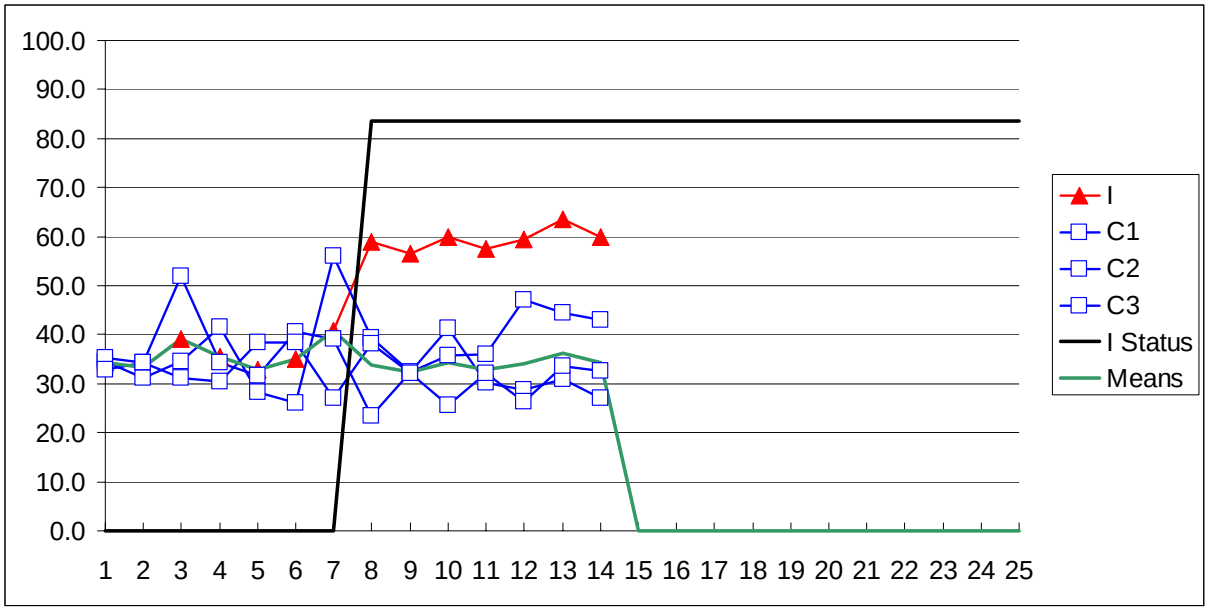

Figure 4. Example of a "scenario", using the ARR design and data, for estimating power. The values shown are the means for the 4 streams for 14 sampling times (times 1, 13 and 14, in the original data, were omitted because missing data did not allow all upstream-downstream comparisons to be calculated). In this example, the impact (indicated by the "I Status" line) has been arbitrarily started between Times 7 and 8 and the values for the impact stream before this time are equal to the *average of the control streams* and after this time are equal to the control averages plus 75% (the control average is shown by the "Means" line). The y-axis for the graph is Bray-Curtis dissimilarity (between randomly paired upstream and downstream replicates).

Calculating power for variations on the ARR design

The approach used here follows Underwood (1993): see in particular his Table 9 where he outlines the procedure for calculating power in MBACI designs similar to that described here. Note the following points:

- In general in ANOVA, the null hypothesis is tested by comparing estimates of two "sources of variation"—by dividing a numerator Mean Square by a denominator Mean Square—and determining whether the resulting F-ratio could have been observed by chance (*i.e.* whether or not $P<0.05$).
- Power is determined by making estimates of these two sources of variation and calculating the probability that the null would be rejected, given the magnitude of these two sources.
- Which sources of variation is used as the numerator, and which as the denominator, depends on the particular null being tested (also, which factors are fixed, etc.; see Table 3).
- For obvious reasons, the magnitude of these estimates has a major effect on the power of the test.
- For some tests, the denominator Mean Square is the "Residual", representing variation among replicates (see Table 3). This must be estimated from the data but its magnitude will not be affected by altering other sources (*i.e.* when exploring the consequences of impacts of different magnitudes).

- For other tests, the denominator Mean Square is the "T(BA) × L" (see Table 3): this is based on variation from time to time at each stream (or site). This is estimated from the means for each stream at each time (*i.e.* the values in Figure 4) and its magnitude *may* be altered when exploring the effects of different impacts. If this value is unrealistically small, then the power for some tests may be overestimated.

Table 3. Design of the MBACI analysis (following Underwood 1993). The df are calculated assuming 4 streams, sampled 7 times before, and after, the impact. The "MS den" column indicates which source of variation would be used as the denominator for testing the significance of that particular row (note that Underwood (1993) describes a sequential testing procedure which may provide more powerful tests but which considerably complicates the testing process). The sources of variation italicised are of particular importance for hypotheses about impacts.

#	SOURCES	df	#den	MS den
1	Before-After = BA	1		No test
2	Times(BA) = T(BA)	12	9	T(BA) x L
3	Locations = L	3	9	T(BA) x L
4	*I v C = IC*	*1*	*9*	*T(BA) x L*
5	Among C = C	2	9	T(BA) x L
6	BA x L	3	9	T(BA) x L
7	*BAxIC*	*1*	*9*	*T(BA) x L*
8	BAxC	2	9	T(BA) x L
9	T(BA) x L	36	16	Residual
10	T(bef)xL	18	16	Residual
11	*T(bef)xIC*	*6*	*16*	*Residual*
12	T(bef)xC	12	16	Residual
13	T(aft)xL	18	16	Residual
14	*T(aft)xIC*	*6*	*16*	*Residual*
15	T(aft)xC	12	16	Residual
16	Residual	112		
17	Total	167		

Procedure

Workbook notes. Using the workbook to estimate power is described below.

Using the workbook, the procedure is relatively simple.

- First, an estimate of the MS Residual is required. The easiest way to obtain this is by analysing a representative data set. For the BC dissimilarity data examined here, this estimate is 55.48. This value is entered into the appropriate cell on the "Setup" page.
- Second, a representative "scenario" is created by stipulating means for all sampling times for all streams: these values are entered into the appropriate cells on the "Model" page. The easiest way to start this process is by using the observed means. (The BC means are on the "Means" page and there are "macro" buttons on the "Model" page to facilitate copying and changing these.)
- Third, the other basic aspects of the design (number of sampling times, significance level (α), etc.) are entered into the appropriate cells on the model page.

- Fourth, and if necessary, the "Power" page is tidied (by deleting unwanted entries) and the appropriate column for the results is entered into the appropriate cell on the "Setup" page.
- Fifth, the "Calculate power" button on the "ANOVA power" page is pushed. The entered means are analysed giving an ANOVA table on the "ANOVA" page and power on the "ANOVA power" page: the latter results are also copied to the "Power" page.
- Finally, the scenario is changed—by introducing an impact—and the process repeated.

Critical points

- The procedure described by Underwood (1993) involves sequential testing which cannot be easily replicated here. Using the results on the "ANOVA" page in the workbook may very well not always give the same result as following his sequential procedure. Further, tests of some sources of variation may not be entirely appropriate, or relevant, if other sources are significant. Underwood (1993) must be understood if the workbook and these notes are to be used to give reliable results.
- In particular, the estimates provided by the workbook are based on the "default" ANOVA model for the ARR design (interpreted as described above). As noted earlier in this document, in some circumstances the "default" model may be adjusted, making it possible to do more powerful tests. The sequential procedure described by Underwood (1993) incorporates such adjustments.
- To simplify the calculations involved, the workbook assumes that the impact starts half way through the monitoring period: thus, if there were 10 sampling times, the impact is assumed to have started between times 5 and 6. This results in a balanced design with the same number of observations before and after the impact. Calculating power in other situations (*e.g.* with the impact starting between times 9 and 10) is rather more complicated[9]. There are, however, an endless number of possible alternative "impact scenarios" and it is impossible to explore them all. Estimating power in complex monitoring situations, such as this one, is always likely to be an indicative, rather than exact, process. For this reason, the "half way" limitation, although regrettable, may be acceptable.

Examples

The table below (Table 4) shows estimates of power under three scenarios. These results were obtained, following the steps above, by first using the observed means, then altering the means for the impact stream. The procedure is illustrated in more detail in Appendix 2.

Results for the first scenario ("Observed means") for the "BA × IC" test—which tests for a long-term change from before to after at the impact stream versus the controls—indicates low power (10%). This is not surprising as little long term change in the behaviour of the impact stream, relative to the controls, is evident (Figure 3).

Setting means for the impact stream equal to the average of the control streams results in the power of all tests involving Impact–Control comparisons (all sources with "IC") being 0, as there are no differences. The power to detect a large long-term change after the impact (Scenario c) is high (77%), although this will be a slight overestimate as the pattern of means used has reduced "T(BA) × L" variance somewhat (these values are not shown here but are in the workbook).

[9] (And, frankly, beyond my abilities.)

Table 4. Estimates of power of tests on BC dissimilarities under different scenarios: (a) “Observed means” = all means at observed values (as in Figure 3); (b) “Impact = control” = all means for the impact stream equal to the average of the controls at that time; (c) “Impact = 1.75 × control” = means for the impact stream *after* the impact equal to 1.75 times the average of the controls at that time (as in Figure 4). (Note that these results are in included the original spreadsheet.)

SOURCE	**Observed means**	**Impact = control**	**Impact = 1.75 × control**
Before-After=BA	No test	No test	No test
Times(BA)=T(BA)	2%	0%	0%
Locations=L	28%	26%	92%
I v C=IC	*17%*	*0%*	*77%*
Among C=C	24%	35%	35%
BA x L	24%	26%	92%
BAxIC	***10%***	***0%***	***77%***
BAxC	25%	36%	36%
T(BA) x L	97%	79%	79%
T(bef)xL	90%	83%	83%
T(bef)xIC	*8%*	*0%*	*0%*
T(bef)xC	92%	92%	92%
T(aft)xL	81%	21%	22%
T(aft)xIC	*69%*	*0%*	*0%*
T(aft)xC	52%	52%	52%

References

Downes BJ, Barumta LA, Fairweather PG, Faith DP, Keough MJ, Lake PS, Mapstone BD, Quinn GP (2002) *Monitoring Ecological Impacts: Concepts and Practice in Flowing Waters*. Cambridge University Press, Cambridge.

Keough MJ, Mapstone BD (1995) *Protocols for Designing Marine Ecological Monitoring Programs Associated with BEK Mills.* National Pulp Mills Research Program, Technical Report No. 11. CSIRO, Canberra.

McGuinness, KA (2002) Of rowing boats, ocean liners and tests of the ANOVA variance homogeneity assumption. *Austral Ecology* in press.

Quinn GP, Keough MJ (2002) *Experimental Design and Analysis for Biologists*. Cambridge University Press, Cambridge.

Stewart-Oaten A, Bence JR (2001) Temporal and spatial variation in environmental impact assessment. *Ecological Monographs* **71**:305–39.

Underwood AJ (1981) Techniques of analysis of variance in experimental marine biology and ecology. *Oceanographical & Marine Biological Annual Reviews* **19**:513-605.

Underwood AJ (1990) Experiments in ecology and management: their logics, functions and interpretations. *Australian Journal of Ecology* **15**:365-389.

Underwood AJ (1991) Beyond BACI: experimental designs for detecting human environmental impacts on temporal variations in natural populations. *Australian Journal of Marine & Freshwater Research* **42**:569-587.

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

Underwood AJ (1993) The mechanics of spatially replicated sampling programs to detect environmental impacts in a variable world. *Australian Journal of Ecology* **18**:99-116.

Underwood AJ (1994) On Beyond BACI – sampling designs that might reliably detect environmental disturbances. *Ecological Applications* **4**:3-15.

Underwood AJ (1997) *Experiments in Ecology: Their Logical Design and Interpretation Using Analysis of Variance*. Cambridge University Press.

Winer BJ, Brown DR & Michels KM (1991) *Statistical Principles in Experimental Design* (3rd ed). McGraw-Hill, New York.

Appendix 1: Other analyses

Analysis of Numbers of Species

As noted earlier, there is no factor "Reach" in analyses of dissimilarity because this variable is derived from comparisons of upstream and downstream samples: each upstream–downstream pair provides one value (or observation). Standard analyses of other variables—such as the abundances of individual taxa or number of species—would include the factor "Reach" because such analyses would also need to compare upstream and downstream reaches to test for impacts. Specifically, such analyses would test if *the difference between upstream and downstream* samples—in, say, number of species—changed from before to after the impact *at the impact site* relative to control sites.

Including "Reach", however, would increase the number of factors in analyses from three to four, making analyses, and calculations of power, more complex. It would also complicate comparisons of the power of tests on different variables because the degrees of freedom of tests for impacts would differ between three and four factor designs.

A simple way to resolve these problems is to analyse, not the "raw", say, observed numbers of species in upstream and downstream samples *but the differences between the upstream and downstream samples*. In other words, a procedure similar to that used for the Bray-Curtis dissimilarities is employed. Analyses of the dissimilarities are based on comparisons of randomly paired upstream and downstream samples. Analyses of differences between randomly paired samples in other variables—such as the abundances of individual taxa or number of species—could be done just as easily.

Analyses of the simple differences will, however, be complicated by the fact that the data will, inevitably, contain both positive and negative differences: the latter make it difficult to use the transformations normally employed to limit variance heterogeneity (but see McGuinness 2002). This problem can be resolved by analysing, not the *difference*, but the *percentage* that the downstream sample (of each pair) constitutes of the total number of species, or abundance, etc. For instance, analyses of the number of species in the samples would proceed by analysing the percentage of the total number of species comprised by the downstream sample:

$$\text{Downstream\%} = 100 \times \frac{\text{number in downstream sample}}{\left(\text{number in downstream} + \text{number in upstream}\right)}$$

The resulting analysis (Table 5) is similar in form to that derived from dissimilarity calculations.

It is important, however, to recognise that such analyses are *not* testing exactly the same hypotheses as four-factor analyses done by including "Reach". The three-factor analyses test hypotheses about *differences in percentages* whereas the four-factor analyses test hypotheses about *differences in absolute numbers*. Any impact which causes the upstream and downstream sites to differ[10] should, if it is sufficiently large, be detected by both analyses but the two types of analyses may differ in power.

[10] More precisely: "any impact which causes the differences between upstream and downstream sites to change from before to after the impact".

Table 5. ANOVA on percent of total numbers of species at downstream site (see text for further details) for 14 times at the four streams. The first 7 times were, arbitrarily, regarded as the "before impact" samples. The "den" column indicates which term was used as the denominator for that F-ratio.

SOURCE	SS	df	MS	den	F	p-level
1-Before/After=BA	33.9	1	33.9	--	--	--
2-Times(BA)=T(BA)	731.1	12	60.9	5	1.17	0.34
3-Locations=L	177.7	3	59.2	5	1.14	0.35
4-BAxL	*587.7*	*3*	*195.9*	*5*	*3.78*	*0.02*
5-T(BA)xL	*1867.6*	*36*	*51.9*	*6*	*1.60*	*0.01*
6-Residual	3623.6	112	32.4			
7-TOTAL	7021.5					

The significant terms in this analysis are the "BA × L" and "T(BA) × L" interactions (Table 5). The latter term indicates significant short-term (time to time) temporal variation in each stream. In other words, mean downstream% varies from time to time on each stream *but* all streams do not show the same pattern of change (Figure 5). The system is again "interactive" (*sensu* Underwood 1993) with the four streams showing significantly different behaviour through time.

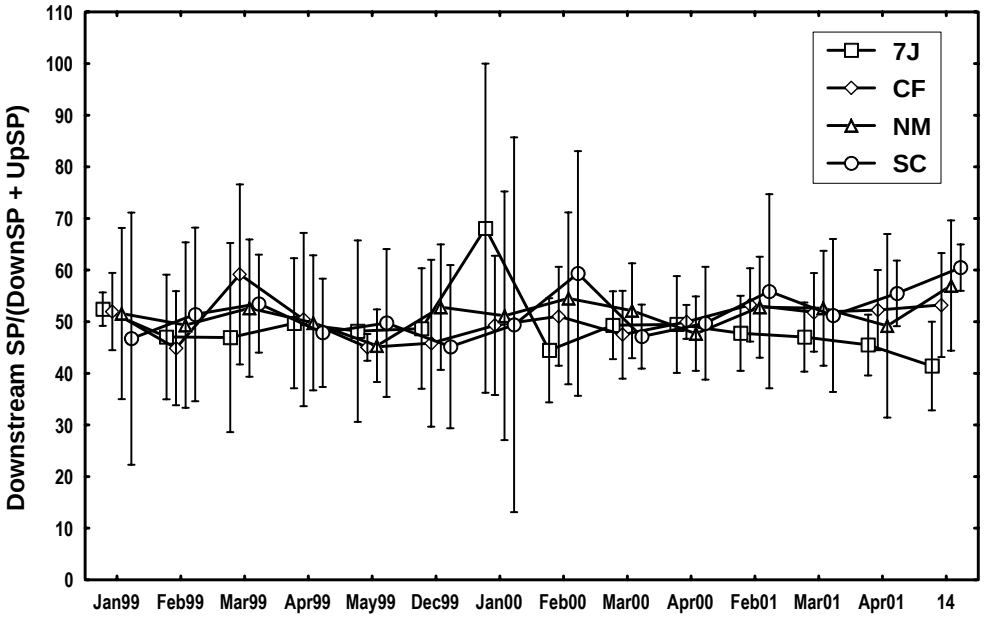

Figure 5. Mean downstream% for replicates (randomly paired) at 14 times on each stream. The values shown are the means for the 4 streams for 14 sampling times.

The significant "BA × L" interaction indicates that all streams *do not* show the same trends after the "impact" as before (remembering that there is no actual "impact" in these data). This indicates that the longer-term trend differs among streams (Figure 6). This is not obvious in Figure 5 but is apparent when the means for each stream "before" and "after" the "impact" are examined (Figure 6): downstream% decreases on stream 7J, but changes little on the other three streams.

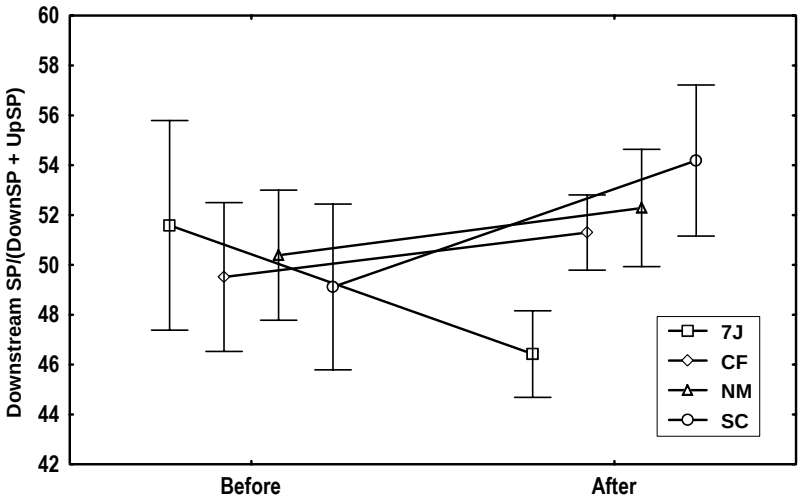

Figure 6. Mean downstream% for the 4 streams “before” and “after” the impact. (Note that the distinction is arbitrary as there was no impact.)

Appendix 2: Examples

Example 1

Objective

Assume that the objective is to calculate the power of the test for a long-term change, coincident with the impact, where dissimilarity at the impact sites increases 70% relative to controls. Assume the following conditions:

- Change: dissimilarity at the impact sites increases 70% relative to controls
- Replication: 3 samples at 4 streams (3C; 1I) at 3 times before and after (6 total times).
- Significance level: use standard $\alpha = 0.05$.

Procedure

Proceed as follows (after starting the spreadsheet) on the "Setup" page:

1. On the "Setup" page, enter the significance level (0.05), total number of sampling times (6) and number of replicates (3; in the cells with white text on a RED background).
2. Also on this page, enter the "Residual MS" which has been previously calculated from analyses of pilot data (55.48 for the dissimilarity data used here).
3. Finally, if desired, alter the column (on the "Power") page where results will be copied for later reference. (Note that the "Power" results page records some basic information about each analysis but *does not* record the means used; this would be too cumbersome.)

Next, on the "Model" page, do the following:

4. Clear the existing means by pressing the "Clear means" button.
5. Copy in the means to be used to calculate power. In this case we will start with the existing dissimilarity means (press the "Copy 'means' data" button).
6. Delete the means after Time 6 using standard *Excel* actions (this is not strictly necessary as the means for Time 7 onwards would not be used anyway, but it limits confusion).

Now, still on the "Model" page, we need to create the desired "impact". There are many possible ways of doing this. Two obvious, and plausible, scenarios are to (1) use the existing impact means for Times 4, 5 and 6 but increase each of these by 70%; or (2) use the average of the controls for Times 4, 5 and 6 but increase each of these by 70% (or do both). For illustration, we'll adopt the second option.

7. Pressing the "Copy control mean to I" button calculates the mean of the 3 controls at each time and copies this to the Impact stream.
8. Increase the (new) impact means for Times 4, 5 and 6 by 70% (use a calculator or copy the values to another spreadsheet, increase them, then copy them back). The average of the controls at Time 4 is 35.1: multiplying by 1.7 gives 59.7, which is entered into that cell. Values for the next two means are 55.8 and 59.7.

Finally, we move to the "ANOVA Power" page and calculate power.

9. On the "ANOVA Power" page, press "Calculate power".

10. The power for each test is displayed in the "Power" column (yellow on BLUE, fourth from the right). (Note that, as emphasised elsewhere in this document, the power displayed is calculated using the "default" test of the hypothesis: more powerful tests may be possible if the procedures of Underwood 1993 are followed.)
11. The power for the "BAxIC" test—the test for a long-term change at the impact site, coincident with the impact—is 76% (which is not bad).

It may be worthwhile running this sample example, but using option (1) above: using the existing impact means (but increasing each by 70% after the impact). Steps 1 to 6 are the same as before.

7. In this case do not press the "Copy control mean to I" button, as we want to use the existing impact means as the starting point for the "impact scenario".
8. Increase the impact means for Times 4, 5 and 6 by 70% to 64.4, 55.9 and 70.4, respectively.

Finally, we move to the "ANOVA Power" page and calculate power.

9. On the "ANOVA Power" page, press "Calculate power".
10. The power for each test is displayed in the "Power" column (yellow on BLUE, fourth from the right).
11. The power for the "BAxIC" test is now 70% (which is still not bad).

The power has decreased slightly: why? Using the existing impact means as the starting point for the "impact scenario" increases *short-term temporal variability among locations* (the "T(BA)xL" term) slightly and this reduces, slightly, the power of the test for long-term changes.

Example 2

Objective

Assume that the objective is now to calculate the power of the test for a long-term change, coincident with the impact, where dissimilarity at the impact sites increases 35% relative to controls. A secondary objective is to determine how many sampling times are needed to give power = 70% (*i.e.* the same as above). Assume the following conditions:

- Change: dissimilarity at the impact sites increases 70% relative to controls
- Replication: 3 samples at 4 streams (3C; 1I), *initially* at 3 times before and after (6 total times).
- Significance level: use standard $\alpha = 0.05$.

Procedure

Proceed as follows (after starting the spreadsheet) on the "Setup" page:

1. On the "Setup" page, enter the significance level (0.05), total number of sampling times (6 to start with) and number of replicates (3; in the cells with white text on a RED background).
2. Also on this page, enter the "Residual MS" which has been previously calculated from analyses of pilot data (55.48 for the dissimilarity data used here).
3. Finally, if desired, alter the column (on the "Power") page where results will be copied for later reference.

Next, on the "Model" page, do the following:

4. Clear the existing means by pressing the "Clear means" button.
5. Copy in the means to be used to calculate power. In this case we will start with the existing dissimilarity means (press the "Copy 'means' data" button).
6. Delete the means after Time 6 using standard *Excel* actions (this is not strictly necessary as the means for Time 7 onwards would not be used anyway, but it limits confusion).

Using option (1) enter the impact means increased by 35%.

7. In this case do not press the "Copy control mean to I" button, as we want to use the existing impact means as the starting point for the "impact scenario".
8. Increase the impact means for Times 4, 5 and 6 by 35% to 51.1, 44.5 and 55.6, respectively.

Finally, we move to the "ANOVA Power" page and calculate power.

9. On the "ANOVA Power" page, press "Calculate power".
10. The power for each test is displayed in the "Power" column (yellow on BLUE, fourth from the right).
11. The power for the "BAxIC" test is now 53% (which is not great).

Then attempt the second objective.

1. On the "Setup" page, enter an increased number of sampling times—try 14 here (just a guess, which happens to match the available number of samples).

Then jump to Steps 3–5 and do these as before.

7. In this case do not press the "Copy control mean to I" button, as we want to use the existing impact means as the starting point for the "impact scenario".
8. Increase the impact means for Times 8 to 14 by 35% (61.1, 34.1, 59.6, 52.5, 39.9, 56.0, 72.2).
9. On the "ANOVA Power" page, press "Calculate power".
10. The power for each test is displayed in the "Power" column (yellow on BLUE, fourth from the right).
11. The power for the "BAxIC" test is now 62%.

We need to increase the number of sampling times. The easiest way to do this is to start using the means from Time 1 again: that is, for Time 15, use Time 1; for Time 16, use Time 2, and so on. (Other plausible ways of extending the series could be devised: for instance, by randomly selecting means from the existing series to extend it.)

1. On the "Setup" page, enter an increased number of sampling times—try 28 here (just a guess, which happens to be double the available number of samples).

Then jump to Steps 3–5 and do these as before.

7. We need to extend the series. Simply select the existing means for Times 1 to 14, for all streams, and copy them at Time 15 to extend the series to Time 28.
8. Increase the impact means for Times 15 to 28 by 35%.

9. On the "ANOVA Power" page, press "Calculate power".
10. The power for each test is displayed in the "Power" column (yellow on BLUE, fourth from the right).
11. The power for the "BAxIC" test is still about 62%.

This result might seem odd. It is arises because extending the series, using the existing means for the impact stream, has increased short-term variability: as a consequence, power did not increase. This emphasises the importance of variability in estimates of power and reinforces the need to consider these as "ball park" estimates in this situation (in an experiment with clearly defined treatments, this would not usually be the case).

Extending the series (using the method above, of copying means from Time 1), gave a power of 62% for 14 times, 57% for 20 times[11], 62% for 28 times and 73% for 42 times. A power of about 70% would, therefore, require some 30 to 40 sampling times *but this is highly dependent upon the pattern of variability*. And note (again) that more powerful tests might well be possible if the step-wise procedure of Underwood (1993) is followed.

[11] Note that the impact stream at the last sampling time, 14, had a particularly (and unusually) high dissimilarity value. The presence of this value, increased by 35%, increases temporal variability.

Appendix 3: Summary of response to comments on first draft

Response of the consultant, Dr McGuinness, to comments on an earlier draft from Dr Chris Humphrey, ***eriss***

1. *It would be useful to provide text and worked example for a simple one control vs single impact stream scenario*

As discussed in the e-mail, I've reconsidered this issue. The document now contains a special section ("The special case of two streams") which discusses this scenario *and recommends against adopting it*. Several of the reasons for this are presented in the report (I expanded the discussion under "Implications" and introduced a new section on the "two stream design"). I'll add a few other comments here:

- It is not simple for two reasons. First, in the "two stream" scenario, the "streams" factor should probably be considered a fixed factor (rather than a random factor as it is in the MBACI design). This changes the entire way in which power is calculated and makes those calculations more complex (in this multi-factorial situation). In particular, in the current design (as I've interpreted it) power is calculated just using the central F-distribution (as shown in Underwood's papers). The two stream case would require some calculations with this but some with the non-central F-distribution (which is not a native Excel function). Further, the Underwood design requires multiple controls and, as the spreadsheet has been constructed for this MBACI design, much of it stops working if there is only one control.
- This means that incorporating estimates of the "two stream" scenario would not be a simple process. It would involve altering a fair amount of the code in the workbook to get it to work with the "single control" case AND to sometimes use the non-central F-distribution. I can see this taking QUITE a long time, and I've already spent more time on this part of the job than I thought I would (due, I freely admit, to underestimating the complexity of the task; something I'll not do again in a hurry!). I'm reluctant to take on this additional task.
- The most important reason, however--and, ultimately, the one that matters--is that I consider the "two stream" design to be a very poor choice for a monitoring program (and that is putting it mildly!). I regard the analyses I did in the report as only "preliminary" (after all, I just looked at two things and didn't use "years" as a factor, although I don't think that this would affect the results). EVEN so, the results illustrate the problems which could arise if a "two stream" design were adopted (there is text in the report which specifically discusses this; although it could be altered/removed if you don't think it relevant). With a "two stream" "single control" approach it would be far too easy for spurious "impacts" to be detected.

2. *It would also be very useful to provide text and worked example for the situation where sample size - in this case temporal replication because streams are already allocated - is to be determined for given effect size, alpha and beta (20% for latter say).*

I've added examples in Appendix 2 (to avoid cluttering up the main text).

3. *Some additional text describing the extreme difficulty in estimating power where temporal replication, before and after, is unequal. Not an issue with the actual analysis of data though. This is the most likely scenario in monitoring. Again, can this (power) be determined by manual iteration [much as 2] and if so, can this be described and if necessary, explained by a worked example?*

I've modified, and expanded, the text under "Procedure" to discuss these issues more fully. Estimating power with unbalanced numbers of samples before and after is, frankly, beyond my abilities. Sorry—can't help with this one.

4. *describe more fully perhaps, implications of significant interactions that were found in the analysis where the data are split in two equal time periods. Does this "matter" and what are the limitations if any?*

This is now discussed more fully under "analysis of differences in community structure".

5. *p.11, dot-point, "third", includes significance level; assume this is alpha. Beta as well?*

Yes; significance level is alpha (clarified in the text). Beta is 1 – power (*i.e.* 1 *minus* power; expressed as a proportion). So beta is an output, not an input.

6. *text under table 5 should refer to differences not dissimilarity I think. Again as for 4, what is the significance of this?*

I've fixed the error in the text and expanded the explanation (which now includes figures).

7. *while obvious, it should be mentioned that in univariate [non-dissim] approaches where MBACI and not MBACIP are employed, the paired site controls within a stream are not independent and so the 'P' design is the most suitable - perhaps ...*

I added comments relating to this issue to the "Reach" bullet (under "The ARR Design") and in a foot note.

8. *would be useful to be more explicit about the lowest level of replication, logged data, pairwise combinations of dissims – probably already in but may need rearranging(??)*

I've added a short paragraph on "Replication in this design" which discusses this more explicitly than was the case in the original document.

9. *could be useful to incorporate further fixed vs random variable discussion in the light of the Downes et al (2002) book. Has profound implications. Emphasising 'worst-case' is important (for random) – text could be strengthened though haven't checked to see whether this is already the case.*

I have expanded considerably the text under "Implications of the design used".

10. *data in the two figures (dissim plotes) not an exact match*

The explanation for the difference was simple when I remembered/realised what it was. Results for the three control streams are the same in both graphs. Results for the impact stream in the second graph (originally Figure 2) after the "impact" showed an increase, reflecting one particular scenario. This much was obvious. What I forgot (even though it said it in the original caption) was that the means for this stream *before the impact* were also altered: to be equal to the mean of all three control streams. (This scenario results in no difference between the impact stream and the average control stream before the impact.) Thus, there was no error. I have added some explanatory text to the caption for the first graph to explain the difference.

Other notes

I've removed reference to the "other workbook" as I only included this in the previous draft to describe why I had altered my approach.