Southern California Regional Monitoring Project

The Reference Envelope Approach to Impact Monitoring

author: Robert W. Smith
EcoAnalysis, Inc.

under contract to: Southern California Coastal Water Research Project

sponsored by: U.S. EPA, Region IX
grant No.: X-009904-01-0
Section 104(b)(3) - Clean Water Act
NPDES Program - Regional Monitoring Project

March, 1995

EcoAnalysis

Southern California Regional Monitoring Project

The Reference Envelope Approach to Impact Monitoring

author: Robert W. Smith
EcoAnalysis, Inc.

under contract to: Southern California Coastal Water Research Project

sponsored by: U.S. EPA, Region IX
grant No.: X-009904-01-0
Section 104(b)(3) - Clean Water Act
NPDES Program - Regional Monitoring Project

March, 1995

EcoAnalysis

TABLE OF CONTENTS

ABSTRACT

Statistical tests used to differentiate environmental impacts from naturally occurring changes should ideally be based on pertinent data from before and after the onset of the impacting activity. Because it is often difficult to obtain sufficient data before the impact, we explore methods that can be used to test for impacts when only after-impact data are available. We first discuss how a proper statistical test needs to incorporate replicate, spatial, and time by space interaction variance components in the background error variance. In order to obtain a background error variance containing these variance components, a sampling design with randomly located reference stations is required. When there is no spatial autocorrelation in the indicator (dependent variable) values, reference stations located systematically rather than randomly can also be appropriate.

The proposed tests are based on the assumption that, in the absence of an impact, the indicator values at a potentially impacted station (an impact station) will appear as if they were indicator values drawn from the population of reference stations. Two types of tests are proposed. The first type involves t statistics testing the null hypothesis that the impact station mean equals the mean of the reference station distribution. The second type of test defines a statistical interval where impact stations with indicator values outside the interval bounds are considered as impacted. The null hypothesis for this type of test is that the indicator value at the impact station is equal to a chosen percentile of the population of reference stations, and as such, allows for more conservative assumptions regarding the relationship between the impact station and the population of reference stations. Here we propose the use of a tolerance interval and a Hampel outlier identifier. We express both the first and second types of tests as statistical interval bounds defining the edge of a "reference envelope". Impact stations with indicator values outside the edge of the envelope are considered impacted, and those with indicator values inside the envelope are considered unimpacted. All the proposed tests are based on the assumption that the distribution of reference stations approximates a normal distribution. Other critical assumptions are reemphasized in the final discussion.

We used systematically-sampled marine benthic data from the Southern California Bight to compute the values of 13 indicators that are sensitive to organic enrichment from sewage outfalls in the area. Using these indicators, we tested for non-normality of the reference station distribution and for spatial autocorrelation. In

addition, variance components for the indicators were estimated for input into a simulation model used to evaluate the performance of the tests. In the absence of spatial autocorrelaton, the simulation results showed that the actual type-1 error of the tests was generally close to the nominal type-1 error, indicating that the tests properly incorporate the variance components. In the presence of spatial autocorrelation, the actual type-1 error tended to be below the nominal type-1 error, indicating a loss in test sensitivity.

We applied the methods to the benthic data to detect impacts from the sewage outfalls. Two types of results output are demonstrated. The first type displays, for all indicators and in a single figure, the position of all stations in relation to the reference envelope edge for a single statistical test. The second type of display focuses on a single indicator at a time, showing the envelope edges from all tests and the indicator values in a single table. The problem of pseudoreplication with statistical tests detecting treatment effects in space is discussed, and it is shown how the proposed tests avoid pseudoreplication by incorporating variance from spatial differences in the background error variance.

INTRODUCTION

Often a primary objective of environmental monitoring programs is to estimate the environmental effects of a particular anthropogenic activity. Fulfilling this objective requires the ability to differentiate naturally occurring biological or physical changes from changes due to the activity in question. With this goal in mind, it is important that the sampling and statistical designs of the monitoring program are up to this task. Often, the sampling designs include sampling locations close to the anthropogenic activity in question (e.g., a dump site or sewage outfall), and one or more locations farther from the activity that are assumed to be unaffected (by the activity). Here we refer to the potentially impacted sampling locations near the activity as *impact stations*, and sampling locations in an area assumed to be unaffected by the activity as *reference stations*. We refer to the pertinent anthropogenic activity as an *impact*, with the realization that we are only talking about a *potential* negative impact, since often activities will have no impact or at least no negative impacts. We refer to a parameter measured to quantify the biological or environmental changes of interest as an *indicator*.

A direct comparison of indicator values at an impact station with the indicator values at a reference station is not necessarily sufficient to differentiate changes due to the impact from natural changes, since the indicator values at the two stations could differ naturally even without the presence of the impact in question (Hurlbert 1984). More complex monitoring designs, involving samples at reference and impact stations at multiple times both *before* and *after* the onset of the impacting activity, are better able to differentiate among natural and anthropogenic changes (Eberhardt 1976, Green 1979, Skalski and McKenzie 1982, Bernstein and Zalinski 1983, Stewart-Oaten et al. 1986, Millard and Lettenmaier 1986, Faith et al. 1991, Green 1993, Underwood 1994).

Obtaining sufficient data before the onset of the activity is often difficult (Osenberg et al. 1994) or impossible, so other approaches utilizing only *after-impact* data need to be developed. Methods commonly applied to after-impact data include various types of pattern analysis where the patterns of changes in the indicator values are correlated with proximity to the impact in space and/or time (Smith and Greene 1976, Stull et al. 1986, Eberhardt and Thomas 1991, Osenberg et al. 1994). This approach has been very informative in understanding the patterns of change in the vicinity of an impact, but with increasing distance from the impact, the indicator gradients caused by the impact become more subtle, and it is not always clear at which point the impact disappears,

i.e., reference or natural conditions exist. In fact, for some monitoring designs, all the sampled locations are potentially impacted, so there are no directly comparable data from the program indicating how the indicators might appear in natural, unaffected locations.

In this paper, we suggest and evaluate statistical methods that may be useful in differentiating changes due to an impact from changes due to natural background variability when *no (or insufficient) before-impact data are available*. These methods require that the sampling design include an effectively random sample of multiple reference stations (from an underlying reference station population or distribution) that represent natural conditions, and can be used as a basis of comparison with potentially impacted conditions.

Before describing the statistical techniques, we discuss the different sources of natural variability that can contribute to the variance of the reference station distribution. This information is important, since valid statistical tests must properly incorporate these sources of variability.

We then describe two categories of methods, each with different null hypotheses reflecting different levels of assumptions concerning the relationship between the reference and impact stations. Methods in the first category assume that all stations, including the impact station, originate randomly from the same distribution of reference stations. Here we propose modified t statistics to test the null hypothesis that the mean indicator value at the impact station is equal to the mean of the reference station distribution. The second category of methods allows for more uncertainty concerning the relationship between the reference and impact stations. Here one can test for worst-case scenarios in which the impact station originated from the tail (in the direction of impact) of the distribution of reference stations. These methods include a tolerance interval and an outlier identifier, and evaluate the null hypothesis that the impact station mean is equal to a particular *percentile* of the reference station distribution.

Methods in the second category directly define one-sided statistical intervals (Vardeman 1992). The t statistics in the first category can also be expressed as a statistical interval, so with both approaches we reject the null hypothesis of no impact when the mean indicator value at an impact station is outside the computed interval bound for the particular test. Thus, each test defines what we call a "*reference envelope*". Unimpacted stations and reference station will tend to be found inside the envelope, and impacted stations will tend to be found outside the envelope. The border between the inside and outside of the envelope (the "envelope edge") is defined by the bound of the statistical interval from one of the proposed tests.

To illustrate and evaluate the proposed methods, we utilize benthic marine data from the Southern California Bight to compute parameters (variance components) for input into a simulation model that evaluates the performance of the various statistical tests. We also use these data to demonstrate applications of the methods.

The techniques were developed as part of a project to consider possible approaches to the analysis of data from a planned regional monitoring program in the Southern California Bight (EcoAnalysis et al. 1993, SCCWRP and EcoAnalysis 1993). This motivated our choice of marine benthic monitoring data from this area. We emphasize, however, that the proposed techniques would be applicable wide range of environmental applications.

METHODS

Variance Components

To set the stage for the introduction of specific statistical tests, we first identify various sources of natural background variability that need to be considered in designing and evaluating a proper statistical test. Each different source of variability is called a *variance component* (Sokal and Rohlf 1981, Searle et al. 1992). If these sources of natural background variability can be quantified, then the statistical test can compare the background variability with the variability between reference and impact stations. The variability between the impact and reference stations will need to be sufficiently greater than the expected natural background variability for us to confidently conclude that there might be actual changes due to the impacting activity.

To describe the sources of variance that are pertinent to the present situation, we first focus on the variance components expected to affect the *variance of the means of reference stations*, since this variance represents the natural background variability among unimpacted loctions. We define this variance of the reference station means as $S^2_{\bar{Y}}$, which actually is an estimate of the variance of the underlying distribution or population of reference stations at a single point in time. For simplicity, we describe the variance components and use simulated examples in term of reference station *pairs*, but the same concepts apply to more than two stations.

Replicate (small-scale spatial) variance component (σ_R^2)

When sampling at a station with one or more replicates, the measured indicator values will, of course, usually vary among the replicates. Thus, we would expect computed means for sets of replicates taken from the same station at the same time to differ (Figure 1). The variance due to this replicate or small-scale spatial variability will be referred to as σ_R^2. The expected value of the variance *in the reference station means* due solely to the replicate variance is

$$E(S_{\bar{X}}^2) = \frac{\sigma_R^2}{r}, \qquad (1)$$

where r is the number of replicates taken at a station, $\bar{X}$ is the mean indicator value (of the replicates) at a station, and $S_{\bar{X}}^2$ is the computed variance of the station means over multiple sampling events from the reference station population. The *expected value,* signified by *E(v)*, is the average value of the parameter v that would occur with a very large number of sampling events.

Figure 1 shows an example application of equation 1, where $\sigma_R^2 = 100$ with two replicates at each station. We would expect that the variance of the station means would approach $100 / 2 = 50$ over multiple samplings or simulations. In fact, the average variance of the station means over the simulations in Figure 1 is about 57, which is approaching the expected value. With 1000 simulations (instead of the 5 simulations in Figure 1), the mean variance equaled 51, which is much closer to the expected value.

Spatial variance component (σ_S^2)

As the geographic positions of two reference stations diverge, we would expect the *underlying* mean values for some indicators to differ at the two stations. We define the *underlying* value of a parameter as the actual value the parameter would have if the entire population of interest were measured instead of being sampled. The different underlying indicator values at separate locations will normally be associated with differences in habitat, history, and proximity to conditions directly or indirectly affecting the indicator values.

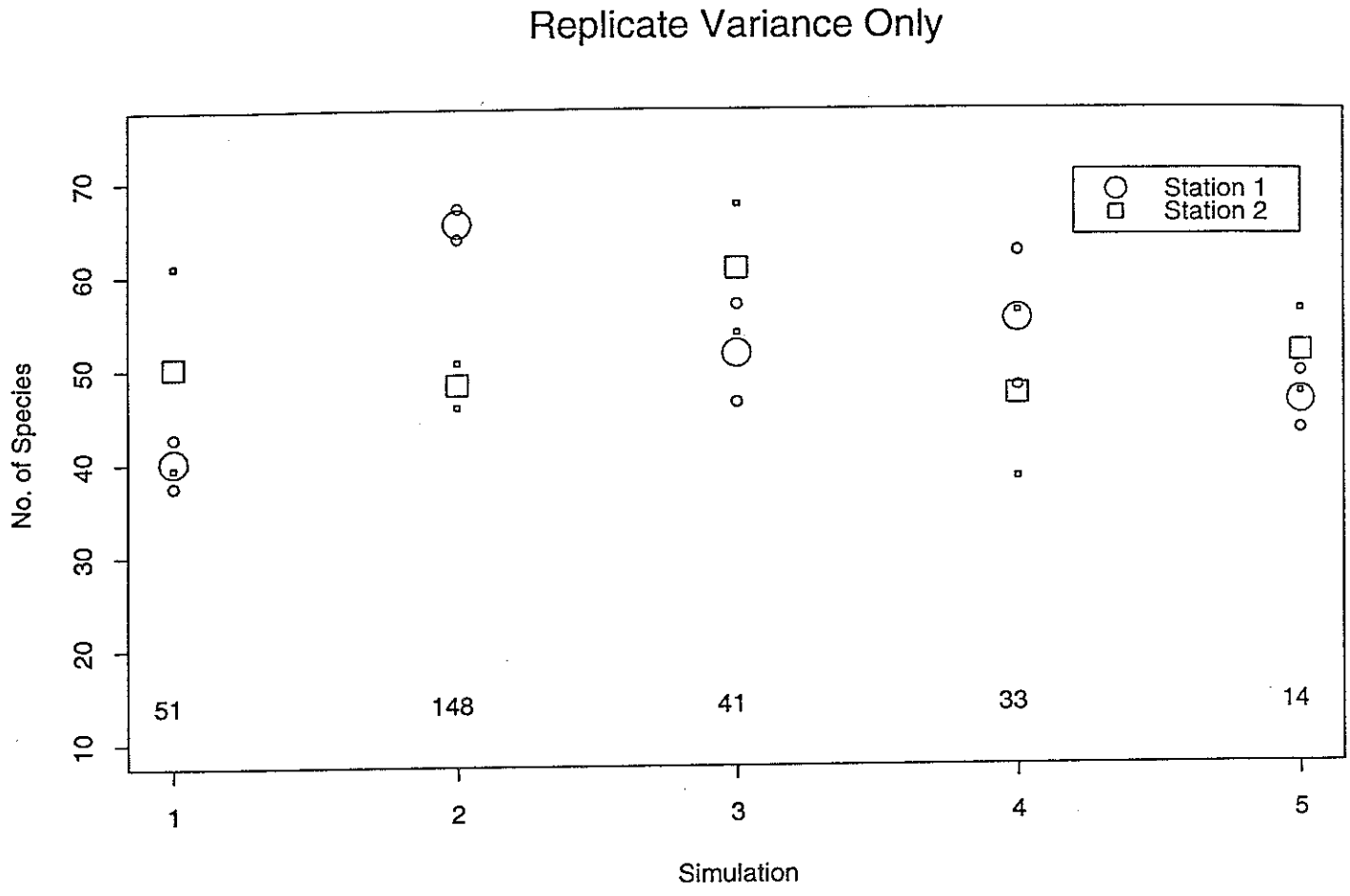

Figure 1. Results of five simulations with two stations with two replicates each, illustrating replicate variance. The indicator values (number of species) for all replicates for all stations were drawn randomly from a normal distribution with a mean of 50 and a variance of 100 ($\sigma_R^2 = 100$ and $r = 2$). The small symbols represent the replicate values and the large symbols represent the mean values of the replicates for each station. The variances of the two station means for each simulation are shown immediately above the X axis. The average variance of the station means over the five simulations is about 57, compared to the expected value of 50.

Here we refer to σ_S^2 as the variance among station means due to geographic separation of the stations. Figure 2 exhibits the results of simulations containing both σ_R^2 and σ_S^2. The expected value of the variance *of the reference station means* due to the spatial and replicate variances is

$$E(S_{\bar{X}}^2) = \sigma_S^2 + \frac{\sigma_R^2}{r}. \qquad (2)$$

In Figure 2, $\sigma_S^2 = 25$ and $\sigma_R^2 = 100$, with two replicates at each station. Using equation 2, we would expect that the variance of the station means would approach $25 + 100 / 2 = 75$ over multiple simulations. The average variance of the station means over the five simulations in Figure 2 is about 73, which is already close to the expected value of 75.

The analyst needs to carefully consider the contribution of varying habitat to the spatial variance component. There are two considerations of importance here. The first is the *sensitivity* of the statistical tests. If an indicator responds to a particular habitat characteristic that varies widely in the reference area, the spatial variance component will be relatively high for that indicator. As the spatial variance component gets larger, the statistical tests necessarily become less sensitive, since the spatial variance is part of the background variability against which the impacts will be compared. Therefore, if we want a sensitive statistical test, we could either choose an indicator that does not respond to the habitat variability in the reference area, or we could choose reference stations that do not vary much for the habitat characteristic in question.

The second issue related to spatial variability concerns the *validity* of the statistical tests. A statistical test that confuses habitat differences with impacts may be sensitive, but it would be sensitive to the wrong thing, and would be invalid. To prevent this, we would at times want habitat differences to contribute to the background variability of the statistical test, so that when we compare the reference stations to an impact station, natural habitat differences do not become confused with an impact. To include this contribution of habitat variability in the spatial variance component, the reference stations should be sampled from an area that includes the amount of habitat variability that we think might exist between our impact station and any of our reference stations.

The relative "amount" of habitat variability that should be included in the spatial variance component should vary with the degree of uncertainty regarding the habitat differences between impact and reference locations. If the habitat characteristics in the reference area and the impact area are similar, then we would want to

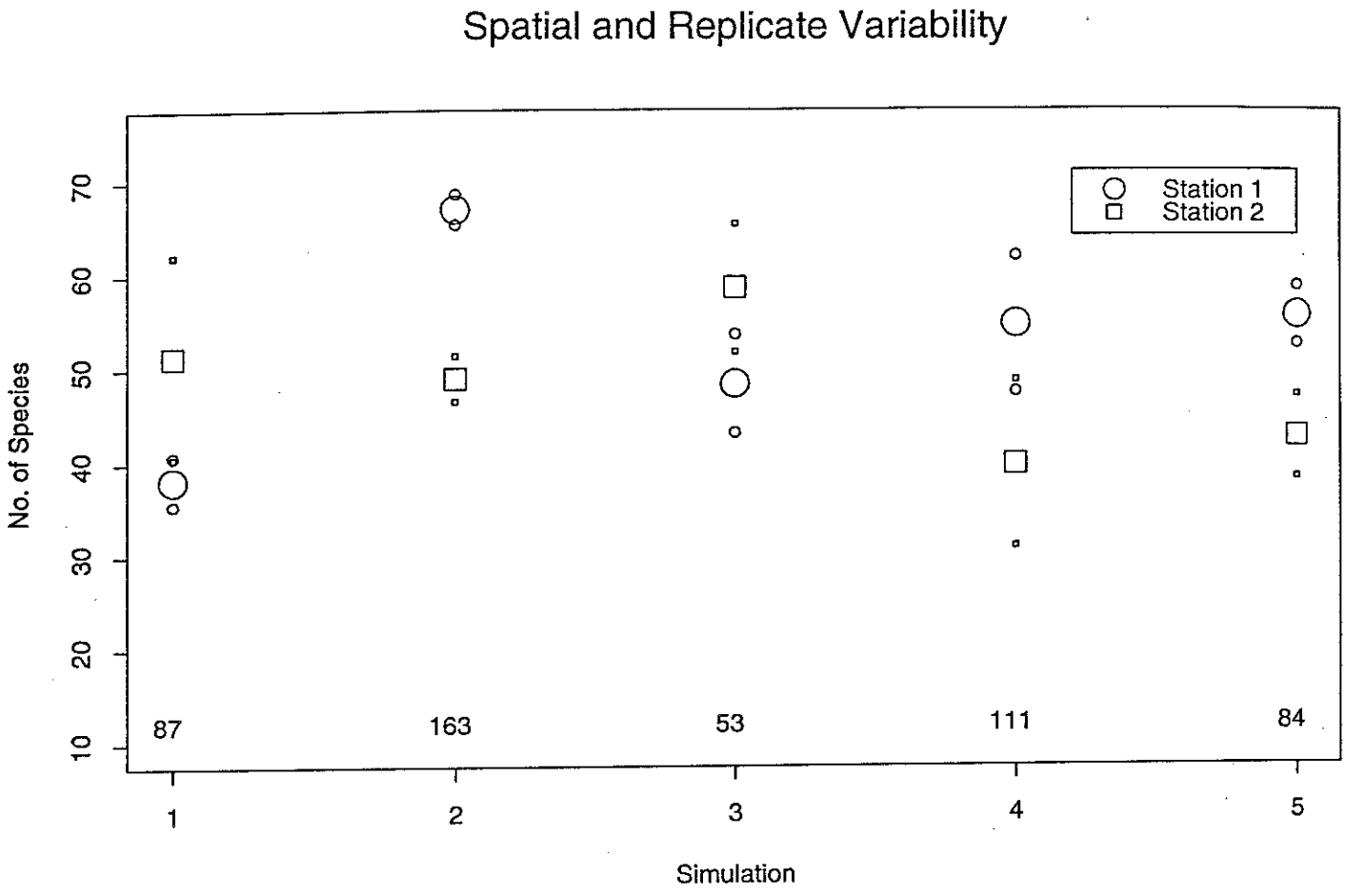

Figure 2. Results of five simulations with two stations illustrating spatial and replicate variances. The underlying mean values for each station were first drawn randomly from a normal distribution with a mean of 50 and a variance of 25 ($\sigma_S^2 = 25$). Then the replicate indicator values for a station were drawn randomly from a normal distribution with a mean equalling the underlying station mean from the first step, and a variance of 100 ($\sigma_R^2 = 100$ and $r = 2$). The small symbols represent the replicate values and the large symbols represent the mean values of the replicates for each station. The variances of the two station means for each simulation are shown immediately above the X axis. The average variance of the station means over the five simulations is about 73, compared to the expected value of 75.

minimize the habitat variability in the reference area to obtain the most sensitive statistical tests. Unfortunately, the impacting activity often alters the habitat, so we cannot be sure the of the habitat characteristics of the impact area that we are trying to match in the reference area. Or we may not be able to even find matching habitats in the reference area, but are only able to sample a range of habitats that appears to "cover" the habitat in the impact area. Finally, we may not have sufficient data or knowledge concerning the pertinent habitat differences in the impact and reference areas, or we may not know the relationship between our indicators and habitat changes. The point here is that as we become more uncertain about the relevant habitat differences in the reference and impact areas, we need to hedge against confusing habitat with impact by including more habitat variability in the spatial variance component. This is accomplished by sampling reference stations from an area with habitat differences covering the range of habitats that we think could differentiate the impact area from the reference area.

In summary, we need to balance the needs for test sensitivity and test validity when choosing indicators and reference stations. It should be noted that in this discussion, we are referring only to habitat characteristics to which the indicator in question is sensitive, since habitat features that do not affect our indicator values are irrelevant.

As we show in the results section, it is sometimes possible to mathematically remove the effects of a habitat characteristic from the indicator data in order to increase the sensitivity of the statistical test, prevent the confusion of impacts and habitat, and decrease violations of test assumptions.

Temporal variance components (σ_T^2 , σ_{ij}, and σ_{TxS}^2)

As noted earlier, $S_{\bar{X}}^2$, is actually an estimate of the variance of the underlying distribution of reference stations. When we consider temporal variance component, we need to distinguish between two types of temporal variability. One type of temporal variance will increase the variance of the distribution of reference stations (an interaction variance), and the other (a covariance) will "move" the whole distribution in concert so that the mean of the distribution changes without changing the variability of the distribution (at a point in time).

The underlying mean indicator value at a reference station will tend to vary over time. We call this the temporal variance σ_T^2. The effect of this temporal variability on the variance among the indicator means from a pair of stations sampled at about the same time will vary depending on the extent to which the indicator means

covary over time at the two stations. For example, let's say that a change of 10 indicator units occurs from time 1 to time 2 at two stations. If the direction of the change is the same for both stations, then the variance among the two station means will be the same at both times 1 and 2 (adding a constant to two values will not change the variance of the values), and the temporal variance will have no effect on the variance of the station means at time 2. On the other hand, if the direction of change of 10 units was in opposite directions at the two stations (indicator value increased by 10 at one station and decreased by 10 in the other), then the variance of the station means at time 2 would be *increased* by the temporal variance. The expected proportion of the temporal variance that contributes to the variance of two station means at any one time will depend on the similarity of the pattern of temporal changes at the two stations. This similarity of magnitude and direction of changes over time at two stations is quantified as the *covariance* among the station means over time. We call this covariance σ_{ij}, where the ij subscript indicates that the covariance over time is for stations i and j.

The degree to which two stations do *not* covary over time is measured as

$$\sigma^2_{TxS} = \sigma^2_T - \sigma_{ij}, \qquad (3)$$

where σ^2_{TxS} is the *time by space interaction* variance associated with the two stations in question. For simplicity, we are assuming that the temporal variance equals σ^2_T at both stations. Here we are subtracting σ_{ij} from the temporal variance because a positive σ_{ij} moves the mean of the distribution of reference stations (i.e., stations move in concert), but does not increase the variance of the distribution of reference stations. When considering the replicate, spatial, and temporal sources of variability,

$$E(S^2_{\bar{Y}}) = \sigma^2_{TxS} + \sigma^2_S + \frac{\sigma^2_R}{r}. \qquad (4)$$

Thus, it is the σ^2_{TxS} component of the temporal variance that affects the variance of the reference station means at one point in time. If the underlying mean indicator values at two stations track perfectly over time, then $\sigma_{ij} = \sigma^2_T$, and from equation 3, $\sigma^2_{TxS} = 0$, i.e., the temporal variance will not affect the variance of the station means. Or, if $\sigma_{ij} = 0$, the indicator values at the two stations are independent, and from equation 3, $\sigma^2_{TxS} = \sigma^2_T$, i.e., all the temporal variance will contribute to the variance of the station means. Negative covariance will increase the station

variance to a degree even greater than the temporal variance, and positive covariance will affect the station variance to a degree less than the temporal variance.

It is interesting to note that some form of the time by space interaction variance is the proper background error variance for the models that incorporate data both before and after the onset of the impacting activity (see references in the introduction). The background error for the statistical tests proposed in this paper (see equation 4 and the tests below) includes the time by space interaction variance *in addition to the spatial variance* (both models include replicate variance). This implies that that the proposed tests will tend to be less sensitive than the tests utilizing before- and after-impact data, since the background error for the proposed tests will usually be larger. This is the price that we pay for not having before-impact data.

Figure 3 shows five simulations of station pairs sampled over time. The magnitude of σ_{ij} (and therefore σ^2_{TxS}) varies for each simulation. The values of σ^2_S, σ^2_T and σ^2_R are the same for each simulation, so the decreasing variance of the station means from Figure 3a to 3e is partially due to the increasing value of σ_{ij} used in the simulations. Figure 3 visually confirms the relationship between covariance over time and the variance of the station means.

We use Figure 3d as an example of an application of equation 4. In the simulation,

$$\sigma^2_T = 500, \sigma^2_S = 400, \sigma^2_R = 80, \text{ and } \sigma_{ij} = 375.$$

Using equation 3, we compute

$$\sigma^2_{TxS} = \sigma^2_T - \sigma_{ij} = 500 - 375 = 125,$$

and from equation 4,

$$E(S^2_{\bar{X}}) = 125 + 400 + \frac{80}{2} = 565.$$

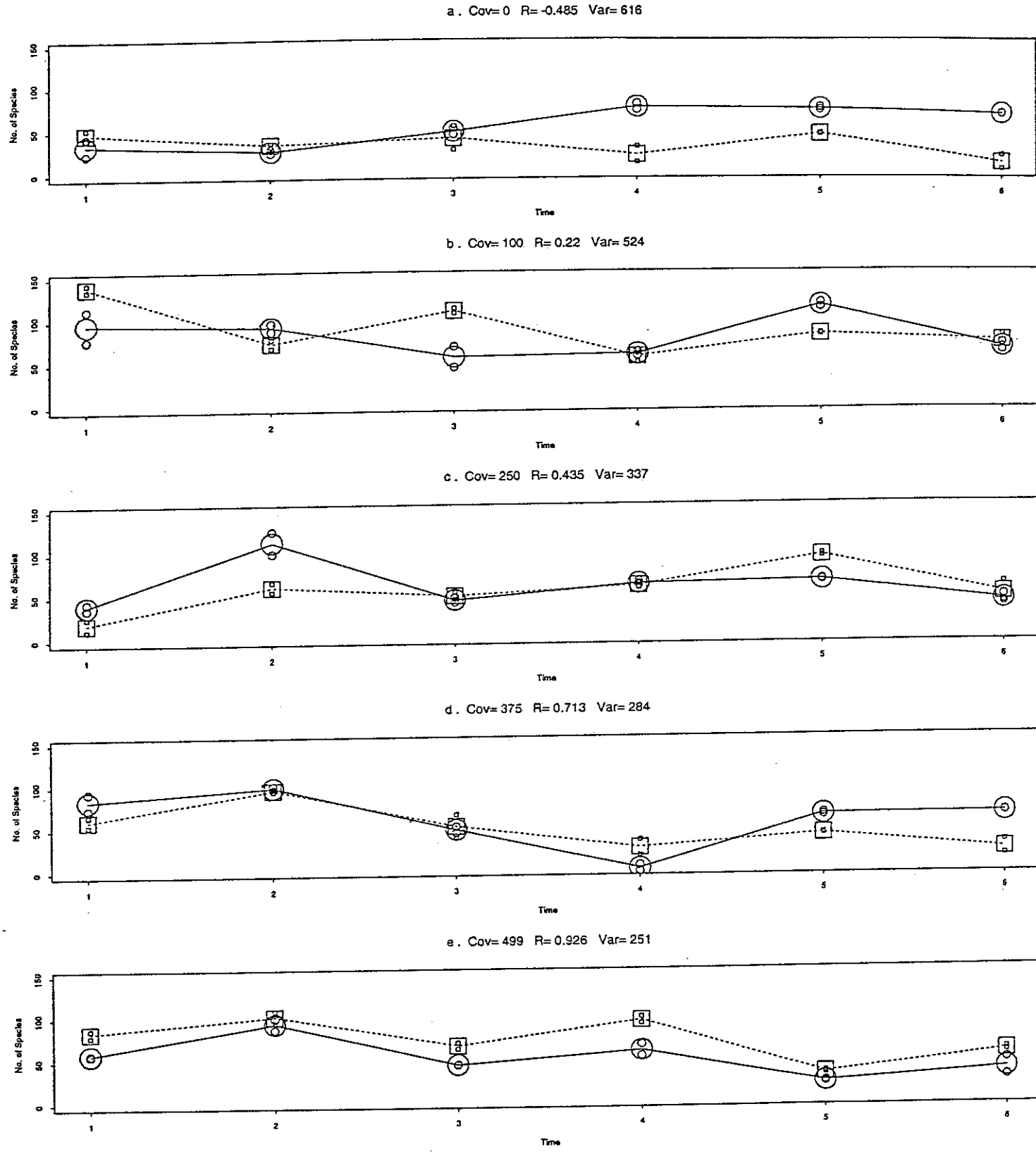

Figure 3. In a-e, simulations with increasing amounts of covariance over time are shown. For each simulation, the two underlying station mean for all times were first established by drawing randomly from a normal distribution with a mean of 50 and a variance of 400 ($\sigma_S^2 = 400$). The underlying mean for each station at each time was determined by drawing from a multivariate normal distribution, which is characterized by the underlying station means (from the first step) and the values for σ_T^2 and σ_{ij}, for each station pair. In all simulations, $\sigma_T^2 = 500$, and σ_{ij} (labeled Cov) is varied from zero in 3a to 499 in 3e. Finally, two replicate indicator values were simulated for each station by drawing from a normal distribution with an underlying mean as determined above, and a variance of 80 ($\sigma_R^2 = 80, n = 2$). For each plot, R is the correlation between the two stations over time, and Var is the average variance of the station means at one time for the simulation. Note that as the covariance (Cov) increases, the variance of the means (Var) decreases, and the means of the two stations track over time to a greater extent.

Statistical Tests

The objective of the proposed statistical tests is to evaluate the general null hypothesis that an *impact station could have originated from the population of reference stations*. In other words, in the absence of sufficient data at the impact station before the onset of the impacting activity, we are assuming that the impact station, without the presence of any impact, would resemble a reference station. Given this assumption, it is obvious that the reference stations should be chosen to represent or cover the conditions that would be present in the impact area if the impacting activity had never occurred.

All the proposed tests are to be applied to data gathered more or less at the same point in time, although in some cases it may be feasible to pool variances from data within different sampling times. The methods are developed as parametric tests with distributional assumptions of normality where statistical inferences are made.

To help clarify the computational formulae for the proposed tests, we will present sample computations based on the data in Table 1. Table 2 summarizes the proposed statistical tests.

Table 1. Example data with four reference stations and an impact station. Some summary statistics (defined below) are $\overline{\overline{X}}_C = 51.25, \overline{X}_I = 34.0, S^2_{\overline{X}_C} = 65.75, S^2_{R_C} = 4.75, S^2_{R_I} = 2, m = 4,$ $MED = 49,$ and $r = r_C = r_I = 2.$

Station	Replicate 1	Replicate 2	Mean ($\overline{X}$) of Replicates	Replicate Variance	$\lvert\overline{X} - MED\rvert$
Reference					
1	50	48	49.0	2.0	0.0
2	40	44	42.0	8.0	7.0
3	60	63	61.5	4.5	12.5
4	54	51	52.5	4.5	3.5
Impact	35	33	34.0	2.0	15.0

Table 2. Summary of the proposed statistical tests. We are assuming a one-tailed null hypotheses where the indicator in question is expected to decrease with impact. Where the indicator is expected to increase with impact, replace $\leq$ with $\geq$ in the null hypothesis. The assumptions listed are not exhaustive, but include the assumptions that differentiate a test from other tests.

Approach	Test	Null Hypothesis	Assumptions
1. Comparison of means	Test 1 (Standard t test)	Reference station mean≤impact station mean	$\sigma_S^2 = 0, \sigma_{TxS}^2 = 0$
	Test 2	Mean of reference station distribution ≤impact station mean	Equal replication and equal replicate variance at all stations.
	Test 3	"	Equal replication and equal replicate variance at all reference stations.
	Test 4	"	Equal replicate variance at all reference stations.
2. Comparison of impact station mean with percentile (p) of reference station distribution	Tolerance Interval	pth percentile of reference station distribution ≤impact station mean	Equal replication and equal replicate variance at all stations.
	Hampel Outlier Identifier	"	"

Approach 1 - Comparison of means

The four tests in this category all define t statistics of the general form

$$t = \frac{\overline{X}_C - \overline{X}_I}{\sqrt{S^2_{\overline{X}_C - \overline{X}_I}}}, \qquad (5)$$

where $\overline{X}_C$ and $\overline{X}_I$ are the indicator means representing the reference and impact stations, respectively, and $S^2_{\overline{X}_C - \overline{X}_I}$ is the estimated variance of the difference in the means in the numerator when the null hypothesis is true. For the reference station, we use the subscript C (for "Control") to avoid confusion with the R subscript for replicates. For the purposes of the discussion, we expect that the indicator values will decrease when an impact is present, and we will want to perform a 1-tailed test, i.e., when the computed $t \geq t_{\alpha,df}$, we will reject the null hypothesis. The $t_{\alpha,df}$ is the critical Student t value for a nominal type-1 error level of α and df degrees of freedom. The proposed tests vary in the manner in which the numerator and denominator of equation 5 are computed.

To define any one of the tests in this category in terms of a statistical interval, the lower bound of the interval can be expressed as

$$L = \overline{X}_C - t_{\alpha,df}\sqrt{S^2_{\overline{X}_C-\overline{X}_I}}, \qquad (6)$$

which is a rearrangement of equation 5 with the critical $t_{\alpha,df}$ substituted for t, and solving for $\overline{X}_I$. Here L is the maximum value of $\overline{X}_I$ that will lead to a rejection of the null hypothesis, or in other words, indicator values less than or equal to L will be outside the reference envelope. Equation 6 would be used when we expect the indicator values to decrease with an impact. When the indicator value increases with impact, an upper bound (U) would be computed instead by replacing the minus sign in equation 6 with a plus sign. The values for $\overline{X}_C$ and $\sqrt{S^2_{\overline{X}_C-\overline{X}_I}}$ in equation 6 will vary for the different tests.

Test 1 - Standard t test contrasting a single reference station with an impact station

The common t-test or one-way analysis of variance comparing a single reference station with an impact station is often applied to test the null hypothesis of no impact. The value for $S^2_{\overline{X}_C-\overline{X}_I}$ is computed from the *replicate* variances taken at the stations being compared, and the expected value for $S^2_{\overline{X}_C-\overline{X}_I}$ is *assumed* to be

$$E(S^2_{\overline{X}_C-\overline{X}_I}) = \frac{2\sigma^2_R}{r},$$

where σ^2_R and r are the replicate variance and the number of replicates, respectively. This follows from equation 1 and the fact that the expected variance of a difference is the sum of the variances of the differenced values (when the observations are independent). Here we are assuming that the underlying replicate variance is the same at both stations being compared, so the sum will simply be twice the expected value for a single station. Thus, the final test statistic is

$$t = \frac{\overline{X}_C - \overline{X}_I}{\sqrt{\frac{2\hat{\sigma}^2_R}{r}}}, \qquad (7)$$

where $\overline{X}_C$ is the mean of the replicates at a single reference station, $\overline{X}_I$ is the mean of the replicates at the impact station, and $\hat{\sigma}^2_R$ is a pooled estimate of the within-station replicate variance. A little "hat" (^) above a symbol indicates that the value is an estimate.

When spatial and time by space interaction variance components are present in the data (i.e., $\sigma^2_S > 0$ and/or $\sigma^2_{T\times S} > 0$), the t ratio in equation 7 will be inappropriate. For a proper test, the expected value of the t^2 ratio should be 1 when the null hypothesis is true. Since the numerator of the t^2 ratio will contain all three variance components, but the denominator will only contain the replicate variance component, the expected value of the t^2 ratio will be larger than 1 when the null hypothesis is true. In this case, the computed t value will tend to be inflated beyond that expected by chance, leading to rejection of the null hypothesis when the null hypothesis is actually true at a rate well above the nominal type-1 error of the test. The type-1 error is the rate at which we expect to falsely reject the null hypothesis when it is actually true. For example, when we choose $\alpha = .05$ for the t test, we only expect to falsely reject the null hypothesis about 5% of the time, but as we will demonstrate in the results, our actual rate of false rejections could be much higher.

Test 2 - t statistic assuming equal replication and equal replicate variances at all stations

This test and subsequent tests incorporate all the pertinent variance components described in equation 4. This particular test assumes equality of the underlying replicate variances at all stations, and also equal number of replicates at each station. To obtain variances with the proper variance components, we will need to sample at least two reference stations along with the impact station. From the means of the reference stations, we directly compute $S^2_{\overline{X}_C}$ as the variance of the station means. This computed value of $S^2_{\overline{X}_C}$ will contain all the variance components described in equation 4 (since all these variance components potentially contribute to the variability of the reference station means). We want to compare the impact station mean with the *mean of the reference station distribution* (rather than any particular reference station as with test 1). Such a test would generally be described as

$$t = \frac{\overline{\overline{X}}_C - \overline{X}_I}{\sqrt{S^2_{\overline{\overline{X}}_C - \overline{X}_I}}}, \qquad (8)$$

where the double bars above X_C indicate that we are comparing the *mean of the reference station means* with the mean of the replicates at the impact station. To derive the denominator of the test, we first note that

$$E(S^2_{\overline{\overline{X}}_C - \overline{X}_I}) = E(S^2_{\overline{\overline{X}}_C}) + E(S^2_{\overline{X}_I}). \qquad (9)$$

The expected value of the first term on the right of equation 9 is

$$E(S^2_{\overline{\overline{X}}_C}) = \frac{E(S^2_{\overline{X}_C})}{m}, \qquad (10)$$

where m is the number of reference stations. This leaves us to estimate $S^2_{\overline{X}_I}$, the second term on the right in equation 9. We are treating the impact station as if it were randomly drawn from the population of reference stations, and when the null hypothesis is true, the impact station would continue to appear as a reference station as far as indicator values are concerned. If this is the case, the expected variance of the impact station mean would be the *same as the expected value of the variance of a reference station mean*, which is $E(S^2_{\overline{X}})$, described in equation 4. From this assumption and equations 9 and 10, we obtain

$$E(S^2_{\overline{\overline{X}}_C - \overline{X}_I}) = \frac{E(S^2_{\overline{X}_C})}{m} + E(S^2_{\overline{X}_C}), \qquad (11)$$

where $E(S^2_{\overline{X}_C}) = E(S^2_{\overline{X}})$, with the C subscript added to emphasize that the variance is estimated from the reference station means. Given equation 11, our test statistic becomes

$$t = \frac{\overline{\overline{X}}_C - \overline{X}_I}{\sqrt{\frac{S^2_{\overline{X}_C}}{m} + S^2_{\overline{X}_C}}}. \qquad (12)$$

As noted above, $S^2_{\overline{X}_C}$ is simply the computed variance of the reference station means. The degrees of freedom used to obtain the critical $t_{\alpha,df}$ value from the t distribution is *m-1*. Applying the data in Table 1 to equation 12, we have

$$t = \frac{51.25 - 34.00}{\sqrt{\frac{65.75}{4} + 65.75}} = 1.90.$$

With $t_{.05,3} = 2.353$, we accept the null hypothesis of no impact. From the perspective of the reference envelope, we compute $L = 51.25 - 2.353(9.0657) = 29.92$ as the edge of the reference envelope (equation 6). Since the impact station mean (34.0) is greater than L, we conclude that the impact station is within the reference envelope and not impacted as far as this indicator is concerned.

The following two tests, tests 3 and 4, are modifications of test 2 that allow for varying levels of replication among the stations and different replicate variance at the impact station. In principle, the tests are the same as test 2, so the reader not interested in the tedious details of the test modifications should skip over the sections for tests 3 and 4.

Test 3 - t statistic assuming equal replication and equal replicate variances at the reference stations

This test simply extends test 2 by relaxing the requirement that the replication and replicate variance at the impact station match the replication and replicate variances at the reference stations. The test statistic is

$$t = \frac{\overline{\overline{X}}_C - \overline{X}_I}{\sqrt{\frac{S^2_{\overline{X}_C}}{m} + S^2_{S,TxS} + \frac{S^2_{R_I}}{r_I}}}, \qquad (13)$$

where $S^2_{R_I}$ is the computed replicate variance for the impact station, and r_I is the number of replicates at the impact station. $S^2_{S,TxS}$ is an estimate of the sum of the spatial and time x space variance components, which, following equation 4, can be computed as

$$S^2_{S,TxS} = \hat{\sigma}^2_{TxS} + \hat{\sigma}^2_S = S^2_{\overline{X}_C} - \frac{S^2_{R_C}}{r_C}, \qquad (14)$$

where $S^2_{R_C}$ is the pooled within-station replicate variance for the reference stations, and r_C is the number of replicates at each reference station. What we are doing with this test is breaking down the variance of the impact station mean (which is $S^2_{\overline{X}_C}$ in equation 12) into its separate variance components, and substituting the replicate

variance from the impact station for the replicate variance from the reference stations. The degrees of freedom used to obtain the critical $t_{\alpha,df}$ value from the t distribution is $m-1$.

Applying the data in Table 1 to equation 13, we have

$$t = \frac{51.25 - 34.00}{\sqrt{\frac{6575}{4} + 6575 - \frac{475}{2} + \frac{2}{2}}} = 1.92,$$

and we reach the same conclusion as with test 2.

Test 4 - t statistic with variation in replication level at the reference stations

This test is an extension of test 3 in that the number of replicates at the reference stations can vary. The test is more complicated because we need to adjust the reference station variances and mean of means for the unequal replication at the reference stations. The test statistic is

$$t = \frac{\overline{\overline{X}}'_C - \overline{X}_I}{\sqrt{\frac{1}{\sum_{i=1}^{m} \frac{1}{S^2_{\overline{X}_i}}} + S^2_{S,T\times S} + \frac{S^2_{R_I}}{r_I}}}, \qquad (15)$$

where

$$S^2_{S,T\times S} = \hat{\sigma}^2_S + \hat{\sigma}^2_{T\times S} = \frac{N(m-1)(MST - S^2_{R_C})}{N^2 - \sum_{i=1}^{m} r^2_{C_i}}, \qquad (16)$$

with *MST* as the between-reference station mean square in a standard unbalanced one-way ANOVA, N is the total number of replicates at all reference stations, and r_{C_i} is the number of replicates at the *ith* reference station. The computational formula for *MST* is

$$MST = \frac{\sum_{i=1}^{m} r_{C_i}(\overline{X}_i - \overline{\overline{X}}_C)^2}{m-1}, \qquad (17)$$

with $\overline{X}_i$ as the mean of reference station *i*. Equation 16 is derived from the fact that

$$MST = \hat{\sigma}_R^2 + \bar{r}(\hat{\sigma}_S^2 + \hat{\sigma}_{TxS}^2),$$

where $\bar{r}$ is the effective mean number of replicates at the reference stations, computed as

$$\bar{r} = \frac{1}{m-1}(N - \frac{\sum_{i=1}^{m} r_{C_i}^2}{N}). \tag{18}$$

The adjusted estimate of the variance of the mean of reference station i is computed as

$$S_{\bar{X}_i}^2 = S_{S,SxT}^2 + \frac{S_{R_C}^2}{r_{C_i}}. \tag{19}$$

Finally, the weighted mean of the reference station means is

$$\bar{\bar{X}}'_C = \frac{\sum_{i=1}^{m}(\bar{X}_i / S_{\bar{X}_i}^2)}{\sum_{i=1}^{m}(1/S_{\bar{X}_i}^2)}, \tag{20}$$

with $\bar{X}_i$ as the mean of reference station i. Here we are giving more weight to the reference stations with lower variance of the mean. The approximate degrees of freedom for the critical $t_{\alpha,df}$ is

$$df = \frac{(S_{S,SxT}^2 + S_{R_I}^2/r_I)^2}{S_{S,SxT}^4/(m-1) + S_{R_I}^4/(r_I^2(r_I-1))}. \tag{21}$$

Applying the data in Table 1 to equation 15, we have $MST = 131.5$ (equation 17),

$$S_{S,SxT}^2 = \frac{8(3)(131.5-475)}{64-16} = 63.375, \qquad \text{(from equation 16)}$$

and

$$S_{\bar{X}_i}^2 = 63.375 + \frac{475}{2} = 65.75 \qquad \text{(from equation 19)}$$

for all four reference stations (all i), since there are the same number of replicates at each station. With equal replication at all reference stations $\bar{\bar{X}}'_C = \bar{\bar{X}}_C = 51.25$ (equation 20). Finally,

$$t = \frac{51.25 - 34.00}{\sqrt{\frac{1}{.06084} + 63.375 + \frac{2}{2}}} = 1.92,$$

and

$$df = \frac{(63.375 + 2/2)^2}{\frac{63.375^2}{3} + \frac{4}{4}} = 3.093 \approx 3,$$

leading us to conclusions similar to tests 2 and 3.

Approach 2 - Comparison of the Impact Station Mean With a Percentile of the Reference Station Distribution

The null hypothesis for the methods in this category is that the mean indicator value at the impact station is within a chosen percentile (p) of the distribution of reference stations. The direction of change that an indicator is expected to take in the presence of an impact will determine whether we will choose a p on the lower or upper tail of the reference station distribution. For example, if we expect an indicator value to *decrease* with an impact, we might choose $p=10$, in which case our null hypothesis would be that the impact station mean is greater than or equal to the *10th* percentile of the reference station distribution. Similarly, if we expect the indicator value to *increase* with impact, we might choose $p= 90$, and our null hypothesis would be that the mean of the impact station is less than or equal to the *90th* percentile of the reference station distribution.

With approach 1, we treated the impact station as a random selection from the reference station distribution, and due to the shape of the normal curve, it is more likely that the impact station mean is closer to the mean of the reference distribution and less likely that it is from the tails of the distribution. Tests with approach 2 allow for a weaker set of assumptions regarding the relationship between the reference station distribution and impact station. We may want to assume a worst case scenario where the impact station originated from a particular percentile in the *tail* of the reference station distribution (in the direction of impact), and the methods in approach 2 allow for this weaker assumption. The more extreme the chosen percentile, the weaker or more conservative the assumption about the original state of the impact station will be. This is because when we choose an extreme

percentile, we are allowing for the fact that the impact station might have originated as an outlier member of the reference station distribution (in the direction of impact), and for this reason alone the station might appear impacted when in fact there is no impact. At times, the consequences of identifying an impact can be ominous (large cleanup costs, litigation, etc.), so a regulator might want only to identify impacts for which a strong case can be made. The argument for impact will be strongest when an impact is detected even when a very conservative assumption has been made (by choosing an extreme percentile). It could be argued that a test from the first approach could be used here instead, with an α level adjusted to suit how strong an assumption we are willing to make. The methods from approach 2, on the other hand, may be more defensible and comprehensible, since the level of conservatism of the test is an explicit part of our null hypothesis.

Both tests in this category directly compute statistical interval bounds that we will use as edges of the reference envelope, where impact stations with indicator values outside the envelope edge (interval bounds) will be considered impacted. Besides specifying a percentile value p, we also specify a nominal type-1 error rate α, which is the proportion of the time that we expect an impact station originating from the *pth* percentile of the reference station distribution to appear by chance outside the computed interval bound. The interval bounds for both tests have the general form

$$U = M + g_{n,p,\alpha}V$$

for an upper interval bound when we expect the indicator to increase with impact, or

$$L = M - g_{n,p,\alpha}V$$

for a lower interval bound when we expect the indicator to decrease with impact. Here, M is a measure of central tendency (mean or median) of the reference station distribution, V is a measure of the variability (standard deviation or median deviation) of the reference station distribution, and $g_{n,p,\alpha}$ is a critical table value that provides U and L with the expected statistical properties for given n, p and α values. Note that these equations are similar in form to the familiar confidence intervals of the mean from a sample, where, for example, the lower bound is

$$L = \overline{X} - t_{\alpha,df}S_{\overline{X}}.$$

Here $\overline{X}$ is the sample mean, $S_{\overline{X}}$ is the standard deviation of the sample mean (standard error), and $t_{\alpha,df}$ is the critical Student t value. The following methods all assume that the underlying replicate variances and the numbers of replicates at all stations are equal.

Tolerance Interval

When we expect our indicator values to *decrease* with impact, and we choose the *pth* percentile ($0<p<50$) for our null hypothesis, we compare the mean indicator value at the impact station with the lower bound of a one-sided tolerance interval, which is computed as

$$L = \overline{\overline{X}}_C - g_{1-\alpha,100-p,m} S_{\overline{X}_C} \qquad (22)$$

(Hahn and Meeker 1991, Vardeman 1992). We reject the null hypothesis when the impact station mean is less than L. When we expect our indicator values to *increase* with impact, and choose a p value ($50<p<100$) for the null hypothesis, we compare the mean indicator value at the impact station with the upper bound of a one-sided tolerance interval, which is computed as

$$U = \overline{\overline{X}}_C + g_{1-\alpha,p,m} S_{\overline{X}_C}, \qquad (23)$$

where we reject the null hypothesis when the impact station mean is greater than U. $\overline{\overline{X}}_C$ is the mean of the reference station means, $S_{\overline{X}_C}$ is the standard deviation of the reference station means, α is the nominal type-1 error of the test, and m is the number of reference stations. The values for $g_{1-\alpha,100-p,m}$ and $g_{1-\alpha,p,m}$ can be found in Hahn and Meeker (1991, Table A.12) or Gilbert (1987, Table A3). (Both tables use $p/100$ instead of p). The g values, derived in Odeh and Owen (1980), are based on the noncentral t distribution. Since we are using $S_{\overline{X}_C}$ in the tolerance interval computations, the tolerance interval bounds should incorporate all the pertinent variance components (equation 4).

When applying the data in Table 1 to equation 22, we use $p=10$, $\alpha=.05$, $g_{1-.05,100-10,4}=4.162$, and

$$L = 51.25 - 4.162\sqrt{65.75} = 17.50.$$

The impact mean (34.0) is greater than L, so we accept the null hypothesis that the impact station is not impacted and is within the reference envelope. Note that the interval bound for the tolerance interval (*17.50*) is a fair amount lower than that for test 2 (*29.92*). This is due to the more conservative assumption associated with a tolerance interval with $p=10$, i.e., with the tolerance interval, we are assuming that the impact station originated from tenth percentile of the reference distribution, but with test 2 we are assuming that the impact station originated as a random selection from the reference station distribution, where it is less likely that we would obtain an observation as marginal as the tenth percentile by chance.

Hampel outlier identifier - standardization 3

Davies and Gather (1993) propose techniques for identifying outliers in a data sample. An outlier is defined as a data observation that appears to be aberrant given the assumed properties of the distribution from which the observations were drawn. In a sense, our comparison of the impact station with the distribution of reference stations is an exercise to see if the impact station is an outlier to the reference station distribution. Davies and Gather (1993) proposed a technique, based on ideas in Hampel (1985), that performed well as an outlier identifier in their simulations. The method, called the Hample identifier, standardization 3 is conceptually similar to a tolerance interval. When we expect the indicator to decrease with impact, a lower bound of an interval is defined as

$$L = med(\overline{X}) - g_{m+1,p,\alpha} mad(\overline{X}), \qquad (24)$$

where we declare sample means less than L as outliers, or in our application, impact station means less than L are outside the reference envelope and assumed impacted. Here, $med(\overline{X})$ is the median of the sample, which includes the reference station means plus the impact station mean, p is a percentile used in the same manner as with the tolerance interval, m is the number of reference stations, α is the nominal type-1 error with the same meaning as with the tolerance interval, and $mad(\overline{X})$ is *the median absolute deviation* of the sample, defined as

$$mad(\overline{X}) = med(|\overline{X}_1 - med(\overline{X})|, |\overline{X}_2 - med(\overline{X})|, \ldots, |\overline{X}_{m+1} - med(\overline{X})|). \quad (25)$$

$\overline{X}_1$ represents the mean of the first reference station, $\overline{X}_2$ the mean of the second reference station, and so on, with $\overline{X}_{m+1}$ being the mean of the impact station. For a few α and p values, Davies and Gather (1993) give formulae for computing the $g_{m+1,p,\alpha}$. We have written a computer program that will estimate a $g_{m+1,p,\alpha}$ for any combination of α and p. The program, written in C++, is available from the author upon request.

When we expect the indicator values to increase with impact, an upper bound of the interval is defined as

$$U = med(\overline{X}) + g_{m+1,p,\alpha} mad(\overline{X}). \quad (26)$$

Sample means greater than U would be considered outliers outside the reference envelope.

Equations 25 and 26 are similar in form to the tolerance interval bounds. Using medians and deviations instead of means and standard deviations provides a certain amount of "robustness" to the method in that the outlier values will not distort the computations so their identification becomes more difficult. This method is designed to detect multiple outliers, so in a typical application, one would input a sample of data, and observations not fitting the dominant distribution would be identified as outliers. Here we are proposing a more restricted use of the method, mainly, we use it as a more robust alternative to a tolerance interval. Rather than testing for multiple outliers in a single test, we use the reference station indicator mean values plus a dummy indicator mean value for a hypothetical impact station, and compute a single interval bounds that will apply to all the impact stations. If we expect the indicator value to decrease with impact, the dummy indicator value at the impact station is less than $med(\overline{X})$, or if we expect the indicator value to increase with impact, the dummy indicator value is greater than $med(\overline{X})$. Beyond this, as long as the dummy indicator value is not too close to the median, the exact value chosen for the dummy indicator value will not affect the computations.

It should be pointed out that we have changed the terminology of Davies and Gather (1993) to be consistent with our terminology. Specifically, $p = 2\alpha_N$, $m+1=N$, $\overline{X} = X_N$ and $\alpha = 2\alpha$ in this paper and

Davies and Gather, respectively. Davies and Gather proposed a two-tailed interval, which we have converted to one-tailed intervals by using $2\alpha_N$ and 2α instead of α_N and α.

When applying the data in Table 1 to equation 24, we use $p=10$, $\alpha=.05$, $g_{4+1,10,.05}=9.7$, $MED=49$, $MAD=3$, and

$$L = 49 - 97(3) = 19.9.$$

The impact mean (34.0) is greater than L, so we accept the null hypothesis that the impact station is not impacted and is within the reference envelope. Note that the interval bound (19.9) is closer to the tolerance interval bound (*17.5*) than to the test 2 bound (*29.92*). This is expected, since the tolerance interval and the Hampel identifier are associated with the same more conservative null hypothesis.

Test Evaluations and Applications

We use a simulation model to evaluate the proposed statistical tests and illustrate their properties. The input to the simulation model includes variance components estimated from real-world test data. The same test data are also used to demonstrate the use of the methods to detect impacts.

Test data

The emphasis of the application with the test data is the detection of impacts on benthic communities by sewage outfalls. In the Southern California Bight, sewage outfalls have been one of the more important human activities as far as potential environmental impacts are concerned. The most extensive benthic sampling in the Southern California Bight has occurred at around 60 meters in depth, due the fact that the major sewage outfalls in the area discharge at this depth.

In addition to the dischargers' monitoring programs, the Southern California Coastal Water Research Project (SCCWRP) has on three occasions sampled the benthos outside the areas covered by the outfall monitoring programs. In 1977, the 60-meter Control Survey (Word and Mearns 1979) included 71 benthic stations from Point Conception to the Mexican Border (Figure 4). In 1985 (Thompson et al. 1987), a subset of these stations were

sampled (stations 4, 5, 8, 11, 13,15, 50,52,54,57,60,61, and 71), and in 1990 a smaller subset of the 1985 stations was sampled (stations 13, 15, 50, 52,60, 61, and 71).

The data used to estimate the variance components included all three SCCWRP surveys plus the available outfall monitoring data for 1985 and 1990. From the SCCWRP surveys, we only utilized stations that we considered good reference stations, given previous analyses of the data. The stations eliminated due to possible impacts from nearby outfalls, harbors, or oil seeps were stations 1-3, 6, 9, 10, 23-49, 63-69. The outfall monitoring programs contributing data included the City of Los Angeles (Hyperion outfall), the County Sanitation Districts of Los Angeles County (Whites Point outfall), County Sanitation Districts of Orange County, and the City of San Diego (Figure 4). The exact positions of the outfall monitoring stations, not shown in Figure 4, can be found in EcoAnalysis et al. (1993). We confined our analysis to data taken from between 50 and 70 meters depth, since almost all the replicated data occurred in this depth range, and replicated data was required to estimate the replicate variance component. All replicated stations had five replicates, and were from the outfall monitoring programs. None of the SCCWRP survey stations were replicated.

Indicators

Thirteen indicator parameters were computed for each station replicate (Table 3). All these indicators can be sensitive to organic enrichment and associated effects produced by sewage outfalls. Prior to all analyses, some of the indicators were transformed to remove dependence between the station mean and the replicate variance (Table 3). If this dependency were not removed, assumptions of equal replicate variance and additive effects in the proposed methods would be seriously violated.

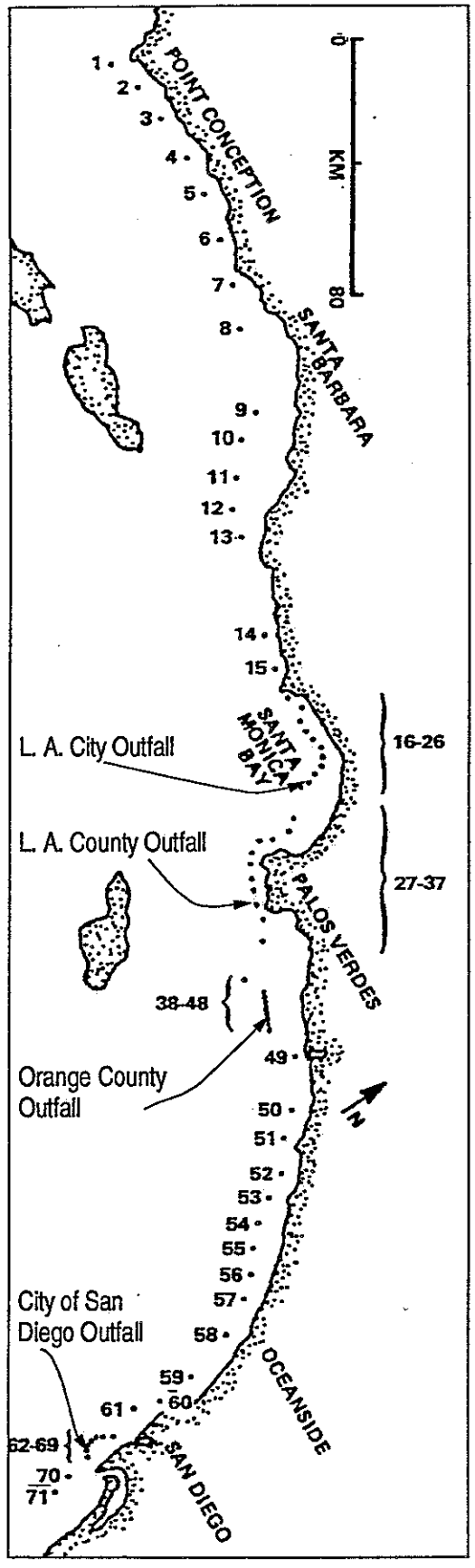

Figure 4. Station locations for the SCCWRP reference surveys (after Figure 1 in Word and Mearns, 1979)

Table 3. The thirteen indicators computed for each station replicate. Taylor's power law was used to determine power transformations (last column) for indicators showing dependence between the station mean and the station replicate variance (Elliott 1977). For example, for a power transformation of .4, $y=x^{.4}$, where y is the transformed data value used in the analyses, and x is the raw data value.

Indicator	Symbol	Reference	Power Transformation
Gleason Richness Diversity	D	Margelef (1958)	
Shannon-Wiener Diversity	H'	Pielou (1969)	
Index 5	I5	Smith and Bernstein (1985) Bernstein and Smith (1986)	
Infaunal Trophic Index	ITI	Word (1978, 1980)	
Evenness Diversity	J'	Pielou (1969)	
No. of Crustacean Species	N_CRUS		.6
No. Species in ITI group 1	N_ITI1	Word (1980)	.7
No. of Species	NSPEC	Pearson and Rosenberg (1978)	
Total Abundance of Amphiodia	T_AMP	Word (1980)	.4
Total Abundance of Echinoderms	T_ECH	Word (1980)	.4
Total Abundance of ITI group 1	T_ITI1	Word (1980)	.4
Total Polychaete Abundance	T_POLY		.2
Total Abundance	TOTAB	Pearson and Rosenberg (1978)	.3

Estimation of variance components

To produce realistic input to the simulation model used to evaluate the statistical methods, we estimated the spatial (σ_S^2) and temporal (σ_T^2, σ_{ij},σ_{TxS}^2) variance components from SCCWRP reference stations that were sampled on more than one occasion (using the VARCOMP procedure, SAS 1990a). We excluded the outfall monitoring data from these estimates since these variance components could be especially influenced by nearby outfalls. Only a single grab was taken at SCCWRP reference stations, so we had to estimate the replicate variance from replicated stations in the outfall monitoring data. Here, we used a regression approach (the generalized additive model technique; Hastie and Tibshirani 1990, Chambers and Hastie, 1993) to model the relationship between the replicate variance and the sediment grain size (% silt-clay) and outfall influence (ITI indicator). The regression model was then used to predict replicate variances for the subset of reference stations used to estimate the spatial and temporal variance components.

We next formed seven groups of reference stations that we consider separately in our simulations evaluating the proposed methods (Table 4). Each of the seven groups was chosen to represent different situations that might affect the behavior of the proposed tests. We then used the general additive model regression technique

to predict the variance components for each group of reference stations, based on the differences in sediment size and distances between stations. The regression model parameters were based on the variance component estimates computed at the SCCWRP reference stations sampled on more than one occasion (see previous paragraph).

Table 4. The seven groups of reference stations to be used in the simulations.

Group	Stations	Representing
1	4, 5, 7, 8, 11-15	North of Santa Monica Bay, remote from any outfalls, fairly wide range of sediment sizes
2	16-22	Santa Monica Bay, closer to outfall, generally finer sediments
3	50-62	Southern area, remote from outfalls, sediment gradient
4	70, 71	Very coarse sediments, outliers for some indicators
5	4, 5, 7, 8, 11-22	All northern stations, heterogeneous
6	4, 5, 7, 8, 11-22, 50-62	All stations but outlier stations 70, 71
7	50-62, 70, 71	All southern stations, including outliers

In our results, we present the average variance components for each group as a percent of the total of all the variance components. This puts all variance components on a common scale without loss of information, since only the relative magnitudes of the different variance components are important. In the simulations we also used the average variance components for each group, except for the replicate variances, where we used the predicted replicate variances for each separate station.

It should be emphasized that these variance component estimates are only very rough average values, and we would be hesitant to claim that they would apply to any one specific set of stations. On the other hand, at the very least we would want our proposed methods to perform well for average conditions. As such, the main purpose of computing variance component estimates is to provide inputs to our simulations testing the performance of the proposed methods.

Testing assumptions

We use the simulation model (described in the next section) to test the robustness of the various techniques to violations of some of the assumptions of the methods. We do this by violating the assumptions to varying degrees in the simulations, and observing the effect on the test results. To make these simulations relevant,

we need to have some idea of the extent to which some of the assumptions might be violated in practice. Accordingly, we use the reference station data to test for violations of some of the more important assumptions of the tests.

Normality

For each group (Table 4), we examined histograms of indicator values and applied the Shapiro-Wilk test (Shapiro and Wilk 1965) to evaluate the normality of the data (SAS 1990b, UNIVARIATE procedure).

Random sampling

For most monitoring programs, the station locations are not chosen randomly. At times it is more important to obtain data at specific key locations or to efficiently obtain a certain amount of coverage of an area. Other types of analysis, e.g., pattern analysis, often perform better with a more systematic sampling pattern. Fulfilling these requirements is more difficult with random sampling, so we most often encounter transects or grids in environmental studies, with the spacing between stations equal or varied in some way to fulfill other requirements. The test data used to compute the variance components is obviously more systematic than random (Figure 4).

Nonrandom sampling patterns of reference stations can potentially violate a key assumption of the proposed methods, which is that the locations of the reference station locations are randomly chosen. When systematic sampling is performed instead of random sampling, but the population (of reference stations) is *in random order* (Gilbert 1987) or *quasi-random* (Barnett 1991), then we can use the systematic data for statistical inference in the same manner as random data without biasing the variance estimates used in the tests. The population is in random order when the following three criteria are satisfied for the reference area.

1. There are no trends in indicator values within the area sampled.
2. There are no natural strata where the indicator values are locally elevated or depressed.
3. There is no correlation between differences in indicator values and distances between stations.

An example will show why these criteria make sense. Let's say that we randomly draw 100 data values from a single normal distribution and store these numbers in a vector of the 100 values. We can think of the successive positions in the vector as successive sampling positions along a transect in space. Because there is no relationship between location in the vector and data value, we will obtain a random subsample of ten values whether we choose values from ten random positions in the vector, systematically pick every eleventh value in the vector , or even take the first 10 values in the vector. Actually, in this situation there is no way to obtain a nonrandom subsample if we do not look at the data values when choosing a sampling position in the vector.

A variation of this example will show the effect of violating the criteria. Let's say we rearrange the data values in our vector so that they are in ascending order from the first vector position to the last. This arrangement would violate the first criterion in that it would produce a "spatial trend" along the vector, and also violate the third criterion in that there would be strong correlation between difference in position along the vector and difference in data values. Now, if we used the first ten values in the vector as our subsample, we would only obtain the ten lowest data values, which is obviously a highly biased sample. If we systematically select every eleventh position in the vector (positions 1, 12, ...,89,100), and compute the mean and variance of the subsample, we will obtain an unbiased mean, but the variance of the chosen data values will tend be *larger* than the variance from a sample chosen from random positions in the vector. This is because the systematic sample will always contain the highest and lowest values, whereby this would not necessarily be true for a random sample. If we were to repeat the whole process (randomly selecting the 100 data values from a normal distribution, rearranging the data from low to high values in the vector, and systematically sampling along the vector) a large number of times, the *variance of the mean* of the multiple systematic subsamples would be *lower* than the variance of the mean that would be obtained from multiple random subsamples. This is true even though the individual systematic subsample means are unbiased. The reason for this result is that the systematic subsamples will be more "balanced" than random subsamples, in that the systematic subsamples will tend to contain more equal numbers of high and low values, which will tend to cancel each other out in the mean computations.

These examples demonstrates some ways that systematic or nonrandom selection of sampling locations could affect estimates of both the variance of the population and the variance of the population means. The proposed statistical techniques depend on reasonable estimates of these variances, so we need to evaluate the

degree to which our sampling design fulfills the three criteria, and when the criteria are not met, we need to determine the effects of this on our test results.

The correlation between differences in indicator values and distance between stations (criterion 3 above) is called *spatial autocorrelation* (Cliff and Ord 1981, Jumars et al. 1977, Jumars 1978, Sokal and Oden 1978), and is commonly measured by an index called Moran's I, which is

$$I = \left(\frac{n}{W}\right) \frac{\sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} w_{ij} z_i z_j}{\sum_{i=1}^{n} z_i^2}, \qquad (27)$$

where n is the number of stations, and $z_i = x_i - \bar{x}$, with x_i being the indicator value for station i and $\bar{x}$ the mean of all stations. The w_{ij} values are weights that convey the spatial information by giving more weight to station pairs that are closer together in space. In this application, we used

$$w_{ij} = \frac{1}{d_{ij}},$$

where d_{ij} is the distance in km between stations i and j. Finally,

$$W = \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} w_{ij}.$$

The expected value of I when the null hypothesis of no spatial autocorrelation is true is

$$E(I) = \frac{-1}{n-1}.$$

To evaluate the probability that the null hypothesis was true for a given computed value of I, we used a randomization technique that involved building a null distribution of I values by randomly shuffling the indicator values among the stations 2000 times, and for each shuffle, computing the value of I. The probability of the actual I value occurring by chance was obtained by comparing the actual I value to the null distribution.

The I index will be sensitive to violations to all three criteria (for random order), since trends and natural strata in the reference area will cause at least some stations in relatively close proximity to have similar indicator values. For each of the seven groups of reference stations, we computed the spatial autocorrelation for each

indicator. Where significant spatial autocorrelation was found, we included the autocorrelation in the simulation model to show the effect of the autocorrelation on the statistical test results.

It should be emphasized that gradients, strata, and autocorrelation are only a potential problem when the data are not randomly sampled (de Gruijter and ter Braak 1990, Legendre 1993). If we had sampled randomly, and found high autocorrelation among the data values, the autocorrelation would not invalidate the statistical tests. In addition, even if the sampling is random, it is still important that the reference area sampled is *representative* of what the impact locations might look like in the absence of the impact. Finally, if systematic sampling is used, the analyst needs to be aware that spatial cycles in indicator values can bias the means and variance estimates if the spacing between stations is a multiple of the spatial period of the cycles. However, in practice, such cycles are usually unlikely.

The simulation model

Figure 5 summarizes the simulation model used to evaluate the proposed statistical tests. Each simulation produces a data matrix (X values in Figure 5) that is applied to each of the proposed statistical tests. For each matrix, *the null hypothesis is true*, since the randomly generated X values contain no changes due to impact. By repeating the simulations many times (30,000), we can determine the long-term behavior of the different statistical tests.

The method of evaluating the long-term behavior of the statistical tests was different for the two approaches. For all the t statistics computed with the *first approach*, we rejected the null hypothesis in a simulation when the test indicated that the impact station was significantly different from the mean of the reference station distribution. If a particular test is performing well, the proportion of rejections of the null hypothesis in all simulations should approximate the nominal type-1 error of the test (we used $\alpha = .05$ for all tests).

In the *second approach*, both methods define statistical intervals that are supposed to cover the *pth* percentile of the reference station distribution $1-\alpha$ proportion of the time, or, in other words, the statistical interval defined should fail to cover the *pth* percentile α proportion of the time. Since we know the actual variance components used in the simulation, we can compute the underlying variance of the reference station distribution

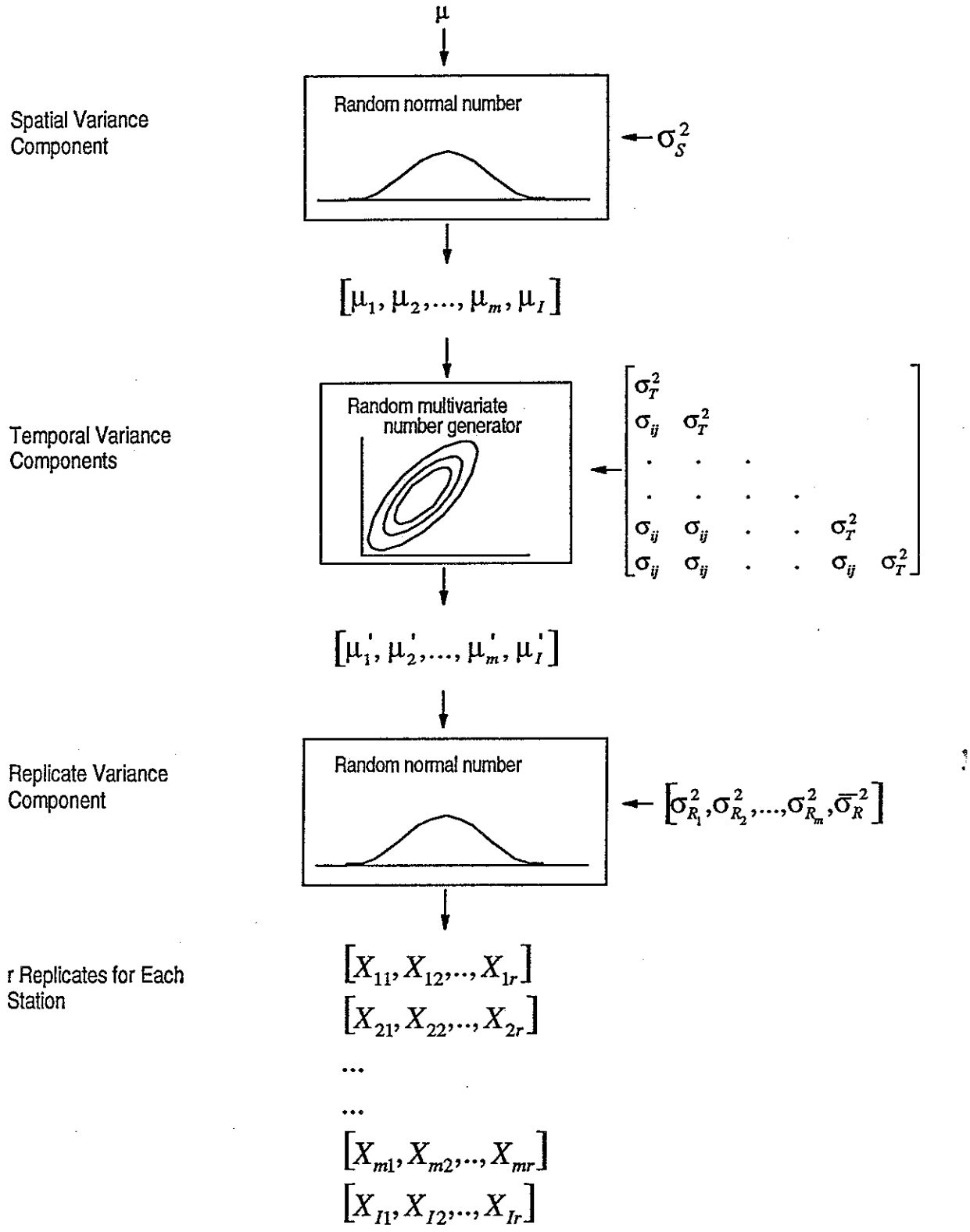

Figure 5. Diagram of the simulation model used to test the proposed procedures. Starting at the top, the spatial variance component (σ_S^2) and the overall mean (μ) are input to a univariate normal random number generator to produce indicator means for each of the m reference stations and the impact station (subscript I). These means and a variance-covariance matrix are input to a multivariate random number generator (Johnson 1987) to add the temporal variances to the means. In the variance-covariance matrix, the mean temporal variance for the stations is in the diagonal, and the mean covariance over time for the station pairs are in the off-diagonal. Finally, the resulting means are input to a univariate normal random number generator to create r replicate indicator values for each station, using the estimated $\sigma_{R_i}^2$ for each station. The X values constitute the data matrix for a single simulation.

(using equation 4) and find the indicator value for the actual *pth* percentile. We can then count the proportion of simulations in which the computed interval bound did not cover this indicator value for the *pth* percentile. This proportion should approximate the nominal type-1 error level (α) if the test is performing well. Note that these statistical tests are in effect testing the null hypothesis that the underlying mean of the indicator value at the impact station is *at worst* equal to the *pth* percentile of the reference station distribution. When we reject this null hypothesis, we say there is an impact. If the underlying mean of the impact station is exactly equal to the *pth* percentile, we should reject the null hypothesis of no impact α proportion of the time.

To test specific violations of test assumptions, we altered the model accordingly. For example, we varied the number of replicates among the different reference and impact stations, or we increased the variability of the replicate variances among the stations.

To include spatial autocorrelation in the simulations we modified the model as follows.

1. Instead of generating the underlying station means in the first step (i.e., $\mu_1, \mu_2, \ldots, \mu_m$ in Figure 5) with a random normal number generator, means reflecting the actual station means were used. These means were rescaled to have a variance equal to σ_S^2 and a mean equal to the actual mean of the station means. This configuration of means will retain the spatial autocorrelation present in the original data.
2. With the spatial autocorrelation present, the expected value of the true spatial variance component is not the original σ_S^2, but some other variance that would be realized if we had sampled randomly instead of systematically. To estimate the true spatial variance component, we modeled a random sampling procedure of a transect with the actual station spacing and including the underlying mean values (described in step 1) at the stations. We used a cubic spline technique to model the relationship between the mean indicator value and the spatial position. The cubic spline will predict an indicator value for any spatial position along the transect by interpolating indicator values between successive spatial positions with a smooth curve. Where there were very large distances between successive station (e.g., from station 22 to 50), we used only a small spatial gap, in order to minimize interpolations between such stations. We did not want to interpolate across the large spatial gaps because these gaps contained non-reference conditions.

3. To model random sampling, we randomly selected m spatial locations along the transect, and used the cubic spline to obtain indicator values for the chosen locations. The variance of the m indicator values was computed.
4. Step 3 was repeated 2000 times. The new spatial variance component was computed as the mean of the 2000 variances computed in step 3.
5. This new spatial variance component was input to a univariate random normal number generator to generate the mean for the impact station in the first step (μ_I), and also used in equation 4 to find the variance of the reference station distribution from which the percentiles are computed (for the evaluation of the approach 2 tests).

RESULTS

The raw indicator values for the reference stations are presented in Appendix A (Table A1). This table should be helpful in observing the nature of the nonnormality and spatial autocorrelation indicated below.

Normality Tests

Table 5 shows the results of normality tests for seven groups of reference stations. When including all stations (column 2), three indicators are associated with a probability less than *.10*. This nonnormality is mainly due to the presence of the outlier stations 70 and 71, since the probabilities for the same indicators are much higher when stations 70 and 71 are removed (column 3). The only other group of stations showing much nonnormality is all the southern stations (next to last column). Again this nonnormality is due to stations 70 and 71, since the probabilities increase when these two stations are removed (last column). In summary, these results indicate that there is no compelling evidence that our assumption of normality is inappropriate for the present data. The nonnormality observed for some of the indicators is removed when the outlier stations 70 and 71 are excluded.

Table 5. Probability from the Shapiro-Wilk tests for normality for seven groups of reference stations (using the 1977 SCCWRP reference stations). Low probabilities indicate evidence of nonnormality. Probabilities less than .10 are emboldened. Reference station group numbers are shown in parentheses in column headings (see Table 4).

Indicator	All	All but 70-71 (6)	All N (5)	N of SMB (1)	SMB (2)	All S (7)	All S but 70-71 (3)
D	**.09**	.69	.55	.84	.40	**.02**	.87
H	.64	.20	.12	.18	.29	.45	.40
I5	.37	.38	.32	.32	.83	**.09**	**.07**
ITI	.11	.26	.28	**.09**	.23	**.01**	.84
J	.56	.32	.22	.58	.29	.81	.61
N_CRUS	.11	.16	.37	.39	.49	.35	.72
N_ITI1	**.05**	.39	.93	.80	.97	**.05**	.30
NSPEC	**.02**	.69	.46	.51	.42	**.01**	.74
T_AMP	.54	.72	.74	.12	.61	**.01**	.52
T_ECH	.77	.87	.95	.28	.54	.17	.44
T_ITI1	.67	.48	.56	.63	.39	.41	.14
T_POLY	.74	.59	.95	.92	.89	**.03**	.37
TOTAB	.66	.17	.31	.41	.77	.91	.96

Spatial Autocorrelation

The last two columns in Table 6 contain the spatial autocorrelation results. In the south, the results for reference station group 3 (stations 50-62) suggest that a sediment gradient (see Table A1) may be causing the autocorrelation observed in about one half of the indicators. The same group of stations with stations 70 and 71 added (reference group 7) shows autocorrelation in all but one indicator. Stations 70 and 71, which are quite different from the other southern stations, comprise a separate stratum that greatly increases the autocorrelation.

In the north, all but one indicator exhibit autocorrelation when Santa Monica Bay is included with the more northerly stations (reference station group 5). It appears that Santa Monica Bay forms a natural stratum that causes the autocorrelation. When the northern stations (group 1) and Santa Monica Bay (group 2) are analyzed separately, the spatial autocorrelation is much lower.

When spatial autocorrelation is present, one way to possibly eliminate the autocorrelation is to drop stations so the distance between adjacent stations increases. We found that we could eliminate just about all the spatial autocorrelation from the entire set of reference stations when no two reference stations were closer than 20 km apart. Unfortunately, this process reduced the number of reference stations from 31 to ten, so this is not an attractive option.

Table 6. (Following pages) The percent type-1 error from the simulations for each proposed test, the estimated percent variance components, and spatial autocorrelation (SA) results. The nominal type-1 error used in the tests was α=5%, so tests performing well will show a simulation type-1 error around 5%. **Test1-test4** are the t statistics from the first approach, **Tol Int** and **Hm3** are the tolerance interval and Hampel Standardization 3, respectively from the second approach. The **I** column is the Moran *I* spatial autocorrelation index, and **P(I)** is the probability of obtaining the Moran *I* value by chance alone. When *P(I)<.05* (bold type), the autocorrelation in the original data was included in the simulation model. All simulations were run with one replicate at all stations, except for tests 1, 3, and 4, which were run with two replicates at all stations. Results are presented separately for each reference station group (see Table 4). The indicator symbols are explained in Table 3. See text for discussion of the pattern of results.

Reference Station Group 1 (North of Santa Monica Bay)

	% Type-1 Error						% Variance Components				SA	
Indicator	Test1	Test2	Test3	Test4	Tol Int	Hm3	σ_{ij}	σ^2_S	σ^2_{TxS}	σ^2_R	I	P(I)
D	10.9	4.9	5.7	5.2	5.2	6.2	26	24	27	23	-.01	.2770
H	28.8	5.1	5.0	5.2	5.2	6.4	4	52	40	4	-.21	.9900
I5	32.7	4.9	5.2	5.2	4.9	6.2	2	47	47	3	-.15	.8355
ITI	8.0	5.1	6.4	5.1	5.2	6.3	46	17	15	22	-.03	.3365
J	38.6	5.1	5.0	4.9	5.0	6.4	1	56	42	1	-.09	.5835
N_CRUS	12.1	5.0	5.6	5.1	5.0	6.0	41	23	21	14	.06	.1180
N_ITI1	10.3	4.9	5.9	5.2	4.8	6.4	28	24	24	23	.03	.1150
NSPEC	3.7	5.1	12.9	9.7	5.0	6.5	48	5	7	40	.07	.0570
T_AMP	22.6	5.1	5.1	5.1	5.2	6.4	20	39	34	7	-.11	.6775
T_ECH	19.7	4.9	4.9	4.9	5.0	6.2	24	38	29	9	-.11	.6450
T_ITI1	15.7	**3.1**	**3.2**	**3.6**	**1.5**	**3.8**	27	31	28	13	.11	**.0350**
T_POLY	31.5	**4.2**	**4.1**	**4.1**	**3.5**	**5.5**	10	32	55	3	.10	**.0335**
TOTAB	20.0	**3.8**	**4.1**	**4.3**	**2.9**	**5.0**	11	29	49	11	.19	**.0055**

Reference Station Group 2 (Santa Monica Bay)

	% Type-1 Error						% Variance Components				SA	
Indicator	Test1	Test2	Test3	Test4	Tol Int	Hm3	σ_{ij}	σ^2_S	σ^2_{TxS}	σ^2_R	I	P(I)
D	7.7	4.9	7.1	5.6	5.1	6.4	46	3	28	23	-.02	.1840
H	25.3	5.1	5.0	5.0	5.0	6.4	7	29	58	7	-.01	.1115
I5	32.5	4.8	5.2	5.2	5.1	6.7	3	47	46	3	-.08	.4155
ITI	7.9	4.9	6.9	5.5	5.1	6.4	39	20	16	25	-.03	.2185
J	37.6	**4.8**	**4.7**	**4.3**	**3.8**	**5.8**	2	40	57	2	.07	**.0315**
N_CRUS	9.2	5.0	6.0	4.9	4.9	6.3	48	12	23	17	.00	.1215
N_ITI1	8.4	5.0	6.8	5.5	5.2	6.4	28	11	30	30	-.11	.6515
NSPEC	2.8	5.2	15.6	12.5	4.8	6.6	65	0	5	31	-.01	.1310
T_AMP	22.7	**3.6**	**3.6**	**3.3**	**2.3**	**5.0**	24	36	33	7	.05	**.0340**
T_ECH	20.0	**3.6**	**3.5**	**3.1**	**1.9**	**4.7**	28	37	26	8	.05	**.0245**
T_ITI1	15.0	5.1	5.5	5.2	5.1	6.6	36	28	24	12	.01	.1090
T_POLY	33.3	5.0	4.9	4.9	4.9	6.5	15	33	50	2	-.06	.3260
TOTAB	17.0	5.1	5.1	5.0	5.1	6.4	17	20	49	15	-.01	.1495

Reference Station Group 3 (Southern, excluding 70, 71)

	% Type-1 Error						% Variance Components				SA	
Indicator	Test1	Test2	Test3	Test4	Tol Int	Hm3	σ_{ij}	σ^2_S	σ^2_{TxS}	σ^2_R	I	P(I)
D	9.3	4.5	4.8	4.5	3.4	5.2	34	12	29	24	.11	.0045
H	26.8	3.7	3.7	3.7	1.9	3.7	5	41	49	6	.12	.0040
I5	32.3	5.0	5.1	5.1	5.1	6.2	3	46	48	3	-.09	.8440
ITI	8.2	4.2	5.4	4.1	3.2	5.0	47	17	14	22	.05	.0385
J	38.2	3.8	3.5	3.4	1.5	4.1	1	48	49	1	.09	.0140
N_CRUS	11.8	5.1	5.5	5.1	5.0	6.1	44	19	22	15	.04	.0550
N_ITI1	9.3	3.9	4.7	4.3	2.9	4.8	31	16	26	27	.10	.0105
NSPEC	3.5	4.7	13.1	10.1	4.4	6.2	54	2	6	37	.08	.0160
T_AMP	22.3	4.9	5.1	5.1	5.1	6.3	23	35	35	7	.00	.1815
T_ECH	19.9	4.8	5.1	4.9	4.9	6.3	27	35	29	9	.01	.1445
T_ITI1	15.6	5.1	5.3	5.1	4.9	6.3	31	28	28	13	-.04	.3910
T_POLY	32.2	5.0	5.2	5.1	4.8	6.0	13	33	52	3	-.05	.5285
TOTAB	18.7	4.7	5.1	5.0	4.8	6.1	13	28	47	12	-.11	.9525

Reference Station Group 4 (70,71)

	% Type-1 Error						% Variance Components				SA	
Indicator	Test1	Test2	Test3	Test4	Tol Int	Hm3	σ_{ij}	σ^2_S	σ^2_{TxS}	σ^2_R	I	P(I)
D	2.8	5.1	15.1	13.6	12.4	7.7	49	0	27	24		
H	25.1	4.7	5.3	6.0	11.4	7.6	9	18	65	8		
I5	31.5	3.7	3.9	4.5	8.7	7.0	2	46	49	3		
ITI	6.3	3.6	7.9	7.5	9.1	5.9	43	19	14	24		
J	37.7	4.1	4.2	4.8	10.2	8.1	2	36	60	2		
N_CRUS	5.8	3.8	9.3	9.0	9.4	6.6	46	11	25	18		
N_ITI1	2.9	5.1	15.1	13.7	12.6	6.9	37	0	30	33		
NSPEC	2.8	5.3	14.9	13.4	12.6	7.6	64	0	5	31		
T_AMP	19.9	2.7	4.0	3.7	6.7	6.8	23	35	35	7		
T_ECH	15.1	1.1	1.3	.9	2.9	4.7	27	36	28	9		
T_ITI1	11.1	1.9	2.8	2.7	4.7	4.6	35	28	25	12		
T_POLY	32.4	3.7	4.0	4.5	9.9	7.0	14	32	51	3		
TOTAB	13.9	4.7	7.1	7.0	11.7	8.3	19	8	55	18		

Reference Station Group 5 (All northern)

	% Type-1 Error						% Variance Components				SA	
Indicator	Test1	Test2	Test3	Test4	Tol Int	Hm3	σ_{ij}	σ_S^2	$\sigma_{T\times S}^2$	σ_R^2	I	P(I)
D	12.1	1.9	2.0	1.7	.3	1.6	25	28	25	21	.26	.0010
H	28.9	1.0	1.0	.9	.0	.7	3	56	37	4	.31	.0005
I5	32.3	3.0	3.1	3.2	.7	2.8	2	49	45	3	.14	.0065
ITI	8.3	5.4	6.3	6.1	4.6	6.1	42	18	16	23	.32	.0005
J	38.3	1.2	1.3	1.1	.0	1.0	1	58	40	1	.26	.0005
N_CRUS	12.9	2.5	2.4	2.1	.5	1.9	41	25	20	14	.18	.0010
N_ITI1	11.1	3.1	3.4	2.9	1.1	3.1	25	28	24	22	.09	.0325
NSPEC	4.2	3.2	7.8	5.4	1.5	3.4	49	6	6	39	.25	.0005
T_AMP	23.2	4.4	4.3	4.8	2.1	4.1	19	41	33	7	.14	.0085
T_ECH	20.5	4.1	3.9	4.2	1.5	3.5	23	40	28	9	.13	.0110
T_ITI1	15.8	4.8	5.1	5.0	5.2	6.1	26	33	28	13	.05	.0975
T_POLY	32.2	2.7	2.8	2.7	.8	2.7	11	32	54	3	.15	.0075
TOTAB	19.3	2.6	2.9	2.9	.9	3.2	11	28	50	11	.13	.0045

Reference Station Group 6 (All stations excluding 70,71)

	% Type-1 Error						% Variance Components				SA	
Indicator	Test1	Test2	Test3	Test4	Tol Int	Hm3	σ_{ij}	σ_S^2	$\sigma_{T\times S}^2$	σ_R^2	I	P(I)
D	12.4	3.3	3.3	3.0	.5	1.8	20	35	24	21	.16	.0025
H	30.0	3.0	3.2	3.0	.2	1.4	3	58	34	4	.16	.0035
I5	32.1	5.2	5.1	5.1	5.1	6.2	3	50	44	3	.04	.0840
ITI	8.7	5.5	6.8	6.6	5.4	7.4	37	19	18	25	.21	.0005
J	39.9	3.4	3.6	3.4	.4	2.7	1	59	39	1	.15	.0040
N_CRUS	13.7	3.0	3.2	2.8	.4	1.3	39	28	19	14	.12	.0085
N_ITI1	12.2	3.9	4.4	4.0	1.5	3.2	20	32	26	22	.08	.0345
NSPEC	4.3	3.7	7.3	5.2	1.7	3.2	40	9	7	44	.17	.0020
T_AMP	21.9	5.1	5.3	5.6	3.3	5.2	14	41	37	8	.17	.0020
T_ECH	20.4	4.9	5.1	5.3	2.9	4.6	16	40	34	10	.17	.0035
T_ITI1	16.0	4.3	4.3	4.5	2.1	4.6	17	32	37	14	.14	.0045
T_POLY	31.8	3.2	3.2	3.1	.9	2.9	10	33	54	3	.18	.0015
TOTAB	20.0	3.2	3.1	3.1	.8	3.2	8	31	50	11	.21	.0005

Reference Station Group 7 (All southern)

	% Type-1 Error						% Variance Components				SA	
Indicator	Test1	Test2	Test3	Test4	Tol Int	Hm3	σ^2_{ij}	σ^2_S	$\sigma^2_{T\times S}$	σ^2_R	I	P(I)
D	10.8	4.2	4.6	4.3	2.7	4.5	30	21	27	22	.18	.0005
H	27.4	3.3	3.2	3.3	1.0	3.0	4	50	42	5	.17	.0005
I5	32.4	3.8	3.9	4.0	1.6	4.3	3	48	46	3	.09	.0045
ITI	8.2	4.4	4.9	3.8	2.8	6.8	45	18	15	22	.15	.0005
J	39.0	3.6	3.2	3.2	1.1	3.3	1	55	43	1	.13	.0015
N_CRUS	11.9	3.7	4.1	4.0	2.0	3.8	43	21	22	15	.10	.0065
N_ITI1	10.2	3.9	4.4	4.2	2.3	4.1	28	23	24	24	.18	.0005
NSPEC	3.8	5.1	12.0	8.7	4.3	5.7	52	4	6	37	.17	.0005
T_AMP	23.0	2.2	2.2	2.2	.4	4.6	21	39	33	7	.12	.0010
T_ECH	20.3	2.7	2.6	2.7	.5	4.3	26	39	27	8	.13	.0010
T_ITI1	16.3	3.5	3.6	3.7	1.6	3.8	29	31	27	12	.06	.0260
T_POLY	32.0	4.0	4.2	4.6	2.4	4.4	12	32	53	3	.09	.0085
TOTAB	19.2	4.9	5.0	5.0	5.2	6.4	12	28	48	11	-.03	.4210

Simulation Results

Table 6 includes some of the simulation results and the variance components used in the simulations (expressed as percents). Where spatial autocorrelation was found (P(I)<.05), the autocorrelation was included in the simulation model (emboldened results). The simulations in Table 6 included the same number of replicates at all stations. The replicate variance at the stations varied a little, since a separate replicate variance was predicted for each station. We now discuss the pattern of results in Table 6 for each test.

Test 1

Test 1, a standard t test comparing a single reference station with an impact station and utilizing only the replicate variance, generally has a type-1 error well above the nominal type-1 error level of 5%. At the upper extreme, when the replicate variance is relatively low compared to the spatial and time by space interaction variance components, the type-1 error can get close to 40% (e.g., see results for J, evenness diversity for all groups). At the lower extreme, the type-1 error of the test is closest to 5% when the replicate variance is high in relation to the spatial and interaction variance components. For example, for all the reference station groups, the number of species (NSPEC) has relatively high replicate variance, and the type-1 error ranges from 2.8 to 4.3%. This pattern of results was expected, since the test does not incorporate the spatial and interaction variance

components, and these components contribute less to the variance of the reference station means when the replicate variance is relatively high.

The results show that test 1 should be avoided unless it is known that the replicate variance is large relative to the spatial and time by space interaction variance components.

Tests 2, 3, and 4

These three tests incorporate all the expected variance components. The simulations included equal replication and relatively small variation in replicate variance among the stations, so when no spatial autocorrelation is present, we would expect these tests to perform well, since none of the assumptions of either test are greatly violated. Our initial discussion excludes reference station group 4, which includes only two reference station and could not be measured for spatial autocorrelation.

Where no spatial autocorrelation was detected, the type-1 error for test 2 is in all cases close to the nominal rate of 5%. Tests 3 and 4 also perform well, except when the replicate variance is high relative to the spatial and time by space variance components (e.g., see NSPEC for groups 1-4, 7). This is apparently due to the fact that these two tests estimate a value for $S^2_{S,T\times S}$, and when the replicate variance is relatively high, the probability of a negative value for the $S^2_{S,T\times S}$ increases (equations 14 and 16). When a negative estimate occurs, a value of zero is used in the test, since a variance cannot be negative. The resulting underestimation of $S^2_{S,T\times S}$ causes a smaller denominator and the null hypothesis will tend to be rejected more often (inflating the type-1 error a bit). The output from the simulation model confirmed that the number of negative $S^2_{S,T\times S}$ estimates was high when the type-1 error was inflated.

When spatial autocorrelation is present, these tests in general tend to produce a type-1 error that is below the nominal rate of 5%. At the worst, when the spatial variance component is quite high relative to the interaction and replicate variance components, the type-1 error rates can get as low as 1% (e.g., see H for reference station group 5). This pattern is expected, since the autocorrelation is contained in the spatial variance component, and we expect the systematic sampling in the presence of autocorrelation to inflate the variance of the reference station

means. This inflated variance will tend to inflate the denominator of the t statistic and reduce the type-1 error level below the nominal level.

Tolerance interval, Hampel identifier

Without spatial autocorrelation, all these methods perform well. In the presence of spatial autocorrelation, the type-1 error can be lower than the nominal rate, again depending on the relative size of the spatial variance component. In some cases, the Hampel identifier seems more robust to the effects of the autocorrelation. For example, with reference station group 1, the average type-1 error for the last three indicators is 2.6 and 4.8% for the tolerance interval and Hampel identifier, respectively.

Reference station group 4

The results for this group seem to be more variable, probably due to the fact that there are only two stations in the group. The tolerance interval type-1 error is inflated when the spatial variance component is relatively low. These results suggest that more than two reference stations will provide more stable type-1 error levels.

Variations in the replicate variance and replication level

Test 2, the tolerance interval, and Hampel identifier all assume that the replicate variance and the replication level are equal at all stations. Tests 3 and 4 have relaxed assumptions on the replication level, but still assume equal replicate variances at the reference stations. We ran additional simulations to test the robustness of the various methods when these assumptions were violated. We also ran simulations for some tests to see the effect of including more replicates at all stations.

Since these assumptions and additional tests all involve the replicate number or replicate variance, the maximal effect will be observed where the replicate variance is relatively high compared to the spatial and interaction variance components. For this reason, we ran additional simulations with the number of species (NSPEC) in reference group 3, a situation with relatively high replicate variance (Table 7).

The type-1 error for test 1 increases as the replication increases (Table 7). This would be expected since the higher replication decreases the relative contribution of the replicate variance and accentuates the influence of the missing spatial and interaction variance components.

With more replicates, the type-1 error for tests 3 and 4 gets closer to 5% (Table 5, rows a-d). The larger number of replicates improves the results since fewer negative estimates of $S^2_{S,T\times S}$ are produced. The improvement is much more rapid with test 4, which gets close to 5% with only three replicates, while test 3 does not get as close until there are five replicates. As demonstrated in the methods section, using the data in Table 1, the computed t values will be the same for tests 3 and 4 when the numbers of replicates are the same at all stations. Since the number of replicates is the same for both tests in rows a-d, the different results for test 4 must be due to the estimate of the degrees of freedom (equation 21).

When there are five replicates at the impact station and only one or two replicates at the reference stations (Table 7, row f), the assumption of equal replication at all stations is violated for all but tests 3 and 4. The type-1 error for tests 3 and 4 is still high due to the negative variance estimates discussed above. The remaining tests show a somewhat depressed type-1 error level, except for the Hampel identifier, which seems to be more robust to this assumption. The depression of the type-1 error would be due to the fact that the variance of the impact station mean is lower than expected, due to the extra replication at the impact station (i.e., in equation 4, the value for r is greater for the impact station than for the reference stations).

Table 7. The results of additional simulations with variations of replicate numbers and replicate variances using the number of species (NSPEC) for reference group 3, which is associated with relatively high replicate variance. The first four rows (a-d) show the effects of increasing the replicate numbers on tests 1, 3, and 4. Row e shows the effect of five replicates at the impact station but only one or two replicates at the reference stations. Row f shows the effect of unequal replicate numbers among the reference station (varying between 1 and 5, with 5 replicates at the impact station). In row g, the replicate variance at the impact station is double the average replicate variance at the reference stations, and in row h, the replicate variance is one half that at the reference stations. In row i, the replicate variance at the reference stations varies among the stations (alternating double and half and equal the replicate variance used in Table 6). In rows g-i, five replicates were used at all stations for tests 3 and 4.

Simulation	Test1	Test2	Test3	Test4	Tol Int	Hm3
Approx. equal rep variance						
a. 2 reps, all stations	3.5	4.5	13.1	10.1	4.1	5.6
b. 3 reps, all stations	5.7		9.0	5.9		
c. 4 reps, all stations	7.1		7.8	5.1		
d. 5 reps, all stations	8.2		6.4	4.8		
e. 5 reps, Impact station	6.2	1.2	11.9	9.6	4.0	6.3
f. 1-5 rep reference, 5 reps Impact		1.5	5.8	6.3	1.6	4.2
Varying rep variance						
g. Impact double rep variance		9.3	6.2	4.6	3.8	5.0
h. Impact half rep variance		1.7	7.3	5.6	4.0	6.2
i. Reference rep variance varying		4.2	7.0	5.0	4.5	6.6

In Table 7, rows g and h, the replicate variance at the impact station is double and one half, respectively, times the replicate variance at the reference stations. This situation violates an assumption for all methods except tests 3 and 4. The tolerance interval and Hampel standardization 3 seem relatively robust to this violation, while test 2 is more sensitive. As expected, tests 3 and 4 perform well, but with test 4 closer to 5%.

Row i in Table 7 involves varying the replicate variance among the reference stations. None of the methods seem particularly sensitive to the varying replicate variances used in the simulation.

It should be emphasized that we chose a worst-case situation to test the robustness of the methods (Table 7). In cases where the replicate variance is lower relative to the spatial and interaction variance components, these methods will be more robust than demonstrated here. This makes sense since the violations considered all involve variability in station means due to the replicate variance, and when the replicate variance is a smaller proportion of the variance of the station means, the violations of these assumptions will become less important.

Application With Test Data

We now apply the proposed methods to the 1977 SCCWRP data set. It should be emphasized that the purpose of the following analyses is not to thoroughly analyze the data, but only to demonstrate how one might approach an analysis with the proposed statistical techniques. The results shown only apply to the status of the benthos in 1977. Since that time, all the sewage dischargers have greatly improved the quality of their effluents, with corresponding changes in the benthos (e.g., Stull et al. 1986). We also want to avoid the implication that the methods are only applicable to benthic infaunal data.

Patterns of Change in Indicator Values With Expected Impacts

When applying these techniques to actual data, it is important that the analyst understand the behavior of the chosen indicators. First of all, when the expected type of impact occurs, the direction and pattern of change in indicator values should be known. Figure 6 summarizes the relationship between the indicators and the expected outfall gradient in the area. For indicators increasing with impact, we will be using the upper bounds (U values) of the statistical intervals for the reference envelope edge (e.g., I5), and for those generally decreasing (most indicators), we will be using the lower bounds (L values).

Note that about half the indicators initially increase and then decrease along the outfall gradient (we analyze these as if they were decreasing with impact). These will not be the more sensitive indicators, since we will not be measuring the effects of the impact until the indicator values decrease below levels in the reference area. By this point, we are fairly far along the outfall gradient.

Other indicators (e.g., T_AMP, T_ECH, T_ITI1) decrease rapidly as we progress away from reference conditions. These indicators will be very sensitive to an impact, but will not be as useful for quantifying the degree of impact beyond a certain point, since these indicators level out fairly quickly with increasing impact.

The I5 and ITI indicators tend to follow the outfall gradient in a more linear manner, and as such may be more useful quantifying the degree of impact over the entire range of impacts. There is some circularity in this argument since we used ITI to quantify the outfall gradient in the first place (in Figure 6). However, our experience has shown this to be generally true.

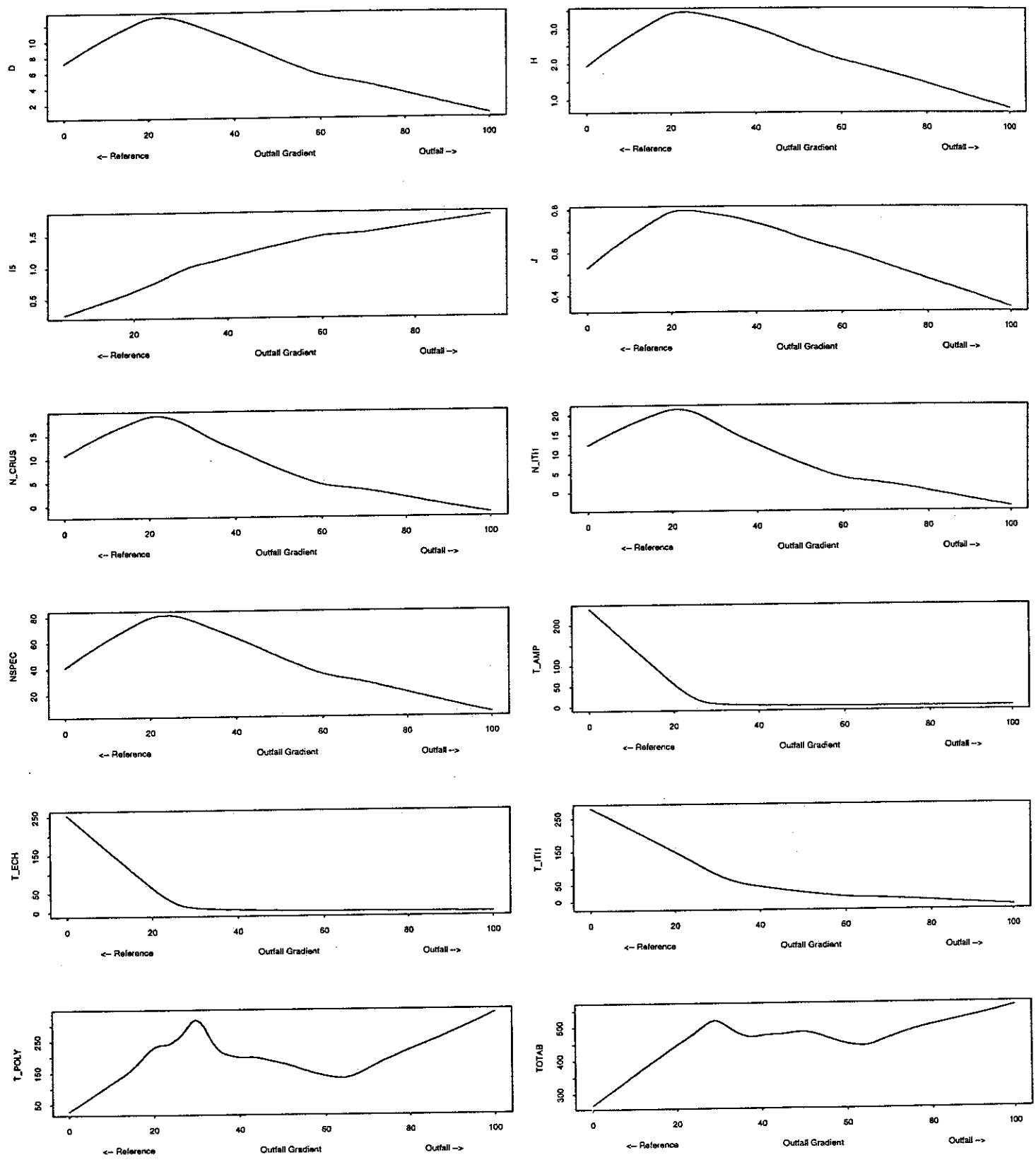

Figure 6. Smoothed trend of the indicator values along the outfall gradient in the Southern California Bight. The outfall gradient is quantified as 100-ITI, where ITI is the infaunal trophic index (Word 1978). LOWESS smoothing (Chambers and Hastie 1993) was applied to all the data to produce the curves.

The patterns observed for the species diversity measures and total abundance match the patterns described by Pearson and Rosenberg (1978) concerning the effects of organic enrichment on benthic communities.

Relationships Between Indicators and Habitat

In the Southern California Bight, the natural benthic communities on the mainland shelf vary mainly with depth and sediment grain size (Jones 1969). With the test data used in the application, all stations were taken at 60 meters depth, so changes in indicator values from variation in depth were not a concern. However, within the data sample, there is a very wide range in sediment grain size, which ranges from 10 to 95% silt-clay (Tables A1 and A2 in Appendix A). Figure 7 shows the relationships between the sediment grain size and the indicator values.

We perform two separate analyses illustrating different approaches to habitat variation and display of results. In the first analysis, we use the linear relationships between the indicators seen in Figure 7 to mathematically remove the effect of sediment size from the data, and use the adjusted data in the statistical tests. When the effect of sediment size is removed from the data, we reduce the contribution of varying sediment size to the spatial component of the background variability, while at the same time we minimize the probability that we will confuse an impact with differences in sediment size in the reference and impact areas. Thus, as discussed in the section on the spatial variance component, we are increasing the test sensitivity without sacrificing test validity. In addition, if the spatial autocorrelation found in the data is at least partly due to sediment gradients, then removing the effect of sediment size can remove some of the spatial autocorrelation. Removal of spatial autocorrelation from the reference stations will tend to produce more sensitive statistical tests, since the autocorrelation tends to depress the type-1 error rate below the nominal type-1 error level (Table 6), indicating an overprotective test.

A second analysis is performed where we do not adjust the indicators for sediment size. In this case, we will need to carefully consider the choice of reference stations so that differences in sediment size are not confused with impact.

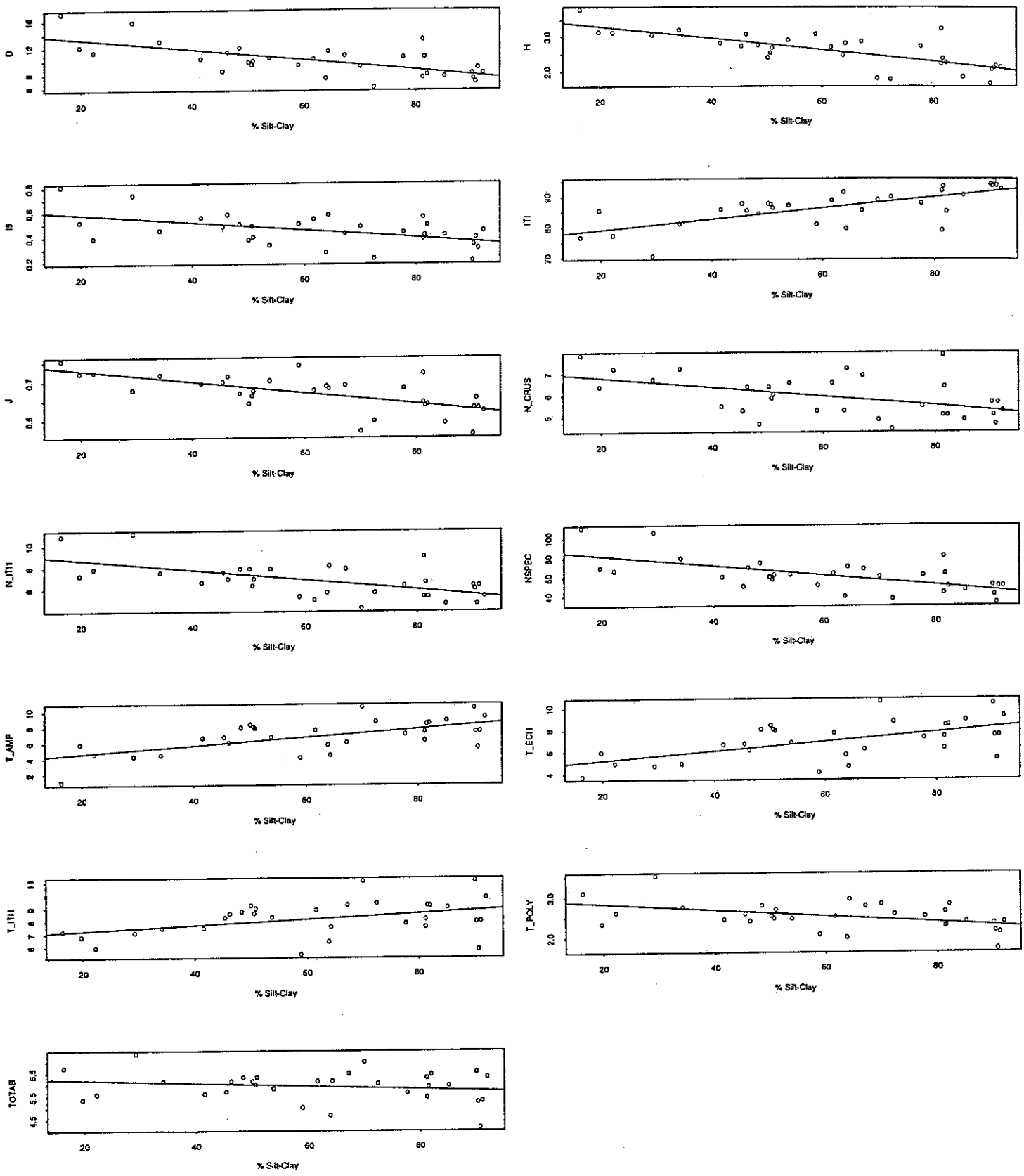

Figure 7. The relationship between the indicators and the sediment grain size (% silt-clay) at the reference stations. The trend lines were produced with linear regression. Some of the indicators were transformed as indicated in Table 3. The regression slopes for all indicators except TOTAB are significantly different from zero ($\alpha < .05$).

Analysis 1

Using the reference stations only, we used simple regression to fit a linear relationships between percent silt-clay and each indicator (Figure 7). The resulting regression equations were then applied to all the 1977 SCCWRP data, and the regression residuals were used as the data values in the subsequent analyses. The residual indicator value at a station includes the variation not explained by the sediment grain size at that station. This procedure assumes that we do not consider a change in sediment size as an impact.

We reran the spatial autocorrelation analyses after removal of the sediment size effects (Table 8). When comparing these results with the original analyses (Table 6), the most dramatic change observed was the reduction in spatial autocorrelation for reference station group 3, which includes the stations above San Diego in the south. This is the area where we suspected that a sediment gradient (Table A1) was causing the autocorrelation. The removal of the sediment size effects has eliminated the sediment gradient and the related autocorrelation. Since reference station group 3 is not associated with any problems from spatial autocorrelation, we use reference station group 3 to define the reference envelope in this analysis.

Table 8. Spatial autocorrelation results after removal of the effect of sediment size. The probability associated with the Moran I value is shown (same as *P(I)* in Table 6). See Table 4 for the membership of the reference station groups.

	Reference Station Group					
Indicator	1	2	3	5	6	7
D	**.0090**	**.0280**	.7230	**.0245**	**.0475**	**.0155**
H	.2920	.0515	.4285	**.0495**	**.0230**	.2465
I5	.4825	.3530	.8160	.3150	.5800	.1660
ITI	.3000	**.0450**	.3545	**.0105**	**.0190**	.1535
J	.6260	.1260	.4165	.2915	.0150	.4725
N_CRUS	**.0245**	**.0495**	.5070	**.0265**	**.0240**	.3325
N_ITI1	.0595	.3065	.5785	.2455	.3300	**.0085**
NSPEC	**.0080**	**.0195**	.6630	**.0200**	**.0340**	**.0135**
T_AMP	.3870	.2175	.8490	.4545	**.0030**	.1065
T_ECH	.4515	.1860	.8790	.4325	**.0020**	.1845
T_ITI1	.1035	.1850	.9470	.2350	**.0045**	.6760
T_POLY	**.0340**	.1700	.2170	.0520	**.0025**	.2050
TOTAB	**.0095**	.0695	.6210	**.0190**	**.0005**	.5940

Using the data from reference group 3, we computed the envelope edge (interval bounds, equation 6) for each indicator using test 2 (equation 12). We eliminated the data for the T_POLY and TOTAB indicators, since they would be difficult to interpret, due to the multimodal pattern along the outfall gradient (Figure 6). Test 2 was chosen over tests 3 and 4, since there was only a single replicate at all stations, and tests 3 and 4 require at least two replicates. To conserve space and avoid redundancy, we only show the results for a single test (test 2). In the second analysis, we show output where the results for the different statistical tests can be directly compared for an indicator.

In order to present the analysis results in a concise form, we rescaled and centered the indicator data to reflect the position of each data point in relation to the reference envelope edge. Here we rescaled the residuals data for each indicator to a scale from 0 to 100, and then subtracted the (rescaled) value of the reference envelope edge from each rescaled data value. Positive rescaled data values are inside the reference envelope (acceptance of the null hypothesis of no impact), and negative values are outside the envelope (indicating a probable impact). The more negative the value, the greater the impact. With the data in this form, the results for all indicators can be shown on a single figure (Figures 8-11).

In Figures 8-11, the indicators are ordered on the vertical axis by expected sensitivity to an impact, which was determined by examination of the patterns of the indicators along the outfall gradient (Figure 6). The most sensitive indicators are toward the bottom, and the least sensitive indicators are toward the top of the vertical axis. Successive figures from Figure 8 to 11 show the stations in the vicinity of individual outfalls, beginning with the outfall showing the least impact (Figure 8) to the outfall showing the greatest impact (Figure 11). Figures 8-11 allow us to observe the patterns of all the indicators at once, and provide a broader perspective on the nature and extent of probable impacts.

Analysis 2

Here we are not adjusting the indicator data for the influence of sediment size, so we need to consider whether our statistical tests will tend to confuse sediment size differences with impacts. We note that the sediment sizes in the vicinity of the outfalls range from 20 to 79 percent silt-clay (Table A2). If our chosen reference stations

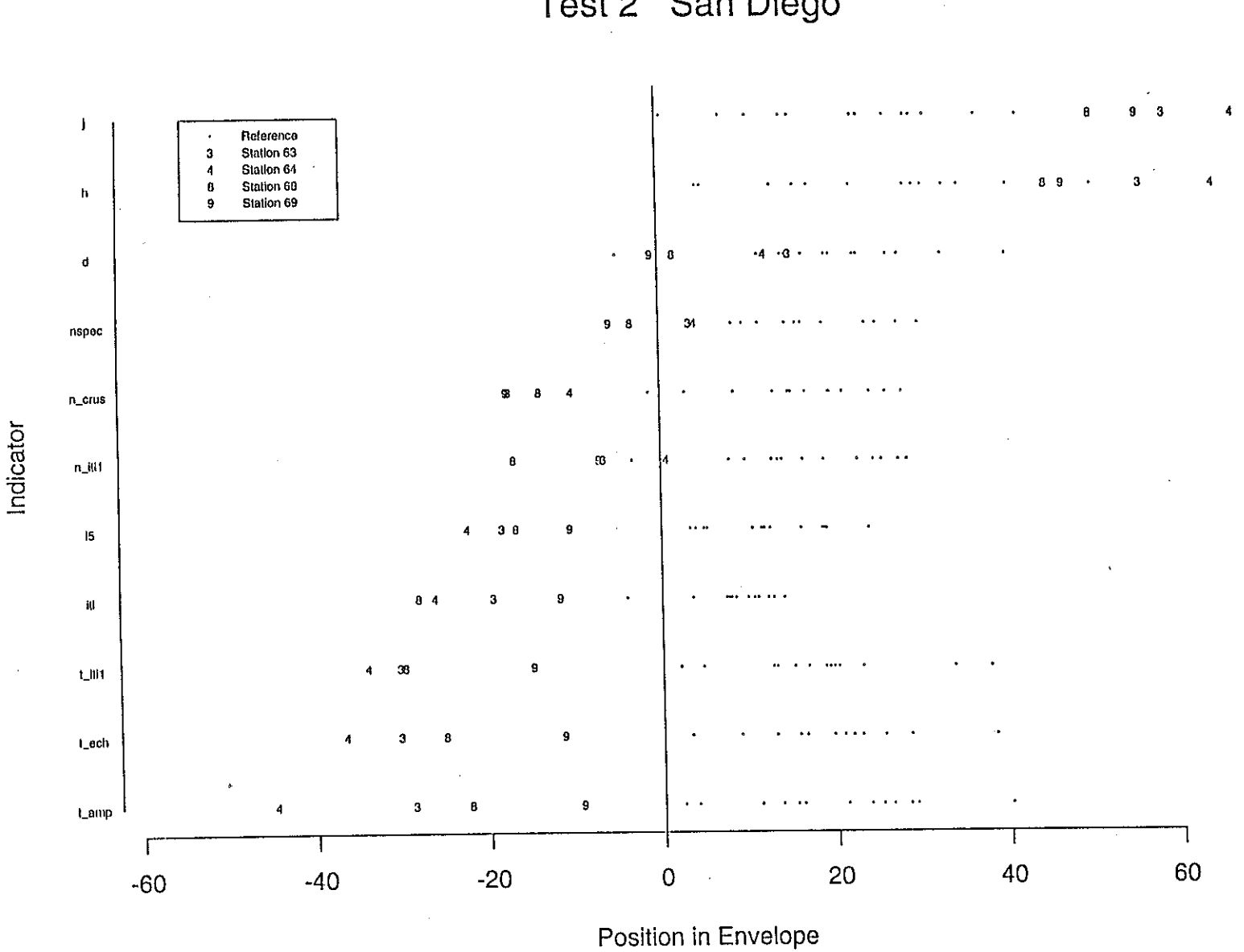

Figure 8. Results of the first analysis for the impact stations in the vicinity of the San Diego outfall. Negative indicator values are outside the reference envelope (impacted), and positive indicator values are inside the reference envelope (unimpacted). The symbols for the stations are the second digit of the station number.

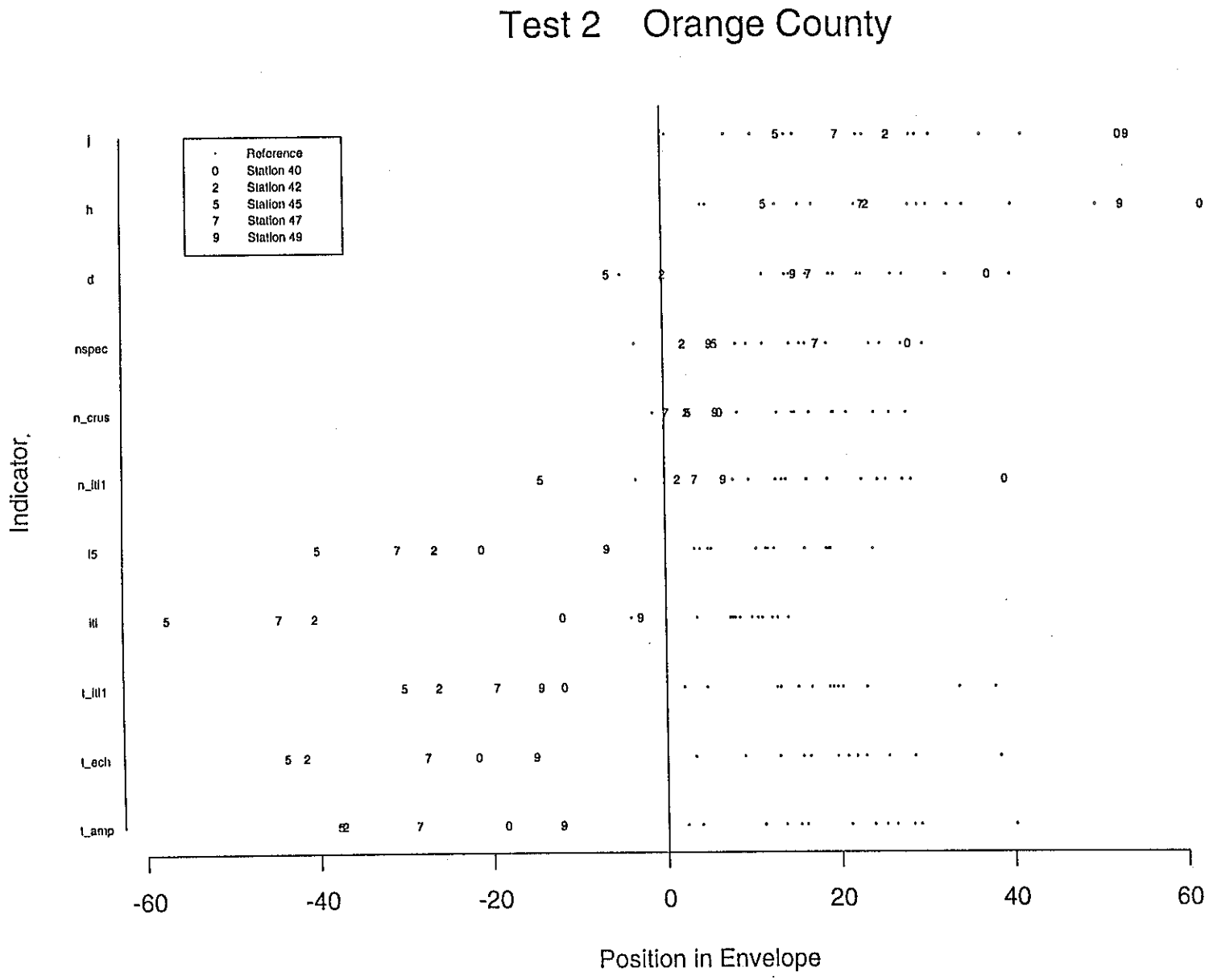

Figure 9. Results of the first analysis for the impact stations in the vicinity of the Orange County outfall. Negative indicator values are outside the reference envelope (impacted), and positive indicator values are inside the reference envelope (unimpacted). The symbols for the stations are the second digit of the station number.

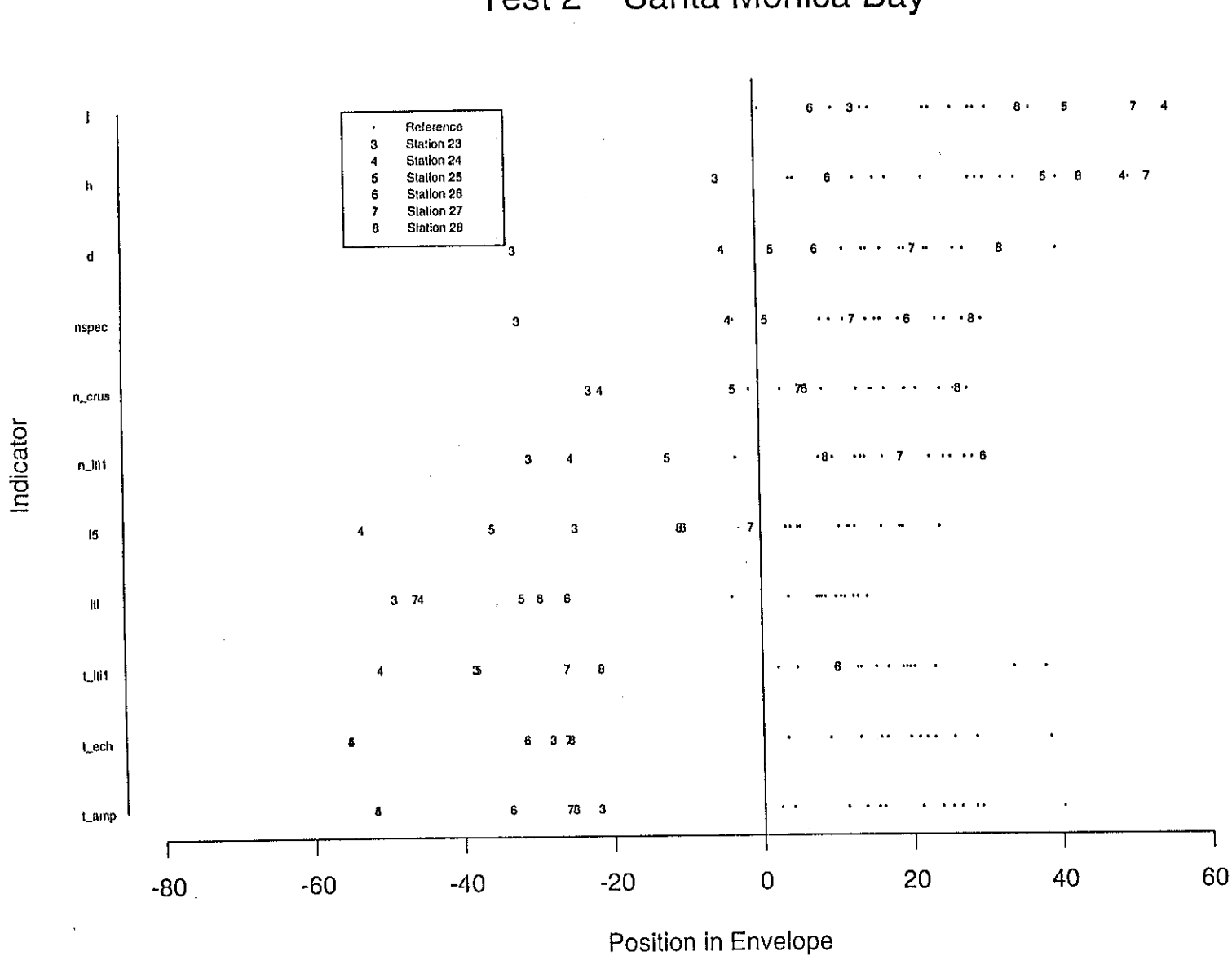

Figure 10. Results of the first analysis for the impact stations in the vicinity of the City of Los Angeles outfall in Santa Monica Bay. Negative indicator values are outside the reference envelope (impacted), and positive indicator values are inside the reference envelope (unimpacted). The symbols for the stations are the second digit of the station number.

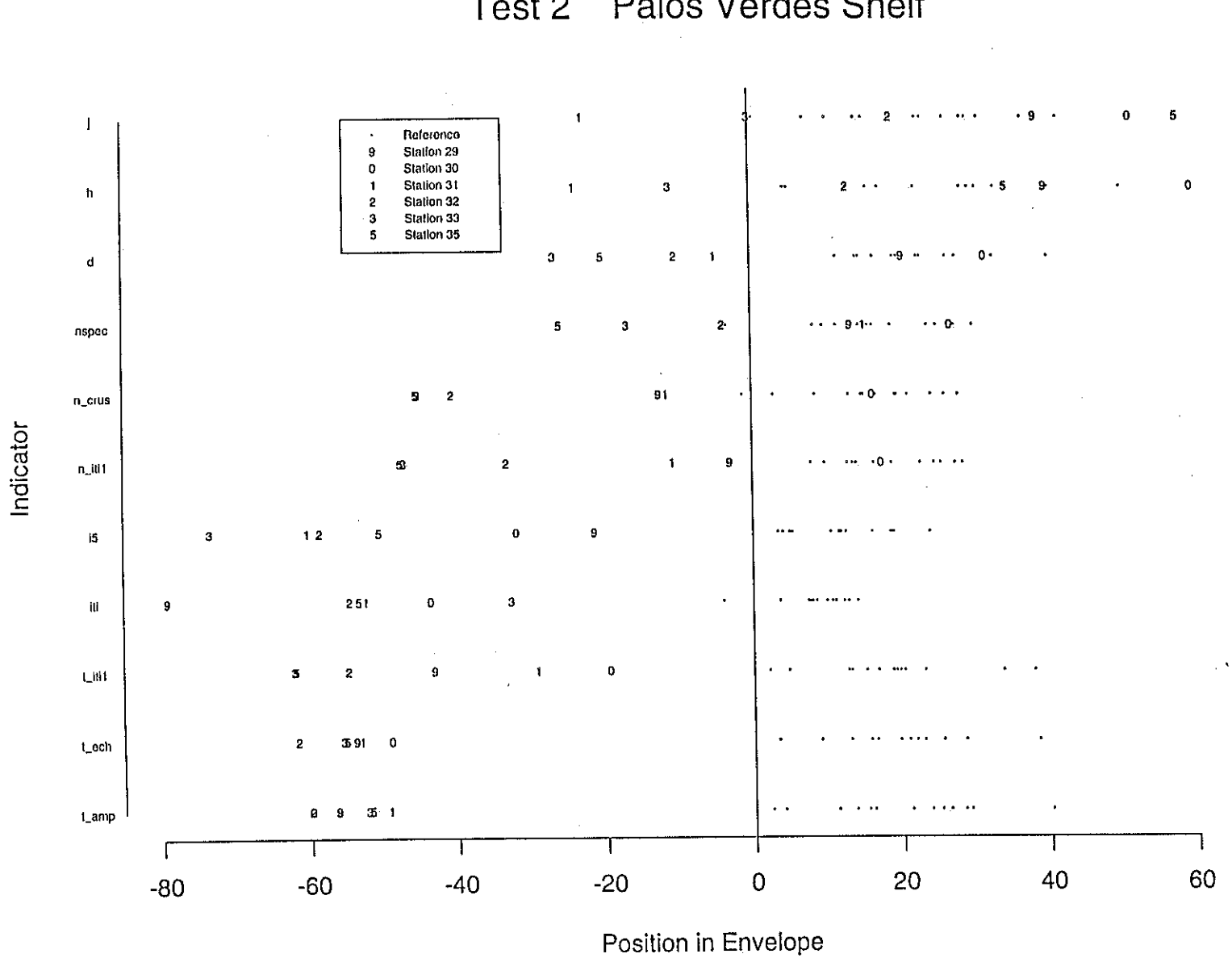

Figure 11. Results of the first analysis for the impact stations in the vicinity of the Los Angeles County outfall on the Palos Verdes Shelf. Negative indicator values are outside the reference envelope (impacted), and positive indicator values are inside the reference envelope (unimpacted). The symbols for the stations are the second digit of the station number.

all had fine sediment (say 85-100% silt-clay), or very coarse sediment (say 0 to 15% silt-clay), then our statistical tests would tend to confuse sediment differences with impacts, since the impact and reference stations will cover nonoverlapping ranges of sediment size. To prevent this, we want to include the effect of sediment size differences in the spatial variance component (i.e., include sediment effects in the background variability). This can be accomplished by choosing reference stations with a mix of sediment sizes. Accordingly, reference station group 1 (north of Santa Monica Bay), with a range of sediment sizes ranging from 20 to 81 percent (Table A1), was chosen for this analysis.

In this analysis, we computed the reference envelope edges (interval bounds) for test 2, the tolerance interval, and the Hampel identifier. For the last two methods, we used four different values of p. We could not utilize tests 1, 3 and 4, since there was no replication at the stations. Here we display the results in a manner that allows us to focus on a *single indicator* (Table 9) at a time. This choice of display has nothing to do with the type of analysis, but is included to show another useful way that results can be displayed. When displaying the results as in Table 9, it is easy to compare results from different statistical tests and p values (or even different α values, if desired). We can see that test 2 is the most sensitive of the tests, with the tolerance interval and Hampel identifier being more conservative, as expected. The Hampel identifier is consistently more conservative than the tolerance interval (with p held constant).

Another advantage with this type of display is that we can see what percentiles of the reference station distribution the underlying mean at the impact station would need to equal for us to falsely declare an impact. For example, if we were to declare that station 45 was impacted, we would be confident that this was a good decision, since for us to be wrong, the underlying Index 5 mean for station 45 would have been *by chance* greater than the 99.9th percentile of the reference distribution, according to the Hampel identifier. Or, if we were using the tolerance interval, the underlying Index 5 mean for station 45 would have been by chance greater than the 99.99th percentile of the reference distribution. Since these are unlikely chance events, we would gain confidence in a decision to declare station 45 as impacted, given the values of the indicator in question.

Table 9. Results with a single indicator (Index 5) in the second analysis. The Index 5 values are sorted in ascending order, since the index values increase with impact. Interval bounds (envelope edges) for different statistical tests and p values are shaded. For the station types, REF=reference, SD=San Diego outfall area, OC=Orange County outfall area, SMB=City of Los Angeles outfall area in Santa Monica Bay, and PV=County of Los Angeles outfall area on the Palos Verdes shelf.

Statistical Test	Index 5 (I5)	Station Type	Station
	.20	REF	21
	.23	REF	19
	.24	REF	18
	.26	REF	4
	.27	REF	20
	.34	REF	17
	.35	REF	14
	.37	REF	7
	.41	REF	13
	.44	REF	22
	.48	REF	8
	.50	REF	5
	.50	REF	12
	.51	REF	11
	.53	REF	16
	.53	REF	15
	.56	SMB	27
Test 2	.68		
	.71	OC	49
	.74	SD	69
Tolerance Interval p=90	.75		
	.81	SMB	28
Tolerance Interval p=95	.82		
	.82	SMB	26
	.87	SD	63
	.87	PV	29
	.88	SD	68
Hampel Identifier p=90	.91		
	.93	SD	64
	.94	OC	40
Tolerance Interval p=99	.96		
Hampel Identifier p=95	1.01		
	1.02	OC	41
	1.03	OC	39
	1.07	OC	42
	1.07	OC	48
	1.08	PV	30
	1.10	SMB	23
Tolerance Interval p=99.9	1.12		
Hampel Identifier p=99	1.19		
	1.19	SMB	25
	1.19	OC	47
	1.19	OC	44
Tolerance Interval p=99.99	1.25		
	1.25	OC	46
	1.26	OC	43
Hampel Identifier p=99.9	1.34		
	1.34	OC	45
	1.43	PV	35
	1.47	SMB	24
Hampel Identifier p=99.99	1.51		
	1.51	PV	32
	1.60	PV	31
	1.80	PV	33

DISCUSSION

The detection of impacts in environmental monitoring is a serious undertaking, since undetected impacts may harm the environment, or frequent detection of false or ecologically insignificant impacts can lead to a waste of resources and discredit important environmental protection programs. We have defined and evaluated statistical tests that could be used to test for impacts when no before-impact data are available. A large percentage of monitoring programs lack sufficient before-impact data, so there is ample need for such techniques.

Related Approaches

Underwood (1989), Keough and Quinn (1991), and Smith (1991) emphasize the need for multiple reference locations when testing for impact or treatment effects. These approaches also include sampling before and after the onset of the impacting activity. In this paper, we are proposing that even in the absence of before-impact data, we can still make inferences concerning differences between the impact location and an appropriate reference area, since the variation among the reference locations can be used to estimate the natural background variability among stations separated in space. Our approach would not be as sensitive as a design incorporating before and after impact data, but it often will be preferable to the application of an inappropriate test (e.g., test 1) or no test at all.

The idea of a reference envelope delimiting a population of reference locations is conceptually similar to the multivariate approaches of Bloom (1980) and Hughes et al. (1990). Clarke and Green (1988) emphasize the need to include a "relevant level of between-area variation" in the background error term of the statistical test for impact when comparing spatially separated locations.

Boswell et al. (1994) introduce the idea of the "crystal cube", where one "side of the cube" conceptually corresponds to the "edge of the reference envelope" for a single indicator. A side of the cube is defined with a one-sided prediction interval for a single future observation (Whitmore 1986, Hahn and Meeker 1991, Vardeman 1992). The formula for the bound of such a prediction interval is identical to the formula for the reference envelope edge when using test 2 (equation 12 rearranged in the form of equation 6). However, the bound defined by our test 2 would differ from the prediction interval bound of Boswell et al. (1994). The background error variance computed for their test will contain different variance components, due to the fact that their variance is

computed from multiple samples in both space and time. (The variance in test 2 , $S^2_{\bar{X}_c}$, is computed from samples taken at one only period in time.) The Boswell et al. (1994) sampling plan involves sampling a *single* randomly-located reference station per year, so multiple years of sampling would often be required before a sufficiently powerful test is available. Additional power could be obtained by sampling multiple reference stations each year. Davis (1994, p 837) derives a compuational formula for the pertinent variance when multiple reference stations are sampled at each time.

Which approach, the "reference envelope" or "crystal cube", is preferable will depend on the situation. If there is only one effective sampling period (and at least two reference stations), then the reference envelope approach is preferred since multiple sampling periods are required for the crystal cube approach. Beyond this situation, the better choice will depend on the relative power or sensitivity of the two approaches, which in turn will depend on the number of effective sampling times, the number of reference stations, and the relative magnitudes of the spatial, temporal, and replicate variance components.

As far as variance components are concerned, the main difference between the two approaches is the manner in which the temporal variance is utilized. With the reference envelope, the interaction between time and space (σ^2_{TxS}) contributes to the background error variance, where with the crystal cube approach, the "parallel time-to-time component" (i.e., σ_{ij}) contributes to the background error variance (Davis 1994).

Multiple Comparisons

The prediction intervals discussed above (including our test 2) involve setting a bound on a single future observation, but frequently we will be interested in applying the same bound to several separate observations (impact stations). This suggests that when applying the same prediction interval bound to several impact stations, the proportion of impact stations found outside the bounds will tend to be inflated due to the multiple comparisons. Hahn and Meeker (1991, p 63) give tables and a (conservative approximate) formula for prediction interval bounds for more than one future observation. The formula corrects for the multiple comparisons by applying the Bonferroni inequality to the nominal type-1 error level, i.e., α / m is used instead of α for the critical $t_{\alpha,df}$, where m is the number of comparisons applied to the bound. The *tolerance interval*, on the other hand, is designed so that

all future observations are within the computed bounds with the nominal error rate (Vardeman 1992), so no adjustments are needed for multiple comparisons when using a single indicator. Incidentally, there is no reason that the background error variance used in the crystal cube approach could not be used with a tolerance interval.

Multiple comparisons will also be involved when any of the statistical tests are applied to multiple indicators. Jones (1984), and Davis (1994) discuss some approaches to adjusting tests for multiple comparisons. The price paid for adjusting the statistical tests for multiple comparisons is a loss of power for the individual tests. There is still controversy over the extent to which tests should be adjusted. For example, Saville (1990), in reference to multiple pairwise comparisons of treatment groups, proposed that no adjustments for multiple comparisons be applied when using the tests for generating hypotheses (See also letters generated by Saville's article in Holland 1991, Lea 1991, and Saville 1991). Most often, regulators will judge the severity of impacts with a weight of evidence approach, which could be thought of as a hypothesis generation process involving evaluation of *patterns* of test results and information from multiple sources. For this reason, we did not apply any adjustments for multiple comparisons in the applications of the reference envelope with the benthic data

Pseudoreplication and the Background Error Variance of the Tests

In the classic paper by Hurlbert (1984), the term *pseudoreplication* is defined "in analysis of variance terminology, as the testing for treatment effects with an error term inappropriate to the hypothesis being considered". A clear example of pseudoreplication is the use of the standard t test (test 1) when testing for impacts. The error term of this test is based on small-scale spatial (replicate) variance, but we are comparing (reference and impact) stations that can separated by relatively large distances (compared to the small-scale spatial differences), so it would not be surprising that such a test would often detect "impacts", when in fact the test is only detecting differences in the underlying indicator means at two geographically-separated, unimpacted stations. The error term of the test contains no information on the natural variability of indicator values on the larger geographic scale. This is consistent with the greatly inflated type-1 error of this test in our simulations.

In order to avoid pseudoreplication in manipulative ecological experiments with a spatial dimension, Hurlbert (1984) states that a treatment area should not be spatially segregated from a control or reference (nontreatment) area, but instead replicate treatment and reference locations should be *interspersed* in space to

avoid confusing spatial differences with treatment differences. The logic behind this recommendation is that, even without a treatment effect, any two areas in space will almost certainly differ somewhat in their underlying mean dependent variable (indicator) values. In a pseudoreplicated ecological experiment we would subject multiple randomly-chosen locations within one area to a treatment, and also randomly choose multiple untreated locations within a separate area. We would then compare the mean indicator values in the two areas using a statistical test with the aim of discovering whether there are any treatment effects. When we reject the null hypothesis, we may only be detecting differences in underlying indicator means in the two areas rather than a real treatment effect. Even when there are only small differences in the underlying means for the two areas, we increase our chances of detecting this difference as a treatment effect as we increase the number of sampling locations within the two areas. However, if the treatments are *interspersed* in space (instead of segregated), and we compare treatment locations with the untreated locations, we will not be confusing treatment effects with spatial differences, since we are no longer also comparing two segregated areas in space.

Our proposed general approach to detecting impacts could be thought of as an ecological experiment with the impacting activity serving as a treatment applied to a small area in space (the impact station). To detect a treatment effect (impact), we compare the impact station with a larger (unimpacted or untreated) reference area, from which we obtain a sample of multiple reference stations. Even though we might have multiple impact stations, we are considering the impact stations *one at a time* in our statistical tests. We are only considering a single impact station at a time because impacting activities often set up strong spatial gradients in indicator values in the impact area, and we will usually have an interest the condition of individual locations along the gradient (in contrast to only being interested in the condition of the entire area). In addition, these indicator gradients in the impact area would greatly inflate our background error variance if we used the variance between different impact stations in the estimation of the background error variance. This would make our test much less sensitive to the impact.

Our approach clearly involves comparing two segregated areas in space with a statistical test, and as such, we need to consider whether our tests are subject to serious problems from pseudoreplication. Multiple impact (treatment) stations cannot be interspersed in space with multiple reference stations, since the location of the impacting activity is usually fixed, and by necessity the typical sampling design involves a reference area

segregated from an impact location (Stewart-Oaten et al. 1986, Eberhardt and Thomas 1991). At times, the impact area could be spatially interspersed within a larger reference area. This may reduce the pseudoreplication, but with only one impact area, there is still some segregation.

When we have segregated reference and impact areas, a useful design would involve a random sample of multiple reference areas. From the variance of the means of the reference areas, we could estimate a background error variance that would reflect the variability in mean indicator values among segregated areas in space. With such a design, we would not expect pseudoreplication to be a problem, since the background error variance incorporates expected differences among the areas, and we will not tend to confuse impacts with spatial differences.

Since we are only considering a single impact station at a time, we are concerned with the underlying mean indicator value in a limited subarea of the entire impact area. From this perspective, our pseudoreplication problem becomes a problem of distinguishing actual impacts from underlying differences between "limited subareas" of space. If we consider each of our reference stations to be a sample (of one) from a subarea of the larger reference area, then the background error variability for our proposed statistical models will incorporate expected variability among the subareas, and pseudoreplication should not be a problem. What we call a reference area in the statistical tests is a matter of scale, and when we consider individual impact stations in our statistical tests, the relevant scale involves smaller areas that can be represented by individual stations rather than relatively large areas.

One of the properties of a pseudoreplicated statistical test comparing two segregated areas for treatment effects is that any underlying mean indicator differences existing between the segregated reference and treatment areas will be detected as an impact if we just sample a sufficient number of stations within both areas (when the null hypothesis is true). This is because the background error term for such a test would include the sum of the variance of the means for the two areas, and as the number of stations sampled within the areas increases to a large number, the variance of the means will tend toward zero. Recall that the variance of a mean is S^2/n, where n is the number of stations sampled, and S^2 is the variance of the stations, and as n increases, S^2/n diminishes toward zero. As the error term of the test tends toward zero, more minute underlying differences in the areas sampled will tend to be detected as impacts. However, with the proposed tests, the error variances will *not* tend

toward zero as the number of stations sampled increases. With tests 2-4, the error variance tends toward the underlying variance of the reference station distribution (rather than zero) as the number of reference stations sampled increases (see equation 12). For the tolerance interval, the measure of (background error) variability used to compute the interval bounds is the estimated variance of the reference station means (equations 22 and 23), and as we sample more reference stations, this estimated variance also tends toward the underlying variance of the reference station distribution. (The Hampel identifier is similar in concept to the tolerance interval). Thus, to detect an impact with the proposed tests, the difference between the impact station mean and the mean of the reference station means in a survey must exceed differences that would be predicted from the estimated variance of the reference station distribution. This is why we emphasize that the reference stations be chosen from an area that includes at least a level of physical and geographic variation that might exist between a reference station and the impact station. When this is the case, the background error variability of the tests should contain variability commensurate with the spatial differences of the segregated impact and reference locations, and this situation will prevent the tests from confusing spatial differences with impacts. The scale of the desired physical and geographic variation in the reference area will depend somewhat on the indicators utilized, since different indicators can vary in their sensitivity to different environmental factors and geographic separation.

The removal of the effect of sediment size on the indicator values in our first analysis demonstrated how we can use relevant information to decrease the background error variance due to habitat (sediment size) differences (which are part of the spatial variability), while at the same time lower the probability an impact will be confused with habitat differences between impact and reference locations. This type of procedure can enhance both the sensitivity and the validity of the statistical tests.

Test Assumptions

The simulations show that the proposed tests generally work well and the test applications have produced reasonable results. It is not surprising that the simulation model produced good results, since it incorporated the basic assumptions of the statistical tests. However, when these basic assumptions are violated in practice, the results of the statistical tests could potentially be very misleading. Therefore, it is imperative that the analyst

thoroughly consider the assumptions associated with the use of the proposed techniques. We discuss the more important assumptions in more detail below.

Random Sampling of Reference Stations

When the positions of the reference station locations are not chosen at random, we need to consider the effect of our sampling design on the statistical tests. Most commonly, we will have some sort of systematic sampling design. When the spatial autocorrelation in the indicator values is absent or relatively low, even with a systematic sampling design, the tests will perform as if the station location were randomly chosen. As the degree of (positive) spatial autocorrelation increases, the tests become more conservative as far as type-1 error is concerned, i.e., the actual type-1 error tends to be lower than the nominal (α) type-1 error. This fact also implies that the tests are less sensitive to impacts. Consequently, when we suspect that spatial autocorrelation is present for an indicator, and an impact is detected, the autocorrelation should not cast doubt on the conclusion of impact. This is because the autocorrelation makes it more difficult to detect an impact, but in spite of this, an impact was still detected. On the other hand, if an impact was not detected, this may be partly due to the autocorrelation.

When applying inferential statistics to data with spatial structure (and non-random sampling), the computations can in some cases be modified to adjust for spatial autocorrelation (Griffith 1978, Cliff and Ord 1981, Legendre et al. 1990, Legendre 1993, Kleinn 1994). Future research is needed to determine how the proposed tests could be adjusted for autocorrelation.

Indicator Values at the Impact Station Under the Null Hypothesis

Our tests 2-4 assume that the (unimpacted) impact station would appear as a random selection from the reference station distribution. Thus, it is assumed that the impact station most frequently would have an underlying mean indicator value closer to the mean of the distribution, and less frequently have underlying mean values toward the tails of the distribution. The tolerance interval and Hampel outlier identifier also assume that the impact station originated from the reference station distribution, but we make a more pessimistic assumption that the impact station originated from the tail of the distribution toward an indication of an impact. This allows for more caution in our assumptions regarding the unimpacted state of the impact station. Such caution would be warranted

when the costs of impact are very high and we would want to be confident that any identified impact had a high probability of being an actual impact.

Normality of Reference Station Distribution

When we have evidence that the reference station data is non-normal to an extent that it would distort the statistical test results, it is often possible to transform the data to obtain a more normal distribution (Box and Cox 1964). With our test data, we transformed some of the indicators to obtain replicate variances that were not dependent on the station mean. Although these transformations were for a different purpose, they will often also make the distribution of station means more normal.

When transformations are insufficient for this purpose, nonparametric approaches can sometimes be utilized. There are nonparametric tolerance intervals (Hahn and Meeker 1991), but to be useful for the present application, a fairly large sample size (number of reference stations) would be needed. More recently, computational methods requiring smaller sample sizes have been proposed (Vangel 1994). Randomization (Edgington 1987) or bootstrap (Efron and Tibshirani 1993) methods could possibly be applied to the proposed tests, but this is an area for future research.

If we have reason to believe that the reference station indicator values follow some known distribution other than the normal distribution, we could generate the appropriate g values for the Hampel outlier identifier using simulations with that distribution. It may be possible to do something similar with tolerance intervals.

Variance Estimates from Multiple Spatial or Temporal Strata

We may be able to substantially increase the degrees of freedom by pooling data from multiple surveys and/or strata in order to estimate the background error variance. This approach brings with it the assumption that the variance among the reference stations is the same within all strata. When considering all the temporal and spatial variance components that can affect the magnitude of this variance, we realize that in practice it may be difficult to assume that all these spatial and temporal components are constant across strata, or that they change over time or space in ways that the differences cancel out to produce the same overall variance. The extent of the problems with pooling will probably depend on the application, and if pooling over strata is being considered, the associated assumptions should be carefully evaluated.

ACKNOWLEDGMENTS

We would like to thank those providing valuable comments on the text, including Brock Bernstein and Laura Riege. We also appreciate the helpful input supplied by Allan Stewart-Oaten, Brock Bernstein, Jim Zalinski, Steve Ferraro, Neil Willits, Roger Green, Don Stevens, Brian Melzian, Jeff Cross, and Bruce Thompson. Allan Stewart-Oaten suggested the approach used in tests 2-4. Mary Bergen had the difficult job of standardizing the taxonomy of the test data. The contribution of test data from the Southern California Coastal Research Project (SCCWRP), the City of Los Angeles, the County Sanitation Districts of Los Angeles and Orange County, and the City of San Diego is much appreciated. The seed for the basic idea behind the proposed methods was planted during interactions with Stuart Hurlbert several years ago. The project was partially funded by the U.S. EPA Region IX (Grant # X-009904-01-0 to SCCWRP), under the direction of Janet Hashimoto and Terry Fleming. The material in this report does not necessarily reflect the opinions of the U.S. EPA. All errors and omissions are the sole responsibility of the author.

REFERENCES

Barnett, V. Sample Survey Principles and Methods. 1991. Oxford University Press, New York.

Bernstein, B.B., and R. W. Smith. 1986. Community approaches to monitoring. IEEE Oceans '86 Conference Proceedings: 934-939.

Bernstein, B.B., and J. Zalinski. 1983. An optimum sampling design and power tests for environmental biologists. Journal of Environmental Management **16**:35-43.

Bloom, S.A. 1980. Multivariate quantification of community recovery. Chapter 6 In: The Recovery of Damaged Ecosystems. J. Cairns, Jr. (ed), Ann Arbor Science Publications Inc.: 141-151.

Boswell, M.T., J.S. O'Connor, and G.P. Patil. 1994. A crystal cube for coastal and estuarine degradation: Selection of endpoints and development of indices for use in decision making. Chapter 24 in Handbook of Statistics, G.P. Patil and C.R. Rao, eds. Elsevier Science B. V.: 771-790.

Box, G.E.P., and D.R. Cox. 1964. An analysis of transformations. Journal of the Royal Statistical Society, Series B **26**:211-252.

Chambers, J.M., and T.J. Hastie. 1993. Statistical Models in S. Chapman and Hall, New York, NY, USA.

Clarke, K.R. and R.H. Green. 1988. Statistical design and analysis for a 'biological effects' study. Marine Ecology Progress Series 46: 213-226.

Cliff, A.D., and J.K. Ord. 1981. Spatial Processes. Poin Limited, London, England.

Davies, L., and U. Gather. 1993. The identification of multiple outliers. Journal of the American Statistical Association 88(423): 782-792.

Davis, C.B. 1994. Environmental regulatory statistics. Chapter 26 in Handbook of Statistics, G.P. Patil and C.R. Rao, eds. Elsevier Science B. V.: 817-865.

de Gruijter, J.J., and C.J.F. ter Braak. 1990. Model-free estimation from spatial samples: A reappraisal of classical sampling theory. Mathematical Geology, 22(4): 407-415.

Eberhardt, L.L. 1976. Quantitative ecology and impact assessment. Journal of Environmental Management 42: 1-31.

Eberhardt, L.L. and J.M. Thomas. 1991. Designing environmental field studies. Ecological Monographs 61(1): 53-73.

EcoAnalysis, SCCWRP, and Tetra Tech. 1993. Analyses of ambient monitoring data for the Southern California Bight. Report to U.S. EPA, Wetlands, Oceans and Estuaries Branch, Region IX, San Francisco, CA, 94105.

Edgington, E.S. 1987. Randomization Tests. Second Edition. Marcel Dekker, New York, NY, USA.

Efron, B., and R.J. Tibshirani. 1993. An Introduction to the Bootstrap. Monographs on Statistics and Applied Probability 57. Chapman and Hall, New York, NY, USA.

Elliott, J.M. 1977. Some Methods for the Statistical Analysis of Samples of Benthic Invertebrates. Freshwater Biological Association Scientific Publication No. 25.

Faith, D.P., C.L. Humphrey, and P.L. Dostine. 1991. Statistical power and BACI designs in biological monitoring: Comparative evaluation of measures of community dissimilarity based on benthic macroinvertebrate

communities in Rockhole Mine Creek, Northern Territory, Australia. Australian Journal of Marine Freshwater Research 42: 589-602.

Gilbert, R.O. 1987. Statistical Methods for Environmental Pollution Monitoring. Van Nostrand Reinhold Co., New York. 320 pp.

Green, R.H. 1979. Sampling Design and Statistical Methods for Environmental Biologists. Wiley-Interscience - John Wiley & Sons, New York: 257 pp.

Green, R.H. 1993. Application of repeated measures designs in environmental impact and monitoring studies. Australian Journal of Ecology 18: 81-98.

Griffith, D.A. 1978. A spatially adjusted ANOVA model. Geographical Analysis 10(3): 296-301.

Hastie, T.J. and R.J. Tibshirani. 1990. Generalized Additive Models. Chapman and Hall. New York. 336 pp.

Hahn, G.J. and W.Q. Meeker. 1991. Statistical Intervals. A Guide for Practitioners. A Wiley-Interscience Publication. John Wiley & Sons, Inc. New York. 392 pp.

Hampel, F.R. 1985. The breakdown points of the mean combined with some rejection rules. Technometrics 27: 95-107.

Holland, B. 1991. Comments on Saville. The American Statistician 45(2): 165.

Hughes, R.M., T.R. Whittier, C.M. Rohm, and D.P. Larsen. 1990. A regional framework for establishing recovery criteria. Environmental Management 14(5): 673-683.

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

Johnson, M.E. 1987. Multivariate Statistical Simulation. John Wiley & Sons, Inc. New York.: 230 pp.

Jones, D. 1984. Use, misuse, and role of multiple-comparison procedures in ecological and agricultural entomology. FORUM: Environmental Entomology 13: 635-649.

Jones, G.F. 1969. The benthic macrofauna of the mainland shelf of Southern California. Allan Hancock Monographs in Marine Biology No. 4: 219 pp.

Jumars, P.A. 1978. Spatial autocorrelation with RUM (Remote Underwater Manipulator): vertical and horizontal structure of a bathyal benthic community. Deep-Sea Research, 25: 589-604.

Jumars, P.A., D. Thistle, and M.L. Jones. 1977. Detecting two-dimensional spatial structure in biological data. Oecologia (Berl.) 28: 109-123.

Keough, M.J. and G.P. Quinn. 1991. Causality and the choice of measurements for detecting human impacts in marine environments. Australian Journal of Marine and Freshwater Research 42: 539-554.

Kleinn, C. 1994. Comparison of the performance of line sampling to other forms of cluster sampling. Forest Ecology and Management 68: 365-373.

Lea, P. 1991. Multiple confusions. The American Statistician 45(2): 165-166.

Legendre, L. 1993. Spatial autocorrelation: trouble or new paradigm. Ecology 74(6): 1659-1673.

Legendre, P., N.L. Oden, R.R. Sokal, A. Vaudor, and J. Kim. 1990. Approximate analysis of variance of spatially autocorrelated regional data. Journal of Classification 7(1): 53-75.

Margalef, D. R. 1958. Information theory in ecology. General Systems 3: 36-71.

Millard S.P. and D.P. Lettenmaier. 1986. Optimal design of biological sampling programs using the analysis of variance. Estuarine, Coastal and Shelf Science 22: 637-656.

Odeh, R.E., and D.B. Owen. 1980. Tables for Normal Tolerance Limits, Sampling Plans, and Screening. Marcel Dekker, Inc. New York.

Osenberg, C.W., R.J. Schmitt, S.J. Holbrook, K.E. Abu-Saba, and R. Flegal. 1994. Detection of environmental impacts: Natural variability, effect size, and power analysis. Ecological Applications 4(1): 16-30.

Pearson, T.H. and R Rosenberg. 1978. Macrobenthic succession in relation to organic enrichment and pollution of the marine environment. Oceanography and Marine Biology Annual Review 16: 229-311.

Pielou, E.C. 1969. An Introduction to Mathematical Ecology. Wiley-Interscience. New York: 286 pp.

SAS Institute Inc. 1990a. SAS/STAT User's Guide, Version 6, Fourth Edition. Volume 2. SAS Institute Inc., Cary, NC.

Saville, D.J. 1990. Multiple comparison procedures: The practical solution. The American Statistician 44(2): 174-180.

Saville, D.J. 1991. Reply to Holland and Lea. The American Statistician 45(2): 166-167.

SCCWRP and EcoAnalysis. 1993. The reference envelope approach in regional monitoring off Southern California. Results of a workshop sponsored by U.S. EPA, Region IX. August 30, 1993. Draft Report.

Searle, S.R., G. Casella, and C.E. McCulloch. 1992. Variance Components. John Wiley and Sons, New York. 510 pp.

Shapiro, S.S. and M.B. Wilk. 1965. An analysis of variance test for normality (complete samples). Biometrika 52: 591-611.

Skalski, J.R. and H. McKenzie. 1982. A design for aquatic monitoring programs. Journal of Environmental Management 14: 237-251.

Smith, M.P.L. 1991. Environmental impact assessment: the roles of predicting and monitoring the extent of impacts. Aust. J. Mar. Freshwater Res. 42: 603-614.

Smith, R.W. and B.B. Bernstein, 1985. Index 5: A multivariate index of benthic degradation. Report prepared for NOAA, under contract to Brookhaven Nat. Lab.: 118 pp. Available from authors at EcoAnalysis Inc., 221 E. Matilija St., Ojai, CA 93023.

Sokal, R.R. and Oden, N.L. 1978. Spatial autocorrelation in biology. I. Methodology II. Some biological applications of evolutionary and ecological interest. Biological Journal of the Linnean Society, 10: 199-228; 229-249.

Sokal, R.R. and F.J. Rohlf, 1981. Biometry. 2nd edn. W.H. Freeman and Co., San Francisco. 859 pp.

Stewart-Oaten, A., W.W. Murdoch, and K.R. Parker. 1986. Environmental impact assessment: "pseudoreplication" in time?. Ecology 67(4): 929-940.

Stull, J.K., C.I. Haydock, R.W. Smith, and D.E. Montagne. 1986. Long-term changes in the benthic community on the coastal shelf of Palos Verdes, Southern California. Marine Biology 91: 539-551.

Thompson, B.E., J.D. Laughlin, and D.T. Tsukada. 1987. 1985 reference site survey. SCCWRP Tech. Rept. 221: 50 pp.

Underwood, A.J. 1989. The analysis of stress in natural populations. Biological Journal of the Linnean Society 37: 51-78.

Underwood, A.J. 1994. On beyond BACI: Sampling designs that might reliably detect environmental disturbances. Ecological Applications 4(1): 3-15.

Vangel, M.G. 1994. One-sided nonparametric tolerance limits. Communications in Statistics - Simulation 23(4): 1137-1154.

Vardeman, S.B. 1992. What about the other intervals? The American Statistician 46(3): 193-197.

Whitmore, G.A. 1986. Prediction limits for a univariate normal observation. The American Statistician 40: 141-143.

Word, J.Q. 1978. The infaunal trophic index. S. Calif. Coastal Water Res. Project (SCCWRP) Annual Report: 19-39.

Word, J.Q. 1980. Classification of benthic invertebrates into Infaunal Trophic Index feeding groups. S. Calif. Coastal Water Res. Project (SCCWRP) Biennial Report 1979-1980: 103-121.

Word, J.Q. and A.J. Mearns, 1979. 60-meter control survey off Southern California. S. Calif. Coastal Water Res. Project (SCCWRP), TM 229: 58 pp.

APPENDIX A. RAW INDICATOR VALUES FOR THE SCCWRP 1977 SURVEY DATA.

Table A1. Raw indicator values for the SCCWRP reference stations in 1977 survey. Table 1 describes the indicator symbols. The last column is the percent silt-clay at the station.

Station	d	h	i5	iti	j	n_ crus	n_ iti1	nspec	t_ amp	t_ ech	t_ iti1	t_ poly	totab	% s/c
North of Santa Monica Bay														
4	11.3	3.12	.26	77	.75	27	19	66	45	54	87	122	310	22
5	10.3	2.82	.50	86	.69	17	15	60	113	115	152	88	316	42
7	9.3	3.07	.37	81	.78	16	11	51	33	34	67	37	218	59
8	12.1	3.14	.48	85	.74	22	17	69	83	87	120	70	273	20
11	13.0	3.25	.51	79	.74	31	23	81	97	99	156	126	460	81
12	11.4	2.82	.50	79	.66	27	20	70	41	46	155	218	430	64
13	12.9	3.20	.41	81	.73	27	18	79	42	53	152	156	426	34
14	10.8	2.87	.35	85	.68	25	19	68	89	97	260	161	508	67
15	11.2	3.07	.53	85	.73	22	16	69	89	93	216	80	424	46
Santa Monica Bay														
16	10.2	2.71	.53	89	.65	23	10	63	161	165	233	99	426	62
17	10.5	2.38	.34	93	.58	22	15	63	201	204	256	63	376	82
18	7.2	2.08	.24	94	.56	15	13	41	146	151	178	46	251	91
19	8.8	2.18	.23	94	.56	18	14	50	149	154	179	41	259	92
20	6.7	2.13	.27	95	.61	13	9	33	66	66	80	14	116	91
21	7.4	2.47	.20	91	.67	16	12	39	78	78	104	31	173	64
22	8.4	2.72	.44	88	.70	16	18	50	118	118	197	114	333	45
Southern														
50	7.4	2.22	.32	92	.59	15	11	43	148	148	188	60	289	81
51	7.9	2.26	.45	85	.58	15	11	50	207	207	254	171	498	82
52	8.0	2.13	.41	93	.55	16	11	50	258	259	299	74	467	92
53	8.0	1.67	.17	94	.42	18	14	51	348	349	413	68	523	90
54	7.6	1.85	.41	91	.48	14	9	46	228	233	245	77	383	85
55	6.0	1.78	.15	90	.49	12	12	37	223	224	269	111	404	72
56	9.1	1.80	.46	89	.44	14	8	60	363	365	409	179	651	70
57	10.4	2.72	.32	88	.66	17	14	61	132	142	170	100	321	78
58	9.9	2.68	.34	86	.65	20	16	62	174	178	240	139	467	51
59	9.7	2.40	.30	88	.59	22	19	60	203	204	257	104	427	50
60	9.4	2.53	.43	87	.63	19	14	57	182	184	218	92	392	51
61	10.4	2.90	.29	87	.70	23	19	62	117	122	198	89	357	54
62	11.9	2.75	.41	84	.64	13	19	74	180	183	226	168	463	48
Far South - very coarse sediments														
70	15.8	3.05	.67	70	.65	24	31	106	38	48	135	545	765	29
71	17.2	3.79	.75	77	.81	31	30	110	1	27	139	287	570	16

Table A2. Raw indicator values for SCCWRP non-reference stations in 1977 survey. The last column is the percent silt-clay at the station. Stations closest to the outfalls in the different areas are marked with an asterisk.

Station	d	h	i5	iti	j	n_ crus	n_ iti1	nspec	t_ amp	t_ ech	t_ iti1	t_ poly	totab	% s/c
Northern - oil present														
1	12.5	3.26	.74	63	.76	28	16	75	31	42	73	77	364	20
2	16.1	3.48	.83	77	.74	33	22	113	22	45	292	543	1033	13
3	12.4	3.03	.69	69	.69	22	23	81	102	110	147	194	627	31
6	10.9	3.24	.59	86	.79	14	17	60	65	68	102	56	223	43
9	6.2	2.60	.18	90	.77	7	7	29	33	33	50	18	90	95
10	7.8	3.06	.57	75	.83	13	6	40	30	31	37	48	148	94
Santa Monica Bay - near outfall														
23	6.9	2.40	1.10	53	.66	9	7	39	3	3	11	69	244	21
24	7.6	3.16	1.47	60	.82	7	5	47	0	0	6	232	414	51
25 *	8.3	2.91	1.19	66	.74	12	8	51	0	0	20	190	401	51
26	10.3	2.53	.82	66	.59	17	23	75	1	4	158	779	1331	34
27	8.9	2.92	.56	63	.74	13	14	51	26	29	53	104	269	70
28	12.4	3.17	.81	66	.73	24	15	78	8	12	48	245	510	42
Palos Verdes Shelf - near outfall														
29	8.2	2.47	.87	49	.64	7	7	47	1	2	22	51	276	79
30	10.3	3.13	1.08	64	.75	17	14	64	0	2	70	195	451	67
31	8.0	1.55	1.60	56	.37	10	9	65	0	0	33	1939	3034	46
32	5.8	2.05	1.51	58	.56	2	2	39	0	0	6	422	708	67
33 *	5.2	1.75	1.80	66	.49	2	1	36	0	0	1	351	875	52
35	5.9	2.82	1.43	56	.84	2	1	29	0	0	1	60	111	51
Near harbor														
36	12.1	3.28	.77	71	.76	22	20	75	5	12	67	285	448	16
37	14.4	2.92	.80	77	.64	23	29	95	0	4	183	568	689	27
38	18.6	3.92	.79	80	.81	39	29	126	0	18	237	384	834	10
Orange County - near outfall														
39	11.3	3.35	1.03	75	.79	18	19	70	52	54	100	204	457	29
40	12.0	3.42	.94	76	.80	15	23	72	24	24	84	200	369	53
41	10.1	3.02	1.02	67	.73	16	14	63	11	11	43	190	451	53
42	8.9	2.73	1.07	60	.68	15	13	57	1	1	37	179	537	41
43	10.6	2.52	1.26	54	.58	18	17	74	0	2	41	272	1008	31
44	8.8	2.35	1.19	51	.57	14	12	62	2	3	40	223	1012	23
45 *	9.6	2.76	1.34	49	.65	17	11	69	0	0	22	608	1202	23
46	9.0	2.65	1.25	53	.65	14	11	60	0	0	41	323	718	25
47	12.0	3.00	1.19	55	.69	16	16	78	1	4	43	196	600	24
48	12.0	3.34	1.07	65	.78	10	21	74	1	3	88	198	441	20
49	9.7	3.23	.71	80	.81	15	13	54	36	37	75	88	237	52
San Diego - near outfall														
63	8.8	3.11	.87	74	.81	7	8	47	16	17	39	108	182	63
64 *	8.6	3.32	.93	71	.86	9	10	48	2	9	31	117	235	62
68	8.4	3.07	.88	68	.79	9	7	48	16	17	34	129	270	50
69	7.2	2.92	.74	78	.79	7	8	40	56	57	81	69	226	62