Washington State Department
of Ecology
Publication No. 13-03-035

Development of Puget Sound Benthic Indicators

Report to the Washington State Department of Ecology

J. Ananda Ranasinghe
Eric D. Stein
Melanie R. Frazier
David J. Gillett

Southern California Coastal Water Research Project

Technical Report 755 - August 2013

Development of Puget Sound Benthic Indicators

Report to the Washington State Department of Ecology

by

J. Ananda Ranasinghe[1]
Eric D. Stein[1]
Melanie R. Frazier[2]
David J. Gillett[1]

[1] *Southern California Coastal Water Research Project, 3535 Harbor Blvd., Suite 110, Costa Mesa, CA 92626, USA*

[2] *U.S. Environmental Protection Agency, Western Ecology Division, 2111 SE Marine Science Drive, Newport, OR 97365, USA*

SCCWRP Technical Report 755
Washington State Department of Ecology Publication No. 13-03-035

August 2013

DEPARTMENT OF ECOLOGY
State of Washington

SCCWRP
Established 1969

Washington State Department of Ecology Publication Information

Development of Puget Sound Benthic Indicators
Publication No. 13-03-035

Development of Puget Sound Benthic Indicators is available at
https://fortress.wa.gov/ecy/publications/SummaryPages/1303035.html.

The Activity Code for this project is 01-900.

Authors and Contact Information

J. Ananda Ranasinghe[1]
Eric D. Stein[1]
Melanie R. Frazier[2]
David J. Gillett[1]

[1]Southern California Coastal Water Research Project, 3535 Harbor Blvd., Suite 110, Costa Mesa, CA 92626, USA
[2]U. S. Environmental Protection Agency, Western Ecology Division, 2111 SE Marine Science Drive, Newport, OR 97365, USA

For more information contact:

Communications Consultant
Environmental Assessment Program
Washington State Department of Ecology
P.O. Box 47600
Olympia, WA 98504-7600
Phone: 360-407-6764

EXECUTIVE SUMMARY

The Puget Sound Ecosystem Monitoring Program (PSEMP), formerly known as the Puget Sound Ambient Monitoring Program and the Puget Sound Assessment and Monitoring Program (PSAMP), has collected benthic macrofaunal samples since 1989, but has yet to make full use of that data because of challenges in interpreting what constitutes abnormal deviation from an expected biological assemblage. In other regions of the country, such interpretation is facilitated by use of benthic indices that remove much of the subjectivity and provide a simple means for communicating complex information to managers. The goal of this project is to develop, calibrate, and validate several benthic indices for Puget Sound data and determine whether they perform well enough to justify using them in assessments of Puget Sound benthic condition.

Data from five Sound-wide benthic surveys were combined and used to (1) assess the number and distribution of benthic macrofaunal assemblages in Puget Sound, (2) calibrate five benthic indices that were originally developed and applied elsewhere for use in Puget Sound, (3) create a validation data set to evaluate performance of the calibrated Puget Sound benthic indices, based on best professional judgment (BPJ) of expert benthic ecologists, and (4) evaluate the performance of the calibrated Puget Sound benthic indices. The data used for the assessment included benthic macrofaunal species abundance data and habitat data from 1,023 site events from 1989 to 2008, which were segregated into a 983 sample calibration data set, and a 40 sample validation data set. The calibration data were used for benthic index development, while the validation data, which were independent of the calibration data, were used to validate and evaluate the calibrated benthic indices.

The Puget Sound Benthic Assemblage

A single benthic macrofaunal assemblage was identified in Puget Sound. As a result, development of a single set of habitat-related benthic indices should suffice for assessing benthic condition in Puget Sound.

The benthic macrofaunal assemblage was identified by hierarchical cluster analysis of macrobenthic species abundance data from 601 sites selected from the 1,023 available site-events. These sites were not affected by poor sediment chemistry or associated with areas of known toxicity. Groups of samples with similar species composition were identified by the cluster analysis and tested for differentiation based on five habitat variables: sediment grain size, depth, salinity, latitude, and longitude. Habitat-related sample groups were identified by sequentially examining splits in the cluster analysis dendrograms, to assess whether each split reflected habitat differentiation. Although five habitat-related sample groups were identified with significant Mann-Whitney-Wilcoxon differences in habitat variables across dendrogram splits, five lines of evidence indicated that the sample groupings are sub-assemblages of a single naturally occurring Puget Sound benthic assemblage rather than five distinct assemblages. The main reason was that few abundant species occurred exclusively in single groupings; instead, many abundant species occurred in large percentages of the samples in many of the groupings.

Benthic Index Development and Calibration

We developed and calibrated five benthic indices that have been used in other regions of the country using data from Puget Sound: the Benthic Response Index (BRI), AZTI Marine Biotic Index (AMBI), Relative Benthic Index (RBI), Benthic Quality Index (BQI) and a RIVPACS model-based O/E (observed over expected) index. Each index was screened for initial acceptability by evaluating its independence from the influence of habitat variables, such as salinity, bottom depth, and substrate type. None of the five indices were strongly affected by habitat variables and therefore all were retained for the validation phase.

Selection of Validation Samples, Based on Experts using Best Professional Judgment to Assess Benthic Condition

The five benthic indices were validated using expert best professional judgment (BPJ). The BPJ approach was used because we were not certain that the environmental gradients represented by the available data represented the full range of possible conditions that could be encountered. The BPJ approach has been shown to be an acceptable approach under these circumstances. We provided species composition and abundance data from 40 sites in Puget Sound to six independent benthic experts, who evaluated the sites in terms of four condition categories and ranked them from best to worst condition. We evaluated (1) the magnitude of the condition gradient identified by the experts and (2) the level of agreement among the experts, to determine the suitability of the data for evaluating and validating Puget Sound benthic indices. Following two iterations of ranking by the experts, there was agreement on 17 of the 40 potential validation samples, which were then selected for inclusion in Puget Sound benthic index validation data set.

Benthic Index Validation

The performance of the five calibrated benthic indices was evaluated by comparing the rank order of index values to the median rank order assigned by the experts for the 17 validation samples selected during the BPJ exercise. Spearman correlation coefficients for four of the five indices were >0.75; only the O/E index failed to meet this minimum validation criterion. The four benthic indices that adequately evaluated the rank order of the validation samples can likely be used for benthic assessments in Puget Sound; they are the Benthic Response Index (BRI), AZTI Marine Biotic Index (AMBI), Relative Benthic Index (RBI), and Benthic Quality Index (BQI).

Establishing Thresholds for Benthic Assessments

Three sets of assessment thresholds were established for evaluation. Threshold recommendations were based on the accuracy with which a set of thresholds identified benthic community condition consistent with the categories established for the validation data. The three sets of assessment thresholds included (1) developer assessment thresholds which were based on principles established by the original benthic index developers, and (2) two sets of non-developer thresholds based on statistical (kappa and category) optimization of assessment thresholds applied and tested on the 17 sample validation data set. Category optimized thresholds performed best and were selected for benthic assessment use.

Evaluation of Index Performance

The performance of each benthic index by itself and all possible index combinations was assessed by comparing it to the consensus expert condition assessment in three ways:

1. Station classification accuracy: the accuracy with which an index differentiated benthos identified by the six experts as altered;
2. Categorical classification accuracy with respect to four categories established for index calibration. This is more challenging than status classification because it requires finer discrimination of the same benthic responses among a larger number of categories; and
3. Bias in category designation: the sum of differences between index or index combination category and the consensus categorical classification of the experts when categories are expressed numerically.

Based on these criteria the benthic quality index (BQI) was identified as the best performing index. Two three-index combinations (AMBI, BQI and RBI, and BQI, BRI and RBI) performed to a slightly inferior level than the BQI. These three measures are suitable for incorporation in routine benthic assessments of Puget Sound.

Recommendations

Based on these results, adoption of a two phase strategy is recommended:

- The first phase is to build on the results of this study by preparing guidance and documentation to support routine benthic monitoring in Puget Sound.
- The goal of the second phase is to confirm the results of this study for broader application:
    - Expanding confidence in the BQI above and beyond the 17 validation samples that are the basis for the present study, by sampling sites of known poor, intermediate, or pristine condition to test and potentially improve the assessment methods proposed in this report.
    - Exploring the second (AMBI, BQI and RBI index combination) and third (BQI, BRI and RBI index combination) choice assessment methods identified in the present study, also with high potential for success, in order to confirm the accuracy of the methodology with a view to using them if it is necessary, desirable, or expedient.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

Benthic macrofauna are good indicators of the "health" of the marine environment as reflected by conditions of the surface sediments. Because they are relatively sessile and spend most of their lives in the sediment, they are good integrators of condition over time. Benthic indices are often used to "summarize" information on overall community composition and abundance and provide simple, management oriented, measures of benthic community condition. Benthic indices are often used in coastal and estuarine habitats to assess the effects of physical disturbance, organic loading, and chemical contamination (Pearson and Rosenberg 1978, Bilyard 1987, Dauer *et al.* 2000). Benthic indices (Weisberg *et al.* 1997; Van Dolah *et al.* 1999; Borja *et al.* 2000, 2003; Smith *et al.* 2001; Diaz *et al.* 2004; Rosenberg *et al.* 2004; Muxika *et al.* 2005; Marques *et al.* 2009; Pinto *et al.* 2009; Ranasinghe *et al.* 2009) summarize complex benthic species composition data and provide a simple quantitative scale of community condition that facilitates interpretation in a management context (Thompson *et al.* 2012).

Successful development of benthic indices requires addressing several technical challenges. Benthic species composition and abundances vary naturally from habitat to habitat, and definitions of reference condition and measurements of deviation from reference should vary accordingly. The first challenge, therefore, is to accurately identify, in the region of interest, natural habitat factors that influence benthic species composition sufficiently to affect benthic indices. Once habitat-related benthic assemblages are identified, the next challenge is dividing available data for each assemblage into independent calibration and validation data sets. For successful benthic index development, both habitat-specific data sets should include samples representing the entire disturbance gradient from undisturbed to severe disturbance, which is often a substantial challenge. The calibration data usually include a large number of samples to account for within habitat-related benthic assemblage variability and small-scale spatial heterogeneity in the benthic indices that are calibrated. The validation data often include a smaller number of samples, which are used to evaluate the accuracy of benthic indices developed using the calibration data. For effective validation, the relative condition and condition category of each of the validation samples must be accurately characterized. In early benthic index development efforts, this challenge was met by using sediment chemistry and sediment toxicity data. Subsequently (Weisberg *et al.* 2008; Ranasinghe *et al.* 2009; Teixeira *et al.* 2010, 2012), as in the present study, validation samples were ranked and categorized by expert benthic ecologists using best professional judgment (BPJ).

Benthic macrofauna have been sampled in Puget Sound since 1989 by the Washington State Department of Ecology. Along with measures of sediment chemistry and sediment toxicity, they have been included in comprehensive sediment monitoring for the Puget Sound Ecosystem Monitoring Program (PSEMP), formerly known as the Puget Sound Ambient Monitoring Program (PSAMP; Dutch *et al.* 2009). Although benthic macrofaunal samples have been collected for PSEMP for over two decades, the benthos data have not been used to their full effectiveness because they lack interpretational context. Benthic invertebrate surveys produce a complex list of species that occur at a site, and it can be difficult to determine what constitutes abnormal deviation from an expected biological assemblage. Index-based approaches to summarizing such data have facilitated the use of benthic macrofauna as indicators of sediment condition in other marine and estuarine systems (Weisberg *et al.* 1997; Hyland *et al.* 1999, 2003;

Bergen *et al.* 2000; Dauer *et al.* 2000; Summers 2001; Diaz *et al.* 2004; Borja and Dauer 2008). While reducing complex biological data to a single value has disadvantages, the resulting indices remove much of the subjectivity associated with interpreting data. The indices also provide a simple means for communicating complex information to managers (Dauer *et al.* 2000, Hale *et al.* 2004, Bilkovic *et al.* 2006).

Several studies have described reference ranges for benthic metrics in Puget Sound (Striplin Environmental Associates Inc. 1996, Striplin Environmental Associates Inc. and Roy F. Weston, Inc. 1999, MER Consulting 2000) and reviewed potential benthic index approaches that could be used (Striplin Environmental Associates, Inc. 2003). While these previous studies set the stage, they stopped short of developing benthic indices for Puget Sound assessments. Here we present the first calibration of such indices for Puget Sound, doing so for five index types that have been applied successfully in other parts of the country.

In this report, we document the process we used to calibrate, validate, and evaluate benthic indices for use in Puget Sound benthic assessments:

- In Section 2, we describe the available Puget Sound benthic data, and divide them into independent calibration and validation data sets.
- In Section 3, we determine that the benthic organisms of Puget Sound comprise a single habitat-related benthic assemblage; therefore, a single calibration of each benthic index should suffice for assessing benthos throughout Puget Sound.
- In Section 4, we calibrate five benthic indices to Puget Sound, using the 983 sample calibration data.
- In Section 5, we use an iterative Delphi process based on the Best Professional Judgment of six expert benthic ecologists to select samples for benthic index validation and determine their benthic condition. Seventeen samples from the 40 sample validation set were selected for benthic index validation and evaluation, based on the level of agreement among the experts.
- In Section 6, we evaluate and validate the five benthic indices calibrated in Section 4, based on the 17 validation samples that were selected in Section 5 and identify the best performing index and index combinations.
- Our conclusions and recommendations are presented in Section 7.

2. STUDY AREA AND DATA

2.1 Study Area

Puget Sound is a major coastal resource that is ecologically, culturally, and economically important. It is one of the largest marine estuaries on the west coast of the United States, covers 7,252 km^2 with inland marine waters, has 4,023 km of shoreline, and has an average depth of 137 m (U.S. Environmental Protection Agency 2007). Pacific Ocean waters mix with freshwater from over 10,000 rivers and streams that flow into the Sound from the surrounding Cascade and Olympic mountains. The deep, cold, tidal waters and warmer, shallow estuaries are home to an abundance of marine plant and animal life. Over 4 million humans reside in the watersheds that drain into Puget Sound.

2.2 Data Sources

We identified projects that collected benthic species abundance, habitat, sediment chemistry, and sediment toxicity data synoptically from sampling sites in Puget Sound (Table 2-1). Next, we acquired the data, evaluated them for methodological consistency, normalized them for units of measure, and compiled them in a database. If available, we included data about habitat conditions such as depth, bottom water salinity, sediment grain-size distribution, sediment contaminant concentrations, and toxicity to amphipods or other organisms. If multiple benthic samples were collected on a site visit, we only included species abundance data from the first sample in the database.

Benthic samples were collected with 0.1-m^2 Van Veen grab samplers and sieved through 1-mm sieves. Only samples penetrating at least 5 cm into the sediment with no evidence of sediment disturbance (e.g., washout or slumping) were processed. Material retained on the screen was preserved in 10% sodium borate buffered formalin. In the laboratory, samples were rinsed and transferred from formalin to 70% ethanol 3 to 14 days after collection. Organisms in the samples were sorted into taxonomic categories and identified and enumerated by experienced taxonomists. Taxonomic inconsistencies among programs were eliminated by cross-correlating the species lists, identifying differences in nomenclature, and resolving discrepancies by consulting the taxonomists from each program.

Table 2-1. Data sources for calibration and validation samples.

Project	Period	Reference	Number of Samples		
			Calibration	Validation	Total
EMAP	1999, 2004	USEPA 2004	24	6	30
PSAMP/NOAA	1997-1999	Long *et al.* 2003	283	15	298[1]
PSAMP Spatial	2002-2008	Dutch *et al.* 2009	214	7	221
PSAMP Temporal	1989-2008	Dutch *et al.* 2009	402	12	414
Urban Waters	2007-2008	Dutch *et al.* 2009	60	0	60
Total			983	40	1,023

[1] 300 samples were collected for PSAMP/NOAA, but data for two azoic samples were not included in the present study

All 1,023 samples in the database were analyzed to determine the habitat-related benthic assemblage of Puget Sound (Section 3). Of the 1,023 samples, 983 were used to calibrate five benthic indices to Puget Sound data (Section 4). The other 40 samples were used in a "best professional judgment" study to create a validation data set (Section 5). Based on the level of agreement among benthic experts in Section 5, 17 samples were selected for benthic index validation and evaluation (Section 6).

2.3 Selection of Index Calibration and Validation Data

Validation of benthic index performance based on data independent of those used to calibrate the indices is necessary to assure the accuracy of condition assessments based on benthic indices (Borja and Dauer 2008, Borja *et al.* 2009). To ensure independence, we selected 40 validation samples from the 1,023 sample Puget Sound sediment quality database and withheld them from the benthic index calibration data. The validation samples were selected systematically in three steps in an effort to ensure that a wide range of benthic conditions were represented:

1. The species abundance data for the 1,023-sample database were analyzed by principal coordinate ordination;
2. A disturbance gradient in the principal coordinate space was identified as a vector joining the centroids of "uncontaminated" and "contaminated" sites, based on available sediment chemistry and sediment toxicity data; and
3. The 1,023 samples were ordered along the disturbance gradient and 40 samples were selected at regular intervals, starting with the second sample at the disturbed end of the gradient.

While it is generally accepted that current models of benthic response do not discriminate between chemical contamination and other sources of stress (Borja *et al.* 2003), this approach ensured that a range of benthic conditions was represented in the calibration and validation data.

3. THE HABITAT-RELATED BENTHIC MACROFAUNAL ASSEMBLAGE OF PUGET SOUND

3.1 Introduction

Benthic indices are intended to identify the location of sites along disturbance gradients. However, the species composition and abundances used to calculate benthic indices respond to habitat as well as anthropogenic and other disturbance gradients. Therefore, changes in species composition and abundance due to habitat differences have the potential to interfere with the performance of benthic indices. Individual species are typically distributed in complex ways along environmental gradients, but the combined result is often a series of identifiable assemblages that partition available habitat along gradients of a few variables (Boesch 1973, Orloci 1975, Boesch 1977, Whittaker 1978, Smith *et al.* 1988, Bergen *et al.* 2001, Llansó *et al.* 2002, Hyland *et al.* 2004, Ranasinghe *et al.* 2012a, Thompson *et al.* 2013). Compartmentalizing habitat-related biological variability improves the ability of benthic indices to detect disturbance-related effects on biota and increases sensitivity by increasing signal to noise ratios. The typical strategy for reducing habitat "noise" interference with benthic index performance is to separate benthos into distinct assemblages and calibrate indices separately for each assemblage. This is typically the first step in benthic index development.

Furthermore, defining reference conditions for benthic indices requires identifying the habitat variables that are most important in structuring biological assemblages, assessing whether natural breaks in species composition and abundance rise to the level of biological assemblage differences, and determining the threshold values of these variables that result in natural breaks in biological distributions are necessary components of defining reference conditions (Hughes *et al.* 1986, Bald *et al.* 2005). The subjectivity of decisions about whether breaks in biological distributions rise to the level of a distinct assemblage can be minimized by basing decisions on high fidelity and exclusivity of dominant species. Fidelity is the frequency of occurrence of a species in samples of a potential assemblage, while exclusivity is the abundance of a species in a potential assemblage relative to its total abundance in all samples; both measures are usually expressed as percentages.

In Puget Sound, naturally occurring habitat-related benthic assemblages have not been identified for the purpose of developing benthic indices and defining undisturbed reference conditions, although benthic macrofauna have been monitored consistently since 1989. Llansó *et al.* (1998) studied Puget Sound benthic communities sampled from 1989 to 1993, but did not segregate unaffected and adversely affected stations. As a result, stress effects were not distinguishable from habitat effects. Furthermore, spatial coverage and habitat heterogeneity of the study were limited because the data included only revisits to 76 fixed stations. Ranasinghe *et al.* (2012a) included Puget Sound in a study describing assemblages along the west coast of the United States, but the low sample density necessitated by the large spatial scale may not have comprehensively defined assemblage patterns in Puget Sound.

We used data from likely unaffected sites sampled in Puget Sound, including recent monitoring programs with broader spatial coverage and spatially random sampling designs that increase sampled habitat heterogeneity to: (a) identify the reference benthic assemblages and sub-assemblages that occur naturally in Puget Sound, (b) identify the habitat factors that are

associated with these assemblages and sub-assemblages, and (c) evaluate the effects of inter-annual benthic variability at sampling sites in relation to sub-assemblage affinity.

3.2 Methods

We used hierarchical cluster analysis of macrobenthic species abundance data to identify the benthic assemblages that occur naturally in Puget Sound and the habitat factors that structure them. These analyses were based on 1,023 benthic samples from five sound-wide regional projects conducted between 1989 and 2008 (Table 3-1). Four of the projects used probability-based spatial sampling designs, so that all Puget Sound sub-tidal habitats were included in the sampling frame.

Habitat data that were collected with most samples included sediment grain size distribution (percent fines), bottom depth, and bottom water salinity measurements. All data were evaluated for methodological consistency and normalized for units of measure.

Because our objective was to define natural groupings of samples with similar species composition, data from potentially contaminated sites were eliminated prior to analysis, based on sediment chemistry and toxicity characteristics of the sediment. These characteristics included whether the sediment chemicals exceeded Washington State Sediment Quality Standards, whether significant sediment toxicity was detected, and whether stations were classified as "urban" or "harbor." Data from 10 unusually depauperate samples containing less than ten organisms or fewer than ten taxa were also eliminated from the data. After eliminating potentially contaminated sites, data from 601 samples remained for analysis (Table 3-1).

Groups of samples with similar species composition were identified by hierarchical cluster analysis and the groups were tested for habitat differentiation using non-parametric statistical methods. Hierarchical cluster analysis was used because it is a classification analysis that sequentially segregates samples based on similarity of species composition and abundance. The resulting splits separating sample groups are easily linked to differences in measured habitat factors, if habitat effects exist. There are many examples of the successful application of this technique in marine and estuarine benthic zonation studies (e.g., Boesch 1973, Boesch 1977, Bergen *et al.* 2001, Ranasinghe *et al.* 2012a). Q-mode cluster analyses were conducted using flexible sorting of Bray-Curtis dissimilarity values with $\beta = -0.25$ (Bray and Curtis 1957, Lance and Williams 1967, Clifford and Stephenson 1975). Prior to cluster analysis, the influence of dominant species was reduced by cube-root transformation of species abundances. Nodal analysis (two-way table) interpretation was facilitated by standardization of abundances by the species mean for samples with abundances greater than zero (Smith 1976, Smith *et al.* 1988). The step-across distance re-estimation procedure (Williamson 1978, Bradfield and Kenkel 1987) was applied to dissimilarity values higher than 0.80 to reduce the distortion of ecological distances caused by joint absences of a high proportion of species; distortion occurs due to the non-monotonic truncated joint species distribution. Prior to cluster analysis, species occurring only at one site were eliminated.

Habitat-related assemblages were identified by sequentially examining splits in the cluster analysis dendrogram, starting with the first split and proceeding along branches, to assess whether each split reflected habitat differentiation. Habitat differentiation was defined as:

(1) a significant ($p < 0.05$) Mann-Whitney-Wilcoxon difference in median for any of five habitat variables (percent fine (<63 µ grain size) sediments, bottom depth, bottom salinity, latitude, and longitude) between the two sample groupings defined by the dendrogram split, and (2) accurate segregation of more than 90% of the samples in the split according to criteria based on significant habitat variables. Probabilities were not adjusted to account for multiple testing because we were interested only in controlling the comparison-wise error rate.

For each habitat-related assemblage, abundant and characteristic taxa were identified as those with a mean assemblage abundance >100 per 0.1-m^2 sample and either exclusivity >80% or fidelity >50%. Exclusivity was calculated as the abundance of a taxon in assemblage samples, expressed as a percentage of its total abundance in all samples. Fidelity was calculated as the frequency of occurrence of a taxon in assemblage samples, expressed as a percentage.

The effect of small spatial scale heterogeneity and annual variability on assemblage fidelity was assessed using results of the same cluster analysis for 270 samples from 50 stations that were revisited in multiple years. The relative magnitude of small spatial scale assemblage variability and stability over time were evaluated relative to sub-assemblage (cluster group) membership of the samples, and by measuring the percentage of samples from a site that occurred next to each other in the dendrogram.

3.3 Results

Statistically significant ($p < 0.05$) Mann-Whitney-Wilcoxon test differences for bottom water salinity, bottom depth, fine sediments, latitude and longitude were detected across the four dendrogram splits labeled in Figure 3-1 (Table 3-2). Split 1 and Split 2 were significantly different for three of the five habitat variables, while Split 3 and Split 4 were significantly different for four. None of the splits was significant for all five habitat variables. Medians for percent fines, depth, salinity and latitude were significantly different across three of the four splits, while medians for longitude were significant across two splits. Sample grouping locations are presented in Figure 3-2.

We addressed five questions to determine whether the biota of these groups constitute distinct assemblages or whether they are sub-assemblages of a single naturally occurring Puget Sound benthic assemblage: Are (1) exclusivity and (2) fidelity of abundant and dominant benthic organisms to the sample groupings sufficient to justify designation as an assemblage? Are (3) classification accuracy of the groupings by habitat variables and (4) differences in taxa richness and abundances among the groupings high enough to justify designation as assemblages? (5) How many assemblages do the nodal (two-way table) analysis results indicate?

Exclusivity was low for almost all the taxa (Table 3-3), in contrast to large numbers of exclusivity values >90% of many dominant taxa to single groupings in similar studies (e.g., Ranasinghe *et al.* 2012a), indicating relatively small differences between sub-assemblages. Similar macrobenthic taxa were characteristic of the five identified groupings, supporting the view that the groupings are sub-assemblages of a single Puget Sound assemblage. Only one of nine abundant taxa, and five of 31 taxa comprising the top-ten sub-assemblage abundance dominants had exclusivity >90% for a sub-assemblage (Table 3-3).

Many dominant taxa had fidelity values >50% for multiple groupings, confirming that similar macrobenthic taxa were characteristic of the five groupings and further supporting the view that they are sub-assemblages of a single Puget Sound assemblage. Six of nine abundant taxa had fidelity >50% for multiple sub-assemblages, including sub-assemblages on either side of Split 1 for four of the six; Split 1 was the primary split in the dendrogram. Ten of the 31 top-ten sub-assemblage abundance dominants had fidelity >50% for three or more sub-assemblages, and 11 had fidelity >50% for none of the sub-assemblages (Table 3-4).

None of the habitat criteria classified samples across any of the four splits with habitat segregation accuracy ≥90% (Table 3-5), an indication that biota in the groupings were not clearly separable by habitat variables and another indication that the major splits in the species abundance dendrogram represent habitat-related sub-assemblages of a single Puget Sound benthic assemblage. Samples across Split 1 classified with 86.0% accuracy into fine and coarse sediment sub-assemblages at a threshold of 60% fine sediments. At Split 3, the fine sediment sub-assemblage segregated into deep and shallow sub-assemblages with 87.5% accuracy at a depth threshold of 60 m. Although percent fine sediment medians, means, and distributions differed across Split 2 and Split 4 samples, there was complete overlap and meaningful segregation was not possible. The significant salinity, latitude, and longitude distributions also overlapped, and meaningful segregation across splits was not possible.

Mean taxa richness and total abundance varied, at most, by a factor of two among sample groupings (Table 3-6) in contrast to generally larger multipliers reported by Llansó *et al.* (2002a), Hyland *et al.* (2004) and Ranasinghe *et al.* (2012a). Thus community parameters also varied less among the groupings than would normally be expected if they constituted two or more assemblages.

Nodal (two-way table) analysis of the dendrograms also supported inclusion of the entire data set in a single assemblage (Figure 3-3). In the figure, abundances of species are presented with darker symbols representing greater abundance. The 601 samples are distributed horizontally across the page in the same order as the station dendrogram (Figure 3-1) and the 994 species in the data are arranged along the vertical axis in the same order as in the species dendrograms (not shown). The presence of many dark horizontal striations extending across the entire figure indicates the general dominance of universally abundant taxa. The absence and poor definition of distinct vertical blocks corresponding to the sub-assemblages depicts the relative uniformity of the single assemblage present in the data set. The coarse sediment sub-assemblages (on the right) are more abundant and diverse, and the fine sediment sub-assemblages (to the left) are characterized by the absence of taxa rather than the presence of unique, abundant, taxa.

Habitat criteria separating samples across splits were associated primarily with differences in sediment composition and depth (Table 3-5), although there were differences among groupings associated with all five habitat variables (Figure 3-4, Table 3-2). None of the criteria included salinity, latitude, or longitude. The groupings were distributed throughout Puget Sound (Figure 3-2). As indicated earlier, these differences in habitat factors did not result in biological species abundances differences of sufficient magnitude to support designation as multiple separate assemblages.

For Split 1, Split 3, and Split 4, there was an association between sub-assemblage sediment composition and depth. Fine sediment sub-assemblages were generally associated naturally with the deeper branch of the split and coarse sediment sub-assemblages with the shallower branch. However, five coarse (15.7 to 27.3% fines) sediment samples collected at a depth of 268 to 270 m at Station PST-0026 from 1989-1994 were exceptions, and demonstrated that sediment grain size was a more important determinant of biological sub-assemblage composition than depth.

There was a high level of consistency in the biota sampled at stations revisited in multiple years. Only 2 of 270 revisited station samples were classified in a different sub-assemblage from other samples from the same station, and even then, only in adjacent sub-assemblages within the same primary fine sediment (A) or coarse sediment (B) split. For stations sampled on three or more occasions, 196 of 224 revisited station samples (87.5%) were located adjacent to each other on the dendrogram. Six of eight station groups with different numbers of revisits averaged >85% adjacent samples on the dendrograms (Table 3-7).

3.4 Discussion

Five lines of evidence indicate that the Puget Sound benthic populations should be regarded as a single assemblage, which implies that segregation of Puget Sound into multiple habitats is not necessary for benthic index development. The species composition and abundances of benthic assemblages often vary naturally with habitat, and definitions of reference condition and measurements of deviation from reference should vary accordingly, if species composition and abundance differences among sample groupings are of sufficient magnitude. In the present study, multiple lines of evidence indicate that differences are insufficient to designate distinct assemblages, and all Puget Sound benthic biota from potentially uncontaminated stations were designated a single assemblage. The designation of a single Puget Sound habitat-related benthic assemblage facilitates benthic index development and benthic assessments by eliminating the need for segregation of these data by benthic habitat type, reducing the necessary effort and potentially increasing confidence in index development results by avoiding decreases in the numbers of data that inevitably result from data segregation.

The results of our analysis are consistent with those of other macrobenthic assemblage analyses, indicating that sediment grain size and bottom depth are among the primary habitat determinants structuring benthic assemblages at local spatial scales (Llansó *et al.* 1998, Van Dolah *et al.* 1999, Bergen *et al.* 2001, Llansó *et al.* 2002, Hyland *et al.* 2004, Ranasinghe *et al.* 2012a, Thompson *et al.* 2013). Our Puget Sound assemblages are similar to those of Llansó *et al.* (1998b). Our results also reflect differences due to the large geographic scales and presence of freshwater inputs from large rivers in many of those previous studies. Latitude and salinity, which are often the primary physical factors differentiating assemblage composition at large spatial scales, were replaced by sediment grain size, and, to a lesser extent, depth at the local Puget Sound scale. This is consistent with the findings of Weisberg *et al.* (1997) and Ranasinghe *et al.* (2012a) that substrate differences differentiate assemblages in higher salinity waters, and have less effect where salinity is low. Puget Sound has a relatively stable high salinity regime that reflects the influence of the Pacific Ocean, despite the consistently high rainfall in the region.

Our finding that bottom depth differentiated the Puget Sound fine sediment sub-assemblage into deep and shallow components (Table 3-3) differs from that of Ranasinghe *et al.* (2012a), who found that bottom depth was important only as a modifier of sediment grain-size differences in Puget Sound. Our results indicate that depth splits the fine-grained sediment sub-assemblage further into deep (A1) and shallow (A2) components at a depth of 60 m, while Ranasinghe *et al.* (2012a) found that depth only modified grain size effects at depths less than 40 m. This difference is due largely to differences in sampling intensity between the large spatial scale Ranasinghe *et al.* (2012a) study and the present, locally more intense study. At the reduced sampling intensity necessitated by the larger spatial scale, data indicated that depth was important only between 0 and 40 m as a modifier of sediment grain size related assemblage determination. A different picture emerges due to the increased sampling intensity of the present study, which includes samples from a greater extent of Puget Sound. Although the large coast-wide data correctly identified the major factors that determine assemblage composition, the increased sampling intensity and spatial coverage of the present study more accurately characterizes the relationships.

The species groups were demoted to sub-assemblages of a single Puget Sound assemblage in the present study, due to a paucity of exclusive species and lower habitat classification accuracy in the spatially broader and higher intensity data sets available for the current analysis. This illustrates a challenge inherent in identifying assemblage differences across cluster analysis dendrogram splits and evaluating their importance. As in previous studies, we recognized the importance of sediment grain size as the primary habitat determinant of assemblage composition in Puget Sound.

In contrast, Ranasinghe *et al.* (2012a) recognized the biota as distinct assemblages, based on assemblage exclusivity >85% of 10 species and 91.9% habitat classification accuracy in their coast-wide study. Each assemblage described by Ranasinghe *et al.* (2012a) included multiple taxa with exclusivity >90%, with 100% exclusivity often observed. More than half the abundant and characteristic taxa in each assemblage had high exclusivity, with >80% of the abundance of those taxa occurring in that assemblage alone. Only 4 of 69 characteristic taxa were abundant in more than one assemblage. In contrast, exclusivity values in the present study were low, and strings of multiple exclusivity values >90% were absent from the sub-assemblage columns (Table 3-3), indicating a paucity of species with high affinity for single assemblages.

Despite the differences in habitat variables among the major Puget Sound sample groupings, the biological species abundance differences among them were insufficient to justify designating them as distinct assemblages. Rather, they are correctly viewed as sub-assemblages of a single Puget Sound assemblage. The similarity of the biota was clearly demonstrated by the low biological exclusivity of dominant taxa for the groupings and the high biological fidelity of dominant species for multiple groupings, including many groups on either side of the primary biological grouping split in Puget Sound. The nodal analysis results showed only weak, if any, grouping-species associations and differences in numbers of taxa and species abundances between the groupings were small, confirming the lack of substantial biological differences among the groupings. Furthermore, it was not possible to develop habitat criteria that were able to segregate 90% of the samples in adjacent groupings, indicating that it was likely that although there were statistical differences in habitat variables, they were not reflected by corresponding

differences in biota. The low level of biological differences indicates that the habitat differences have only a small effect on the biota as a whole, and need not be taken into account during benthic index development. The sample groupings are sub-assemblages of a larger single Puget Sound assemblage.

Table 3-1. Data sources. PSAMP: Puget Sound Ambient Monitoring Program; NOAA: National Oceanic and Atmospheric Administration; WEMAP: Western Environmental Monitoring and Assessment Program.

Project	Period	Samples		Reference
		Unaffected	Total	
PSAMP Temporal	1989-2008	286	414	Dutch *et al.* 2009
PSAMP Spatial	2002-2008	128	221	Dutch *et al.* 2009
PSAMP/NOAA	1997-1999	159	298[2]	Long *et al.* 2003
Urban Waters	2007-2008	2	60	Dutch *et al.* 2009
WEMAP	1999, 2004	26	30	USEPA 2004
Total		**601**	**1,023**	

Table 3-2. Ranges of values for bottom salinity, depth, fine sediments, latitude, and longitude for samples across splits in the dendrogram (Figure 3-1). Bolded numbers indicate significant ($p < 0.05$) differences in median across the dendrograms splits that were identified by Mann-Whitney-Wilcoxon tests. Fine sediments pass through a 0.63 µ sieve.

Split	Sub-Assemblage	N	Salinity (psu)	Depth (m)	Fines (%)	Latitude (Degrees)	Longitude (Degrees)
1	A	271	**12.0-34.0**	**2.4-250.0**	**18.6-100**	47.096-49.003	123.130-122.255
	B	330	**10.0-35.0**	**2.1-270.0**	**0.9-97.5**	47.004-48.991	123.249-122.243
2	B1 & B2	317	15.0-35.0	2.1-270.0	**1.0-97.5**	**47.004-48.991**	**123.207-122.243**
	B3	13	10.0-33.0	4.5-133.0	**0.9-20.2**	**47.388-48.839**	**123.249-122.264**
3	A1	130	**12.0-34.0**	**19.2-250.0**	18.6-98.9	**47.327-48.267**	**123.130-122.255**
	A2	141	**20.0-33.0**	**2.4-225.0**	23.8-100	**47.096-49.003**	**123.079-122.469**
4	B1	144	**15.0-35.0**	**5.0-270.0**	**2.6-97.5**	**47.218-48.991**	123.207-122.243
	B2	173	**17.0-35.0**	**2.1-268.0**	**1.0-94.1**	**47.004-48.984**	123.166-122.286

[2] 300 samples were collected for PSAMP/NOAA, but data for two azoic samples were not included in the present study

Table 3-3. Sub-assemblage exclusivity of dominant taxa. The ten most abundant taxa in each sub-assemblage are included. Exclusivity is the abundance of a taxon in sub-assemblage samples expressed as a percentage of its total abundance in all samples. Abundant and characteristic taxa were identified as those with a mean assemblage abundance >100 per 0.1-m^2 sample and exclusivity >80% or fidelity >50%. Exclusivity values >80% are presented in bold font.

Taxon	Higher Taxon	Mean Abund. ($0.1\ m^{-2}$)	Sub-assemblage Exclusivity (%)				
			A1	A2	B1	B2	B3
Axinopsida serricata	Mollusca: Bivalvia	415.9	29.6	9.2	47.5	13.5	0.2
Nutricola lordi	Mollusca: Bivalvia	376.6	0.0	8.1	11.3	78.3	2.3
Euphilomedes carcharodonta	Arthropoda: Ostracoda	212.4	0.3	1.6	29.1	68.9	0.0
Amphiodia spp.	Echinodermata: Ophiuroidea	210.2	0.7	49.6	12.4	37.3	0.1
Aphelochaeta spp.	Annelida: Polychaeta	160.0	1.5	7.5	24.0	66.9	0.1
Macoma carlottensis	Mollusca: Bivalvia	152.1	64.7	9.5	23.3	2.5	0.0
Rochefortia tumida	Mollusca: Bivalvia	142.8	0.9	12.7	14.0	71.6	0.8
Owenia fusiformis	Annelida: Polychaeta	127.5	0.2	0.1	0.8	**98.5**	0.5
Eudorella pacifica	Arthropoda: Cumacea	125.1	15.8	72.2	7.8	4.2	0.0
Euphilomedes producta	Arthropoda: Ostracoda	98.4	28.2	7.8	56.0	7.8	0.1
Mediomastus spp.	Annelida: Polychaeta	88.6	9.2	3.0	27.0	57.6	3.2
Acila castrensis	Mollusca: Bivalvia	82.8	6.7	29.8	37.3	26.2	0.0
Alvania compacta	Mollusca: Gastropoda	72.7	0.0	3.0	7.0	**89.9**	0.1
Prionospio (Minuspio) lighti	Annelida: Polychaeta	71.2	18.8	47.5	10.5	22.9	0.2
Protomedeia grandimana	Arthropoda: Amphipoda	67.2	0.1	**92.3**	2.9	4.7	0.0
Scoletoma luti	Annelida: Polychaeta	66.3	6.2	5.2	62.1	26.5	0.1
Levinsenia gracilis	Annelida: Polychaeta	65.6	17.4	57.0	17.7	7.9	0.0
Phyllochaetopterus prolifica	Annelida: Polychaeta	61.7	0.1	1.4	2.2	**96.3**	0.0
Heteromastus filobranchus	Annelida: Polychaeta	57.5	38.7	14.5	40.3	6.5	0.0
Pholoe species complex	Annelida: Polychaeta	52.3	3.6	65.8	13.8	16.6	0.2
Paraprionospio alata	Annelida: Polychaeta	48.8	13.6	46.6	26.5	13.3	0.0
Spiophanes berkeleyorum	Annelida: Polychaeta	36.3	26.6	17.9	22.6	31.9	1.0
Galathowenia oculata	Annelida: Polychaeta	24.8	4.0	2.9	48.3	33.7	11.2
Sigambra bassi	Annelida: Polychaeta	23.6	36.5	43.7	12.7	7.2	0.0
Polycirrus spp.	Annelida: Polychaeta	19.3	0.8	7.5	47.8	37.7	6.3
Spiophanes bombyx	Annelida: Polychaeta	18.7	0.0	0.1	2.5	28.5	68.9
Exogone lourei	Annelida: Polychaeta	18.0	0.2	0.2	35.4	58.4	5.8
Eudorellopsis integra	Arthropoda: Cumacea	10.5	**99.7**	0.0	0.2	0.2	0.0
Ampelisca cristata	Arthropoda: Amphipoda	9.8	0.0	0.0	0.2	3.2	**96.6**
Tellina nuculoides	Mollusca: Bivalvia	5.8	0.0	0.0	0.3	12.4	**87.4**
Dendraster excentricus	Echinodermata: Echinoidea	2.0	0.0	0.0	0.0	17.9	**82.1**

Table 3-4. Sub-assemblage fidelity of dominant taxa. The ten most abundant taxa in each sub-assemblage are included. Fidelity is the frequency of occurrence of a taxon in a sub-assemblage expressed as a percentage. Abundant and characteristic taxa were identified as those with a mean assemblage abundance >100 per 0.1 m^2 and exclusivity >80% or fidelity >50%. Fidelity values >50% are presented in bold font.

Taxon	Higher Taxon	Mean Abund. (0.1 m^{-2})	Sub-assemblage Fidelity (%)				
			A1	A2	B1	B2	B3
Axinopsida serricata	Mollusca: Bivalvia	415.9	**86.2**	**62.4**	**93.1**	**68.2**	15.4
Nutricola lordi	Mollusca: Bivalvia	376.6	1.5	47.5	**59.7**	**60.7**	**61.5**
Euphilomedes carcharodonta	Arthropoda: Ostracoda	212.4	4.6	22.7	49.3	**71.7**	7.7
Amphiodia spp.	Echinodermata: Ophiuroidea	210.2	23.8	**75.9**	**52.8**	**53.8**	23.1
Aphelochaeta spp.	Annelida: Polychaeta	160.0	29.2	37.6	**68.1**	48.0	23.1
Macoma carlottensis	Mollusca: Bivalvia	152.1	**80.0**	49.6	**61.1**	28.3	0.0
Rochefortia tumida	Mollusca: Bivalvia	142.8	19.2	**66.0**	**71.5**	**84.4**	38.5
Owenia fusiformis	Annelida: Polychaeta	127.5	1.5	2.1	17.4	24.9	46.2
Eudorella pacifica	Arthropoda: Cumacea	125.1	**82.3**	**68.1**	46.5	40.5	7.7
Euphilomedes producta	Arthropoda: Ostracoda	98.4	**53.8**	37.6	**77.1**	30.1	7.7
Mediomastus spp.	Annelida: Polychaeta	88.6	**53.8**	34.8	**66.7**	**88.4**	**84.6**
Acila castrensis	Mollusca: Bivalvia	82.8	33.1	48.9	25.0	26.0	7.7
Alvania compacta	Mollusca: Gastropoda	72.7	1.5	22.0	41.7	**68.2**	15.4
Prionospio (Minuspio) lighti	Annelida: Polychaeta	71.2	**73.8**	**80.1**	**56.3**	**59.0**	15.4
Protomedeia grandimana	Arthropoda: Amphipoda	67.2	1.5	34.8	7.6	12.7	0.0
Scoletoma luti	Annelida: Polychaeta	66.3	44.6	48.9	**85.4**	63.0	15.4
Levinsenia gracilis	Annelida: Polychaeta	65.6	**70.8**	**58.9**	**51.4**	25.4	0.0
Phyllochaetopterus prolifica	Annelida: Polychaeta	61.7	1.5	2.1	8.3	39.3	0.0
Heteromastus filobranchus	Annelida: Polychaeta	57.5	49.2	47.5	47.2	9.2	0.0
Pholoe species complex	Annelida: Polychaeta	52.3	33.1	**68.8**	**54.9**	**57.2**	23.1
Paraprionospio alata	Annelida: Polychaeta	48.8	**76.2**	**82.3**	**69.4**	**50.9**	0.0
Spiophanes berkeleyorum	Annelida: Polychaeta	36.3	**63.1**	**56.0**	**54.9**	49.1	15.4
Galathowenia oculata	Annelida: Polychaeta	24.8	11.5	7.1	38.9	22.5	46.2
Sigambra bassi	Annelida: Polychaeta	23.6	**51.5**	33.3	9.7	12.1	0.0
Polycirrus spp.	Annelida: Polychaeta	19.3	6.2	9.9	44.4	46.2	**61.5**
Spiophanes bombyx	Annelida: Polychaeta	18.7	0.0	0.7	6.9	16.8	76.9
Exogone lourei	Annelida: Polychaeta	18.0	1.5	1.4	27.1	35.8	38.5
Eudorellopsis integra	Arthropoda: Cumacea	10.5	29.2	0.0	0.7	0.6	0.0
Ampelisca cristata	Arthropoda: Amphipoda	9.8	0.0	0.0	0.7	5.8	7.7
Tellina nuculoides	Mollusca: Bivalvia	5.8	0.0	0.0	0.7	5.8	38.5
Dendraster excentricus	Echinodermata: Echinoidea	2.0	0.0	0.0	0.0	2.9	23.1

Table 3-5. Habitat classification accuracy for samples across splits in the dendrogram.

Split	Sub-assemblage	Description	N	Habitat Criteria	Accuracy (%)
1	A	Fine sediments	271	Fines > 60%	86.0
	B	Coarse sediments	330	Fines <= 60%	
2	B1 & B2	Coarse sediments	317	-	*
	B3	Very coarse sediments	13	-	
3	A1	Fine deep sediments	130	Depth > 60m	87.5
	A2	Fine shallow sediments	141	Depth <= 60m	
4	B1	Transitional coarse sediments	144	-	*
	B2	Moderately coarse sediments	173	-	

*: Although medians and distributions of fine sediments differed across Split 2 and Split 4, distribution overlap prevented effective separation.

Table 3-6. Species richness and abundance (mean ± standard error) for each sub-assemblage.

Sub-Assemblage	Description	Samples	No. of Taxa		Total Abundance ($0.1m^{-2}$)
			Overall	Mean ($0.1m^{-2}$)	
A1	Deep central fine sediments	130	339	31.4±1.0	243.4±16.5
A2	Shallow, N and S fine sediments	141	337	30.1±0.9	348.8±19.4
B1	Transitional coarse sediments	144	583	50.3±1.3	485.3±24.1
B2	Shallow coarse sediments	173	826	66.1±2.0	761.4±56.0
B3	Mixed depth coarse sediments	13	247	37.5±6.1	331.2±65.7
	Total	**601**			

Table 3-7. Percentage of temporal replicates from single sites that clustered adjacent to each other on the dendrogram.

Station Visits	Number of Stations	Adjacent Samples at Stations	
		Mean (%)	Range (%)
19	2	65.8	57.9 – 73.7
18	2	88.9	83.3 – 94.4
17	2	94.1	88.2 – 100
14	1	85.7	-
6	12	94.4	66.7 – 100
5	3	86.7	80 – 100
3	5	93.3	66.7 – 100
2	23	60.9	0 – 100
270	**50**		

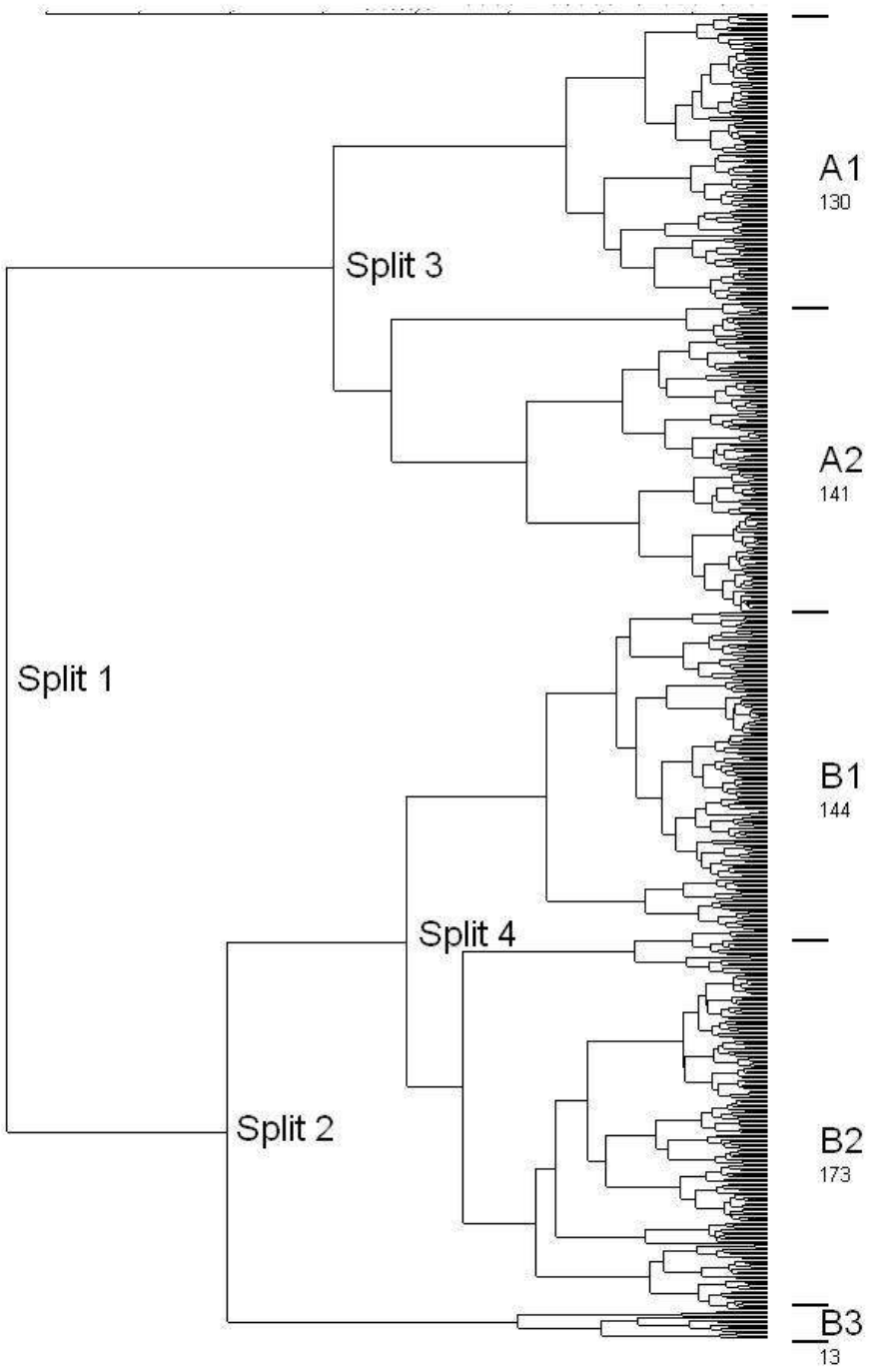

Figure 3-1. Dendrogram showing the habitat-related sub-assemblages (A1-B3) identified by cluster analysis. A1 & A2: Puget Sound fine sediment sub-assemblage; B1, B2, & B3: Puget Sound coarse sediment sub-assemblage. The number of samples for each sub-assemblage is presented under the sub-assemblage letter. Splits 1-4 identify dendrogram branch points referred to in the text and tables.

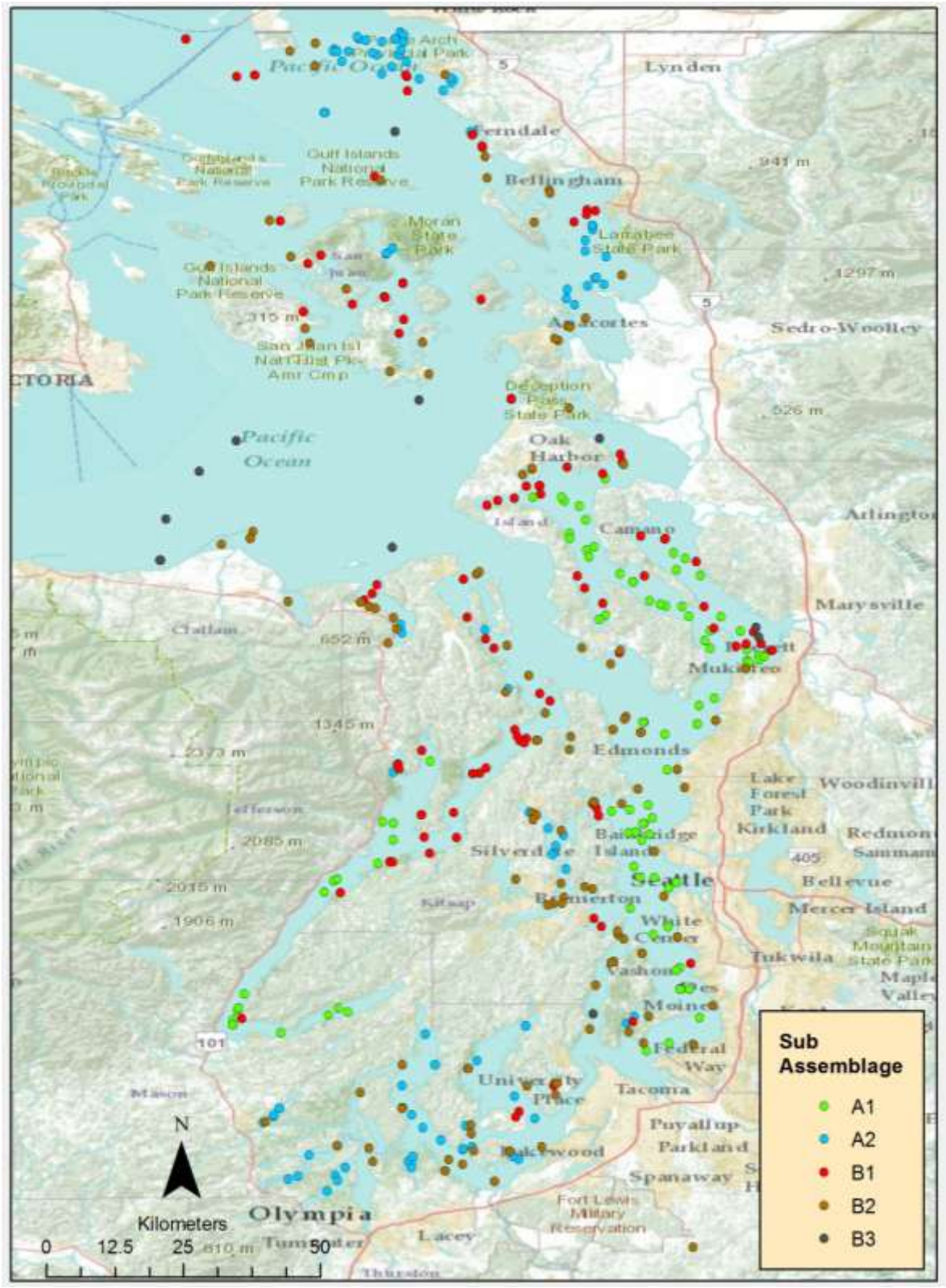

Figure 3-2. Sub-assemblage locations.

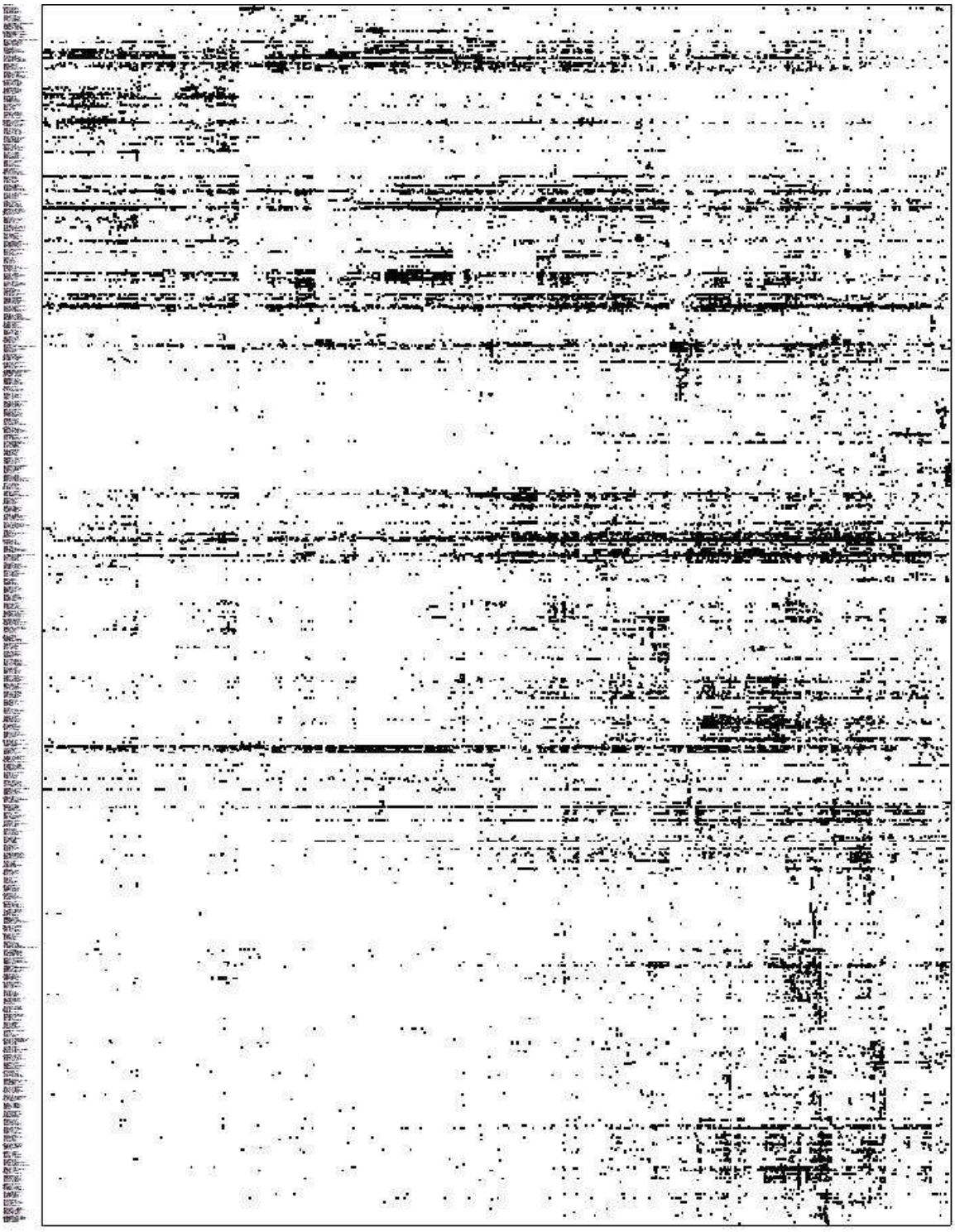

Figure 3-3. Two-way table (nodal analysis) for the cluster analysis with samples on the horizontal axis and species on the vertical axis showing the large number of abundant taxa in the Puget Sound Assemblage that are common to most or all of the sub-assemblages. The size and intensity of the sample-species symbol depicts relative abundance.

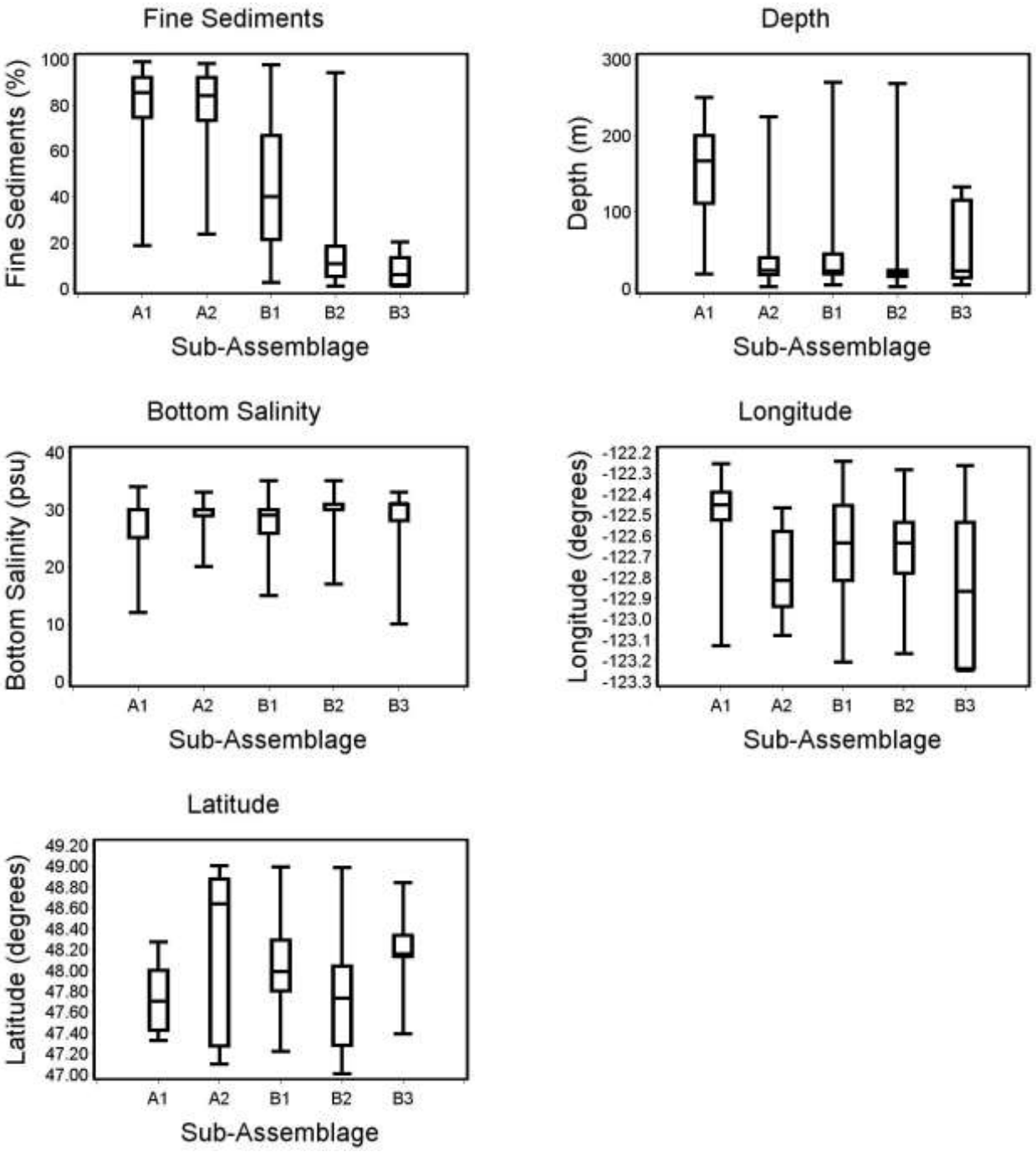

Figure 3-4. Box and whisker plots of habitat variables for each sub-assemblage. Boxes indicate quartiles and medians. Whiskers join the box to the extremities of the range.

4. BENTHIC INDEX CALIBRATION

4.1 Introduction

Five benthic indices previously calibrated for other geographies were re-calibrated for Puget Sound, using the 983 sample Puget Sound calibration data. The five indices were selected because: (a) they had been used successfully elsewhere for regional benthic condition assessments, (b) they were considered to have potential for use in Puget Sound when calibrated with local data, and (c) the data required for calibration of Puget Sound versions of the indices were available. Index calibration involved applying the previously successful calibration procedures to Puget Sound data. The five benthic indices were: (1) the Benthic Response Index (BRI), (2) the AZTI Marine Biotic Index (AMBI), (3) the Relative Benthic Index (RBI), (4) the Benthic Quality Index (BQI), and (5) an Observed over Expected (O/E) Index based on the RIVPACS model.

4.2 Benthic Index Calibration

Benthic Response Index (BRI)

The BRI is the abundance-weighted mean tolerance score of species in a sample (Smith *et al.* 2001, 2003; Ranasinghe *et al.* 2009). BRI tolerance scores are species and assemblage specific, based on the position of species abundance peaks on the disturbance gradient in the target habitat. Higher BRI values are associated with higher pollution and disturbance levels. The index formula is given by the following expression:

$$I_s = \frac{\sum_{i=1}^{n} p_i \sqrt[f]{a_{si}}}{\sum_{i-1}^{n} \sqrt[f]{a_{si}}}$$

where I_s is the BRI value for sample *s*, *n* is the number of species for sample *s*, p_i is the position for species *I* on the disturbance gradient (tolerance score), a_{si} is the abundance of species *I* in sample *s* and *f* is a transformation determined during index calibration. Species in the sample without p_i values are ignored.

BRI calibration involves calculation of tolerance scores for species that occur in the habitat-related benthic assemblage. Previously, the BRI was successfully calibrated and validated for California coastal (Smith *et al.* 2001) and embayment and estuary habitats (Ranasinghe *et al.* 2009).

We calibrated the BRI using the methods of Smith *et al.* (2001, 2003) and Ranasinghe *et al.* (2009). The first step in BRI calibration was identifying a disturbance (pollution) vector in a principal coordinates ordination (PCO) space. The Puget Sound calibration species abundance data were analyzed by PCO and the disturbance vector was identified using disturbed and undisturbed sites selected on the basis of sediment contaminant and sediment toxicity measurements. Chemical evaluations were based on comparison of one or more of 41 sediment contaminants to Washington State Sediment Quality Standards. Toxicity evaluations were based on amphipod survival or sea urchin fertilization test results significantly different from, and less than 80% of, control results, or sand dollar embryo survival and normal morphological development significantly different from, and less than 85% of, control results. The direction of

the disturbance gradient in the PCO space was identified by joining the average position (centroid) of the disturbed ("dirty") sites and the average position of the undisturbed ("clean") sites in the multivariate PCO space.

The second BRI calibration step was calculation of species tolerance scores, which reflect the positions of the abundance peaks of species on the disturbance gradient. Species tolerance scores were calculated for species that occurred in two or more samples in the calibration data sets. An optimization procedure was used to select the data transformations and the maximum number of species occurrences to be used to calculate tolerance scores (Smith *et al.* 2001, 2003; Ranasinghe *et al.* 2009). The objective was to include low abundances in tolerance score calculations only if they contribute signal, rather than noise. The combination of transformations and numbers of occurrences that maximized the Spearman correlation between the disturbance gradient and the optimized BRI were selected. Tolerance scores were calculated for abundance and BRI calculation transformations with exponents of 0, 0.25, 0.33, 0.5, and 1.0 (presence-absence, fourth-root, cube-root, square-root and no transformations). The combination of transformations and species occurrences with the highest Spearman correlation between the disturbance gradient and the optimized index was selected (Table 4-1).

Table 4-1. Parameter values for BRI calibration. The transformation exponent *f* is used for index calculations and the exponent *e* is used to develop species tolerance scores, while *t* is the maximum number of abundances used to determine the position of each species abundance peak on the disturbance gradient (See text and Smith *et al.* 2001).

Parameter	Value
Number of samples	983
e	0.5
f	0.33
t	7
Number of species with tolerance scores	814
Spearman correlation coefficient between optimized index and disturbance vector	0.895

AZTI Marine Biotic Index (AMBI)

The AMBI (Borja *et al.* 2000), like the BRI, is an abundance weighted tolerance score for organisms in a sample. Organisms are classified into one of five Ecological Groups (EG) based on the species tolerance to disturbance and the AMBI is based on abundance percentages of the Ecological Groups in the sample. In contrast, the BRI is based on abundance-weighted species tolerance scores. AMBI values are on a continuous scale from 1 to 7; higher values are associated with higher disturbance (pollution) levels. The AMBI was originally developed for Basque coastal regions of Spain, is widely used in European coastal areas, and its use is spreading worldwide.

AMBI calibration included classifying encountered species into ecological groups (Tables 4-2 and 4-3) and calculating AMBI values. Sufficient information and knowledge was available to classify 841 of the 1065 encountered species. The standard universal AMBI calculation (Borja *et al.* 2000) was applied to all samples. Guidelines for application of the AMBI (Borja and Muxika 2005) recommend only assessing samples where at least 50% of the organisms are assigned to Ecological Groups and exercising care if less than 80% of the organisms are assigned

to Ecological Groups. Only 11 calibration samples did not meet the 50% requirement (Table 4-2), and 22 samples had EG assignments in the 50%-80% "exercise care" range.

Table 4-2. Numbers of samples meeting the AZTI Marine Biotic Index (AMBI) criterion of ecological group (EG) assignment to ≥50% of abundance.

Data	Samples	Samples Meeting AMBI 50% Criterion	Samples Meeting AMBI 50% Criterion (%)	No. of Species	No. of Species with Assigned EGs	Species with Assigned EGs (%)
Calibration	983	972	98.9	1047	837	79.9
Validation	40	40	100.0	528	477	90.3
All	1,023	990	96.8	1,065	841	79.0

Table 4-3. Assignment of abundance (%) to AZTI Marine Biotic Index (AMBI) ecological groups (EGs).

Data	Samples	Abundance Assigned to EGs	EG I	EG II	EG III	EG IV	EG V
Calibration	983	97.9	9.5	38.6	27.4	14.9	9.6
Validation	40	98.8	17.0	35.5	22.9	16.1	8.5
All	1,023	97.9	9.9	38.5	27.2	15.0	9.5

Relative Benthic Index (RBI)

The RBI evaluates benthic condition based on several community parameters (Hunt *et al.* 2001, Ranasinghe *et al.* 2009). Detailed instructions for RBI calculation are provided by Bay *et al.* (2009). The index is scaled from 0 to 1.0 in each habitat by subtracting the lowest value and dividing by the range; thus 0 was the "worst" sample in the calibration data set and 1 the "best." It has been successfully validated in two California embayment and estuary habitats.

We calculated RBI values following the method of Hunt *et al.* (2001), Ranasinghe *et al.* (2009), and Bay *et al.* (2009). The first step in RBI calibration was selecting negative and positive indicator taxa for the Puget Sound assemblage. Then, RBI values were calculated as the weighted sum of (a) four community parameters (total number of species, number of crustacean species, number of crustacean individuals, and number of mollusc species), and abundances of (b) the three positive and (c) the two negative indicator organisms (Table 4-4). For positive indicator taxa, we followed the previous RBI practice of selecting an amphipod, a bivalve, and a polychaete.

Table 4-4. Positive and negative indicator species selected for Relative Benthic Index (RBI) calculations. A: Amphipod. B: Bivalve. P: Polychaete.

Indicator Type	Species
Positive indicator species	*Heterophoxus affinis* (A) *Yoldia seminuda* (B) *Praxillella pacifica* (P)
Negative indicator species	*Axinopsida serricata* (B) *Aphelochaeta* spp. (mainly *Aphelochaeta glandaria* complex (P)

Benthic Quality Index (BQI)

The BQI combines abundance weighted species tolerance scores and biodiversity to assess samples (Rosenberg *et al.* 2004). Species tolerance scores are based on the biodiversity of the samples in which the species occurred. Higher BQI values are associated with lower pollution levels. It has been calibrated for the Norwegian coast.

We calibrated the BQI for the Puget Sound assemblage using the method of Rosenberg *et al.* (2004). First, for each sample in the calibration data, the expected number of species for a subset of 50 individuals was calculated as:

$$ES50_k = \sum_{i=1}^{s}\left[1-\frac{(N_k - N_{ki})!(N_k - 50)!}{(N_k - N_{ki} - 50)!N_k!}\right],$$

where s is the number of species in sample k, N_k is the total abundance of all species in sample k, and N_{ki} is the abundance of species i in sample k. Next, species tolerance scores, $ES50_{0.05i}$, were computed for species that were found in at least three samples as the 5th percentile of the distribution of expected numbers of species for the samples in which the species occurred. Tolerance scores were calculated for 735 species in all Puget Sound. Once species tolerance scores were calculated, the BQI value for each sample k was computed as

$$BQI_k = \left(\sum_{i}^{n}\left(\frac{A_i}{totA}ES50_{0.05i}\right)\right)\left(\log_{10}(S+1)\right),$$

where n is the number of species in the sample with tolerance scores, A_i is the abundance of species i, $totA$ is the total abundance in the sample, and S is the number of species in the sample.

Observed over Expected (O/E) Index

Observed over expected indices are based on the RIVPACS approach (Wright *et al.* 1993, Van Sickle *et al.* 2006, Van Sickle 2008), which assesses benthic condition based on the ratio of the O/E number of benthic species in a sediment sample after controlling for habitat variables. For samples that are unpolluted, the O/E index should be about 1.0. Sites that are polluted are expected to have O/E value significantly less or more than one, indicating a reduction or increase in the number of species present.

We calibrated an O/E index using the methods of Wright *et al.* (1993) and Van Sickle (2006, 2008) with R code available at http://www.epa.gov/wed/pages/models/rivpacs/rivpacs.htm (Van Sickle 2008). First, unaffected samples in the calibration data were clustered to identify site groups based on the taxa that were present. Then discriminant function analysis of habitat variables at the site groups was used to build discriminant functions that can be used to classify future sampling sites into site-groups based on habitat variable values. The O/E index calibration was based on 506 minimally impacted reference samples from the 983-sample calibration data, for which salinity, percent fine sediment, bottom depth, total organic carbon, latitude, and longitude measurements were available. Minimally impacted reference samples for this calibration were selected based on the criteria described in Section 3.2.

The reference data were used to calibrate the model and establish the probable range of O/E values for unimpacted sites. Cluster analysis was used to define site-groups, based on Bray-Curtis dissimilarity values estimated using presence or absence data for 946 taxa. Samples were clustered using an agglomerative hierarchical flexible sorting approach with beta = -0.6 (Lance and Williams 1967, Legendre and Legendre 1998). In order to classify future samples into site-groups, a discriminant function analysis was used to associate the site-groups with potential habitat variables including salinity, sediment particle size (percent fines), sample depth, and sediment total organic carbon. Sample depth and TOC were transformed using natural logarithms prior to inclusion in the model. Several candidate models were explored with varying numbers of site-groups and permutations of habitat variables. To evaluate the candidate models, the O/E value for each sample was calculated by determining the probability of the sample belonging to each site-group based on its habitat variables. This was combined with the probability of each taxon occurring in each site-group to generate an expected taxon list specific to each sample. Only taxa with >50% predicted probability of occurring in a sample were used to estimate O/E values (Van Sickle *et al.* 2007). The optimal model was identified by the lowest Root Mean Squared Error (RMSE) value (Van Sickle *et al.* 2005, Van Sickle 2008):

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}\left(\frac{O}{E}-1\right)^2}{N}}$$

According to this criterion, the final model included 25 biotic groups and the discriminant function model included all the possible habitat variables. A "partitioning around medoids" approach to identifying biotic groups was also tested, but did not improve on the RIVPACS model.

Predictive improvement was also quantified by calculating the reduction in RMSE of the predictive model (i.e., the model built using a discriminant function) from the null model. The r^2 value for the modeled O/E number of taxa in the calibration samples was 0.68, which is considered adequate. Summary statistics for the model are presented in Table 4-5.

The optimal model was then used to obtain O/E values for all the samples.

Table 4-5. Summary statistics for O/E predictive models after excluding outliers (p<0.05; see Van Sickle *et al.* 2005, Van Sickle 2008). O/E: observed over expected species value.

Statistic	Mean	Standard Deviation	Root Mean Square Error
RIVPACS O/E predictive model based on calibration samples	1.03	0.26*	0.26
Null O/E model	1.00	0.31	0.31
Predictive improvement over the null model	-	0.05	0.05

*: Excludes 15 outliers that fell outside site-groups based on habitat variables (p <0.05).

4.3 Independence of Benthic Indices from Habitat Variables

We evaluated the independence of the benthic indices from habitat factors by computing Spearman correlation coefficients to evaluate the strength of the associations between index values and bottom water salinity, sediment grain size (percent fines), bottom depth, latitude and longitude. Ideally, index values are uncorrelated with habitat variables and are indicative of benthic condition, rather than habitat factors.

Overall, the indices were uncorrelated with the five habitat variables. Only 2 of 30 analyses yielded Spearman Correlation coefficients greater than 0.5; most were less than 0.2 (Table 4-6). The two r values >0.5 involved only one habitat variable (percent fines) and the BRI (r = 0.53) and BQI (r = -0.66) benthic indices. Graphical analysis indicated that these two associations were orthogonal to the index axes and definitely not driving the indices.

Table 4-6. Associations between each index and percent fines, depth, salinity, latitude and longitude. Spearman correlation coefficients, probabilities, and numbers of data are presented.

Index	Percent Fines	Depth	Salinity	Latitude	Longitude
BRI	0.53	-0.05	-0.25	-0.05	0.06
	<0.0001	0.118	<0.0001	0.157	0.065
	972	983	919	983	983
AMBI	0.29	-0.21	-0.17	-0.16	0.24
	<0.0001	<0001	<0.0001	<0.0001	<0.0001
	939	950	890	950	950
RBI	-0.27	0.09	0.21	0.17	0.08
	<0.0001	0.0031	<0.0001	<0.0001	0.01
	972	983	919	983	983
BQI	-0.66	0.03	0.21	-0.06	0.20
	<0.0001	0.4144	<0.0001	0.05	<0.0001
	972	983	919	983	983
O/E	-0.04	0.14	0.08	-0.04	0.16
	0.2107	<0.0001	0.0190	0.2162	<0.0001
	918	918	918	918	918

4.4 Associations among the benthic indices

There was general agreement among the indices and there were indications that they were accurately differentiating contaminated and uncontaminated sites, but some correlations were substantially lower than the others. Indices with common polarity (low values indicating "good" conditions and high values indicating "poor" conditions for both indices, or vice versa) had positive correlations; indices with opposite polarity had negative correlations. At the extremes, the polarity of *a priori* contaminated and uncontaminated sites met expectations, with high or low values conforming to index formulation. As previously mentioned, there was substantial overlap between uncontaminated and contaminated sites at the middle of the range.

The strongest associations were observed between the BRI and BQI, BRI and AMBI, BRI and RBI, and RBI and BQI (Figure 4-1), with Spearman correlation coefficients between 0.51 and 0.69. The strongest correlation was between the BRI and BQI, with a Spearman correlation coefficient of -0.69.

4.5 Discussion

All five indices calibrated successfully and were carried forward to the validation phase because they met the two primary calibration criteria. First, none of the indices was strongly driven by habitat factors. Second, none of the indices exhibited strong aberrant behavior clearly inconsistent with contaminated-uncontaminated polarity of the environmental condition gradient.

Index performance is measured primarily at the validation and evaluation stage, based on how well an index evaluates the condition of samples independent of the data used to calibrate it. The primary principle is that index performance should be judged on the basis of performance with respect to independent data, rather than the data used to calibrate it.

It is not possible to say that any index calibrated more successfully than any other. Independent validation data sets should be used to make judgments about the relative performance of individual indices and to select index threshold values that classify sites into different assessment categories of environmental condition. The selection of combinations of one or more indices that perform best and are most efficient for use in routine assessments is addressed in subsequent sections of this report.

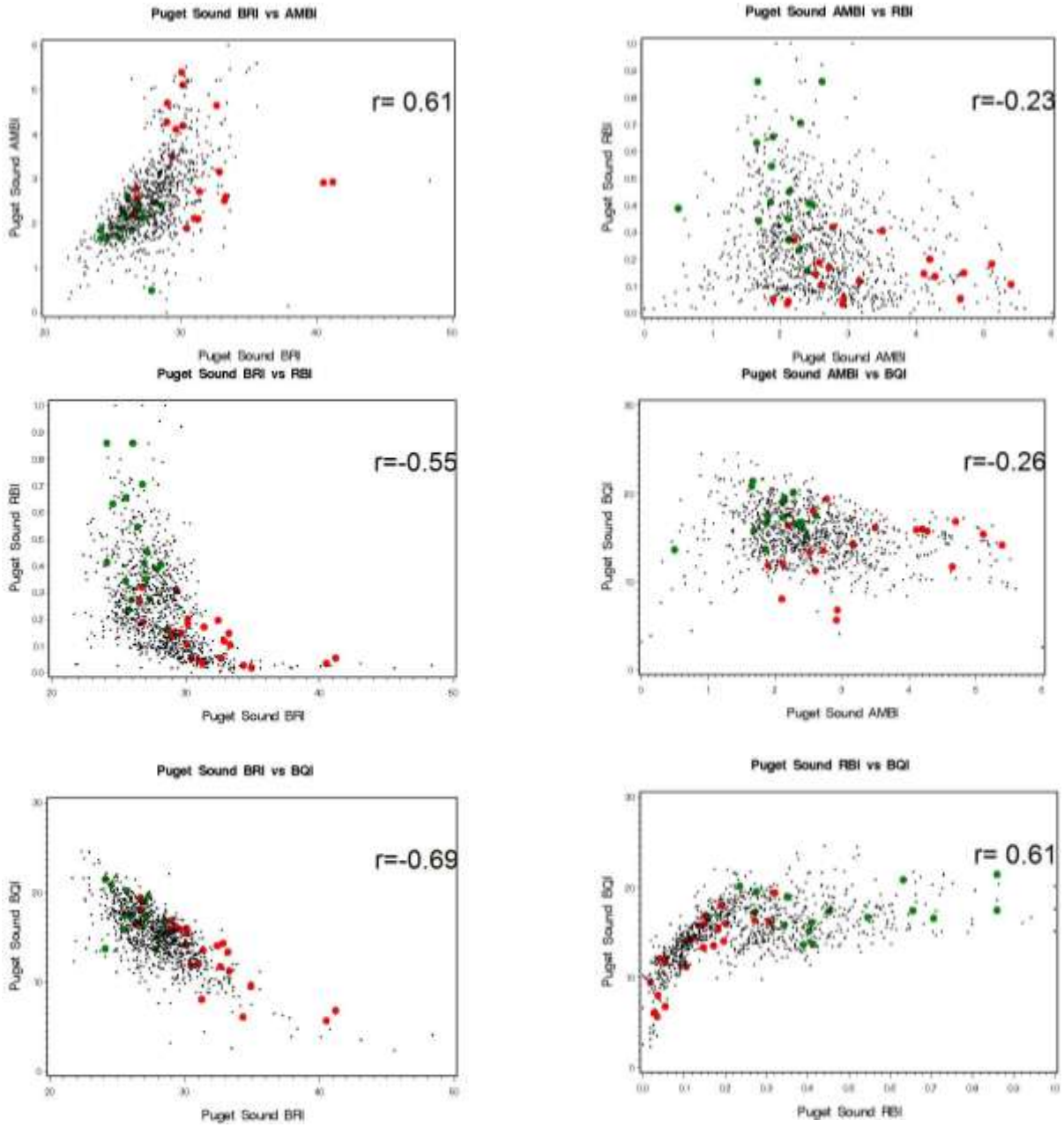

Figure 4-1. BRI and AMBI (top left), BRI and RBI (center left), BRI and BQI (bottom left), AMBI and RBI (top right), AMBI and BQI (center right), and RBI and BQI (bottom right) values for the calibration data. Green and red dots represent "unaffected" and "affected" endpoints (See Section 3.2). R values are Spearman correlation coefficients.

5. SELECTION OF BENTHIC INDEX VALIDATION DATA, BASED ON EXPERTS USING BEST PROFESSIONAL JUDGMENT TO ASSESS BENTHIC CONDITION

5.1 Introduction

Validation of benthic index performance is necessary to assure the accuracy of assessments based on benthic indices (Borja and Dauer 2008, Borja *et al.* 2009, Ranasinghe *et al.* 2012b, Teixeira *et al.* 2012). Validation usually uses data independent of those used for index calibration to evaluate the accuracy with which benthic indices order samples along a disturbance gradient, and assign assessment categories relative to assessment thresholds. Examples of assessment categories are "reference, loss of biodiversity, loss of community function, and defaunation" (Smith *et al.* 2001) and "undisturbed, slight disturbance, moderate disturbance, and high disturbance" (Ranasinghe *et al.* 2009, Teixeira *et al.* 2012).

Establishing the true position of validation samples along a disturbance gradient and their condition relative to assessment categories is potentially problematic. Initial benthic index development efforts (Weisberg *et al.* 1997, Van Dolah *et al.* 1999, Paul *et al.* 2001) relied on measurements of sediment toxicity and sediment contaminant concentrations. However, this approach is not always reliable because chemicals that are present may be tightly bound to sediments and, therefore, unavailable to affect benthic organisms. After exposure to high contaminant concentrations over extended periods of time benthic organisms may adapt and show minimal or no effects. Examples are the lack of effect on southern California mainland shelf benthos of DDT degradation compounds at levels predicted to have serious effects and mercury that occurs naturally in San Francisco Bay sediments. Conversely, chemicals that are not measured, such as recently invented pesticides and fire retardants, may have toxic effects on some benthic organisms. Complicating matters further, physical disturbance due to natural factors such as wave action, strong currents, and underwater landslides following earthquakes may be indistinguishable from anthropogenic effects. Interpreting benthic indices for management purposes is further complicated when the indices are based on different combinations of measures and metrics (Diaz *et al.* 2004, Borja *et al.* 2009).

One potential solution is to apply expert best professional judgment (BPJ) to establish a set of samples that provide a uniform scale for validating any index, but this assumes that there is consensus on benthic community condition classification among experts (Teixeira *et al.* 2010). Application of BPJ often conforms to general models of benthic community responses to stress (Pearson and Rosenberg 1978, Dauer 1993), but different experts often emphasize different aspects or elements of these models, leading to uncertainty regarding the extent to which experts agree in their application of BPJ. Three recent benthic BPJ studies in eight regions produced mixed results, with average correlations between local expert rankings >0.9 in five regions, 0.79 in a sixth region (Weisberg *et al.* 2008, Teixeira *et al.* 2010), and 0.29 and 0.38 in two more regions (Thompson *et al.* 2012).

Here, for 40 Puget Sound samples assessed by benthic experts using BPJ, we evaluate (1) the magnitude of the overall benthic condition gradient, (2) the condition of the samples relative to (a) each other, (b) assessment thresholds, and (3) the level of agreement among the experts, in order to determine the suitability of establishing the set of samples as a standard for validation of

Puget Sound benthic indices. In Section 6, the suitability of five benthic indices for use in Puget Sound assessments is evaluated, based on the validation data established here.

5.2 Methods

Six expert benthic ecologists were provided species composition and abundance data from 40 sites in Puget Sound and asked to determine the condition of the benthic community at each site. They were asked to rank the relative condition of each site from best to worst. They were also asked to assign each site to one of four categories of absolute condition: (1) Undisturbed: a community that would occur at a reference site; (2) Low Disturbance: a community that shows some indication of stress, but could be within measurement error of undisturbed condition; (3) Moderate Disturbance: a community that shows clear evidence of physical, chemical, natural, or anthropogenic disturbance; and (4) High Disturbance: a community with a high magnitude of disturbance.

The benthic experts were also asked to list the attributes of the benthos they used to determine site rankings and condition categories and to rate the importance of the attributes as follows: (1) Very important, (2) Important, but secondary, (3) Marginally important, (4) Useful, but only to interpret other factors, and (5) Very low importance. Attributes that were not used by an expert for site classification were assigned a rank of 6 for the purpose of calculating an average importance of that attribute among experts. The experts were free to use attributes of their choice and no standardized list or guidance was provided. Since all the experts identified indicator species as one of the attributes used in their assessment, they were also asked to list the organisms used as indicator species and to rank the species importance using the same scale. The experts were not asked to differentiate among potential causes for affected condition as it is generally recognized that current models of benthic response to stress do not discriminate between chemical contamination and other sources of disturbance (Borja *et al.* 2003).

The experts were selected to represent a range of affiliations and experience. They are listed in the acknowledgments. Three of the experts were from a state agency that conducts routine marine benthic monitoring and three (including one retired) from municipalities that implement benthic monitoring programs to assess the effect of discharge outfalls. Several of the experts from both affiliations were employed by private consulting companies during some part of their careers. Their experience in benthic monitoring ranged from 25 to 43 years, with an average of 34 years. The primary geographic expertise of all six experts was benthos of the west coast of the United States, although three were more familiar with Puget Sound and three were more experienced in southern California.

The 40 sites were selected systematically from a 1,023-sample Puget Sound sediment quality database in an effort to ensure that a wide range of benthic conditions was represented. For each site, the database included benthic community composition and physical habitat conditions, and sediment toxicity and sediment chemistry data were available for some sites. The samples were selected in three steps:

1. The species abundance data for the 1,023-sample database were analyzed by principal coordinate ordination;

2. A disturbance gradient in the principal coordinate space was identified as a vector joining the centroids of contaminated and uncontaminated sites, based on available sediment chemistry and sediment toxicity data; and
3. The 1,023 samples were ordered along the disturbance gradient and 40 samples were selected at regular intervals, starting with the second sample at the disturbed end of the gradient.

While it is generally accepted that current models of benthic response do not discriminate between chemical contamination and other sources of stress (Borja *et al.* 2003), this approach ensured that a range of benthic conditions was represented in the calibration and validation data. Although chemical contamination was used to ensure that a range of site conditions was included in the assessment, the experts were not provided the chemical data. They were only supplied salinity, sediment grain size (percent fine sediments), bottom depth, and species composition and abundance information. Although community measures such as diversity, dominance, and similarity indices were not provided, the data were distributed in spreadsheets, facilitating calculation by the individual experts of any measures they considered useful. The samples were only identified as being from "Puget Sound" and specific site locations and coordinates were not provided. To further ensure expert evaluations were based only on provided species abundance and habitat data, the samples were randomly numbered simply from one to forty, replacing the original sample identifications that may potentially have been related to the disturbance gradient through project reports.

The level of agreement among experts was initially evaluated based on the number of categories to which samples were assigned by the experts, and the average Spearman correlation coefficient among the experts for rankings of all 40 samples. The initial level of agreement was considered low because only 13 of the 40 (32.5%) samples were assessed in one or two adjacent categories and the average Spearman correlation coefficient among experts for the 40 sample rankings was only 0.34. In a previous benthic BPJ study with 9 experts in two habitats (Weisberg *et al.* 2008), 91.7 and 90.9% of 24 and 11 samples, respectively, were assessed in 1 or 2 adjacent categories, and average correlation coefficients were 0.91 and 0.91, respectively. The maximum average deviation among experts in the two habitats was 11.9 and 11.3%, respectively.

Because the levels of agreement among experts were towards the lower end of values for previous benthic BPJ studies, (1) standards of agreement were set for inclusion of samples in the Puget Sound benthic index validation data, and (2) a Delphi approach (Burns *et al.* 2000, Elwyn *et al.* 2006, Wright and Shannon 2006, Hsu and Sandford 2007) involving interaction among the experts examining the rationale for outlier category assignments was used to generate additional insights and improve agreement about condition categories. Following the interaction, the experts revised their rankings to conform to agreement about evaluation principles achieved during the interaction.

The post-Delphi interaction categories and rankings were compiled, and the numbers of adjacent categories and average ranking deviation from the mean among the experts were calculated for each sample. Samples categorized in one or two adjacent categories, and with an average ranking deviation among experts $\leq$11% were selected for inclusion in the Puget Sound benthic index validation data.

5.3 Results

Initial condition category assignments and sample rankings showed limited agreement among experts, but improved substantially after the Delphi interaction (Table 5-1). The number of potential validation samples meeting both the category and rank criteria for use in index validation increased from six to seventeen.

The number of samples where category was changed after the Delphi interaction ranged from 1 for Expert B, to 4 for Experts C and E, 7 for Expert D, 10 for Expert F, and 14 for Expert A (Figure 5-2). It is apparent from Table 5-2 that almost all the changes were relatively minor, yet they resulted in almost three times as many samples meeting the criteria for inclusion in the validation data after the Delphi interaction (Table 5-1). Most changes involved, after discussion, moving samples from a position close to one side of a category boundary to a position close to, but on the other side of, the same category boundary. Several experts stated that although there was certainty about the condition rank order, there was uncertainty about exactly in which category the sample belonged. An exception was a change of several samples from Category 3 to Category 2 by Expert A, whose primary expertise was in southern California. The expert, who previously downgraded coarse sites where capitellids and oligochaetes occurred, chose to make the change after learning about amphipod tube mats that occur in Puget Sound coarse sediments, effectively creating fine sediment habitats suitable for these organisms.

Even before the Delphi interaction, there was substantial agreement in condition categories assigned by the experts (Table 5-2). At least half of the experts agreed on sample condition category for 37 of the 40 samples. All the experts agreed on condition category for only one sample, but five of the six experts agreed on condition categories for three samples, and four agreed for thirteen samples. On the other hand, three samples were assigned by different experts to all four condition categories, but in each case three or more experts did agree on whether the samples were in “good” or “bad” condition.

Some experts exhibited a tendency to categorize samples more strictly or more liberally than others experts (Table 5-3). For example, five of the six experts assigned samples to all four narrative condition categories (Table 5-3). Expert F initially categorized no samples in the least disturbed reference category and 32 of the 40 samples as “bad” (“moderate disturbance” or “high disturbance” categories). These differences among experts were quantified by ranking them on a gradient of severity from 1 to 6, with 1 indicating a greater tendency to assign sites to the two "good" condition categories. The Delphi process did not substantially change the expert severity rankings (Table 5-3).

The experts used nine criteria, five of which were used by four or more of the experts, to assess the samples (Table 5-4). The three most important criteria were dominance by tolerant indicator taxa, species richness, and taxonomic diversity at levels above family (Table 5-4). Total abundance was used by all of the experts, but many of them ranked this criterion of lesser importance because they used it only to modify assessments when values were very high or very low. The other criteria included the abundance or presence of selected species or higher taxa and two community measures, the Swartz Dominance Index and Pielou’s evenness (J′).

There was considerable consistency in the tolerant indicator taxa identified by the experts, while sensitive taxa generally varied from expert to expert (Table 5-5). The exceptions were echinoderms and crustacea, which were identified by three experts as sensitive taxa at the phylum and class levels, respectively. The polychaetes *Capitella capitata* and *Aphelochaeta* spp., the bivalve *Parvilucina tenuisculpta*, and oligochaetes were most frequently recognized as tolerant indicator taxa. Many of the tolerant taxa were identified at the species or genus level, while sensitive taxa were more often identified at higher taxonomic levels.

Of the four parameters used by five of six experts, taxa richness was most highly correlated with the consensus site rankings of the experts (Table 6). Of the biological parameters, abundance correlated most poorly with the consensus site rankings.

5.4 Discussion

Our results suggest that the 17-sample data set with high agreement among the experts is suitable for evaluating and validating the assessment accuracy of benthic indices in Puget Sound, based on the wide range of condition categories in the data and the high level of agreement among the experts. The range of condition categories assigned with high agreement by the experts included a broad range of benthic conditions. Past validation efforts have been compromised by difficulty in identifying sites that represent relatively pristine, impacted, and mid-range sites.

Our approach of selecting only samples with high agreement among experts for validation of benthic indices and using the Delphi process to increase confidence in expert judgment likely solved a situation that may have been irreconcilable. In the absence of the Delphi adjustments, the initial low level of agreement among the experts would likely have resulted in an index validation that was difficult to defend.

The present study clearly demonstrates that (a) restricting samples used for validation to those with high agreement among the experts, and (b) adopting the Delphi interaction process to improve understanding and increase the level of consensus among the experts is a viable method of validating benthic indices. Reducing the number of samples to those with highest agreement among experts and increasing the number of samples with high agreement by the Delphi interaction serve to improve and increase confidence in the accuracy of the validation. Using the process to (a) select the samples to be used for validation, (b) establish their relative condition, and (c) identify assessment condition category "truth" for benthic index evaluation clearly is more defensible, especially when initial levels of agreement are low due to differing backgrounds or different geographic experience of participating experts.

As Weisberg *et al.* (2008) concluded, consensus expert opinion as an evaluation benchmark may facilitate evaluation of how the indices are performing in assessing sites experiencing intermediate levels of disturbance. This is a more difficult, but more relevant, assessment challenge for indices. The use of expert opinion also provides a benchmark to assess index performance. Index developers have generally identified an index as successful if it correctly differentiated 80% of sites with extreme conditions. A better evaluation benchmark would be an index's ability to classify sites with a level of correlation comparable to that among experts.

It is important that the experts using the Delphi approach reach common views and agreement through their own volition and that there should be no pressure to reach consensus or for failure to do so. Although the experts generally agreed on the most important criteria used for assessment (Table 5-4), they often disagreed on their relative importance, as did the experts in Weisberg *et al.* (2008). The approach we adopted was to initially pick samples where there was disagreement among two or more experts and discuss their species composition. The approach was successful in some cases, increased levels of understanding among the experts, and allowed consensus to be reached for some samples, but not others. One source of apparently irreconcilable differences was a tendency for southern California experts to base their assessments on perceived sensitivity or tolerance of dominant species and Puget Sound experts to assess samples based on community parameters such as Pielou's evenness (J′) and Swartz's Dominance. Where community abundances were unbalanced due to abundances of species that are known to be sensitive in some situations, agreement could not be reached.

Geography-related differences among these two sets of experts in the characterization of species responses to stressors and pollution may also have contributed to differences. For example, *Euphilomedes carcharodonta* was characterized by a southern California expert as a tolerant "halo" species that is abundant surrounding and close to discharge outfalls, but not at the discharge itself. In contrast, a Puget Sound expert characterized this species as common in shallow water with coarser sediments and constant water movement, a scenario more likely for sensitive species.

Most indices include abundance or proportions of sensitive and tolerant taxa as important assessment metrics, which are also important for European assessments (Borja *et al.* 2000, Muxika *et al.* 2005, Dauvin, 2007). For the sensitive and tolerant taxa parameters at least, benthic indices could provide a means of improving upon the experts' assessments because the list of species relied upon by an individual expert is typically limited or is a broad generalization applied to higher-level taxa (e.g., Gammaridea). Every species occurring at a site provides information regarding community condition, and indices that integrate empirical data from many samples to capture information for a larger number of taxa may lead to more accurate assessments. Consensus lists of such taxa are well developed in Europe (Borja and Muxika, 2005).

Table 5-1. Numbers of samples meeting category, rank, and both sets of criteria for inclusion in validation data before and after Delphi interaction. The category criterion was assignment of a maximum of two adjacent categories by all six experts. The ranking criterion was an average ranking deviation from the mean ≤11% for each sample. The total number of potential validation samples was 40.

Samples Meeting Category Criterion		Samples Meeting Rank Criterion		Samples Meeting Both Criteria	
Before	**After**	**Before**	**After**	**Before**	**After**
13	28	9	19	6	17

Table 5-2. Condition categories assigned to samples by benthic ecologists before and after Delphi interaction. Changed categories are colored purple. Column letters represent different benthic experts. # Cats: Range of categories. # Cats Change indicates whether the Delphi interaction changed the range of categories. Condition categories: 1: Undisturbed; 2: Low Disturbance; 3: Moderate Disturbance; 4: High Disturbance. Samples are ordered by increasing median condition ranking before Delphi Interaction.

Before Delphi Interaction								After Delphi Interaction							
A	B	C	D	E	F	# Cats	Sample No.	A	B	C	D	E	F	# Cats	# Cats Change
1	2	2	1	1	2	2	**33**	1	2	2	1	1	1	2	
3	1	1	1	1	2	3	**11**	2	1	1	1	1	2	2	Change
2	1	1	1	1	3	3	**27**	2	1	1	1	1	2	2	Change
2	1	2	1	1	3	3	**21**	2	1	2	1	1	2	2	Change
3	1	2	1	1	3	3	**12**	2	1	2	1	1	2	2	Change
1	1	2	1	1	3	2	**16**	1	1	2	1	1	2	2	
1	2	2	1	1	3	3	**03**	1	2	2	1	1	2	2	Change
2	2	3	1	2	3	3	**35**	2	2	3	1	2	2	3	
3	1	3	1	1	2	3	**39**	2	1	3	1	1	2	3	
1	2	2	1	2	3	3	**06**	2	2	2	1	2	2	2	Change
1	3	1	3	3	2	3	**05**	1	3	1	3	3	2	3	
3	2	2	1	2	3	3	**24**	2	2	2	1	2	3	3	
2	3	1	2	2	2	3	**22**	2	3	2	2	2	2	2	Change
1	3	3	1	3	3	3	**14**	2	3	3	1	3	3	3	
2	3	1	2	3	3	3	**13**	2	2	1	2	2	2	2	Change
2	3	1	2	3	2	3	**08**	2	3	2	2	3	2	2	Change
2	2	2	2	3	3	2	**29**	2	2	2	2	3	3	2	
3	3	3	1	2	3	3	**25**	3	3	3	2	3	3	2	Change
2	2	1	3	3	3	3	**40**	2	2	1	3	2	3	3	
3	2	3	2	2	3	2	**01**	2	2	3	2	2	3	2	
3	3	3	1	2	2	3	**23**	3	3	3	2	3	2	2	Change
2	3	3	1	3	3	3	**31**	2	3	3	2	3	3	2	Change
3	2	3	1	2	3	3	**15**	3	2	3	2	2	3	2	Change
2	3	2	2	3	3	2	**02**	2	3	3	2	3	3	2	
2	3	3	2	3	3	2	**04**	3	3	3	2	3	3	2	
1	4	2	3	4	3	4	**38**	1	4	2	3	4	3	4	
1	3	4	2	2	3	4	**37**	2	3	4	2	2	3	3	Change
1	3	4	1	1	4	4	**28**	2	3	4	1	1	4	4	
2	4	4	2	2	3	3	**07**	3	4	4	2	2	3	3	
3	3	4	1	3	2	4	**26**	3	3	4	3	3	2	3	Change
2	3	3	2	3	3	2	**18**	3	3	3	2	3	3	2	
2	3	2	4	3	4	3	**09**	2	3	2	3	3	4	3	
3	3	2	3	3	3	2	**10**	3	3	2	3	3	3	2	
4	3	4	2	3	3	3	**30**	4	3	4	3	3	3	2	Change
3	3	2	3	3	3	2	**17**	3	3	3	3	3	3	1	Change
2	3	3	3	3	3	2	**34**	3	3	3	3	3	3	1	Change
4	4	4	3	4	3	2	**36**	4	4	4	3	4	3	2	
2	4	3	4	4	4	3	**20**	3	4	3	4	4	4	2	Change
4	4	3	4	4	3	2	**19**	4	4	3	4	4	3	2	
4	4	4	4	4	4	1	**32**	4	4	4	4	4	4	1	

Table 5-3. Numbers of samples categorized as "Good" (Categories 1 and 2) or "Bad" (Categories 3 and 4) by benthic ecologists A thru F before and after Delphi interaction. The Severity Rank ranks the experts on a gradient of severity from 1 to 6, with 1 indicating a greater tendency to assign sites to the two "good" condition categories.

Category	A		B		C		D		E		F	
	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After
1	9	5	6	6	7	5	18	13	9	9	0	1
2	16	20	9	10	13	13	11	14	10	10	8	15
3	11	11	19	18	13	15	7	10	16	16	28	20
4	4	4	6	6	7	7	4	3	5	5	4	4
Good	25	25	15	16	20	18	29	27	19	19	8	16
Bad	15	15	25	24	20	22	11	13	21	21	32	24
Severity Rank	2	2	5	5.5	3.5	3.5	1	1	3.5	3.5	6	5.5

Table 5-4. Criteria used by benthic ecologists to rank and categorize samples. Importance is the average importance for all experts where 1: Very important; 2: Important, but secondary; 3: Marginally important; 4: Useful, but only to interpret the other factors; 5: Very low importance; 6: Not used. N is the number of experts that used the criterion.

Criteria	Importance	Std. Deviation	N
Numeric dominance by tolerant indicator taxa	1.7	1.21	5
Species richness (number of taxa)	2.0	1.26	6
Taxonomic diversity at levels above family (a surrogate for community diversity of ecological function)	2.3	1.86	5
Total abundance; very high or very low levels indicate disturbance	3.2	1.17	6
Presence of sensitive indicator taxa	3.5	2.17	4
Swartz Dominance Index	4.2	2.23	3
Pielou's Evenness (J')	4.3	2.58	2
Diversity of, or dominance by, specific higher level taxa such as sensitive arthropods or echinoderms	4.8	1.83	2
Presence of tolerant indicator taxa	5.2	2.04	1

Table 5-5. Indicator taxa identified by the benthic ecologists. ☼: Identified as both sensitive and tolerant.

Sensitive Taxa		
Genus and Species	**Group**	**N**
Euphilomedes carcharodonta☼	Arthropoda: Ostracoda	1
Euphilomedes product	Arthropoda: Ostracoda	1
Nutricola lordi☼	Mollusca: Bivalvia	1
Galathowenia oculata	Annelida: Polychaeta	1
Amphiodia spp.☼	Echinodermata: Ophiuroidea	1
Ophiura sarsi	Echinodermata: Ophiuroidea	1
Amphipholis squamata	Echinodermata: Ophiuroidea	1
Higher Taxa	**Group**	**N**
Ampharetidae (most species)	Annelida: Polychaeta	1
Maldanidae (most species)	Annelida: Polychaeta	1
Sabellidae	Annelida: Polychaeta	1
Terebellidae	Annelida: Polychaeta	1
Trichobranchidae	Annelida: Polychaeta	1
Arthropoda (most species)	Arthropoda	3
Echinodermata (most species)	Echinodermata	3
Ophiuroidea (other than Ophiuridae)	Echinodermata: Ophiuroidea	1
Gammaridea (most species)	Arthropoda: Amphipoda	1
Cumacea	Arthropoda: Cumacea	1

Tolerant Taxa		
Genus and Species	**Group**	**N**
Capitella spp., *Capitella capitata* species complex	Annelida Polychaeta	4
Aphelochaeta spp.	Annelida Polychaeta	3
Parvilucina tenuisculpta	Mollusca: Bivalvia	3
Armandia spp., *Armandia brevis*	Annelida Polychaeta	2
Dorvillea (Schistomeringos) annulata	Annelida Polychaeta	2
Axinopsida serricata	Mollusca: Bivalvia	2
Macoma carlottensis	Mollusca: Bivalvia	2
Euphilomedes carcharodonta☼	Arthropoda: Ostracoda	2
Mediomastus spp.	Annelida Polychaeta	1
Scoletoma luti	Annelida Polychaeta	2
Prionospio (Prionospio) steenstrupi	Annelida Polychaeta	1
Macoma nasuta	Mollusca: Bivalvia	1
Nassarius mendicus	Mollusca: Bivalvia	1
Nutricola lordi☼	Mollusca: Bivalvia	1
Amphiodia spp.☼	Echinodermata: Ophiuroidea	1
Higher Taxa	**Group**	**N**
Oligochaeta	Annelida: Oligochaeta	3
Capitellidae	Annelida Polychaeta	1
Dorvilleidae	Annelida Polychaeta	1

Table 5-6. Spearman correlation coefficients between selected abiotic and benthic measures in the BPJ samples and site rankings by the benthic ecologists.

Measure	Spearman Correlation Coefficient (n = 40)	Probability
Number of taxa	-0.75	<0.0001
Total abundance	-0.29	0.07
Fine sediments (%)	0.22	0.17
Bottom depth (m)	-0.10	0.52
Salinity	-0.21	0.20

6. BENTHIC INDEX EVALUATION AND OPTIMIZATION

6.1 Introduction

In previous sections of this report, three important preliminary tasks were accomplished. First, in Section 3, a single benthic assemblage was identified in Puget Sound, with the consequence that only one set of assessment thresholds is necessary for Puget Sound benthic assessments. This differed from other assessment areas (e.g., San Francisco Bay and Chesapeake Bay) where different assemblages coexist in different habitats related to salinity or sediment type and it is necessary to use different sets of assessment thresholds for each assemblage. Second, in Section 4, five benthic indices developed in other geographies were calibrated for use in Puget Sound. These indices were shown to be indicative of benthic community condition and are not driven by habitat factors. Third, in Section 5, a set of 17 samples independent of the calibration data were selected for evaluation of the calibration results. The relative condition and condition categories of these samples met stringent standards for agreement among six benthic experts and may be assumed to represent "truth" and are suitable for evaluating the index calibration results.

Completion of these tasks leaves three questions to be answered before one or more of the five calibrated benthic indices can be accurately and efficiently used in routine benthic monitoring:

1. Do the five calibrated benthic indices validate in Puget Sound?
2. What are the correct assessment thresholds that should be applied to benthic index values in order to accurately assign benthic community condition categories in Puget Sound?
3. What is the optimum suite of indices for accurate benthic condition determination? Is it necessary to calculate and apply all the indices that work, or is there a more efficient and less time-intensive alternative?

The goal of this section is to answer these three questions. The specific objectives are to: (1) evaluate the accuracy of benthic indices in ranking benthic condition, (2) determine assessment thresholds for accurate condition category assignments, and (3) identify one or more index combinations that optimize benthic condition category assessment to accurately assign condition categories with as few indices as possible.

6.2 Methods

Validation Data

Index performance was assessed by comparing index results for 17 sites that were not used in index calibration to the consensus assessment of six benthic experts. Each expert was provided information on species abundances and habitat (depth, salinity and sediment grain size) for each site without the site location. The experts were then asked to (1) rank the sites from best to worst condition, and (2) classify each site on the four-category scale of benthic condition to which the benthic indices were calibrated. The experts were initially provided data for 40 sites and 17 sites that met strict criteria of agreement among the experts were included in the validation data. Details of the validation data selection process are provided in Section 5.

The 5 benthic indices were calculated from the benthic species abundance data for the 17 validation samples by applying the calibrations described in Section 4. The BRI values were

calculated as abundance-weighted mean tolerance scores for species in the samples, based on species tolerance scores calculated in Section 4. The O/E values were calculated for the 17 validation samples, using discriminant functions developed during calibration, first to identify the habitat site-group to which a sample belonged, and then to evaluate the observed species in relation to expectations for a minimally disturbed reference site. Although the lowest possible O/E value is 0.0, and the peak or optimum O/E value for undisturbed samples is 1.0, O/E values >1.0 are often encountered and are considered to represent less than optimal condition (Ranasinghe *et al.* 2009). Therefore, for samples with O/E values >1.0, "unwound" O/E values were calculated that were less than 1.0 by the same amount by which the O/E value exceeded 1.0, resulting in unwound O/E values on a scale from 0 to 1. The AMBI index values were calculated by applying species Ecological Group assignments developed during calibration to the standard AMBI equation. The RBI and BQI values were calculated using standard equations (Section 4.2).

Benthic Condition Ranking Evaluation

Benthic index performance was assessed (1) by comparing the rank order of index values to the median consensus expert rank order in the BPJ study (Section 5) using Spearman rank correlation coefficients, and (2) comparing the benthic index values to the BPJ rank order graphically. The objective was to determine whether benthic index values responded to relative benthic community condition as expressed by expert consensus with sufficient accuracy. This determination was dependent on benthic community condition ranking of the samples, and independent of condition category assignments. Experts often agree on condition ranks, but may differ slightly on where, on the disturbance gradient, condition category thresholds should be set.

Associations among the five indices were also evaluated and compared to associations among the experts, using Spearman rank correlation coefficients. For informational and explanatory purposes, Spearman rank correlation coefficients between the sample benthic index values and (1) numbers of taxa, (2) the RBI biodiversity component, and (3) total benthic abundance of the samples were also calculated and evaluated. The RBI biodiversity component is the weighted sum of four community parameters (total number of species, number of crustacean species, number of crustacean individuals, and number of mollusc species).

Index Threshold Scaling

The AMBI, BQI, BRI, and RBI performed with sufficient accuracy in the benthic condition ranking evaluation (above) and were calibrated to the four-category benthic condition scale established in Section 5. The O/E was eliminated because of insufficient accuracy in the benthic community ranking evaluation. The condition categories were (1) Undisturbed − a community that would occur at a reference site for that habitat; (2) Low disturbance − a community that exhibits some indication of stress, but might be within measurement variability of reference condition; (3) Moderate disturbance − a community that exhibits clear evidence of physical, chemical, natural, or anthropogenic stress; (4) High disturbance − a community exhibiting a high magnitude of stress. Moderate and high disturbance communities are considered "altered" because they show clear evidence of disturbance due to anthropogenic or natural stress, while undisturbed and low disturbance communities do not. Altered communities could be due to the effects of one or more types of anthropogenic or natural stress while unaffected communities likely indicate minimal stress of all types.

For each of the 4 indices, developer and non-developer threshold sets were established, and the threshold set that performed best with the 17 sample validation data set was selected. Developer thresholds are those provided during the original construction of the index. Non-developer thresholds were established by applying statistical optimization methods to compare index values and consensus benthic condition categories. For the AMBI, the developer thresholds were the universal thresholds modified slightly to convert the five-category universal AMBI scale to the four condition category scale of this study. The two most disturbed of the five AMBI categories (highly disturbed and extremely disturbed) were combined into the high disturbance category. On the 0 to 7 universal AMBI scale, the developer AMBI classifications were: (1) <1.2: Undisturbed; (2) 1.2 to <3.3: Low disturbance; (3) 3.3 to <5.0: Moderate disturbance; and (4) ≥5.0: High disturbance. BQI developer thresholds are equally spaced along the index range (Rosenberg *et al.* 2004), and developer thresholds were established at 7.92, 13.47, and 19.02 on the 2.36 to 24.58 BQI range of the calibration samples. Developer threshold sets were not calculated for the BRI and RBI because none were specified by the developers.

Two sets of non-developer thresholds were selected for each indicator, based on consensus condition categories assigned by six benthic experts to the 17-sample validation data. One optimization technique was based on maximizing the weighted kappa statistic (Cohen 1960 1968), which measures agreement between indicator and consensus categories beyond that expected by chance. Weights were based on the linear weighting scheme of Cicchetti and Allison (1971), which give "partial credit" according to severity of disagreement. The second set of thresholds was based on maximizing agreement between indicator and consensus categories, with no weighting factors. To find the optimal set of thresholds in each case, weighted kappa statistics and percent agreement were computed for all possible sets of triplicate thresholds.

Benthic Assessment Optimization

Condition category assessments by the benthic indices, and by all possible index combinations, for the 17 validation samples were compared to the consensus expert condition assessment in 3 ways:

1. Status classification accuracy, defined as the accuracy with which an index differentiated benthos identified by the six experts as altered (moderate or high disturbance) from benthos identified as unaltered (undisturbed or low disturbance). This mimics the "good-bad" evaluation approach used in many previously published benthic indicator development efforts.
2. Categorical classification accuracy with respect to the four condition categories established for index calibration (undisturbed, low, moderate, and high disturbance). This is more challenging than status classification because it requires finer discrimination of the same benthic responses among a larger number of categories.
3. Bias in category designation; the sum of differences between index (or index combination) category and the consensus categorical classification of the experts when categories are expressed numerically (Undisturbed = 1, High Disturbance = 4). Positive bias indicates a tendency to score samples as more disturbed than the expert consensus, while negative bias indicates a tendency to score samples as less disturbed. Larger absolute values indicate stronger bias.

Index combinations were evaluated as the median of the numeric categories (Undisturbed = 1, High Disturbance = 4). If the median for the indices in a combination fell between categories, it was rounded to the higher effect category. Comparisons to the experts were performed for each of the three threshold approaches associated with each index, with the best performing thresholds used when combining indices. Category-optimized thresholds were selected for all four indices. Status and category classification accuracy and category bias were calculated for the 17 evaluation samples.

6.3 Results

Benthic Condition Ranking Evaluation

Four of the five benthic indices responded with sufficient accuracy to benthic community condition as expressed by expert consensus, while the O/E index did not. The BQI, RBI, AMBI, and BRI were highly correlated with the median expert ranks and had absolute Spearman correlation coefficient values >0.75, which were of high statistical significance (Table 6-1). In contrast, the Spearman correlation for the (unwound) O/E index was not statistically significant and had an absolute value of only 0.31.

Table 6-1. Spearman correlation coefficients between the median BPJ rank order and benthic indices and other measures.

	Correlation with Median BPJ Rank Order (n = 17)	
	Spearman Correlation Coefficient	**Probability of Rejecting Statistical Significance Null Hypothesis**
Benthic Indices		
BQI	-0.98	<0.0001
RBI	-0.85	<0.0001
AMBI	0.78	0.0002
BRI	0.76	0.0004
O/E unwound (see text)	-0.31	Not significant
Experts		
Expert minimum	0.85	<0.0001
Expert mean	0.93	<0.0001
Expert maximum	0.99	<0.0001
Other Measures		
Number of Taxa	-0.96	<0.0001
RBI biodiversity component	-0.94	<0.0001
Total abundance	-0.33	Not significant

Plots of index values against the 17-sample validation condition gradient (Figure 6-1) also supported the conclusion that the BQI, RBI, AMBI, and BRI reflected community condition accurately, while the O/E index did not. The BRI and BQI were closest to being monotonic, changing gradually across the condition gradient with the lowest variability (least wild swings); the AMBI and RBI were only slightly less monotonic. In contrast, the (unwound) O/E index did not track the gradient well (Figure 6-1).

The strongest correlations with the condition gradient were of similar strength for the benthic indices and the experts, but the correlation coefficient range was tighter for the experts and

looser for the indices (Table 6-1). Two biodiversity-related measures, the number of taxa and the biodiversity component of the RBI, were strongly correlated with the condition gradient (Table 6-1) with statistically highly significant Spearman correlation coefficients of -0.96 and -0.94, respectively. In contrast, the statistically non-significant correlation coefficient for abundance was only -0.33.

Table 6-2. Associations between benthic indices. Spearman correlation coefficients and probabilities of rejecting a statistical significance null hypothesis (italicized) are presented. The "unwound" O/E index is described in the text. N = 17 for all cells.

Benthic Index	AMBI	BQI	BRI	Unwound O/E
BQI	-0.54 *0.025*			
BRI	0.44 *NS*	-0.89 *<0.0001*		
Unwound O/E	0.11 *NS*	0.57 *0.016*	-0.59 *0.013*	
RBI	-0.56 *0.020*	0.72 *0.001*	-0.52 *0.032*	0.27 *NS*

In general, between-index correlations were strong, with absolute values >0.5 for seven of the ten associations (Table 6-2). At -0.89, the strongest correlation was between the BRI and the BQI. The second strongest correlation was between the RBI and the BQI, with a Spearman correlation coefficient of 0.72.

The plots of index values along the validation condition gradient (Figure 6-1) and correlation coefficients among the indices (Table 6-1) showed all five indices performing essentially as designed. The BRI and AMBI had low values at the unaffected (low rank) end of the gradient and higher values at the affected (high rank) end. The BQI, RBI, and O/E indices had high values at the unaffected end of the gradient and low values at the affected end.

Index Threshold Scaling

Category-optimized assessment thresholds were selected for all four adequately performing benthic indices (Table 6-3). Category-optimized thresholds were selected for the two indices with developer thresholds because they had higher categorical classification accuracies of 52.9 and 35.2%, respectively for the AMBI, and 82.4 and 41.1%, respectively, for the BQI. The kappa-optimized and category-optimized threshold sets were almost identical because only one or two index categories varied from consensus categories by more than a single category and, as a result, "penalties" for excess deviation were negligible.

Table 6-3. Threshold values for condition category assignments for the four adequately performing benthic indices.

Condition Category	Benthic Index			
	AMBI	**BQI**	**BRI**	**RBI**
1: Undisturbed	<2.1	>19.6	<25.5	>0.6
2: Low Disturbance	2.1 to <2.4	>17.7 to 19.6	25.5 to <26.3	>0.3 to 0.6
3: Moderate Disturbance	2.4 to <3.6	>14.65 to 17.7	26.3 to <28.3	>0.1 to 0.3
4: High Disturbance	>3.6	≤14.65	≥28.3	≤0.1

Benthic Assessment Optimization

One individual index and two three-index combinations classified the 17 validation samples with category classification accuracy >75% (Table 6-4). These three permutations performed better than the average expert. The individual index, which was the BQI, had status classification accuracy of 100% while the three-index combinations had status classification accuracies of 94.1%. Bias for all three permutations was relatively low. However, bias was three times higher for three-index combination #14 than the other three-index combination (#12) and the BQI (#2). The magnitude of bias was low, and slightly lower for the indices than the experts. Bias was present for 12 of the 15 index permutations, and was positive for 11 of the 15 permutations, indicating a tendency to score samples more disturbed than the expert consensus. Only one of the validation samples (#1) had negative bias.

Ranges of category and status classification accuracy were similar for the indices and the experts (Table 6-4). The BQI and RBI, which are individual indices based on community measures, had higher category and status classification accuracy than the AMBI and BRI, which are based on species tolerance scores. There was no apparent consistent effect of index combinations increasing or decreasing category or status classification accuracy relative to individual indices.

Table 6-4. Classification accuracy and bias for indices and index combinations. Classification accuracy is presented for "altered versus unaltered" status and four condition categories. Each of 17 validation samples was assessed into one of four numeric categories by an index or index combination and compared with consensus categories from an independent assessment by six benthic experts. Bias is the sum of differences between index combination and consensus categories; positive values indicate a tendency to score samples as more disturbed than the expert consensus, while negative values indicate a tendency to score samples as less disturbed. The categories were 1: Undisturbed (Reference), 2: Low Disturbance, 3: Moderate Disturbance and 4: High Disturbance. Categories 3 and 4 were considered clearly indicative of anthropogenic or natural stress and Categories 1 and 2 were not. Index results were combined as the median of the numeric categories; if the median fell between categories, it was rounded up to the higher effect category. Results for the benthic experts are presented to provide context.

No of Indices	#	Indices	Category Accuracy (%)	Category Bias	Status Accuracy (%)
One	1	AMBI	52.9	-1	82.4
	2	BQI	82.4	1	100.0
	3	BRI	47.1	1	70.6
	4	RBI	70.6	1	94.1
Two	5	AMBI, BQI	64.7	0	88.2
	6	AMBI, BRI	58.8	0	76.5
	7	AMBI, RBI	58.8	0	88.2
	8	BQI, BRI	64.7	2	94.1
	9	BQI, RBI	70.6	2	94.1
	10	BRI, RBI	58.8	2	88.2
Three	11	AMBI, BQI, BRI	58.8	1	82.4
	12	AMBI, BQI, RBI	76.5	1	94.1
	13	AMBI, BRI, RBI	58.8	1	82.4
	14	BQI, BRI, RBI	76.5	3	94.1
Four	15	AMBI, BQI, BRI, RBI	56.7	2	88.2
Expert Consensus		Minimum	52.9	-8.0	76.5
		Average	74.5	-2.3	93.1
		Maximum	88.2	0.0	100.0

6.4 Discussion

The present study successfully identified that the BQI, a benthic index, and two three-index combinations (#12 and #14) identified sample condition categories and status of validation samples with higher accuracy than the average benthic expert. This level of performance warrants serious consideration of these tools for use in routine benthic monitoring of Puget Sound. BQI performance was slightly better than performance of the three-index combinations but will likely require roughly a third of the calculations required for three-index combination assessments. Therefore, it should be given serious consideration as a preferred method for conducting benthic assessments in Puget Sound.

The results of this study are based primarily on benthic condition assessments of only 17 samples and additional validation studies to increase confidence in, and perhaps improve, the identified assessment method may be useful. Verification that the recommended assessment methods are consistently accurate by sampling and conducting assessments of areas of known condition could be useful to increase confidence in the assessment protocol. However, finding suitable sites, and especially highly affected sites, in order to confirm the efficacy of these assessment protocols in Puget Sound may be challenging. Sediment chemical contamination and sediment toxicity levels may not be as high as in areas such as east coast harbors.

There are also indications that the highly diverse Puget Sound benthic assemblage may respond slightly differently than assemblages in other areas, with less success for measures of benthic species sensitivity and tolerance, and more success for benthic community measures such as numbers of species. Of the tested benthic indices, the BQI had the best relationship with the validation disturbance gradient (Figure 6-1) with a Spearman correlation coefficient of -0.98 (Table 6-1), which may explain its success. The formula for calculation of the BQI includes two components related to biodiversity, and numbers of taxa were highly correlated with the validation gradient with a Spearman correlation of -0.96 (Table 6-1), also possibly accounting for its performance. The presences of sensitive species (usually in low abundance) and dominance by tolerant species are often considered good indicators of unaltered and altered benthic condition, respectively. However, biodiversity is exceptionally high throughout Puget Sound and each sample contains low numbers of many species, making compilation of a list of sensitive species a difficult and nearly impossible task. In this environment, dominance by tolerant species is obvious only in extreme cases and community measures such as Pielou's evenness and Swartz dominance may be more sensitive to low disturbance levels. The reliance of Puget Sound experts on community measures in contrast to the reliance on indicator species by southern California experts (Section 5) may be an unrecognized consequence of this difficulty. Another factor to consider is that numbers of taxa have been highly correlated with condition gradients in most or all of the benthic best professional judgment studies conducted to date, potentially leading to the conclusion that benthic condition gradients reduce to biodiversity gradients. This relationship could be difficult to resolve because the reason for best professional judgment (BPJ) benthic studies is the difficulty of relating independent measures, such as sediment chemistry and toxicity, to benthic community condition.

One of the successes of this study was the identification of four benthic indices that were indicative of benthic community condition and elimination of another that was not. The most desirable situation for benthic indices are high correlation with condition gradients, and

approximately monotonic progression along the gradient with minimal variability relative to the difference in index values between the unaffected and affected ends of the condition gradient. High variability at the ends of the gradient has the most serious consequences due to increased uncertainty about the actual index value at the extremes of condition. These ends serve to anchor the index in differentiating good vs. bad environmental condition. All four indices had high correlation with the condition gradients, and some indices were less monotonic than the others. However, since indices are intended to reproduce the experience of experts in interpreting benthic data using an objective, repeatable, transparent tool, a better evaluation benchmark is whether an index ranks and classifies sites with levels of correlation and accuracy comparable to that among experts. In this study, the three best performing assessment techniques achieved this objective. One of them requires the calculation of only one benthic index, which raises the prospect of implementation of widespread economical and effective benthic monitoring in Puget Sound going forward into the future.

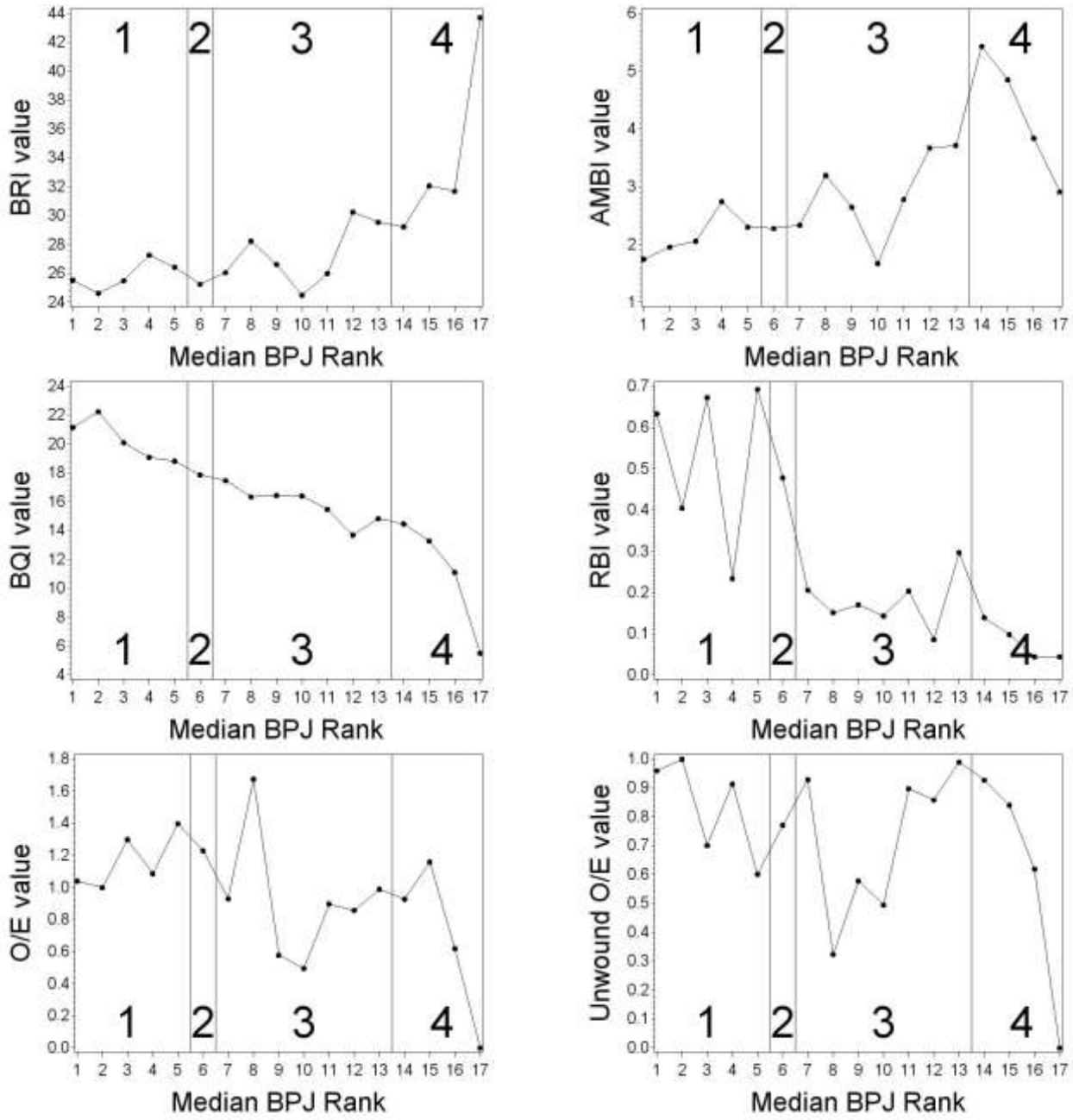

Figure 6-1. Index values along the validation gradient. Vertical lines indicate median categories assigned by experts: 1: Undisturbed, 2: Slight Disturbance, 3: Moderate Disturbance, 4: Severe Disturbance.

7. CONCLUSIONS AND RECOMMENDATIONS

7.1 Conclusions

We reached the following five conclusions, based on the procedures and results documented in this report.

- The benthic and habitat data that are available through the Puget Sound Ecosystem Monitoring Program are of high quality. Sufficient data were available to support identification of the habitat-related benthic assemblage of Puget Sound, segregation of independent benthic index calibration and validation data sets, calibration of Puget Sound benthic indices, and evaluation of their performance.
- The benthic macrofauna of Puget Sound belong to a single habitat-related assemblage. Although five groupings related to sediment grain size and bottom depth were identified, they were considered sub-assemblages of the single Sound-wide assemblage. The main reason was that few abundant species occurred exclusively in single groupings; instead, many abundant species occurred in large percentages of the samples of many of the groupings. The five sub-assemblages were relatively evenly distributed throughout Puget Sound.
- We successfully calibrated five benthic indices for Puget Sound. The indices were not driven by habitat variables and were able to discern sites along a gradient of benthic condition. The indices were generally well correlated with each other. The indices were the Benthic Response Index (BRI), AZTI Marine Biotic Index (AMBI), Relative Benthic Index (RBI), Benthic Quality Index (BQI) and a RIVPACS model-based O/E (observed over expected) index.
- Of the 40 candidate validation samples, the experts agreed on condition category and condition ranking for 17 samples. The samples covered the entire condition gradient identified by the benthic experts and were used (1) for index validation, (2) to determine assessment thresholds for the benthic indices, and (3) to optimize the routine ambient benthic monitoring process.
- Four of the five Puget Sound calibrated benthic indices performed well in evaluating the rank order of the validation samples and it is likely that the indices can be successfully used for benthic assessments. The four well-performing indices were the BRI, AMBI, RBI, and the BQI.
- The BQI assessed benthic condition categories and status more accurately than (1) the average benthic expert and (2) all other tested benthic indices and benthic index combinations. The assessment accuracy of the BQI and the relative ease of benthic assessments requiring the calculation of only one benthic index make it a strong candidate for routine benthic assessments in Puget Sound.
- Two three benthic index combinations performed almost as well as the BQI. One of the combinations (AMBI, BQI, and RBI) demonstrated bias a third of the magnitude of the other, and can be considered a second choice to the BQI, that could also be used if it becomes necessary for some presently unknown reason.

7.2 Recommendations

Our study shows that Puget Sound benthic assessments based on the BQI are more accurate than the average benthic expert and all other benthic indices that were tested. We recommend moving forward with a two-part strategy.

- The first part should build on the results of this study, preparing guidance and documentation to support routine benthic monitoring in Puget Sound.
- The second part of the strategy should confirm the accuracy of the results of this study, with two objectives:
    - Expanding the confidence base above and beyond the 17 validation samples that are the basis for the present study, based on samples from known poor or pristine condition.
    - Exploring the second and third choice assessment methods identified in the present study, which also have high potential for success, in order to confirm the accuracy of the methodology with a view to using them if it is necessary, desirable, or expedient.

8. ACKNOWLEDGMENTS

The study to develop and evaluate indicators of Puget Sound benthic condition was funded by a Section 106 Grant from the U.S. Environmental Protection Agency (USEPA) to the Washington State Department of Ecology. The Puget Soundkeeper Alliance funded a preliminary study that laid the groundwork for the development of benthic indicators by identifying the habitat-related benthic assemblage of Puget Sound. The results of both studies are presented in this report.

The sediment and benthic invertebrate data analyzed for this study were collected from 1989 to 2008 by the Washington State Department of Ecology for the Puget Sound Ecosystem Monitoring Program, including projects in partnership with the USEPA and the National Oceanic and Atmospheric Administration. We acknowledge the contributions of these projects and the efforts and expertise of the field and laboratory personnel who generated those data.

We also thank Margaret Dutch, Lawrence L. Lovell, David E. Montagne, Peter Striplin, Ronald G. Velarde and Kathy I. Welch, who were the benthic experts who participated in the best professional judgment study described in Section 5. Danielle Burnett-Cantrell created Figure 3-2 and provided assistance in data organization and analysis, and Karlene Miller provided editorial assistance. The photograph on the cover, of Commencement Bay and the Port of Tacoma with Mt. Rainier in the background, was taken by Tom Putnam. We are also grateful to staff of the Washington State Department of Ecology Marine Monitoring Unit and the USEPA Western Ecology Division, who reviewed this report.

9. LITERATURE CITED

Bald, J., Á. Borja, I. Muxika, J. Franco and V. Valencia. 2005. Assessing reference conditions and physico-chemical status according to the European Water Framework Directive: A case-study from the Basque Country (Northern Spain). *Marine Pollution Bulletin* **50**:1508-1522.

Bay, S.M., D.J. Greenstein, J.A. Ranasinghe, D.W. Diehl and A.E. Fetscher. 2009. Sediment quality assessment draft technical support manual. Technical Report 582. Southern California Coastal Water Research Project. Costa Mesa, CA.

Bergen, M., D.B. Cadien, A. Dalkey, D.E. Montagne, R.W. Smith, J.K. Stull, R.G. Velarde and S.B. Weisberg. 2000. Assessment of benthic infaunal condition on the mainland shelf of Southern California. *Environmental Monitoring and Assessment* **64**:421-434.

Bergen, M., S.B. Weisberg, R.W. Smith, D.B. Cadien, A. Dalkey, D.E. Montagne, J.K. Stull, R.G. Velarde and J.A. Ranasinghe. 2001. Relationship between depth, sediment, latitude, and the structure of benthic infaunal assemblages on the mainland shelf of southern California. *Marine Biology* **138**:637-647.

Bilkovic, D.M., M. Roggero, C.H. Hershner and K.H. Havens. 2006. Influence of land use on macrobenthic communities in nearshore estuarine habitats. *Estuaries and Coasts* **29**:1185-1195.

Bilyard, G.R. 1987. The value of benthic infauna in marine pollution monitoring studies. *Marine Pollution Bulletin* **18**:581-585.

Boesch, D.F. 1973. Classification and community structure of macrobenthos in the Hampton Roads area, Virginia. *Marine Biology* **21**:226-244.

Boesch, D.F. 1977. A new look at the zonation of benthos along the estuarine gradient. pp:245-266 *in:* B.C. Coull *(eds)*. Ecology of marine benthos. University of South Carolina Press. Columbia, SC.

Borja, Á. and D.M. Dauer. 2008. Assessing the environmental quality status in estuarine and coastal systems: Comparing methodologies and indices. *Ecological Indicators* **8**:331-337.

Borja, Á. and I. Muxika. 2005. Guidelines for the use of AMBI (AZTI's Marine Biotic Index) in the assessment of the benthic ecological quality. *Marine Pollution Bulletin* **50**:787-789.

Borja, Á., J. Franco and V. Perez. 2000. A marine biotic index to establish the ecological quality of soft-bottom benthos within european estuarine and coastal environments. *Marine Pollution Bulletin* **40**:1100-1114.

Borja, Á., I. Muxika and J. Franco. 2003. The application of a marine biotic index to different impact sources affecting soft-bottom benthic communities along european coasts. *Marine Pollution Bulletin* **46**:835-845.

Borja, Á., J.A. Ranasinghe and S.B. Weisberg. 2009. Assessing ecological integrity in marine waters using multiple indices and ecosystem components: Challenges for the future. *Marine Pollution Bulletin* **59**:1-4.

Bradfield, G.E. and N.C. Kenkel. 1987. Nonlinear ordination using shortest path adjustment of ecological distances. *Ecology* **68**:750-753.

Bray, J.R. and J.T. Curtis. 1957. An ordination of the upland forest communities of southern Wisconsin. *Ecological Monographs* **27**:325-349.

Burns, T., M. Fiander and B. Audini. 2000. A Delphi approach to characterising 'relapse' as used in UK clinical practice. *International Journal of Social Psychiatry* **46**:220-230.

Cicchetti, D.V. and T. Allison. 1971. A new procedure for assessing reliability of scoring EEG sleep recordings. *American Journal of EEG Technology* **11**:101–109.

Clifford, H.T. and W. Stephenson. 1975. An Introduction to Numerical Classification. Academic Press. New York, NY.

Dauer, D.M. 1993. Biological criteria, environmental health and estuarine macrobenthic community structure. *Marine Pollution Bulletin* **26**:249-257.

Dauer, D.M., J.A. Ranasinghe and S.B. Weisberg. 2000. Relationships between benthic community condition, water quality, sediment quality, nutrient loads, and land use patterns in Chesapeake Bay. *Estuaries* **23**:80-96.

Diaz, R.J., M. Solan and R.M. Valente. 2004. A review of approaches for classifying benthic habitats and evaluating habitat quality. *Journal of Environmental Management* **73**:165-181.

Dutch, M., V. Partridge, S. Weakland, K.I. Welch and E.R. Long. 2009. Quality Assurance Project Plan: The Puget Sound Assessment and Monitoring Program Sediment Monitoring Component. Washington State Department of Ecology. Olympia, WA. https://fortress.wa.gov/ecy/publications/SummaryPages/0903121.html.

Elwyn, G., A. O'Connor, D. Stacey, R. Volk, A. Edwards, A. Coulter, R. Thomson, A. Barratt, M. Barry, S. Bernstein, P. Butow, A. Clarke, V. Vikki Entwistle, D. Feldman-Stewart, M. Holmes-Rovner, H. Llewellyn-Thomas, N. Moumjid, A. Mulley, C. Ruland, K. Sepucha, A. Sykes, T. Whelan and on behalf of The International Patient Decision Aids Standards (IPDAS) Collaboration. 2006. Developing a quality criteria framework for patient decision aids: online international Delphi consensus process. *BMJ* **333**:417 1-6.

Hale, S.S., J.F. Paul and J.F. Heltshe. 2004. Watershed landscape indicators of estuarine benthic condition. *Estuaries* **27**:283-295.

Hsu, C.-C. and B.A. Sandford. 2007. The Delphi Technique: Making sense of consensus. *Practical Assessment, Research & Evaluation* **12**:1-8.

Hughes, R.M., D.P. Larsen and J.M. Omernik. 1986. Regional reference sites: a method for assessing stream potentials. *Environmental Management* **10**:629-635.

Hunt, J.W., B.S. Anderson, B.M. Phillips, R.S. Tjeerdema, K.M. Taberski, C.J. Wilson, H.M. Puckett, M. Stephenson, W.R. Fairey and J.M. Oakden. 2001. A large-scale categorization of sites in San Francisco Bay, USA, based on the sediment quality triad, toxicity identification evaluations, and gradient studies. *Environmental Toxicology and Chemistry* **20**:1252-1265.

Hyland, J.L., R.F. Van Dolah and T.R. Snoots. 1999. Predicting stress in benthic communities of southeastern U.S. estuaries in relation to chemical contamination of sediments. *Environmental Toxicology and Chemistry* **18**:2557-2564.

Hyland, J.L., W.L. Balthis, V.D. Engle, E.R. Long, J.F. Paul, J.K. Summers and R.F. Van Dolah. 2003. Incidence of stress in benthic communities along the US Atlantic and Gulf of Mexico coasts within different ranges of sediment contamination from chemical mixtures. *Environmental Monitoring and Assessment* **81**:149-161.

Hyland, J.L., W.L. Balthis, M.H. Posey, C.T. Hackney and T.D. Alphin. 2004. The soft-bottom macrobenthos of North Carolina estuaries. *Estuaries* **27**:501-514.

Lance, G.H. and W.T. Williams. 1967. A general theory of classificatory sorting strategies. I. Hierarchical systems. *Computer Journal* **9**:373-380.

Legendre, P. and L. Legendre. 1998. Numerical Ecology. Elsevier. Amsterdam, Netherlands.

Llansó, R.J., S. Aasen and K.I. Welch. 1998. Marine Sediment Monitoring Program - II. Distribution and Structure of Benthic Communities in Puget Sound, 1989-1993. Publication No. 98-328. Washington State Department of Ecology. Olympia, WA. https://fortress.wa.gov/ecy/publications/SummaryPages/98328.html.

Llansó, R.J., L.C. Scott, D.M. Dauer, J.L. Hyland and D.E. Russell. 2002. An estuarine benthic index of biotic integrity for the Mid-Atlantic region of the United States. I. Classification of assemblages and habitat definition. *Estuaries* **25**:1219-1230.

Long, E.R., M. Dutch, S. Aasen, K.I. Welch and M.J. Hameedi. 2003. Chemical Contamination, Acute Toxicity in Laboratory Tests, and Benthic Impacts in Sediments of Puget Sound: A summary of results of the joint 1997-1999 Ecology/NOAA survey. Publication No. 03-03-049. Washington State Department of Ecology. Olympia WA. https://fortress.wa.gov/ecy/publications/SummaryPages/0303049.html.

Marques, J.C., F. Salas, J. Patricio, H. Teixeira and J.M. Neto. 2009. Ecological Indicators for Coastal and Estuarine Environmental Assessment - A user guide. WIT Press. Southampton, England.

MER Consulting. 2000. Peer Review of Ecology's Proposed Benthic Assessment Methods and Endpoints for use in Regulatory Decisions. Seattle, WA.

Muxika, I., Á. Borja and W. Bonne. 2005. The suitability of the marine biotic index (AMBI) to new impact sources along european coasts. *Ecological Indicators* **5**:19-31.

Orloci, L. 1975. Multivariate Analysis in Vegetation Research. Dr. W Junk, Publishers. The Hague, Netherlands.

Paul, J.F., K.J. Scott, D.E. Campbell, J.H. Gentile, C.S. Strobel, R.M. Valente, S.B. Weisberg, A.F. Holland and J.A. Ranasinghe. 2001. Developing and applying a benthic index of estuarine condition for the Virginian Biogeographic Province. *Ecological Indicators* **1**:83-99.

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

Pinto, R.I.C., J. Patricio, A. Baeta, B.D. Fath, J.M. Neto and J.C. Marques. 2009. Review and evaluation of estuarine biotic indices to assess benthic condition. *Ecological Indicators* **9**:1-25.

Ranasinghe, J.A., S.B. Weisberg, R.W. Smith, D.E. Montagne, B. Thompson, J.M. Oakden, D.D. Huff, D.B. Cadien, R.G. Velarde and K.J. Ritter. 2009. Calibration and evaluation of five indicators of benthic community condition in two California bay and estuary habitats. *Marine Pollution Bulletin* **59**:5-13.

Ranasinghe, J.A., K.I. Welch, P.N. Slattery, D.E. Montagne, D.D. Huff, H. Lee, II, J.L. Hyland, B. Thompson, S.B. Weisberg, J.M. Oakden, D.B. Cadien and R.G. Velarde. 2012a. Habitat-related benthic macrofaunal assemblages of bays and estuaries of the western United States. *Integrated Environmental Assessment and Management* **8**:638-648.

Ranasinghe, J.A., E.D. Stein, P.E. Miller and S.B. Weisberg. 2012b. Performance of two Southern California benthic community condition indices using species abundance and presence-only data: Relevance to DNA barcoding. *PLoS ONE* **7(8)**:e40875.

Rosenberg, R., M. Blomqvist, H.C. Nilsson, H. Cederwall and A. Dimming. 2004. Marine quality assessment by use of benthic species-abundance distributions: a proposed new protocol within the European Union Water Framework Directive. *Marine Pollution Bulletin* **49**:728-739.

Smith, R.W. 1976. Numerical Analysis of Ecological Survey Data. Ph.D Thesis. University of Southern California. Los Angeles, CA.

Smith, R.W., B.B. Bernstein and R.L. Cimberg. 1988. Community-environmental relationships in the benthos: Applications of multivariate analytical techniques. pp:247-326 *in:* D.F. Soule and G.S. Kleppel *(eds)*. Marine Organisms as Indicators. Springer-Verlag. New York, NY.

Smith, R.W., M. Bergen, S.B. Weisberg, D.B. Cadien, A. Dalkey, D.E. Montagne, J.K. Stull and R.G. Velarde. 2001. Benthic response index for assessing infaunal communities on the southern California mainland shelf. *Ecological Applications* **11**:1073-1087.

Smith, R.W., J.A. Ranasinghe, S.B. Weisberg, D.E. Montagne, D.B. Cadien, T.K. Mikel, R.G. Velarde and A. Dalkey. 2003. Extending the southern California Benthic Response Index to assess benthic condition in bays Technical Report 410. Southern California Coastal Water Research Project. Westminster, CA.

Striplin Environmental Associates, Inc. 1996. Development of Reference Value Ranges for Benthic Infauna Assessment Endpoints in Puget Sound. Washington Department of Ecology. Olympia, WA.

Striplin Environmental Associates, Inc. 2003. SEDQUAL Analytical Tool Development Support for the Analysis and Interpretation of Benthic Community Data: Status and Recommendations. Publication Number 03-09-090. Washington State Department of Ecology. Olympia, WA.

Striplin Environmental Associates, Inc. and Roy F. Weston, Inc. 1999. Puget Sound Reference Value Project, Task 3: Development of Benthic Effects Sediment Quality Standards. Washington State Department of Ecology. Olympia, WA.

Summers, J.K. 2001. Ecological condition of the estuaries of the atlantic and gulf coasts of the United States. *Environmental Toxicology and Chemistry* **20**:99-106.

Teixeira, H., Á. Borja, S.B. Weisberg, J.A. Ranasinghe, D.B. Cadien, D.M. Dauer, J.-C. Dauvin, S. Degraer, R.J. Diaz, A. Grémare, I. Karakassis, R.J. Llansó, L.L. Lovell, J.C. Marques, D.E. Montagne, A. Occhipinti-Ambrogim, R. Rosenberg, R. Sardá, L.C. Schaffner and R.G. Velarde. 2010. Assessing coastal benthic macrofauna community condition using best professional judgement – Developing consensus across North America and Europe. *Marine Pollution Bulletin* **60**:589-600.

Teixeira, H., S.B. Weisberg, Á. Borja, J.A. Ranasinghe, D.B. Cadien, R.G. Velarde, L.L. Lovell, D. Pasko, C.A. Phillips, D.E. Montagne, K.J. Ritter, F. Salas and J.C. Marques. 2012. Calibration and validation of the AZTI's Marine Biotic Index (AMBI) for Southern California marine bays. *Ecological Indicators* **12**:84-95.

Thompson, B., S.B. Weisberg, A.R. Melwani, S. Lowe, J.A. Ranasinghe, D.B. Cadien, D.M. Dauer, R.J. Diaz, W. Fields, M. Kellogg, D.E. Montagne, P.R. Ode, D.J. Reish and P.N. Slattery. 2012. Low levels of agreement among experts using best professional judgment to assess benthic condition in the San Francisco estuary and delta. *Ecological Indicators* **12**:167-173.

Thompson, B., J.A. Ranasinghe, S. Lowe, A.R. Melwani and S.B. Weisberg. 2013. Benthic macrofaunal assemblages of the San Francisco Estuary and Delta. *Environmental Monitoring and Assessment* **185**:2281-2295.

US Environmental Protection Agency (USEPA). 2004. National Coastal Condition Report II. EPA-620/R-03/002. USEPA Office of Research and Development. Washington, DC.

US Environmental Protection Agency (USEPA). 2007. National Estuary Program Coastal Condition Report. USEPA Office of Research and Development. Washington, DC.

Van Dolah, R.F., J.L. Hyland, A.F. Holland, J.S. Rosen and T.R. Snoots. 1999. A benthic index of biological integrity for assessing habitat quality in estuaries of the southeastern USA. *Marine Environmental Research* **48**:269-283.

Van Sickle, J. 2008. An index of compositional dissimilarity between observed and expected assemblages. *Journal of the North American Benthological Society* **27**:227-235.

Van Sickle, J., C.P. Hawkins, D.P. Larsen and A.T. Herlihy. 2005. A null model for the expected macroinvertebrate assemblage in streams. *Journal of the North American Benthological Society* **24**:178-191.

Van Sickle, J., D.D. Huff and C.P. Hawkins. 2006. Selecting discriminant function models for predicting the expected richness of aquatic macroinvertebrates. *Freshwater Biology* **51**:359-372.

Van Sickle, J., D.P. Larsen and C.P. Hawkins. 2007. Exclusion of rare taxa affects performance of the O/E index in bioassessments. *Journal of the North American Benthological Society* **26**:319-331.

Weisberg, S.B., J.A. Ranasinghe, L.C. Schaffner, R.J. Diaz, D.M. Dauer and J.B. Frithsen. 1997. An estuarine benthic index of biotic integrity (B-IBI) for Chesapeake Bay. *Estuaries* **20**:149-158.

Weisberg, S.B., B. Thompson, J.A. Ranasinghe, D.E. Montagne, D.B. Cadien, D.M. Dauer, D.R. Diener, J.S. Oliver, D.J. Reish, R.G. Velarde and J.Q. Word. 2008. The level of agreement among experts applying best professional judgment to assess the condition of benthic infaunal communities. *Ecological Indicators* **8**:389-394.

Whittaker, R.H. 1978. Classification of Plant Communities. Dr. W. Junk, Publishers. The Hague, Netherlands.

Williamson, M.H. 1978. The ordination of incidence data. *Journal of Ecology* **66**:911-920.

Wright, A. and T. Shannon. 2006. Giving "teeth" to an environmental policy: a Delphi Study at Dalhousie University. *Journal of Cleaner Production* **14**:3-5.

Wright, J.F., M.T. Furse and P.D. Armitage. 1993. RIVPACS: a technique for evaluating the biological water quality of rivers in the UK. *European Water Pollution Control* **3**:15-25.