DRAFT, 25 May 2006

Report on outcome of EnviroLink Project Hawkes Bay Regional Council Advice No. 23

Trend analysis for macroinvertebrates in rivers: Harmonising statistical significance and ecological significance

Outcome of a meeting held at NIWA Hamilton on 1 May 2006

Present

Mr Brett Stansfield	environment scientist	Hawkes Bay Regional Council*
Dr Kevin Collier	aquatic ecologist	Environment Waikato, Hamilton
Mr Graham McBride	statistican/modeller	NIWA, Hamilton
Dr Murthy Mittinity	statistician, Post Doc	NIWA Hamilton
Dr Mike Scarsbrook	aquatic ecologist	NIWA, Hamilton
Dr John Stark	stream ecologist,	Cawthron Institute, Nelson

Background

Statistical hypothesis tests are in common use by water resource management agencies when examining trends in river benthic invertebrates. To date most of those analyses have been based on a relatively short record length (e.g., 8 years, Collier & Kelly 2005), with annual or biannual sampling. The principal outcome of such a test is known as the *p*-value.† A "statistically significant" result is announced if that value is "small" (i.e., less than the "significance level, usually taken as $\alpha = 0.05$). Often the word "statistically" is omitted, creating the impression that what has been detected is somehow environmentally significant.

In fact a statistically significant trend result may be quite different from trends that environmental scientists would view as environmentally significant. To see why, consider a case where a Regional Council has monitored stream sites biannually for 10 years, giving 20 data per site, and that a trend has been "detected" (because *p* was calculated as 0.028, so that the result is statistically significant). Now imagine that their sampling frequency was in fact annual, so that they had only 10 data. The *p* value would most usually then be greater than 0.05, and so a trend would *not* be detected.‡ This demonstrates that *p*-values depend, *inter alia*, on "sample size" (i.e., the number of data available for trend analysis). While they also depend on the variability in those data, the general pattern that emerges for trend analyses§ is that:

A. For small sample size, *p*-values tend to be "large" and so tests will routinely fail to detect environmental important trends.

B. For large sample size, *p*-values tend to be "small" and so tests will routinely detect environmental trivial trends.

---

* Mr Stansfield attended by conference telephone link; his intended travel to Hamilton was thwarted by a slip on SH5.

† Some implementations of such tests don't formally report a *p*-value and compare it with the significance level; rather, they involve comparing a test statistic with a "critical value". The two procedures are entirely equivalent.

‡ For a full discussion of this aspect see McBride (2005).

§ A number of caveats need to be made when discussing this in full, such as the effect of serial correlation, but for general purposes items A and B are defensible.

This feature has been recognized in recent endeavours by Collier (2005) and Stark & Fowles (2005). (Most usually it is not recognized.) A key feature of those reports has been the preparation of a table that seeks to harmonise the two "significances".

These reports also adopt the False Detection Rate (FDR) method when making multiple comparisons (as advocated by McBride 2005). This is a sensible way in which to account for inflating Type I errors (falsely rejecting the hypothesis being tested) when doing many trend tests (e.g., for a number of metrics at a given site, and/or at multiple sites).

The conventional method (typified by "Bonferroni corrections") becomes excessively censoring of information as the number of comparisons is increased, which is an absurd feature.[**] FDR methods completely avoid this problem, by controlling the *proportion* of statistically significant results that may be in error, whereas Bonferroni method guards against the possibility of making *one* Type I error—regardless of the number of comparisons being made. The essence of the FDR procedure is to rank all the *p*-values below the significance level, and then successively step down from the largest until a defined criterion is met. All *p*-values up to this cutoff value are then declared to be statistically significant. This ranking feature makes the FDR a nonparametric procedure. It always produces more statistically significant results than the Bonferroni method, and always less than when using the "$p < \alpha$?" criterion for each and every comparison.

Meeting outcome

A suitable definition of "trend" is needed. We agreed on "A tendency to increase or to decrease over time, in fashion that is meaningful to an environmental professional."

We focused on Table 2 in Collier & Kelly (2005), entitled "Trend classes used to define ecological and statistical significance of relationships for different sample sizes…" This table focuses on Spearman's rank correlation coefficient ("r_s") as the trend assessment statistic—though the same ideas could be applied to other statistics, such as Pearson's linear correlation coefficient, or Kendall's tau (as used by Stark & Fowles).

That table is as follows

	Trend class			
n	**Stable**	**Possible**	**Probable**	**Clear**
5–9	$r_s \le 0.50$	$0.50 > r_s < 0.7$	$0.70 \ge r_s \le r_{s(FDR)}$	$r_s > r_{s(FDR)}$
10–16	$r_s \le 0.50$	$0.50 > r_s < r_{s(\alpha = 0.05)}$	$r_{s(\alpha=0.05)} \ge r_s \le r_{s(FDR)}$	$r_s > r_{s(FDR)}$
>16	$r_s \le r_{s(FDR)}$	NA	NA	$r_s > r_{s(FDR)}$

NA = Not Applicable.

In this table *n* is the sample size, r_s is Spearman's rank correlation coefficient (also called Spearman's rho), $r_{s(\alpha=0.05)}$ is the critical value for r_s for a two-sided hypothesis test that the true value of the coefficient is zero at the 5% significance level (with no

[**] Perneger (1998) argues that employing the Bonferroni method "creates more problems than it solves" and "defies common sense".

FDR correction), and $r_{s(FDR)}$ is the critical value of that coefficient, again at the 5% level, but accounting for the FDR correction.

The numbers 0.50 and 0.70 are values that Dr Collier regards as noteworthy, even in the absence of a statistically significant result.

Modifications to this approach
The meeting noted that while this table does account item A of the *p*-value behaviour, it does not deal with item B (because the value of $r_{s(FDR)}$ in the south-east cell will continue to decrease with sample size). It also mixes up FDR and non-FDR methods, whereas consistency in the use of FDR seems desirable. For technical reasons, the second cutoff should be 25 (cf. 16), some "." signs should be changed to "<", and absolute values of r_s are what should be used. Finally, it was agreed that three (cf. four) "trend classes" would be appropriate: Stable, Possible and Clear.

The *general* form of these modifications is depicted in Figure 1 below. Note that the *particular* form of the proposal cannot be simply displayed on a figure because, as already noted, FDR methods are nonparametric and so the critical values of r_s are unique in every case.

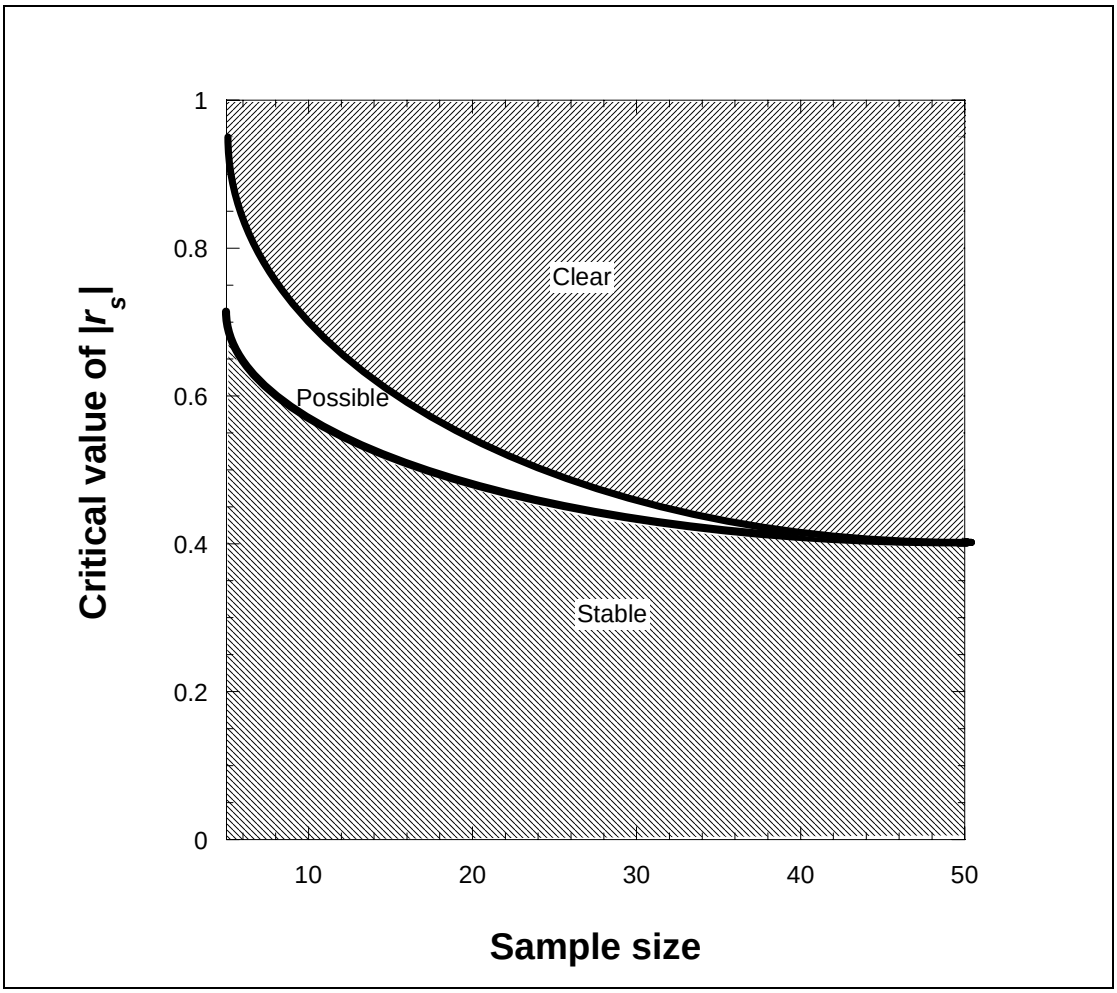

Figure 1. Idealised decision rule scheme for trend analysis using Spearman's rank correlation statistic.

The revised Table is shown below.

	Trend class		
n	**Stable**	**Possible**	**Clear**
5-9	$\|r_s\| \le 0.50$	$0.50 < \|r_s\| \le r_{s(\alpha=0.10)}$	$\|r_s\| > r_{s(\alpha=0.10))}$
10-25	$\|r_s\| \le r_{s(\alpha=0.10)}$	$r_{s(\alpha=0.10} < \|r_s\| < r_{s(\alpha= 0.05)}$	$\|r_s\| > r_{s(\alpha=0.05)}$
>25	$\|r_s\| \le r_{s(\alpha=0.05)}$	$r_{s(\alpha=0.05)} < \|r_s\| < r_{s(\alpha= 0.01)}$	$\|r_s\| > r_{s(\alpha=0.10)}$ but if $n > 40$, $\|r_s\| > 0.4$

$|r_s|$ denotes the absolute value of r_s, to account for the possibility of a downward trend. All critical r_s values should be adjusted by the FDR method. The most accurate table for critical values of r_s is to be found in Zar (1984, and subsequent editions). For example, the table in Snedecor & Cochran (1980) is much less accurate.

Ancillary outcomes

The following points were noted:

- It is always desirable to seek to explain the reasons for trends occurring.
- The mere fact that trends in means or medians have not been detected, even when using the table above, is not necessarily an indication that there are no important changes. It could be that variability in the population is increasing, and that also may be of environmental concern.

Follow-on work

The meeting agreed that this scheme now needs to be trialled on real datasets. Drs Collier, Stark and Mr McBride will attend to that in coming weeks, using existing funding sources. The end result is intended to be a manuscript submitted to an environmental science journal.

References

Collier, K.; Kelly, J. (2005). Patterns and trends in the ecological condition of Waikato streams based on the monitoring of aquatic invertebrates from 1994 to 2005.

McBride, G.B. (2005). *Using Statistical Methods for Water Quality Management: Issues, Problems and Solutions*. Wiley, New York.

Perneger, T.V. (1998). What's wrong with Bonferroni adjustments? *British Medical Journal* **316**:1236–1238.

Snedecor, G.W.; Cochran, W.G. (1980). *Statistical Methods*. 7th ed. The Iowa State University Press. Ames, IA.

Stark, J.D.; Fowles, C.R. (2005). An approach to the evaluation of temporal trends in Taranaki state of the environment macroinvertebrate data. Cawthron Report No. 1135, prepared for Taranaki Regional Council.

Zar, J.H. (1984). *Biostatistical Analysis*. 2nd ed. Prentice-Hall, Englewood Cliffs, NJ.