NOAA Technical Memorandum NOS OR&R 12

Sampling Effort in Assessments of Oil-Spill Impacts to Intertidal Organisms

Douglas A. Coats
Eiji Imamura
Marine Research Specialists
3140 Telegraph Rd., Suite A
Ventura, CA 93003

Allan K. Fukuyama
F/H Taxonomic Services
7019 157th St. SW
Edmonds, WA 98026

John R. Skalski
School of Fisheries
University of Washington
1325 Fourth Ave., Suite 1820
Seattle, WA 98101

Scott Kimura
John Steinbeck
Tenera
P.O. Box 400
Avila Beach, CA 93424

Edited By:

Gary Shigenaka
Rebecca Hoff
NOAA Hazardous Materials Response Division
Office of Response and Restoration
7600 Sand Point Way NE
Seattle, WA 98115

Seattle, Washington

United States Department of Commerce Donald L. Evans Secretary	National Oceanic and Atmospheric Administration VADM Conrad C. Lautenbacher, Jr. USN (Ret.) Under Secretary for Ocean and Atmosphere	National Ocean Service Richard W. Spinrad Assistant Administrator for Oceans Services and Coastal Zone Management

TABLE OF CONTENTS

List of Figures ..... *iv*

List of Tables ..... *ix*

EXECUTIVE SUMMARY **ES-1**

Sampling Decisions **ES-2**

Spatial Variability **ES-3**

Sample Sizes **ES-5**

Maximizing Sampling Resources **ES-8**

CHAPTER 1. INTRODUCTION **1-1**

Monitoring Goals **1-2**

Statistical Considerations **1-3**

The Intertidal Database **1-5**

Applicability Outside of Prince William Sound **1-7**

CHAPTER 2. TREATMENT EFFECTS **2-1**

Species Response **2-3**

Variance Computation *2-3*
Population Changes *2-5*
Estimates of Biological Variability *2-9*
Power Analysis *2-15*
Example Application *2-19*

Community Response **2-21**

Measuring Differences in Community Composition *2-23*
Power Analysis *2-25*

CHAPTER 3. RECOVERY **3-1**

Impact Size **3-1**

Power Analyses **3-2**

Habitat Disturbance *3-2*
Hydrocarbon Exposure *3-7*

TABLE OF CONTENTS
(continued)

CHAPTER 4. CHRONIC EFFECTS 4-1

Chronic Effect Size and Duration 4-2

Power Analyses 4-2

Protothaca Application 4-4

CHAPTER 5. CONCLUSIONS 5-1

Sample Sizes 5-2

Spatial Variability 5-4

Recommendations 5-4

Sample Before a Spill Impacts Intertidal Sites 5-4
Relax Taxonomic Discrimination 5-5
Randomize Quadrat Locations 5-6
Limit the Focus 5-6

CHAPTER 6. LITERATURE CITED 6-1

APPENDIX A. GLOSSARY OF SELECTED STATISTICAL TERMS AND ACRONYMS A-1

APPENDIX B. POWER FORMULATION FOR TREATMENT EFFECTS B-1

Species Response B-2

Test Statistic *B-2*
Power Formulation *B-3*
Variance Estimation *B-5*
Coefficient of Variation *B-7*

Community Response B-8

Test Statistic *B-8*
Power Formulation *B-10*

APPENDIX C. VARIANCE DISTRIBUTIONS C-1

TABLE OF CONTENTS
(continued)

APPENDIX D. POWER CURVES FOR TREATMENT EFFECTS D-1

APPENDIX E. POWER FORMULATION FOR TESTING RECOVERY E-1

Test Statistic E-1

Power Formulation E-2

APPENDIX F. POWER FORMULATION FOR TESTING LONG-TERM STABILITY F-1

Test Statistic F-1

Power Formulation F-2

APPENDIX G. POWER CURVES FOR DETECTING LONG-TERM TRENDS G-1

LIST OF FIGURES

Figure ES.1. Flow chart showing decisions affecting sampling design in an intertidal monitoring program following an oil spill ES-3

Figure ES.2. Sample-size chart showing the number of replicate samples (m) to be collected at n reference and n impact sites to detect a 50% reduction in abundant intertidal populations with a moderate level of natural biological variation. ES-6

Figure ES.3. Number of control and impact sites needed to detect linear departures from parallelism during recolonization at PWS intertidal sites subjected to severe habitat disturbance..... ES-7

Figure 2.1. Variability as a function of population size for PWS intertidal taxa..... 2-10

Figure 2.2. Clumping as a function of population size for PWS intertidal taxa. 2-13

Figure 2.3. Sample-size chart showing the number of replicate samples (m) collected at n sites per treatment that are needed to detect a 50% reduction (100% increase) in abundant intertidal populations with a statistical power $(1-\beta)$ at the one-tailed significance level of $\alpha = 0.1$. The solid curves correspond to different levels of statistical power in an environment with moderate natural biological variation ($CV_W = 1.28$, $CV_B = 0.84$). The dashed curves correspond to the total number of samples to be collected when comparing two-treatments. 2-17

Figure 2.4. Principal infaunal components at Category-1 (reference) sites and Category-3 (oiled and washed) sites from PWS samples collected in (a) 1991 and (b) 1994. 2-24

Figure 2.5. Sample-size chart showing the number of sites (n) per treatment that are needed to detect a treatment effect that causes a separation of C in the treatment means in a (a) one-dimensional, (b) two-dimensional, or (c) three-dimensional ordination..... 2-26

Figure 3.1. Comparison of mean populations at reference and washed sites for six PWS intertidal assemblages: (a) Motile invertebrates, (b) *Fucus*, (c) Total infauna, (d) Annelids, (e) Mollusks, and (f) Crustaceans 3-3

Figure 3.2. Sample sizes needed to detect linear departures from parallelism during the abrupt repopulation event at PWS intertidal sites subjected to hot-water washing..... 3-4

Figure 3.3. Comparison of mean populations at reference and oiled sites for six PWS intertidal assemblages: (a) Motile invertebrates, (b) *Fucus*, (c) Total infauna, (d) Annelids, (e) Mollusks, and (f) Crustaceans 3-8

Figure 3.4. Sample sizes needed to detect linear departures from parallelism during the abrupt repopulation event at PWS intertidal sites subjected to oiling but not invasive cleanup techniques. 3-9

LIST OF FIGURES

(continued)

Figure 4.1. Long-term trend in the difference between average littleneck clam (*P. staminea*) populations at Category-3 washed sites and Category-1 reference sites in PWS..... 4-4

Figure 4.2. Sample-size chart showing the number of sites (n) and the number of replicate samples (m) needed to detect an impact that causes the 11.5% annual increase in intertidal populations over a (a) 5-year and (b) 9-year period with a statistical power ($1-\beta$) at the one-tailed significance level of $\alpha = 0.1$. The curves correspond to different levels of statistical power in an environment with the low natural biological variation associated with littleneck clams (*P. staminea*) ($CV_W = 1.12$, $CV_B = 0.65$). 4-6

Figure 4.3. Power to detect the observed 11.5% increase in littleneck clam populations with 5 samples collected at each of 3 sites as a function of the duration of the monitoring program..... 4-7

Figure C.1. Distribution of Algal Cover (%) and Estimated CVs among Taxa within the Upper Intertidal Zone C-1

Figure C.2. Distribution of Algal Cover (%) and Estimated CVs among Taxa within the Middle Intertidal Zone..... C-2

Figure C.3. Distribution of Invertebrate Cover (%) and Estimated CVs among Taxa within the Upper Intertidal Zone C-3

Figure C.4. Distribution of Invertebrate Cover (%) and Estimated CVs among Taxa within the Middle Intertidal Zone..... C-3

Figure C.5. Distribution of Invertebrate Counts and Estimated CVs among Taxa within the Upper Intertidal Zone C-4

Figure C.6. Distribution of Invertebrate Counts and Estimated CVs among Taxa within the Middle Intertidal Zone..... C-5

Figure C.7. Distribution of Infaunal Counts and Estimated CVs among Taxa within the Lower Intertidal Zone..... C-6

Figure D.1. Sample-size chart showing the number of replicate samples (m) collected at n reference and n treatment sites that are needed to detect a (a) 15% reduction (18% increase) or (b) 25% reduction (33% increase) in sparse intertidal populations with a statistical power ($1-\beta$) at the one-tailed significance level of $\alpha = 0.1$. The curves correspond to different levels of statistical power in an environment with low natural biological variation ($CV_W = 0.49$, $CV_B = 0.00$)..... D-2

Figure D.2. Sample-size chart showing the number of replicate samples (m) collected at n reference and n treatment sites that are needed to detect a (a) 33% reduction (50% increase) or (b) 50% reduction (100% increase) in sparse intertidal populations with a statistical power ($1-\beta$) at the one-tailed significance level of $\alpha = 0.1$. The curves correspond to different levels of statistical power in an environment with moderate natural biological variation ($CV_W = 1.86$, $CV_B = 0.27$) D-3

LIST OF FIGURES
(continued)

Figure D.3. Sample-size chart showing the number of replicate samples (m) collected at n reference and n treatment sites that are needed to detect a (a) 67% reduction (200% increase) or (b) 75% reduction (300% increase) in sparse intertidal populations with a statistical power ($1-\beta$) at the one-tailed significance level of $\alpha = 0.1$. The curves correspond to different levels of statistical power in an environment with high natural biological variation ($CV_W = 3.88$, $CV_B = 0.76$)..... D-4

Figure D.4. Sample-size chart showing the number of replicate samples (m) collected at n reference and n treatment sites that are needed to detect a (a) 25% reduction (33% increase) or (b) 33% reduction (50% increase) in moderately dense intertidal populations with a statistical power ($1-\beta$) at the one-tailed significance level of $\alpha = 0.1$. The curves correspond to different levels of statistical power in an environment with low natural biological variation ($CV_W = 1.03$, $CV_B = 0.19$) D-5

Figure D.5. Sample-size chart showing the number of replicate samples (m) collected at n reference and n treatment sites that are needed to detect a (a) 50% reduction (100% increase) or (b) 67% reduction (200% increase) in moderately dense intertidal populations with a statistical power ($1-\beta$) at the one-tailed significance level of $\alpha = 0.1$. The curves correspond to different levels of statistical power in an environment with moderate natural biological variation ($CV_W = 2.52$, $CV_B = 0.91$) D-6

Figure D.6. Sample-size chart showing the number of replicate samples (m) collected at n reference and n treatment sites that are needed to detect a (a) 75% reduction (300% increase) or (b) 83% reduction (500% increase) in moderately dense intertidal populations with a statistical power ($1-\beta$) at the one-tailed significance level of $\alpha = 0.1$. The curves correspond to different levels of statistical power in an environment with high natural biological variation ($CV_W = 4.44$, $CV_B = 1.73$) D-7

Figure D.7. Sample-size chart showing the number of replicate samples (m) collected at n reference and n treatment sites that are needed to detect a (a) 25% reduction (33% increase) or (b) 50% reduction (100% increase) in abundant intertidal populations with a statistical power ($1-\beta$) at the one-tailed significance level of $\alpha = 0.1$. The curves correspond to different levels of statistical power in an environment with low natural biological variation ($CV_W = 0.76$, $CV_B = 0.32$)..... D-8

Figure D.8. Sample-size chart showing the number of replicate samples (m) collected at n reference and n treatment sites that are needed to detect a (a) 50% reduction (100% increase) or (b) 67% reduction (200% increase) in abundant intertidal populations with a statistical power ($1-\beta$) at the one-tailed significance level of $\alpha = 0.1$. The curves correspond to different levels of statistical power in an environment with moderate natural biological variation ($CV_W = 1.28$, $CV_B = 0.84$) D-9

LIST OF FIGURES

(continued)

Figure D.9. Sample-size chart showing the number of replicate samples (m) collected at n reference and n treatment sites that are needed to detect a (a) 67% reduction (200% increase) or (b) 75% reduction (300% increase) in abundant intertidal populations with a statistical power $(1-\beta)$ at the one-tailed significance level of $\alpha = 0.1$. The curves correspond to different levels of statistical power in an environment with high natural biological variation ($CV_W = 2.35$, $CV_B = 1.57$) D-10

Figure G.1. Sample-size chart showing the number of sites (n) and the number of replicate samples (m) needed to detect an impact that causes a (a) 10% or (b) 15% annual increase in intertidal populations over a 5-year period with a statistical power $(1-\beta)$ at the one-tailed significance level of $\alpha = 0.1$. The curves correspond to different levels of statistical power in an environment with low natural biological variation ($CV_W = 0.76$, $CV_B = 0.32$) G-2

Figure G.2. Sample-size chart showing the number of sites (n) and the number of replicate samples (m) needed to detect an impact that causes a (a) 15% or (b) 20% annual increase in intertidal populations over a 5-year period with a statistical power $(1-\beta)$ at the one-tailed significance level of $\alpha = 0.1$. The curves correspond to different levels of statistical power in an environment with moderate natural biological variation ($CV_W = 1.28$, $CV_B = 0.84$) G-3

Figure G.3. Sample-size chart showing the number of sites (n) and the number of replicate samples (m) needed to detect an impact that causes a (a) 25% or (b) 30% annual increase in intertidal populations over a 5-year period with a statistical power $(1-\beta)$ at the one-tailed significance level of $\alpha = 0.1$. The curves correspond to different levels of statistical power in an environment with high natural biological variation ($CV_W = 2.35$, $CV_B = 1.57$) G-4

Figure G.4. Sample-size chart showing the number of sites (n) and the number of replicate samples (m) needed to detect an impact that causes a (a) 3% or (b) 5% annual increase in intertidal populations over a 10-year period with a statistical power $(1-\beta)$ at the one-tailed significance level of $\alpha = 0.1$. The curves correspond to different levels of statistical power in an environment with low natural biological variation ($CV_W = 0.76$, $CV_B = 0.32$) G-5

Figure G.5. Sample-size chart showing the number of sites (n) and the number of replicate samples (m) needed to detect an impact that causes a (a) 5% or (b) 10% annual increase in intertidal populations over a 10-year period with a statistical power $(1-\beta)$ at the one-tailed significance level of $\alpha = 0.1$. The curves correspond to different levels of statistical power in an environment with moderate natural biological variation ($CV_W = 1.28$, $CV_B = 0.84$) G-6

LIST OF FIGURES
(continued)

Figure G.6. Sample-size chart showing the number of sites (n) and the number of replicate samples (m) needed to detect an impact that causes a (a) 15% or (b) 20% annual increase in intertidal populations over a 10-year period with a statistical power ($1-\beta$) at the one-tailed significance level of $\alpha = 0.1$. The curves correspond to different levels of statistical power in an environment with high natural biological variation ($CV_W = 2.35$, $CV_B = 1.57$) G-7

LIST OF TABLES

Table 1.1. Three parameters that define the quantitative goals of spill assessments 1-3

Table 2.1. Years over which variance estimates were pooled 2-5

Table 2.2. Taxa whose populations were higher during the impact period and declined during the subsequent recovery 2-7

Table 2.3. Summary of CVs by abundance and period 2-11

Table 2.4. CV distribution by assemblage and tidal elevation 2-12

Table 2.5. Species with the anomalously high within-site variability 2-12

Table 2.6. Summary of CVs used in the sample-size calculations 2-16

Table 3.1. Amplitude of the departure of from parallel linear trends and variability about the mean trends 3-6

Table 4.1. Total population change as a function of various annual population increases and study durations presented in this Chapter and in Appendix G 4-3

Table 5.1. Representative sample-size recommendations 5-1

Table B.1. Single classification ANOVA used to estimate variance components B-5

Table B.2. One-way classification of a bivariate ordination using ANODIS B-9

Table E.1. Orthogonal polynomial coefficients for linear trends in annual sampling over periods of two to ten years E-3

Table F.1. Value of the sum of squares term for deviations about the mean year for monitoring programs lasting from two to fifteen years F-3

Executive Summary

"*How many samples do we need?*" and "*Where should we collect them?*" are two basic questions common to all field monitoring programs. Answering these questions becomes more compelling when an accidental oil spill impinges on a coastline and a biological monitoring program must be rapidly implemented to assess initial impacts. This report answers these questions for three intertidal monitoring designs that assess impacts to intertidal populations caused by localized disturbances. Following an oil spill, disturbances can be caused by hydrocarbon exposure or can result from oil-spill cleanup efforts. Also included are sample-size recommendations for long-term monitoring programs designed to assess recovery and lingering chronic effects from an oil spill or other shoreline disturbance.

All other things being equal, sampling effort is largely determined by the inherent variability within the ecosystem. Intuitively, the number of samples needed to reliably discern a given population impact is larger for taxa with highly variable distributions than for taxa with naturally uniform distributions. Namely, it is easier to see slight impact-related changes in a field of nearly uniform measurements than in measurements having a wide natural variability. Because of this interdependence, accurate sample-size determinations are predicated on representative measurements of the inherent spatial and temporal variability in the intertidal populations of interest.

To that end, much of the analysis in this report was focused on obtaining an accurate determination of the inherent population variability within diverse intertidal taxonomic groups. These variability estimates were determined from eleven years of intertidal data collected within Prince William Sound (PWS), Alaska as part of the long-term monitoring of biological recovery conducted by NOAA following the *Exxon Valdez* oil spill. Because sample sizes were computed for many individual taxa, as well as for a wide range in coefficients of variation, the recommendations presented in this report will be applicable to many other intertidal environments and geographic locations. In areas where intertidal variability is thought to be substantially different from the ranges cited in this report, more site-specific sample sizes can be determined by applying the methodology developed in this report to available local historical data, or to data collected during a pilot study.

The design of a field program differs depending on the goals of the monitoring. For example, the optimal number, type, location, and frequency of intertidal sampling in a monitoring program designed to detect the initial acute impacts from a spill, will be markedly different from those of

a recovery assessment. The following three types of intertidal monitoring assessments addressed in this report are common to many major oil-spill monitoring programs.

- Chapter 2 provides sampling guidelines for detecting differences in intertidal populations subjected to different cleanup techniques. Immediately following the *Exxon Valdez* oil spill, intertidal populations subjected to invasive cleanup procedures exhibited greater population reductions than populations exposed to oil but that were subjected to only light cleaning or no cleaning at all. The recommended sampling strategies for detecting the effects of different cleanup treatments are particularly applicable to manipulative field experiments such as the "clearing" experiments currently being conducted by NOAA.

- Chapter 3 recommends sample sizes capable of detecting abrupt recolonization events. Following the *Exxon Valdez* spill, impacted populations remained depressed for approximately two or three years after which populations sharply increased over a period of one or two years. Because of the absence of pre-spill baseline data, these recolonization events were best quantified by differences in population trends at control and impact sites.

- Chapter 4 estimates the sample sizes needed to detect chronic effects by testing for the presence of statistically significant slopes in long-term population trends at impacted sites. Following the PWS recolonization event, lingering chronic effects from the *Exxon Valdez* spill were evident as subtle long-term population trends in the populations of several intertidal taxa.

Sampling Decisions

Very different sampling strategies are needed to detect the three spill-related phenomena described above. Moreover, within a given sampling design, the optimal number of samples differs among intertidal taxa because of inherent differences in their level of spatial and temporal variability. Because of this, an extensive inventory of sample-size charts is provided to cover a wide variety of specific monitoring goals and taxa. During the initial response to an oil spill, decisions based on the information in these charts will help ensure the ultimate success of an intertidal sampling program insofar as meeting its monitoring goals.

The flow chart in Figure ES.1 shows the decision-making process and portrays the inter-relation of the three sample-size analyses presented in this report. The numbers in parentheses reference specific chapters and page numbers in this report. Some general sampling guidelines also emerge from the sample-size estimates that were determined for individual taxa. These sampling recommendations are provided at the bottom of the flow chart and are discussed in the following sections of this summary.

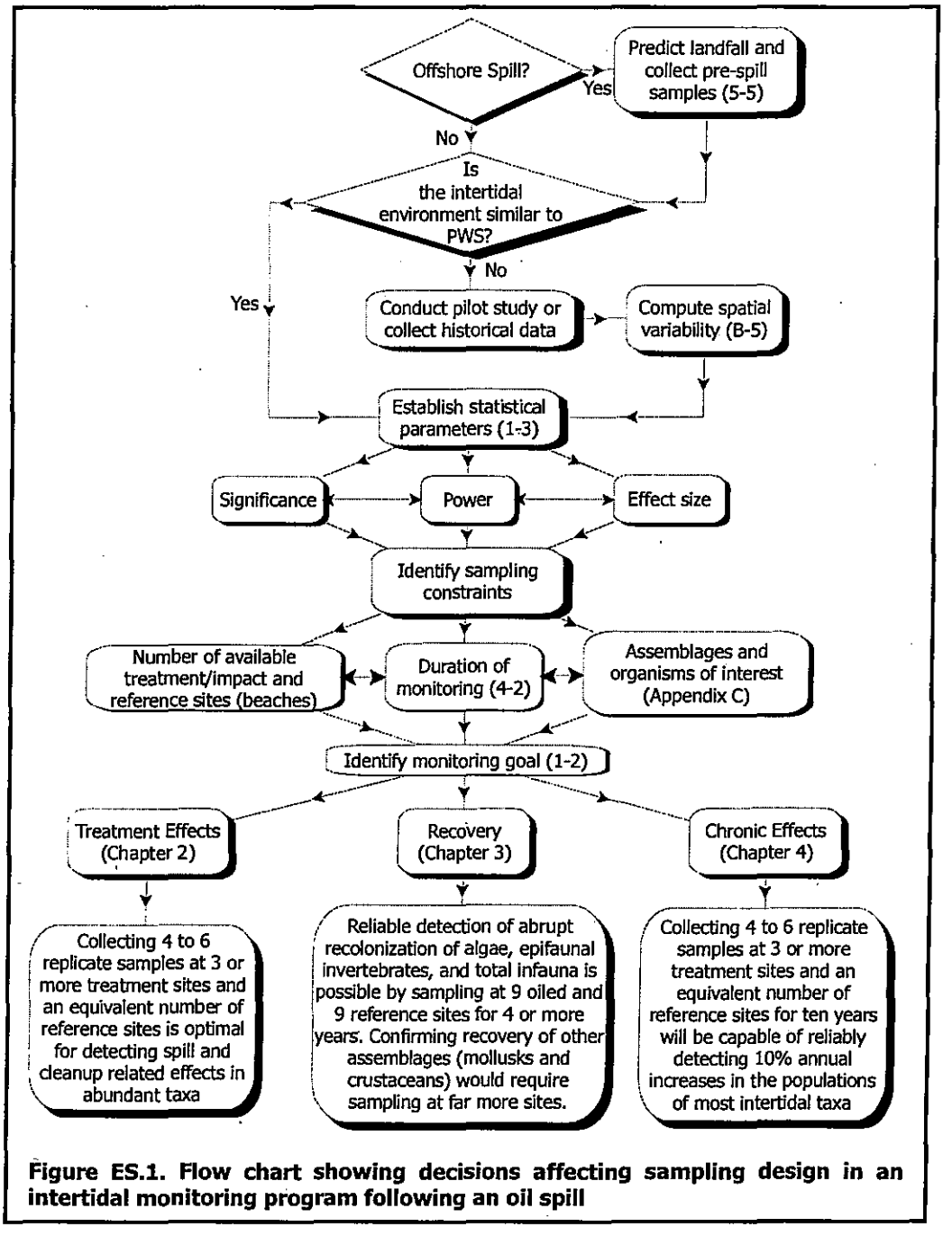

Figure ES.1. Flow chart showing decisions affecting sampling design in an intertidal monitoring program following an oil spill

Spatial Variability

Typical intertidal monitoring programs consist of a number of replicate samples collected at each of several sites or beaches. To assess recovery and long-term chronic effects, this sampling effort is periodically repeated. Samples usually consist of a series infaunal cores or visual enumerations of epibiota within quadrats along a particular tidal elevation. Replicate samples need to be

collected at a number of different beaches or sites, including ones that were impacted by the spill or cleanup treatment, and unaffected ones that can act as control or reference sites.

Consistent with this replicated sampling strategy, optimal sample sizes are dictated by estimates of variability on two spatial scales. Small-scale or "*within-site*" variability is associated with differences in population measurements determined from series of cores or quadrats collected at adjacent locations at a particular site. Larger-scale differences between individual sites are quantified by "*between-site*" measures of variability. In the sample-size analyses described in this report, these two types of variability determine the number of replicate samples (m) that need to be collected at each of n individual sites.

To characterize within-site and between-site intertidal variability, average coefficients of variation were computed for 270 PWS intertidal taxa, both before and after the recolonization event. Several important aspects concerning intertidal variability emerged from the analyses that affect the applicability of sample-size recommendations.

- The sample sizes recommended in this report for optimally determining treatment effects are representative of a large number of taxa, tidal elevations, and effect sizes. Except for a few outlier taxa, average coefficients of variation in the PWS dataset were found to be generally consistent among intertidal assemblages and tidal elevations, and they remained relatively stable before and after the recolonization event. The vast majority of taxa exhibited a marked population increase during the recolonization event, and the 39% of the taxa at impacted sites that were present prior to the recolonization event, exhibited spatial variability in the same range as post-recolonization populations.

- Compared to the influence of tidal elevation and assemblage, the largest differences in average spatial variability were observed among taxa within three general abundance ranges: sparse, intermediate, and abundant. Consequently, sample-size determinations were categorized by population range. Sample sizes determined for the abundant taxa were more reliable than those for sparsely populated taxa. Abundant taxa tended to have lower within-site variability while sparsely populated taxa tended to have lower between-site variability. Additionally, nearly all of the sparsely populated taxa were randomly distributed within the PWS dataset. This suggests that the area sampled by the quadrats and infaunal cores was too small to resolve potential spatial patterns, and that their populations were undersampled. In contrast, intermediate and abundant taxa exhibited a strong tendency to form clumps or aggregates and the spatial variability in these populations was well represented by the variability estimates.

- Optimal sample sizes recommended in this report significantly underestimate the number of samples that would be required to detect impacts to seven highly variable

taxa that exhibit a markedly increased tendency to aggregate or clump (Table 2.5). Characteristics common to these species included a relatively small size, a proclivity to congregate in crevices or in other microhabitats, and brooding of large clutches to an advanced stage of development before release as crawl-away juveniles. As a result, these taxa displayed an inordinately high within-site variability that is not well-represented by the sample-size charts presented in Appendix D of this report.

Sample Sizes

In practice, impacts to individual, sparsely populated taxa are rarely of primary interest. Exceptions might include taxa that are commercially valuable or are designated as environmentally sensitive, threatened, or endangered. Usually, however, widespread impacts to the major intertidal assemblages receive the most attention in monitoring programs. Consequently, the sample-size charts that are most likely to be used in intertidal monitoring programs, are those developed for abundant taxa. General guidelines concerning optimal sample sizes for detecting impacts to abundant taxa are listed at the bottom of the flow chart (Figure ES.1) and are discussed below.

- An optimal intertidal monitoring program for detecting changes in abundant taxa allocates approximately six replicate samples to each site and maximizes the number of sites within the available sampling resources. The shape of the sample-size curves used to detect treatment effects (Figure ES.2) and chronic effects (Appendix G) shows that above a certain point, adding replicate samples (m) within sites has little effect on the statistical power to detect change. Most of the curves for abundant taxa approach a vertical asymptote above $m \approx 6$ and are distinctly vertical above $m = 8$. Similarly, the power curves start to approach a horizontal asymptote below $m = 4$ and are distinctly horizontal below $m = 2$. This suggests that if sampling is planned at three or more sites within each treatment, then at least four, but no more than eight replicate samples should be collected at each site. Similarly, when only one or two replicate samples are being collected at each site, the addition of more sites does little to enhance statistical power. Instead, resources should be directed at increasing the number of replicate samples within each site.

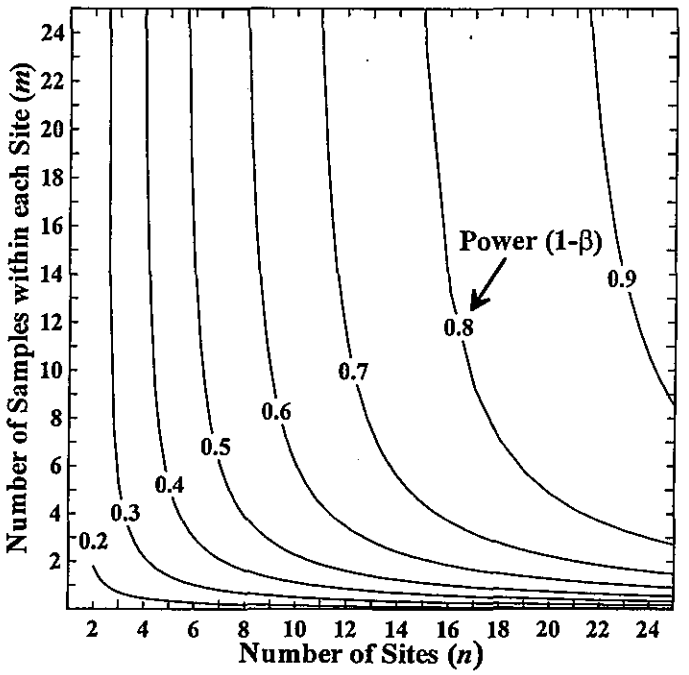

Figure ES.2. Sample-size chart showing the number of replicate samples (m) to be collected at n reference and n impact sites to detect a 50% reduction in abundant intertidal populations with a moderate level of natural biological variation.

- The ability to detect community-wide difference using multivariate analyses at treatment and reference sites is likely to yield low statistical power unless samples are collected at a large number of sites (>10), or unless the treatment effects are large. Additionally, when the number of ordination axes increases, a larger number of samples is required to discern a given separation between treatment groups on the ordination diagram.

- Parallelism (relative-trend) tests are better suited to the detection of recovery in intertidal populations following an oil spill than are direct comparisons of abundance at any particular time. Parallelism tests examine temporal trends in mean abundance at impact sites relative to reference sites. Because they do not assume that mean levels were equal at the reference and impact sites prior to the spill, they can accommodate inherent differences in the carrying capacity between locations within and beyond the spill zone; differences that may have been present before the spill occurred.

- In contrast to the sample sizes needed to detect treatment effects or chronic impacts, the ability to detect recolonization events with parallelism tests varies widely depending on the taxa being tested. Figure ES.3 shows that a marked epibiotic recovery, similar to that experienced at PWS sites subjected to invasive cleaning, can be detected with a four-year monitoring program. Very high power ($1-\beta > 0.98$) for *Fucus* and epifaunal invertebrate assemblages can be achieved by sampling at as little as two reference and two impact sites. In contrast, infaunal populations have much higher variability among sites, making detection of nonparallel trends difficult without sampling at a larger number of sites. In order to achieve a power above 0.7, sampling at a minimum of six reference and six impacted sites would be necessary.

- The striking difference in the ability to detect recolonization among the various intertidal assemblages emphasizes the importance of selecting optimal biological variables to include in a monitoring program designed to assess recovery. The assemblage of concern must not only be exposed to contamination or habitat disturbance, but it must also have the ability to demonstrate recovery within the practical constraints of field sampling. Optimal taxa for monitoring are ubiquitous (abundant), are not extremely clumped or patchy in distribution, and respond uniformly at all sites to the impact. Assemblages with these attributes, such as algae and epifaunal invertebrates, have the greatest likelihood of demonstrating statistically significant effects with modest sampling efforts.

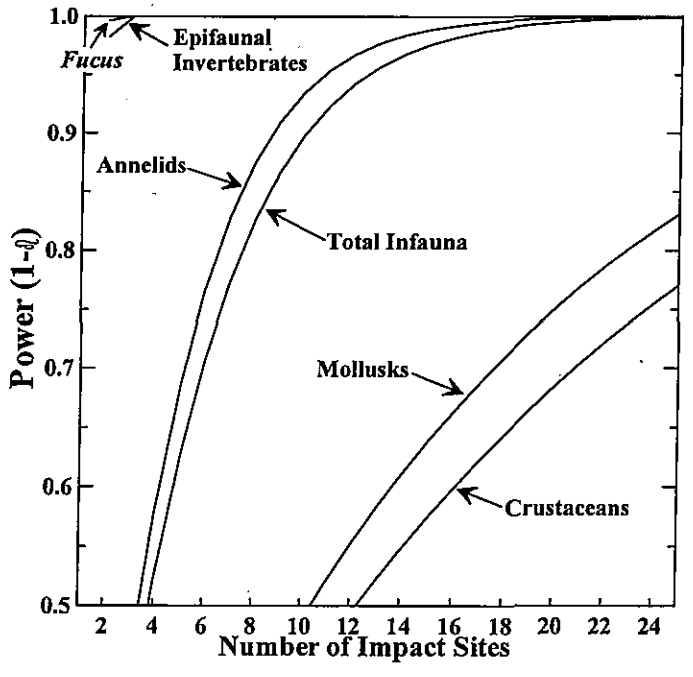

Figure ES.3. Number of control and impact sites needed to detect linear departures from parallelism during recolonization at PWS intertidal sites subjected to severe habitat disturbance.

Maximizing Sampling Resources

This report demonstrates that often a large number of samples must be collected to achieve a even a marginal statistical power to detect changes in intertidal populations. By adhering to traditional intertidal sampling protocols, which are labor intensive and demand the presence of experienced field biologists, there may not be enough time or trained personnel available to collect samples that are sufficient for statistical credibility. However, the sample collection rate can be increased by relaxing some of the traditional field-sampling techniques without unduly sacrificing needed statistical rigor.

- Sample opportunistically before an offshore spill impacts shorelines that are identified as landfall locations using oil-spill trajectory models. Current oil-spill responsiveness and the predictive skill of real-time trajectory modeling now make this feasible in many cases.

- Relax taxonomic resolution of epibiota taxa in the field. Identifying specimens to the lowest taxonomic level in the field is a time-consuming and expensive process. Many studies, most recently Lasiak (2003), have shown that significant differences in marine assemblages apparent at the species level, are often also apparent at family or higher taxonomic levels. Unless individual epibiotic species can be quickly and accurately distinguished visually in the field, they should be enumerated at a higher taxonomic level.

- In multi-year field programs, randomize quadrat locations along transects to avoid establishing and maintaining fixed markers. By specifying certain guidelines concerning spacing and consistency of habitat, the increased statistical power realized by collecting a larger number of additional samples far outweighs any variance reduction that is afforded by fixed sampling locations.

- Identify and limit the goals of the monitoring program at the outset. The availability of impact sites and the resources to sample those intertidal sites quickly becomes apparent following a spill. Consulting the sample-size charts in this report will indicate what monitoring goals are feasible. For example, with less than ten available impact sites, it may not be feasible to identify the future recolonization of mollusk or crustacean populations subjected to invasive cleanup treatments (Figure ES.3). The anticipated duration of sampling also determines the sampling goals. Is it limited to a one-time assessment of treatment effects (Chapter 2), or will multi-year post-spill sampling be conducted to quantify recovery (Chapter 3) and long-term chronic effects (Chapter 4)? Once these questions are answered, consulting the appropriate sample-size charts in this report will help identify the optimal number of replicate samples to be collected at each site.

Chapter 1. Introduction

"How many samples should be collected?" is a perennial question posed by field biologists. This question is particularly pressing when a major marine oil spill impinges on a coastline and impacts sensitive biological communities within the intertidal zone. In the first hours and days after a spill, the initial priorities are containment and cleanup. Nevertheless, field biologists are quickly called upon to design impact assessment studies, often without the benefit of site-specific biological information or access to qualified biostatisticians.

These early sampling-design decisions can have profound consequences for the ultimate performance and validity of the assessment study. Enough intertidal samples must be collected to meet the goals of the monitoring program while the monitoring program itself must be capable of detecting impacts that are both biologically and statistically significant. An undersized study can be a waste of resources if it cannot reliably discern significant impacts and produce useful results. It is equally important, however, to avoid wasting limited resources by collecting too many samples. When destructive sampling is involved, such as with infaunal cores or clearing studies from rocky intertidal areas, a grossly oversized monitoring program may unnecessarily contribute additional damage.

This report provides guidance in this initial decision-making process so that an optimal monitoring program can be quickly established that adequately assesses impacts to intertidal communities after a spill. Although not a substitute for consultation with professional statisticians, these recommendations provide sufficient guidance to safely begin field data collection in a wide variety of circumstances. Subsequently, as more site-specific information becomes available and monitoring priorities become clearer, the statistical design of the monitoring program can be further refined.

Much of the difficulty in sample design arises because there is no simple or universal answer to the question of adequate sample size. A complex series of interrelated issues, driven by biological, environmental, political, logistical, and financial constraints, influence the determination of adequate sample-size. While some of these issues are within the control of the investigator, others relate to the inherent biological variability associated with nature itself. However, once biological variability is established, there are quantitative techniques available to determine adequate sample size. This report determines biological variability from an extensive database of intertidal observations collected in Prince William Sound (PWS), Alaska, as part of the long-term monitoring program conducted after the *Exxon Valdez* oil spill. It uses these

estimates of intertidal variability to provide quantitative guidance for the design of future oil-spill monitoring programs.

Monitoring Goals

One of the ironies of sampling theory is that no single survey design and no single set of sampling size calculations exist that are appropriate for all assessment questions. The optimal sampling design for one aspect of an oil spill assessment may not be desirable for other monitoring objectives. For example, to perform an initial test of impacts, the optimal design would only allocate sites to the extreme conditions of the heaviest oiled sites and the unoiled reference sites. Conversely, to optimally conduct a damage assessment, study sites would need to be evenly distributed across the landscape where they cover a wide range of contamination levels. Hence, the optimal sample allocation for assessing acute effects is the complete opposite of what is needed for assessing damages. Consequently, recommendations for sample size and sample distribution are inextricably linked to the desired study goals.

An oil-spill assessment may have several competing study goals and therefore, consist of several elements or phases. Among the possible elements are:

1) tests for acute impacts;
2) tests for long-term or chronic effects;
3) assessments of initial biological or habitat recovery;
4) assessments of the long-term stability of the recovered system;
5) assessments of damage; and
6) assessments of alternative cleanup or restoration techniques.

All of these study elements have been part of the *Exxon Valdez* oil spill assessment conducted in PWS at one time or another. Each element has its own unique design requirements, performance standards, and sample size requirements.

In any given spill situation, field biologists, in consultation with decision makers, will need to quickly determine the relevant study goals. Most likely, the final study will be a composite of the design elements and sample-size requirements covering several stated goals. The sooner assessment priorities can be identified, and appropriate design and sampling elements can be incorporated in the overall investigation, the more likely it becomes that the monitoring program will achieve its stated goals. Careful consideration of biometrics is crucial at this stage of the spill investigation.

This report focuses on three types of monitoring objectives: (1) testing for initial acute impact effects, (2) assessing abrupt recolonization events that occur a few years after the spill, and (3) evaluating long-term recovery from chronic effects. These kinds of spill impacts to intertidal biota can arise from either hydrocarbon exposure itself or habitat damage caused by the cleanup methods used to remove oil from the intertidal zone. These study elements were selected because they address both short- and long-term study goals, in addition to covering differences between acute-impact damage assessments and investigations of recovery from subtler chronic effects.

Separate chapters provide sample size guidance for each of these three distinct types of studies. Chapter 2 quantifies acute impacts based on statistical tests for differences in mean abundance at impact and reference sites. Chapter 3 quantifies episodes of abrupt repopulation events by testing for departures from parallelism in intertidal populations at reference and impact sites over time. Chapter 4 characterizes weak population trends related to the dissipation of chronic impacts by testing for departures from long-term stability in intertidal populations.

Statistical Considerations

Both qualitative and quantitative goals are part the decision-making process at the outset of an oil-spill assessment study. Qualitative decisions involve the selection of the species and habitats of interest while quantitative goals identify the magnitude of change that is considered important. Both kinds of decisions dramatically influence the overall size of the field investigation; namely, the required number of sites and the number of samples to be collected within those sites. In addition to specifying the magnitude of change that is deemed important, quantitative decisions should also reflect the risks associated with overlooking an important impact.

These quantitative goals are specified with the three statistical parameters listed in Table 1.1. Everything else being equal, the sampling effort is governed by the desired power $(1-\beta)$ to detect a difference $(\geq \Delta)$ between impacted and reference sites at a given statistical significance level (α).

Table 1.1. Three parameters that define the quantitative goals of spill assessments

Parameter	Symbol	Description
Significance	α	The probability of incorrectly finding an important impact when it is in fact, inconsequential
Power	$1-\beta$	The probability of correctly finding an important impact. It is the complement of β, which is the probability of missing a meaningful impact.
Effect Size	Δ	The amplitude of the change in biological properties (impact) that is considered important or meaningful.

Specifying values for α, β, and Δ is not a straightforward process, but the guidance offered in this report will help establish defensible study goals that can be achieved with realistic sampling strategies. In general, Δ should be determined by the size of the change considered to be biologically, economically, or socially important. The other two parameters, α and β, identify the risk of committing two competing types of errors in identifying a change of magnitude Δ (or greater). The risk (α) of a false alarm arises when biological changes of magnitude Δ are mistakenly ascribed to an oil spill or a particular cleanup technique. The parties responsible for the spill or cleanup method would be concerned about setting α too high. Conversely, reducing the risk (β) of missing a meaningful biological impact would be important to the public trustees of the environment.

Ideally, the two types of error would be set equal because, as Skalski (1995) points out, "...it seems reasonable for both parties to bear equal risk." In practice, however, the actual risk levels are less a matter of regulatory policy and more a function of available sampling resources and historical convention. Although not universally adopted by the scientific community, the α-level has been historically set at 0.10, 0.05, or 0.01. These levels are typical of controlled laboratory experiments where the emphasis is on avoiding false claims of an effect and a large number of tests can be easily conducted. In contrast, error levels this low for both α and β are rarely achieved in marine monitoring programs where expensive field surveys are being conducted on highly variable biological communities. In oil spill assessments, the overall number of impact sites is limited by the geographic extent of the spill, and the investigator does not always have the luxury of increasing sampling to achieve small error levels. In practice, α is often set at the highest level (0.1) that is routinely accepted in the scientific literature, while power ($1-\beta$) is reduced and the detectable amplitude (Δ) of impacts that are considered important is allowed to increase.

Cohen (1988) describes some of the trade-offs in the selection of β by looking at the ratio $\frac{\beta}{\alpha}$ to determine the relative seriousness of committing the two types of error. For example, setting $\beta = 0.3$ and leaving $\alpha = 0.1$ means that mistakenly finding an impact is considered three times more serious than mistakenly missing it based on the ratio of the selected error rates. Clearly, setting β too high, for example at or above 0.5, defeats the purpose of the impact assessment because an important impact could be missed one out of every two times. As will be shown in this report, intertidal communities have high natural variability and setting β too low is also

impractical because it leads to unrealistically large sample sizes or unacceptably large amplitudes for detectable biological changes (Δ). Other intertidal studies (Tenera, 1997) have set $\beta = 0.3$ to achieve a detection power $(1-\beta)$ of 0.7. Even then, the size of detectable impacts (Δ) can be large, which leads to a wide range in observed differences that are indeterminate with regard to the presence of an impact.

Biological and societal goals often determine the qualitative decisions concerning which biological communities and habitats are of primary interest. Selection of the taxa and habitats to be studied can have a profound influence on the scope of the field sampling effort. This is because the ability to detect impacts is related to the inherent variability in the biological community of interest. Specifically, the statistical power $(1-\beta)$ of a particular sampling design is related to α and Δ through the variability in the biological parameter being tested for impacts. As the variability increases, the amount of sampling effort required to discern impacts of magnitude Δ, increases. As a species becomes less frequent in the environment and its distribution more patchy, more samples are required to adequately discern tangible differences between impact and control sites. In an assessment study focused on multiple taxa, overall sampling effort is typically driven by the least common taxon that is considered important to monitor.

The distribution of taxa among habitats also directly affects sampling effort. If the taxa that are selected for monitoring are allopatric and do not occupy the same habitat, then the overall study effort is proportionally increased. For example, separate monitoring efforts may be required to assess impacts to taxa endemic to upper versus lower intertidal habitats, cobble versus sandy habitats, or infauna versus epibiota. While the recommendations provided in this report address the sampling effort required for individual intertidal subpopulations, they can be extended to groups of allopatric taxa by summing the sampling effort required for each habitat type.

The Intertidal Database

In 1989, the *Exxon Valdez* accident spilled approximately 11 million gallons of oil in PWS and outer coastal areas of the Gulf of Alaska. About five million gallons impinged on 400 miles of shoreline and became stranded on intertidal habitats (Spies *et al.*, 1996). Oil coated rock surfaces, penetrated into soft sediments, and impacted a wide range of intertidal organisms. Cleaning removed a large amount of stranded oil but also damaged the intertidal environment. High-pressure hot-water washing was particularly destructive (Mearns, 1996).

More than a decade of intensive monitoring at intertidal sites within PWS has provided a detailed characterization of the infaunal and epibiotic distributions over time and space (Coats *et al.*, 1999; Skalski *et al.*, 2001). The sites were exposed to varying degrees of oiling and subsequent invasive cleanup techniques. Three types of intertidal sites were monitored: 1) reference (unoiled), 2) oiled, and 3) oiled with cleaning.

Many populations at impacted sites largely recovered during a large recolonization event that lasted for a period of one to two years beginning around 1990. Recolonization occurred across the full range of intertidal assemblages, including sediment-dwelling infaunal invertebrates, sessile and motile epifaunal invertebrates, and algae. It was also evident at all intertidal elevations sampled. During the recolonization period, most population increases at impacted sites were statistically significant ($p \leq 0.10$) compared to population fluctuations observed at non-oiled control sites. After the initial increase, intertidal populations at impacted sites stabilized with abundance perturbations that tracked those of control sites. Thus, within the resolution of statistical tests applied to abundance, the major intertidal assemblages had largely recolonized impacted sites and achieved equilibrium with ambient environmental conditions by 1993. Subtler spill effects undoubtedly lingered in the intertidal community after this recolonization event, so the ecosystem could not be considered fully recovered. Ongoing chronic effects could still be manifested in unstable age-structures, altered growth patterns, physiological changes, and other effects not reflected in the mean abundance at the impacted and reference sites selected for study by Coats *et al* (1999).

Nevertheless, the large-scale fluctuations in intertidal populations clearly delineated a major recolonization event at sites impacted by the spill. The amplitude of the population increase was larger at oiled sites that were subjected to aggressive cleanup techniques. During recolonization, populations increased by a factor of eleven at sites that were subjected to high-pressure hot-water washing. The average population increase was only a factor of three at oiled sites that were not subjected to invasive cleanup procedures. These results were consistent with other observations of increased damage to intertidal organisms at sites treated with high-pressure hot-water washes. The enhanced recolonization at these sites reflects the increased damage. However, there was no evidence that recolonization was measurably delayed at the oiled sites that received hot-water washing. In fact, timing and duration of the recolonization was remarkably similar across the various impacted sites, tidal elevations, and intertidal assemblages.

In this report, the PWS intertidal database was used to estimate several types of biological variance that are needed to determine the optimal sampling effort in future spills. The sampling

effort is defined by the number of sites, the sampling effort (replication) within those sites, and the duration of sampling needed to accomplish specific study objectives. Using the PWS monitoring data to establish sampling criteria for future spills expands the original intent of the PWS monitoring program; namely, to investigate the long-term recovery of intertidal communities after the *Exxon Valdez* oil spill. With this expanded goal in mind, additional unoiled (reference) sites were purposefully added to the monitoring program in 1998 to better determine the inherent biological variability within intertidal environments. In addition, the observed amplitude of the post-spill recolonization event suggested appropriate choices for size of the acute impacts (Δ) to be used in power analyses that determine sampling size. This amplitude differed depending on whether oil-impacted sites were subjected to aggressive cleaning methods.

Applicability Outside of Prince William Sound

The sample-size recommendations provided in this report are based on variance estimates determined from long-term monitoring of infaunal and epibiotic distributions within PWS. Consequently, they best apply to the design of future spill assessments in the same region. The amplitude of natural intertidal variability is likely to differ in distant locales; thus, the utility of the sample-size estimates would be reduced. However, some of the regional differences in biological variability will undoubtedly be due to differences in population sizes. Larger populations tend to have a higher variance than smaller populations. Normalizing variability estimates by the population size reduces the influence of these differences and extends the applicability of the recommendations reported here.

With this broader applicability in mind, sample size calculations presented in this report are expressed in terms of a coefficient of variation:

$$CV \equiv \frac{\hat{\sigma}}{\hat{\mu}} \qquad (1.1)$$

where $\hat{\sigma}$ is the standard deviation of abundance and $\hat{\mu}$ is an estimate of mean abundance. In the PWS dataset, logarithmic transformation of abundance significantly reduced temporal variations in estimates of mean populations at reference sites, oiled sites, and sites subjected to invasive cleanup techniques. Means computed from untransformed abundance were unduly influenced by the more erratic fluctuations that occurred at sites with higher populations. Logarithmic transformation resulted in time histories of mean abundance that revealed a much clearer pattern of impact and recovery (Coats *et al.*, 1999; Skalski *et al.*, 2001). The transformation was effective because it reduced the dependence of variance on mean population size. In fact, the

variance of a log-normal population distribution is approximately equal to the square of the coefficient of variation computed from raw abundance (x) determined from counts or percent cover:

$$CV^2 \equiv \frac{\sigma^2}{\mu^2} \approx \mathrm{Var}(\ln x) \tag{1.2}$$

Although the CV provides a stable measure of variance across a wide range of population sizes, substantial differences in the spatial distribution of intertidal organisms can limit the utility of the sample-size recommendations outside of PWS. Environments that are significantly more or less heterogeneous than PWS may have a markedly different CVs. Differences in heterogeneity are often manifested in the degree of clumping associated with intertidal assemblages. Marine and freshwater invertebrates tend to follow a negative binomial distribution where organisms form clumps rather than being distributed uniformly over their habitat (Elliot, 1977). The variance of these clumped distributions is given by:

$$\mathrm{Var}(x) = \mu + \frac{\mu^2}{K}, \tag{1.3}$$

where: $\frac{1}{K}$ is an index of the clumping together of individuals in the population. For this distribution,

$$CV(x) = \sqrt{\frac{1}{\mu} + \frac{1}{K}}. \tag{1.4}$$

As clumping increases, CV asymptotes to a constant, $\frac{1}{\sqrt{K}}$, dependent only on the degree of clumping. For sparse, randomly distributed populations, the variance asymptotes to the mean. In many cases however, environmental heterogeneity largely determines the CV. Consequently, extreme environments may produce dispersion patterns and CVs beyond the range observed in the PWS dataset. In those cases, site-specific estimates of CVs should be used in conjunction with the sample-size charts and tables presented in this report. In an effort to extend the utility of the sample-size computations, this report also presents results for CVs computed for the sparse populations observed prior to 1993 at PWS sites impacted by the oil spill.

Chapter 2. Treatment Effects

This chapter specifies the sample sizes needed to detect differences in intertidal biota that have been subjected to different types of physical or chemical treatments. The underlying statistical design consists of a one-way analysis of variance using a number of replicate samples collected concurrently at several sites or beaches. It is most applicable to manipulative field experiments such as the clearing experiments now being conducted under the National Oceanic and Atmospheric Administration (NOAA) auspice as part of the PWS intertidal monitoring program. As part of a post oil-spill assessment, it can lend insight into the efficacy of various cleanup techniques whereby the mean abundance in samples collected at sites subjected to differing levels of treatment is compared. By extension, it can be used to compare mean populations at sites exposed to an oil spill with those of unoiled reference sites. However, as described below, inferences concerning the effects of oiling are weak without additional information.

As part of the PWS monitoring program, the NOAA initiated two field experiments to investigate the recovery mechanisms of intertidal populations exposed to severe habitat disturbance. These manipulative experiments, one for infauna within PWS and another for epibiota within Kasitsna Bay, Alaska, investigated aspects of recovery that were suggested by the long-term monitoring data collected in PWS after the *Exxon Valdez* oil spill. These two studies are currently in progress, and results will be presented in future reports. They were motivated by certain aspects of recovery that were revealed in the PWS data but could not be fully investigated because of the absence of reliable pre-spill data (Coats *et al.*, 1999). These particular manipulative experiments have the added advantage of temporal sampling and include samples collected before treatments were applied. As a result of the added temporal component, the field experiments can more reliably discern subtle temporal effects from the various treatments. In their simplest manifestation, the sample-size recommendations presented in this chapter can be used to design field experiments that are intended to directly compare mean populations at sites exposed to two different cleanup treatments.

This kind of direct comparison is also common in spill-assessment studies and can be used to design post-spill monitoring programs. For example, Peterson *et al* (2001) recently proposed using two-sample tests to compare oiled and reference sites following an oil spill to assess whether the oiled sites returned to the "*innate background levels of the reference sites.*" However, in the absence of other information, Skalski *et al* (2001) advise against using a direct comparison of mean populations at oiled and reference sites to assess recovery after an accidental spill. Instead, statistical evaluations based on temporal changes, such as those

described in the following chapters, can lend more reliable insight into population fluctuations resulting from an accidental oil spill.

Difficulties arise when post-spill populations at a particular point in time are compared without consideration of differences that may have been present before the spill. Specifically, determining impacts or recovery from a comparison of mean intertidal abundance at oiled and unoiled beaches tacitly assumes that the beaches and their intertidal biota were identical prior to the spill. Without data collected prior to the spill, this assumption cannot be confirmed and may lead to erroneous conclusions concerning the extent of recovery within intertidal communities. Some oiled and unoiled beaches, and the intertidal biota that reside on them, were almost certainly different prior to a spill; otherwise, why would some beaches be covered by oil while nearby reference sites were not? Often, the oiled beaches differ in their orientation and exposure to prevailing currents, which results in an observed difference in oil cover. The persistence of oil cover can also differ because of a disparity in the rugosity of rocky shorelines or in sediment grain-size along sand beaches. Differences in these physical characteristics affect the kinds of intertidal biota present along the shorelines. In addition, in a major spill, biogeographic differences in intertidal biota can arise because the only available unoiled reference beaches may lie great distances away from the impacted beaches.

Despite these limitations, sample sizes associated with direct comparisons of population means from a single post-spill sampling event can be of value. At a minimum, they lend practical insight into the influence of spatial variability on the design of monitoring programs that are intended to quantify gross effects from an oil spill or cleanup using a single sampling event. Because sufficient pre-spill data is rarely available, these direct comparisons are often the only means of reliably discerning differences in populations subjected to different levels of hydrocarbon exposure or cleanup treatment. They also lend insight into the design of manipulative experiments where, for example, intertidal populations along several similar beaches are subjected to different physical or chemical treatments, and the differences in mean biological response are then contrasted. In either case, the estimates of intertidal variability derived in this chapter form the basis for more involved assessments of impacts and recovery, some of which are described in other chapters of this report.

The first section of this chapter, entitled Species Response, provides sample sizes for assessing treatment effects that could potentially influence the abundance of a wide variety of taxa residing at various elevations within the intertidal zone, and on both hard- and soft-substrates. As part of the sample-size determination, variability was estimated from the PWS data for 270 individual

taxa residing within three intertidal zones. Some of the trends in the CVs and variances, as well as unusually high variability in certain taxa, are discussed in the following sections. These individual variance estimates are of interest from a biological standpoint, in addition to their value for sampling design.

The second section of this chapter, entitled Community Response, provides guidance on the field sampling effort needed to detect changes in the composition of entire intertidal communities. It is based on power analyses of multivariate community parameters derived from principal component analyses. Changes in overall community composition are often more representative of impacts to intertidal populations, unless a particular species is of interest due to its ecological sensitivity or economic value.

Species Response

Variance Computation

Appendix B formulates the statistical construct used to determine samples sizes for assessing effects based on a direct comparison of mean abundance at sites subjected to two different treatments. The conceptual framework presented in Appendix B is an integral component of the discussion that follows. The number of sites that need to be sampled and the number of replicate samples that need to be collected at those sites are explicit functions of the variability inherent in the biological populations to be sampled. More samples are required to discern differences in populations that are highly variable. Ideally, a site-specific pilot survey would be used to estimate background variability. Applying these preliminary variability estimates in a power analysis would yield optimal sample sizes to be used in the design of a full oil-spill monitoring program. Alternatively, variability computed from the large volume of data collected during the PWS intertidal monitoring study can be used as a preliminary estimate for sample size calculations. Estimating variance components from the PWS data is the subject of this subsection.

Many intertidal impact studies are designed to collect replicate samples at a number of beaches within the region of interest. Ideally, some of the beaches sampled are heavily oiled while others represent reference or control measurements. Sampling at multiple beaches helps to account for differences in the severity of biological impacts on a variety of beaches subjected to the spill. Sample-size determination in this statistical design requires estimation of intertidal variability on two spatial scales. Small-scale or "*within-site*" variability is associated with differences in population measurements determined from epibiotic quadrats or infaunal sediment cores

collected at a series of locations along a particular beach. Larger-scale differences are quantified by "*between-site*" measures of variability. In the power analysis, these two types of variability determine the number of replicate samples (m) that need to be collected at two sets of n sites in order to detect a difference of size Δ at a statistical significance level α, with a statistical power of $1 - \beta$.

As described in Chapter 1, the error rates (α and β) and the size of the change that is deemed significant (Δ), may be set *a priori* by policy or precedent. In practice, these parameters and the power to detect changes are dictated by the number of oiled beaches and sampling resources that are available at the time of the spill. In either case, estimates of biological variability are required to quantify the required sample sizes. Biological variability is best determined empirically from available data. As described in Appendix B, the noncentrality parameter (Φ), which is used to determine sample sizes, can be computed from either of two related estimates of variability: variance ($\hat{\sigma}^2$), as in Equation B.7, or the coefficient of variation ($\hat{CV}_S$), as in Equation B.8. For each taxon or taxonomic group, the within-site ($\hat{\sigma}^2_W$ and $\hat{CV}_W$) and between-site ($\hat{\sigma}^2_B$ and $\hat{CV}_B$) components of these variability estimates were computed from the PWS data using the ANOVA technique described in Appendix B.

In addition, PWS data from two different periods of time were used to compute the variance estimates. One set was representative of sites impacted by oil and was computed from intertidal data collected at oiled sites prior to 1992. The sites selected for the variance computation during the impact period did not include sites that received high-pressure hot-water washing during cleanup. Intertidal populations at those sites were too low to reliably estimate variability. Thus, this first set of variability estimates was indicative of intertidal communities impacted by oiling alone, and not those that were subjected to invasive cleanup techniques. The second set of variability estimates was computed using data collected at all the sites in the years from 1993 through 2000. Most taxa had stabilized by 1993 after experiencing a marked population increase between 1991 and 1992. Thus, variability computed from post-recolonization data collected at all the sites reflected the variability within healthy intertidal populations. Variability computed from data collected at oiled sites between 1989 and 1991 were characteristic of sparse populations associated with oil-spill impacts.

Although variability was computed separately for each year of data, the results were pooled across several years to enhance the reliability of the variance estimates. Because the onset of recovery in the infaunal and epibiotic databases differed slightly, the years over which the

variance estimates were pooled also differed (Table 2.1). Epibiotic populations in the middle intertidal zone began recovering after 1990 while the upper intertidal populations stabilized after 1993 (Coats *et al.*, 1999). Accordingly, variances that are representative of the reduced epibiotic populations impacted by oil were computed by pooling results from 1989 and 1990. Variances representative of unimpacted epibiotic populations were computed by pooling variability estimates determined from data collected from 1994 through 2000.

Table 2.1. Years over which variance estimates were pooled

	Time Span	
Assemblage	Oil Impacted	Unimpacted
Epibiota	1989-1990	1994-2000
Infauna	1990-1991	1993-2000

Because reliable infaunal data was not available in 1989 and because the onset of recovery occurred one year later in the lower intertidal zone, infaunal variances at oil-impacted sites were computed from data pooled across 1990 and 1991. In contrast, most infaunal assemblages in the lower intertidal zone had largely stabilized by 1992, one year earlier than epibiota. Consequently, the variance for unimpacted infaunal populations was estimated by pooling the results applied to data collected at all the sites from 1993 through 2000. The methodology for pooling the variance and CV estimates is described in Appendix B.

Population Changes

Applying the techniques described in the preceding section to the PWS data provides insight into biological variation within the intertidal zone. Estimates of the variation in intertidal populations were computed for 270 infaunal and epibiotic taxa that encompass the lower, middle, and upper intertidal zones. Variation was characterized both in terms of standard deviations as well as coefficients of variation. As described above, two separate periods were also examined: a period immediately following the spill (1989-1991) when populations were impacted by hydrocarbon exposure, and a period (1993-2000) when populations had largely recolonized after the spill.

Intertidal populations for most taxa were sharply reduced after the spill and remained low during the impact period between 1989 and 1991. Those taxa with noticeably depressed populations are evident in the vertical bar graphs shown in Appendix C. At sites that were exposed to oil but not invasive cleanup techniques, intertidal populations increased by approximately three-fold during the recovery period between 1991 and 1993 (Coats *et al.*, 1999). This suggests that the observed effects level from hydrocarbon exposure was approximately $\Delta = -0.67$ within PWS after the *Exxon Valdez* oil spill. For oiled sites within PWS that were subjected to invasive cleanup

techniques, the eleven-fold population increase suggests that the combination of oiling and cleaning caused a much larger change of $\Delta = -0.91$.

Nearly 61% of the taxa examined in the post-recovery (non-impacted) period were completely absent during the impact period at oiled sites that did not experience invasive cleanup. These maximal population differences ($\Delta = -1.0$) are particularly evident in the bar graphs along the centerline of the figures in Appendix C. The large number of epibiotic taxa that were conspicuously absent during the impact period have missing bars in the left-center bar-graph in Figures C.1 through C.6. The number of missing epibiotic taxa is particularly noteworthy when compared to the infaunal population levels shown in Figure C.7, where only seven of the 47 infaunal taxa were completely absent during the impact period.

Twenty-four of the 270 taxa (8.9%) had population levels that were actually higher during the impact period and exhibited a subsequent decrease in abundance during recovery. For some of these taxa, the higher population level during the impact period is consistent with opportunism. Specifically, populations of opportunistic taxa might be higher shortly after a spill because of their tolerance to hydrocarbon exposure and the reduced competition afforded by the elimination of other, more hydrocarbon-sensitive species. Higher population levels during the impact period could have also resulted from decreased predation pressure from predator populations that were slower to recover than prey populations. Regardless of the mechanism, the presence of these taxa with higher populations during the impact period has important implications for the determination of sample sizes. As described in Appendix B, hypothesis tests conducted on populations that could either increase or decrease as a result of impacts, requires the use of two-tailed probability distributions. Two-tailed tests have a markedly lower statistical power and substantially increase the sampling requirements accordingly.

However, for many taxa, the perceived higher population during the impact period may have arisen from sampling uncertainty rather than opportunism or decreased predation pressure. Of the 24 taxa whose populations exhibited higher abundance during the impact period, fourteen had mean abundances that were below 1 count or 1% cover per sample unit. Changes in the populations of these fourteen sparsely populated taxa were not determined with any degree of confidence. The remaining 10 taxa that exhibited measurably higher populations during the impact period are listed in Table 2.2. Of these, the few taxa with markedly higher populations during the impact period stand out in the abundance plots of Appendix D (See for example, *Balanus* in Figure D.4 and *Ligia* in Figure D.5).

Table 2.2. Taxa whose populations were higher during the impact period and declined during the subsequent recovery

				Mean		% Increase (Δ)
Name	Type	Level	Measure	Impact	Non-Impact	
Balanus/Semibalanus	Epifauna	Middle	%	1.292	0.039	3238%
Ligia sp.	Epifauna	Upper	#	1.733	0.056	2998%
Sipuncula	Infauna	Lower	#	1.367	0.164	732%
Elachista fucicola	Algae	Middle	%	1.315	0.548	140%
Phyllodoce sp.	Infauna	Lower	#	3.467	2.329	49%
Gloiopeltis furcata	Algae	Upper	%	1.733	1.220	42%
Ampithoe sp.	Infauna	Lower	#	1.067	0.852	25%
Chthamalus dalli	Epifauna	Middle	%	5.461	4.637	18%
Littorina scutulata	Epifauna	Upper	#	83.200	72.084	15%
Cingula sp.	Infauna	Lower	#	14.800	12.993	14%

The 10 taxa in Table 2.2 are a mixture of taxa such as barnacles and *Gloiopeltis* that are pollution-tolerant, opportunistic, or experienced reduced predation; taxa such as *Ligia* and *Ampithoe* whose population levels are not well determined because of their patchy distribution; and taxa with unknown life histories such as sipunculids and *Cingula*. The higher barnacle populations observed during the impact period are consistent with their life history. The barnacles *Balanus/Semibalanus* and *Chthamalus dalli* are sessile organisms that recruit to open, bare substrata after oil toxicity has declined. These barnacle species rapidly repopulate open substrata with settlement occurring year-round at all tidal levels in other regions such as California and England (Southward, 1967; Highsmith *et al.*, 1996; Morris *et al.*, 1980; Barnes, 1989). This rapid initial recruitment may account for the comparatively high populations observed during the impact period. Subsequent reductions in barnacle populations could have occurred as predators recolonized impacted intertidal areas and competition for space with other settling invertebrates and algae intensified.

Certain algal taxa that exhibited higher abundance during the impact period could also be characterized as tolerant to hydrocarbon exposure or as having experienced reduced predation during the impact period. For example, *Gloiopeltis furcata*, a small, branching red alga (Rhodophyta), and the green string lettuce, *Enteromorpha* sp., a green alga (Chlorophyta), both tend to be highly ephemeral and rapidly repopulate bare substrata after the removal of other organisms (Southward and Southward, 1978; Southward, 1982; Stekoll and Deysher, 1996; Stekoll *et al.*, 1996). Both *Gloiopeltis furcata* and foliose Chlorophyta were identified as early algal colonizers in the PWS intertidal zone after the *Exxon Valdez* oil spill. Their succession is evident in Figure 27 on Page 53 of Coats *et al.* (1999). Similarly, the little Turkish Towel, *Mastocarpus papillatus* (Gigartinales), is a red alga capable of rapidly colonizing from its hardier

alternate life form *Petrocelis*. Stekoll *et al.* (1996) found that the Gigartinales increased rapidly in biomass and abundance at oiled sites following the oil spill.

The higher impact-period populations observed in other taxa cannot be as easily ascribed to opportunism, pollution tolerance, or reduced predation. The Rock Louse, *Ligia* is a highly motile, aggregative isopod that seeks out suitable habitats with the changing tides (Farr, 1978). Its higher abundance in the impact period could have resulted from the fortuitous sampling of cells of organisms within rock crevices and under cobbles. Similarly, the amphipod *Ampithoe* is a tube-dweller often associated with blades of algae (Morris et al., 1980). Little else is known about this crustacean. The higher numbers of *Phyllodoce* during the impact period are puzzling because this genus is thought to be associated with clean water (E. Ruff, personal communication). Sipunculids (*Phascolosoma* and *Themiste*) are deposit feeding infaunal organisms (Rice, 1980). Little, if anything, is known about recolonization of these species following disturbances or their response following an oil spill.

Increases in the abundance of most crustose algae, such as *Ralfsia*, during the impact period were probably partially an artifact of sampling. High population measurements may have resulted from the removal of the *Fucus gardneri* overstory that normally masks the underlying algal layers that adhere close to the substratum. The brown alga, *Elachista fucicola* is an epiphyte typically found on *Fucus* (Abbott and Hollenberg 1976). Consequently, the abundance of this species would be expected to correlate with the abundance of *Fucus*, and the reasons for its higher abundance during the impact period are unclear. It is possible that the increased frequency of *Elachista* during the impact period occurred because *Fucus* tissues were less resistant to colonization by epiphytes because of spill-related stresses. Algal epiphytes often increase on senescent or stressed host tissues.

Estimates of Biological Variability

Estimates of biological variability are needed to compute sample-sizes that are required to achieve the goals of monitoring; namely, the ability to detect effects of a certain size with a specific level of confidence. These variability estimates must be reasonably representative of the actual intertidal communities to be sampled. Preferably, site-specific variability estimates would be established by conducting a pilot study on taxa of interest. While a pilot study may be feasible in the case of an experimental investigation, it is rarely practical immediately after an accidental oil spill, when a comprehensive impact assessment must be designed and executed quickly.

In the absence of a pilot study, estimates of the variability in the intertidal taxa sampled in other regions can act as a surrogate if those estimates are reasonably representative of the biological variability in the region of interest. Variance estimates (σ^2) can be used to compute the noncentrality parameter Φ in Equation B.7, which, in turn, is used to compute the sample-size estimates. However, variance consistently increases with increasing abundance (Figure 2.1a), which makes it difficult to establish a variability estimate (and sample size) representative of a wide range of taxa. This limitation can be partially resolved by recasting Equation B.7 in terms of coefficients of variation (CV) as shown in Equation B.8. CVs are much more stable and a single CV estimate is more representative of a large number of taxa covering a wider range of population sizes.

This is demonstrated by the within-site and between-site CVs plotted in Figure 2.1b and Figure 2.1c. The CVs exhibit only weak trends across the five orders-of-magnitude range in population sizes. The minor CV trends can be categorized into three major abundance ranges. Sparse taxa tend to have lower between-site variability (CV_B in Figure 2.1c) while abundant taxa tend to have lower within-site variability (CV_W in Figure 2.1b). These abundance ranges were based on the sampling units used in the PWS intertidal monitoring program. Infauna were collected using a 15-cm long core that covered a 0.009-m^2 area while epibiota were enumerated within a 0.25-m^2 quadrat. Within these sampling units, sparse taxa were considered to have an average count or percent cover that was less than 0.07, while abundant taxa had average densities that exceeded 3. A further discussion of the definition of sparse taxa is presented in Appendix A, and the basis for the thresholds is described below.

Variability ranges were characterized by the 10th, 50th (median), and 90th percentiles for the within- and between-site CVs as shown by the dashed lines in Figure 2.1bc. These percentiles were chosen to represent low, moderate, and high levels of variability that might be expected when sampling the intertidal environment. An extensive set of power curves based on the CVs for these percentiles are presented in Appendix D and are discussed in subsequent subsections of this chapter.

Table 2.3 shows that the between-site CVs were consistently lower than the within-site CVs. The reduced variability at large spatial scales has direct bearing on the optimal number of replicate samples that should be collected at each site, relative to the total number of sites that need to be sampled. Compared to the difference in variability at large and small spatial scales, the differences between CVs computed from data collected during impact (1989-1991) and non-impact (1993-2000) periods were relatively minor (Table 2.3). Figure 2.1b and Figure 2.1c show the relative consistency in the CV distribution among intertidal communities that were impacted by oil exposure (open circles ○ and squares □) and the intertidal community sampled after recovery (solid circles ● and squares ■). These CVs are relatively consistent despite the absence of many taxa during the impact period. Variance and CVs were indeterminate for a large number of singleton or missing taxa during the impact period. A comparison of the abundance and associated CVs for the individual PWS taxa and tidal levels

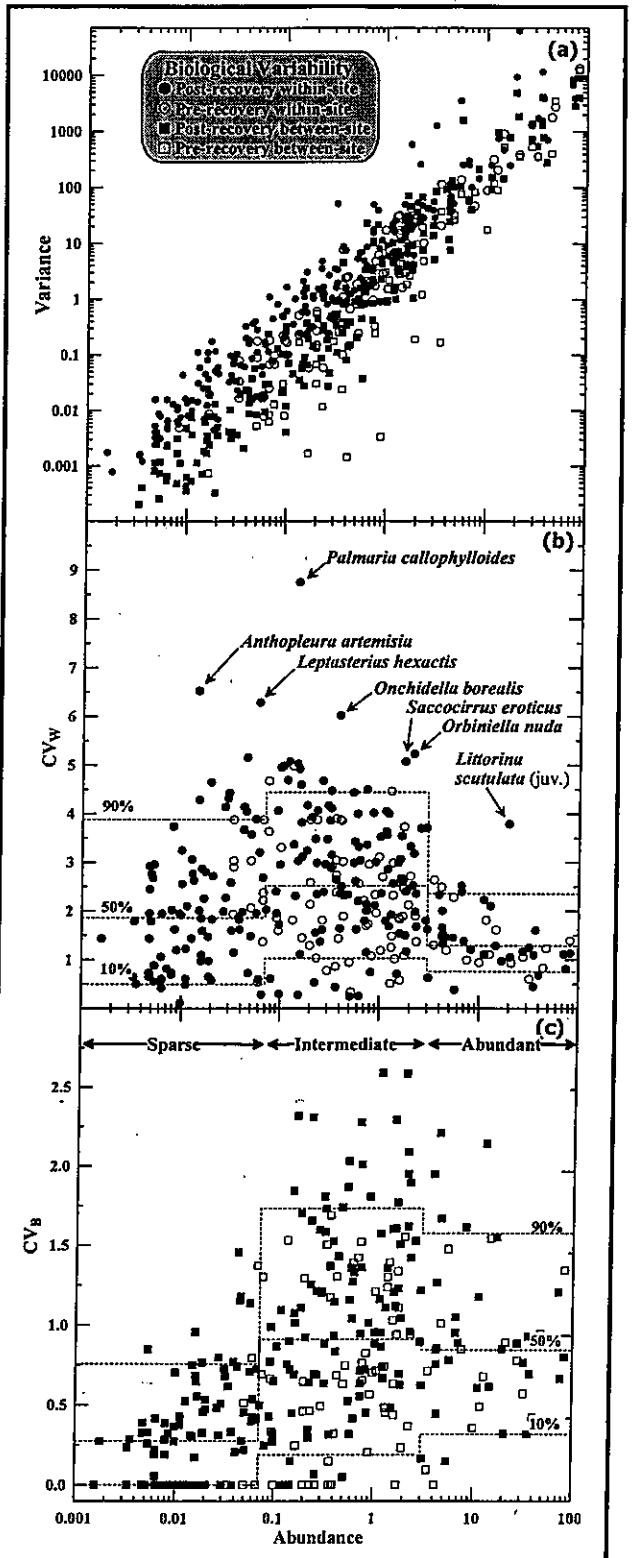

Figure 2.1. Variability as a function of population size for PWS intertidal taxa

Table 2.3. Summary of CVs by abundance and period

	Period	# Taxa	$\hat{e}V_W$ 10%	$\hat{e}V_W$ 50%	$\hat{e}V_W$ 90%	$\hat{e}V_B$ 10%	$\hat{e}V_B$ 50%	$\hat{e}V_B$ 90%
Sparse	Impact	19	0.0	1.9	3.9	0.0	0.0	0.6
	Non-Impact	87	0.6	1.8	4.1	0.0	0.3	0.8
	Pooled	**106**	**0.5**	**1.9**	**3.9**	**0.0**	**0.3**	**0.8**
Intermediate	Impact	60	0.9	2.0	3.9	0.0	0.7	1.4
	Non-Impact	108	1.2	2.9	4.6	0.3	1.0	1.9
	Pooled	**168**	**1.0**	**2.5**	**4.4**	**0.2**	**0.9**	**1.7**
Abundant	Impact	16	0.9	1.2	2.4	0.2	0.7	1.4
	Non-Impact	31	0.7	1.3	2.3	0.3	0.8	1.7
	Pooled	**47**	**0.8**	**1.3**	**2.4**	**0.3**	**0.8**	**1.6**

during impact and non-impact periods are also presented numerically and graphically in Appendix C.

As with the CV consistency between impact periods, the CVs associated with individual assemblages at different tidal elevations exhibited only minor differences (Table 2.4). Compared to the relatively large difference between the within-site and between-site CVs, variability tended to be distributed uniformly across the three assemblages (algae, epifauna, and infauna), three tidal elevations (low, middle, high), and two different measurement types (counts and percent cover). Only epibiotic percent cover at the middle intertidal level exhibited slightly elevated within-site variability at the median and 90th percentiles $\hat{e}V_W$. Nonetheless, the CVs for most individual taxa are well represented by the pooled CVs listed in Table 2.3 and shown by the dashed lines in Figure 2.1bc. Consequently, sample-size calculations based on these levels of variability should be applicable to all but a few of the intertidal taxa encountered in PWS.

Table 2.4. CV distribution by assemblage and tidal elevation

			$\hat{CV}_W$ [a]				$\hat{CV}_B$ [b]			
Type	Level	Measure	# Taxa	10%	50%	90%	# Taxa	10%	50%	90%
Algae	Upper	%	29	0.6	2.1	3.8	12	0.0	0.4	1.2
Algae	Middle	%	73	0.6	2.3	4.1	46	0.3	0.7	1.4
Epifauna	Upper	%	17	1.0	2.3	3.9	11	0.4	1.1	1.7
Epifauna	Middle	%	19	0.9	3.3	4.5	19	0.6	1.0	1.6
Epifauna	Upper	#	18	0.7	2.5	4.2	19	0.1	0.9	1.3
Epifauna	Middle	#	44	0.6	2.1	4.3	30	0.3	0.8	1.5
Infauna	Lower	#	74	0.2	2.2	4.1	78	0.2	1.0	2.0

[a] Sparse and Intermediate Taxa [b] Intermediate and Abundant Taxa

The only exceptions are seven species with unusually high $\hat{CV}_W$. These outliers are clearly evident in Figure 2.1b and are listed in Table 2.5. Even sample-size recommendations based on high-variability ($\hat{CV}_W \approx 4$ for sparse and intermediate abundance and $\hat{CV}_W \approx 2.4$ for abundant taxa) will markedly underestimate the number of samples required to discern effects on these seven species.

Table 2.5. Species with the anomalously high within-site variability

Name	Type	Level	Measure	Mean	CVw	$\frac{1}{K}$ [a]
Palmaria callophylloides	Algae	Middle	%	0.15	8.7	70
Onchidella borealis	Epifauna	Middle	#	0.39	6.0	34
Orbiniella nuda	Infauna	Lower	#	2.15	5.2	27
Saccocirrus eroticus	Infauna	Lower	#	1.75	5.1	25
Leptasterias hexactis	Epifauna	Middle	#	0.06	6.3	23
Littorina scutulata (Juv.)	Epifauna	Upper	#	19.5	3.8	14
Anthopleura artemisia	Epifauna	Middle	#	0.01	6.5	-27

[a] Index of dispersion from Equation 1.4 where increasing values indicate an increased level of clumping

Although an inordinately high $\hat{CV}_W$ excludes only a few species from the sample-size calculations, insight into the reasons for their high variability can shed light on characteristics that would tend to underestimate sample sizes for specific taxa in other intertidal regions. Field biologists should not rely on the sample-size recommendations in this report if an intertidal monitoring program is being designed in a region where many taxa have these high-variance characteristics. Instead, a pilot study should be conducted to determine site-specific variability, and the power curves should be recomputed using the techniques described in Appendix B.

One way to investigate the reasons behind the abnormally high CVs in these selected species is to evaluate their degree of clumping. For most taxa, the overall stability of the CVs with respect

to abundance suggests that the populations of intertidal organisms tend to be distributed in a log-normal fashion. As described in Chapter 1, a log-normal distribution approximates a negative binomial distribution commonly generated by contagious or clumped populations. The dispersion index $\frac{1}{K}$ in Equation 1.4 the represents the level of clumping and can be computed from mean abundance and $\in V_W$. The distribution of most invertebrate taxa tends to be clumped with a dispersion index exceeding zero. A positive dispersion index is indicative of a distribution approximating a negative binomial distribution (Elliot, 1977).

In contrast, strongly negative clumping indices reflect randomly distributed populations that are best represented by a Poisson distribution. If the quadrat or core size is much smaller than the size of clumps, the population will be undersampled, and the perceived density distribution will be random (Elliot, 1977). Figure 2.2 shows for the PWS dataset that taxa with abundances less than approximately 0.07 had strongly negative clumping indices. The abundance threshold for sparse taxa was established at 0.07 because the population distributions of most taxa with lower densities appear to be poorly defined by size of the sampling units used in the PWS monitoring program. Nevertheless, a few of the taxa considered to have intermediate abundances also appear to have been slightly undersampled by virtue of their slightly-negative clumping indices in

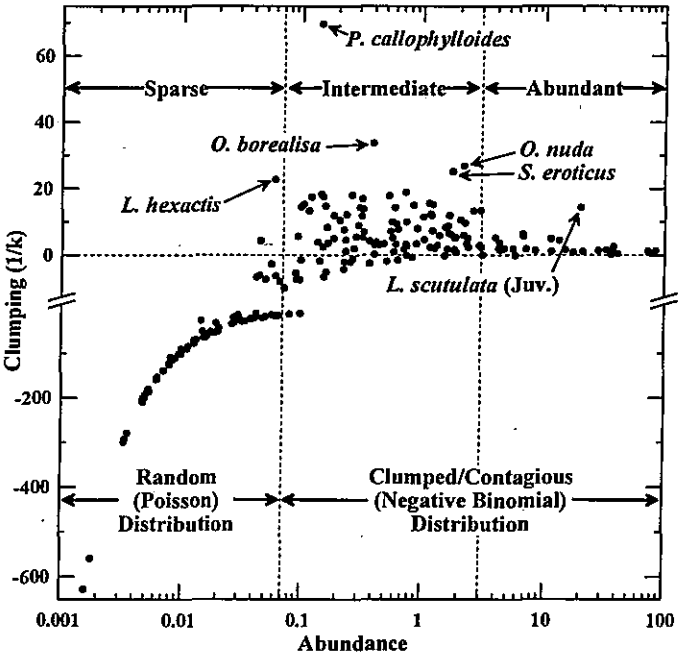

Figure 2.2. Clumping as a function of population size for PWS intertidal taxa.

Figure 2.2. However, for all taxa with mean counts above 3.0, or 3% cover, populations appear to be well-resolved and closely approximate clumped or contagious distributions.

Two of the taxa with an unusually high $\hat{V}_W$ in Table 2.5 were sparsely populated and their elevated variability probably resulted from undersampling. Specifically, the high CV associated with the Moonglow Anemone (*A. Artemisia*) was clearly an artifact of its paucity. This is reflected in its negative clumping index ($\frac{1}{K} = -27$). Similarly, the anomalously high $\hat{V}_W$ associated with the juvenile Checkered Periwinkles (*L. scutulata*) at the middle-intertidal elevation was probably an artifact of sampling, even though their mean abundance of 19.5 and clumping index of 14 were both moderately high. It is likely that their perceived clumping arose as a result of enumeration inconsistencies in the field. If some biologists did not distinguish juveniles (lumping them with the adult populations) while others enumerated juveniles, this inconsistency in identification would artificially increase the apparent patchiness of the juvenile specimens. This was probably the case because juvenile *Littorina scutulata* at the upper tidal elevation, as well as adults at both elevations, did not exhibit an unusually high variability, even though their mean populations were comparable to the juveniles at the middle intertidal elevation. *L. scutulata* has a planktonic larval stage (Behrens Yamada, 1989), and differential settlement could have contributed to its extreme within-site variability.

The remaining five species had dispersion indices that were the highest measured for any of the 270 PWS taxa. Their high variability was caused by a naturally-occurring tendency to form dense clusters or clumps. Characteristics that are often common to species that have an increased tendency to clump include a relatively small size, a proclivity to congregate in crevices or in other microhabitats, and brooding of large clutches to an advanced stage of development before release as crawl-away juveniles. For example, the Six-Rayed Star (*L. hexactis*) is a small, carnivorous sea star that is found in crevices and under rocks. Breeding clusters of a dozen or more stars form under rocks where they brood their young for over a month until they are fully formed. Thus, this seastar had a high clumping index of 23 despite its rare occurrence in the PWS dataset.

The remaining four species with high clumping indices also had an amplified tendency to congregate. Both the highest variability and highest level of clumping was associated with the frilly red ribbon alga *P. callophylloides*. This foliose red alga forms dense patches over relatively large areas in the intertidal zone but can be virtually absent in adjacent areas where rockweed (*Fucus gardneri*) is prevalent. Although the mutual exclusion of these algae may be related to intense competition for light and space, anecdotal observations by the authors of this report

indicate that there were some sites where *Fucus* was absent and *Palmaria* did not appear to be adversely affected by the spill. *Palmaria* is usually found in the lower intertidal zones but appeared to extend its range to higher elevations when *Fucus* was absent. Some of the elevated *Palmaria* variability may have also been an artifact of variations in transect-line elevation. In the middle intertidal zone, transects spanned a comparatively wide elevation range from 3 to 8 ft above mean lower low water.

The three remaining highly-clumped species include the Leather Limpet sea slug (*O. borealis*). It is a herbivore that tends to congregate in rocky crevices and near the holdfasts of seaweed, which could explain its high level of clumping. The polychaete worms *Saccocirrus eroticus* and *Orbiniella nuda* had a similar mean abundance and patchiness. These two polychaetes are not ubiquitously present in sediment cores because they tend to occur in patches within coarse, sandy sediments (Gene Ruff, personal communication).

Power Analysis

Except for the few species with an unusually high degree of clumping, sample-size charts constructed using the pooled CVs presented in Table 2.6 should cover most sampling-design situations in intertidal areas similar to PWS. Power curves are provided in Appendix D for taxa that are sparse, intermediate, and abundant and for taxa that have low, moderate, and high variability within each of the three abundance ranges. The three levels of intertidal variability were estimated from the 10th, 50th (median), and 90th percentiles of the distributions of within-site and between-site CVs shown in Figure 2.1bc. Two sample-size diagrams are included for each of the nine combinations of abundance and variability. They indicate the power needed to detect two different magnitudes of change. The upper plots in Figures D.1a through D.9a show power curves for detecting smaller changes in abundance $(-0.15 \le \Delta \le -0.75)$. The lower plots in Figures D.1b through D.9b show power curves for detecting larger changes $(-0.25 \le \Delta \le -0.8\overline{3})$. The size of the detectable change was different for some of the combinations of variability and abundance because the CVs were different. This was done to plot relatively high power curves $(1-\beta \ge 0.7)$ within a tractable range of replicate samples $(m \le 25)$ and sites $(n \le 25)$.

Table 2.6. Summary of CVs used in the sample-size calculations

			Low Variability (10th Percentile)		Moderate Variability (50th Percentile)		High Variability (90th Percentile)	
Abundance	μ	Clumping	CV_W	CV_B	CV_W	CV_B	CV_W	CV_B
Sparse	$\hat{\mu} < 0.07$	$\frac{1}{K} \sqsupset 0$	0.49	0.00	1.86	0.27	3.88	0.76
Intermediate	$0.07 < \hat{\mu} < 3$	$\frac{1}{K} \approx 0$	1.03	0.19	2.52	0.91	4.44	1.73
Abundant	$3 < \hat{\mu}$	$\frac{1}{K} > 0$	0.76	0.32	1.28	0.84	2.35	1.57

In practice, effects on individual taxa are rarely of primary interest. Exceptions might include a few taxa that are very abundant, of commercial value, or that may be environmentally sensitive, threatened, or endangered. Usually however, widespread effects on the major intertidal assemblages receive the most attention in monitoring programs. Consequently, the sample-size plots that pertain to abundant taxa are discussed in more detail here. If an individual taxon is of interest, the sample-size plots for other abundance categories shown in Appendix D can be used.

Figure 2.3 reproduces the power curves for abundant taxa with moderate variability shown in Figure D.8a. Moderate variability ($CV_W = 1.28$, $CV_B = 0.84$) is typical of most of abundant taxa. For PWS monitoring data, the abundant infaunal taxa consisted of:

- Total infaunal organisms;
- Major taxonomic groups (mollusks, annelids, crustaceans, and nemertean ribbon worms); and
- Prevalent taxonomic aggregates whose species were commonly found in most core samples (gastropods of the *Cingula* and *Fartulum* genera, bivalves in the Montacutidae family, and polychaete worms in the Sigalionidae, and Phyllodocidae families).

Abundant epibiota consisted of:

- Total percent cover of algae and invertebrates, and total invertebrate counts; and
- Rockweed (*Fucus gardneri*), Lichens (*Verrucaria*), Limpets (Lottiidae), Hermit crabs (*Pagurus hirsutiusculus*), Lung Snails (*Siphonaria thersites*), mussels (*Mytilus*), Checkered and Sitka Periwinkles (*Littorina scutulata* and *L. sitkana*), and barnacles (*Semibalanus balanoides* and *Chthamalus dalli*). Although lichens (*Verrucaria*) may account for significant cover, they rarely constitute significant biomass and their identification and quantification vary markedly among different observers.

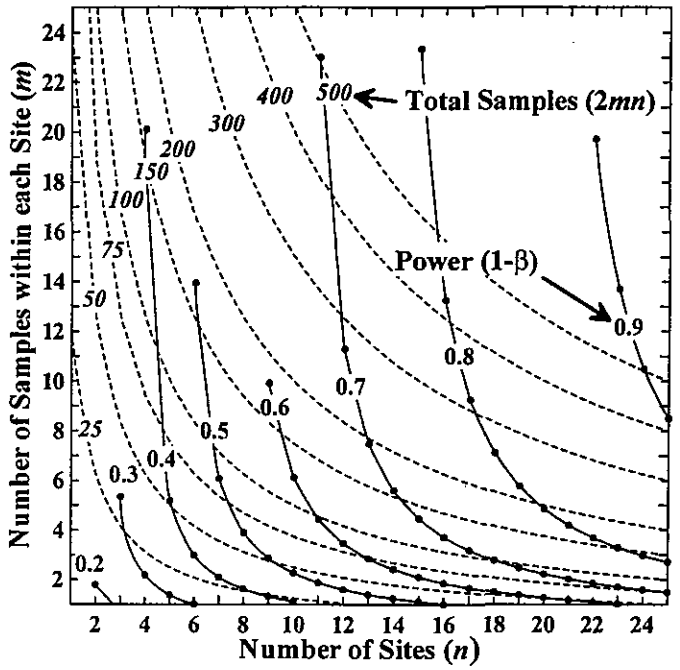

Figure 2.3. Sample-size chart showing the number of replicate samples (m) collected at n sites per treatment that are needed to detect a 50% reduction (100% increase) in abundant intertidal populations with a statistical power ($1-\beta$) at the one-tailed significance level of $\alpha = 0.1$. The solid curves correspond to different levels of statistical power in an environment with moderate natural biological variation ($CV_W = 1.28$, $CV_B = 0.84$). The dashed curves correspond to the total number of samples to be collected when comparing two treatments.

The shape of the power curves for abundant taxa in Figure 2.3 shows that above a certain point, adding replicate samples within sites has little effect on statistical power. In fact, most of the power curves shown in Figures D.7 through D.9 for abundant taxa approach a vertical asymptote above $m \approx 6$ and are distinctly vertical above $m = 8$. The shape of these curves suggests that if sampling is planned at three or more sites subjected to a given treatment, then at least four replicate samples should be collected at each site. However, there is no additional statistical benefit that results from collecting more than eight replicate samples. Similarly, the power curves start to approach a horizontal asymptote below $m = 4$ and are distinctly horizontal below $m = 2$. Consequently, adding additional sites when only one or two replicate samples are being collected at each site does little to enhance statistical power. Instead, sampling resources should be directed at increasing the number of replicate samples. In general, in the design of an intertidal monitoring program to address effects on abundant taxa, approximately six replicate samples should be collected at each site, and any remaining sampling resources should be directed at sampling additional sites.

This recommendation for optimal replicate sample sizes cannot be generalized to assessments of sparse taxa and taxa with intermediate abundance. The families of hyperbolae that form the power curves tend to be less eccentric as abundance decreases. This is evident from a comparison of Figures D.2b, D.5a, and D.8a, which present power curves for detecting 50% reductions in taxa with intermediate variance in three abundance categories. For sparse and moderately abundant taxa, the number of optimal replicate samples varies widely depending on the desired power and number of sites.

The curves shown by the solid lines in Figure 2.3 correspond to various powers to detect a 50% reduction in abundance ($\Delta = -0.5$). Sample sizes were determined at a statistical significance level of $\alpha = 0.1$ for a one-tailed distribution. This corresponds to a 10% risk of a false positive, or a 1-in-10 chance that a reduction in abundance equal to Δ would be found by the monitoring program or experiment, when in fact, effects were negligible. A one-tailed significance level is used because oil-spill impacts to intertidal taxa almost always result in a reduction in their populations. As discussed in previous sections of this chapter, few of the taxa enumerated in the PWS monitoring program had higher impacted populations, and most of those could be ascribed to sampling uncertainty rather than opportunism or spill-related reductions in predation. If the sample-size charts were applied to intertidal organisms that included a large number of potentially opportunistic taxa for which the treatment effects could increase or decrease populations, then the false-positive error rate would double ($\alpha = 0.2$). Accordingly, the sample-size chart shown in Figure 2.3 would also represent the ability to detect a doubling in population ($\Delta = +1$).

Because of how power was formulated in Appendix B, the number of sites or beaches designated by n on the horizontal axis refers to the number of sites within one treatment category. The number of sites within each of two treatment categories is assumed to be equal, so the total number of sites that need to be surveyed to achieve the designated power is actually $2n$. The total number of samples to be collected in this balanced design is then given by $2nm$. The dashed curves in Figure 2.3 correspond to total sample sizes for various combinations of m and n.

Total sample size is an important consideration for infaunal samples where a significant portion of the sampling effort occurs during taxonomic identification in the laboratory after the samples are collected and sieved. Thus in Figure 2.3, collecting 5 replicate samples at 5 impact sites and 5 reference sites (50 samples total) provides approximately the same power (0.4) as collecting 20 samples at each site (150 samples total) with triple the analysis effort. If an analysis budget for 150 samples were available, then a sampling design with 5 replicate samples collected at 14

treated sites and 14 reference sites (140 total samples) would yield a much higher power (0.7) to detect change. This assumes that 14 oiled and reference beaches would be available for sampling and that enough field-sampling resources could be applied to collect the samples at so many different beaches. In contrast, total sample size for epibiotic enumeration, which is largely done in the field, is not always the most important consideration. In the epibiotic case, the ability to mobilize survey teams to multiple sites within narrow tidal windows is often more of a limiting factor. Everything else being equal however, this exercise shows that there is little statistical advantage in increasing the number of replicate samples (m) above 8 if effects on abundant taxa are the focus of the monitoring program. In that case, resources should instead be directed toward sampling at additional beach sites, if they are available.

Example Application

The sample-size charts show the number of replicate samples (m) that need to be collected at n sites per treatment to detect a specific reduction in intertidal populations. Consider the following scenario as an example of how the sample-size charts might be used to determine the number of replicate samples that must be collected within each site.

Suppose a field biologist needed to determine whether the infaunal mollusk population had been negatively impacted by a particular cleanup method that was applied to remove oil from 15 beaches. Further suppose that 15 other oiled beaches were available for survey that were not subjected to the same treatment method and that the biologist knew or assumed that these other beaches had mollusk populations similar to those of the 15 treated beaches prior to cleanup. This establishes the number of treated sites at $n = 15$. Finally, suppose that the various stakeholders agree that reductions in mollusk populations of less than 50% ($\Delta = -0.5$) are not important and that they are willing to risk missing a change this large 30% of the time ($\beta = 0.3$). This corresponds to a power ($1 - \beta$) of 0.7, or a 7-in-10 chance of detecting a 50% reduction in abundance, if in fact such a change had occurred. They further agree that the risk of incorrectly finding a change this large should only occur only 1-in-10 times ($\alpha = 0.1$).

The variability in the infaunal mollusk population within PWS ($CV_W = 1.1$, $CV_B = 0.9$) in Figure C.7) was comparable to moderate variability in abundant taxa. Consequently, Figure 2.3 provides a reasonable estimate of detection power and indicates that the goals of the monitoring program could be achieved with 5 replicate samples collected at each site. Collecting more replicate samples at each site would be relatively unproductive insofar as improving the power to detect change. If the stakeholders required more stringent error rates or there were fewer experimental

sites available, then it would be incumbent upon the field biologist to advise them that the monitoring goals could not be achieved without relaxing the detection limit ($\Delta = -0.5$). For example, from Figure D.8b, detection of a 67% reduction ($\Delta = -0.6\overline{6}$) could be achieved by collecting 5 replicate samples at only 6 treated and 6 reference sites.

This example shows that even with a rather intensive field program, involving monitoring at 30 sites, only large effects can be detected with any degree of confidence or power. Figure C.7 shows that for the *Exxon Valdez* spill, the observed difference in infaunal mollusk abundance between impacted and non-impacted periods was actually relatively small ($\Delta = -0.27$). Thus, detection of effects on mollusk populations from a major spill would be difficult without sampling at many sites. Just to achieve a power of 0.5 for which the impact would be missed half the time, would require sampling at 60 or more sites to detect a 27% reduction in mollusk abundance.

Under the rudimentary statistical design described in this section, effects on mollusks would be difficult to detect without intensive monitoring. However, examination of abundance comparisons plotted in Appendix C shows that many other taxa exhibited much larger reductions due to the effects of oiling. Effects on those taxa would be easy to detect with high level of confidence from the analysis of just a few samples. Also, in the case of a field experiment, the application of different statistical designs could be used to reduce error variance and achieve greater power with fewer samples. For example, the experiments presently being conducted by NOAA include temporal sampling at a number of paired plots where one of the plots was been randomly selected for treatment.

Nevertheless, the example highlights another difficulty with a commonly applied technique for determining recovery after an oil spill. As described at the beginning of this chapter, two-sample *t*-tests are used to assess recovery in intertidal populations by directly comparing impact and control populations at a particular point in time. If the populations are not found to be statistically different, then the populations are assumed to have recovered. However, even if impact and control sites were identical prior to the spill, the example demonstrates that the power to detect small differences is very low unless many sites are monitored. Alternatively, if a large difference is used as the recovery threshold, then intertidal populations may be prematurely deemed to have recovered. Given a realistic sampling effort (60 sites), infaunal mollusks in PWS would be found to have never been impacted, or to have recovered immediately after the spill, before any samples were collected.

As described in the following section and in subsequent chapters, the power to detect effects can be improved through the used of multivariate community indices and through application of alternative statistical designs that take advantage of long-term monitoring to examine the stability of intertidal communities over time.

Community Response

As discussed in the previous section, oil-spill assessments are often primarily concerned with effects on intertidal communities as a whole, although a few individual species may also be of interest because of their economic, societal, or environmental value. However, investigating changes to global community properties, such as total abundance or various diversity indices, dilutes the value of information contained in the response of individual taxa. For example, a dominant taxon may show little change in response to oil exposure while a less abundant taxon may exhibit marked reductions or disappear altogether. Under those circumstances, total abundance may only exhibit a weak response to oil impacts even though the community structure was conspicuously altered.

A variety of strategies have been used to compare intertidal community structures by overcoming obstacles such as how to reduce the influence of rare species and how to simultaneously analyze tens, and sometimes hundreds of different individual taxa. For example, Page *et al.* (1995) limited their analysis to species present in 20% or more of the PWS intertidal samples in their assessment of covariance with grain size, total organic carbon, and wave energy. Gilfillan *et al.* (1995) eliminated all but those species that occurred in 20% or more of the PWS intertidal samples within a habitat or tidal elevation prior to performing univariate hypothesis tests and a canonical correspondence analysis (CCA). This approach eliminated all but approximately 10% of the species for univariate tests, and all but approximately 20% of the species for the CCA. In cases where univariate t-tests or ANOVAs are performed on multiple taxa, application of Bonferroni or some other correction is necessary to control the overall experiment-wise error rate (Sokal and Rholf, 1997: p. 240). Determining which species to include or eliminate can be controversial. Rare species can be problematic to deal with in statistical analysis but biologists debate whether it is advisable to exclude any taxa that may be sensitive to impacts, even if they are few in number.

Multivariate analysis, which simultaneously examines changes in a large number of variables, usually provides a far superior measure of community structure. It distills pertinent information about community structure into a limited number of parameters by reducing redundant information introduced by species whose responses are highly correlated. However, multivariate

analysis in ecological applications is often exploratory or descriptive rather than inferential. One reason for the preponderance of descriptive studies is that field studies typically collect unwieldy amounts of data per sample, such as species abundance and percent cover, along with numerous environmental characteristics. The overparameterization of information then forces investigators to use *post hoc* methods such as cluster analysis, principal components, or correspondence analysis to summarize and categorize the sample observations. Statistical tests often rely on Monte Carlo Permutations to determine the significance of impacts (Coats *et al.*, 1999; pg A-2). Unfortunately, these permutation analyses provide little prognosticative insight into the sample sizes needed to achieve specific power levels. Although many of these *post hoc* methods seek to detect changes in community composition or differences in habitat structure, they are hampered by a lack of formal statistical methods to test hypotheses on multivariate parameters.

This section presents a formal statistical test for community change using the results of principal component or correspondence analysis. These statistical tests can be generalized to a variety of experimental designs and can simultaneously analyze multiple dimensions of a principal components analysis. More importantly, their noncentral distributions can be used in the design of experiments. The tests are based on an analog to analysis of variance (ANOVA). The multivariate tests are characterized by an analysis of distance (ANODIS) where distance refers to separations among individual samples in a multivariate ordination plot. For example, ANODIS can be used to test for differences in community composition by using the location of sample observations in 1, 2, or more dimensions of a principal component analysis (PCA). The statistical tests of significance are based on F-statistics adapted for the analysis of multidimensional data. This parametric approach to data analysis readily permits power and sample-size calculations that are useful in the design of field studies. The statistical formulation of ANODIS is presented in the second section of Appendix B.

Measuring Differences in Community Composition

ANODIS can be used to test whether intertidal communities within samples collected at treated and reference sites could have come from a common multivariate distribution. If the multivariate sample scores from treated and reference sites cannot be distinguished from one another, then the intertidal biota may not have been affected by the treatments, or the statistical power was too weak to discern the small differences between the communities at the two sites. The test can also be applied in experiments or assessments where there are more than two treatment groups, for example, to test for impacts from invasive cleanup methods in addition to impacts from oil.

As described in Appendix B, relevant differences in community composition at sites subjected to different treatments can be measured in terms of a separation index C (Equation B.26). The index measures the separation between the mean community composition at treatment and reference sites within an ordination diagram such as that shown in Figure 2.4. It measures the separation in terms of the number of standard deviations determined from the scatter of observations around each mean. This formulation is convenient for sample-size determinations because test statistics are typically standardized in a similar manner (*cf.* Equations B.6, B.7, and B.8).

For example, in the PCA performed on PWS infaunal data collected in 1991 (Figure 2.4a), the three Category-3 sites that were impacted by invasive cleanup procedures had PCA y-axis values near or above zero as shown by the ■ symbols. Their mean location is indicated by the convergence of the three solid lines. The length of these individual lines is indicative of the within-treatment variability. A similar pattern is presented by the PCA sample scores (indicated by ◇ symbols) for Category-1 sites that were not exposed to oil nor subjected to any treatment. All three of the Category-1 (reference) sites had infaunal communities characterized by negative y-axis values. Again the variability among reference sites is indicated by the length of the dashed lines. The combined within-treatment standard deviation for this ordination was $\hat{\sigma} = 0.45$ while the separation between the means was $D_{\mu_1-\mu_2} = 0.57$. The ratio ($C = 1.25$) indicates that the two means were separated by a slightly more than one standard deviation.

The PCA shown in Figure 2.4b was conducted on infaunal samples collected in 1994 after the period of marked repopulation within the intertidal zone. Sample scores from treated and reference sites overlap and a separation in the means is not visually apparent. The within-treatment standard deviation remained about the same as the 1991 data ($\hat{\sigma} = 0.46$), while the separation in treatment means was negligible ($D_{\mu_1-\mu_2} = 0.15$). This distance constituted only a fraction of a standard deviation ($C = 0.33$). The pattern of intermixed sample scores in 1994 is consistent with a lack of apparent effects. The p-value for this separation is 0.92, indicating that there is little justification for rejecting the null hypothesis of no effects.

Power Analysis

Despite the visually apparent separation in Figure 2.4a, the small sample size of $n = 3$ sites did

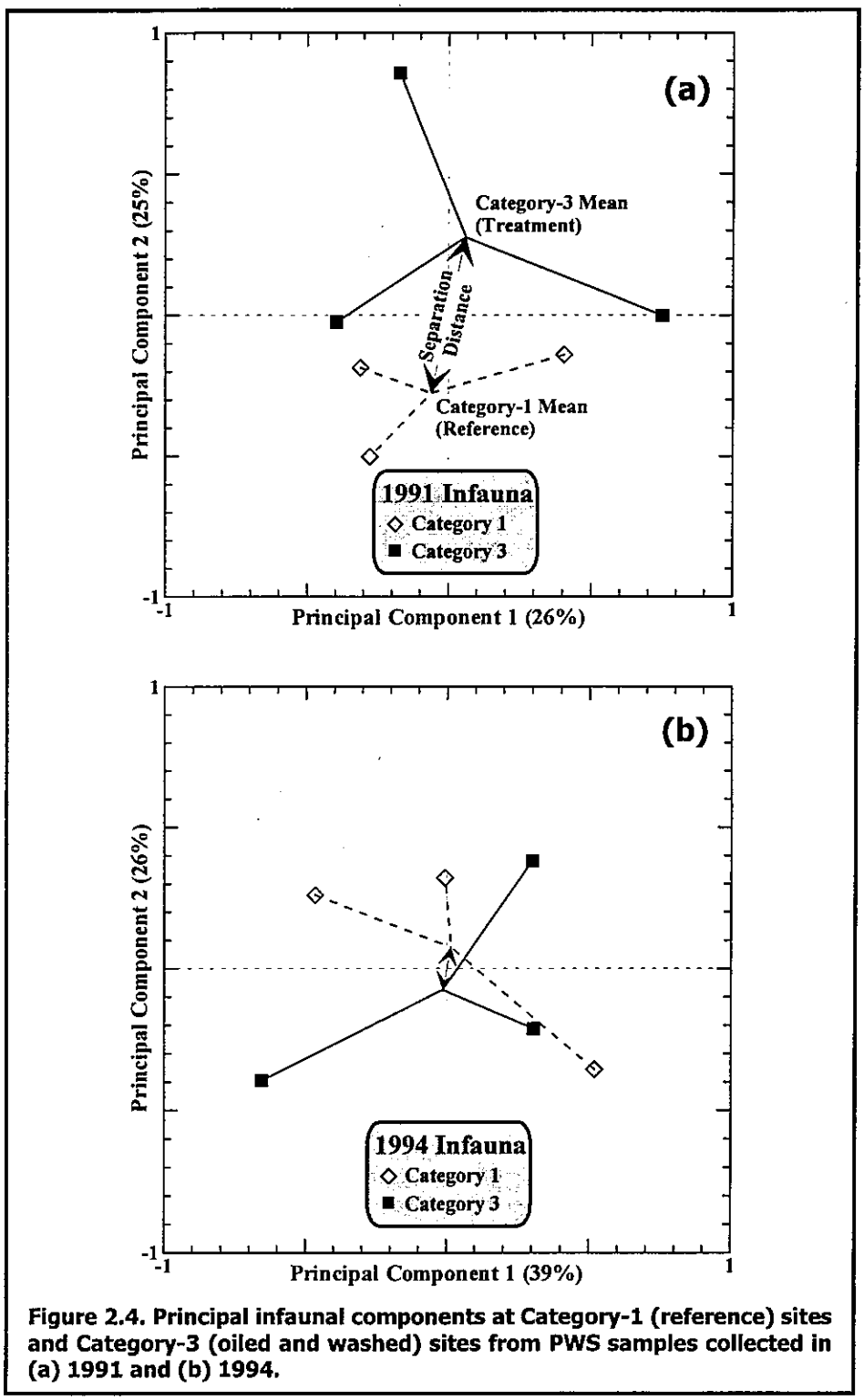

Figure 2.4. Principal infaunal components at Category-1 (reference) sites and Category-3 (oiled and washed) sites from PWS samples collected in (a) 1991 and (b) 1994.

not provide an adequate level of power to detect community changes of magnitude $C = 1.25$. This is evident from the power curves computed from the noncentral parameter in Equation B.27. Power curves are plotted for 1, 2, and 3 ordination axes in Figure 2.5. Figure 2.5b applies to the two-dimensional ordination shown in Figure 2.4a. With $n = 3$ sites and a separation index of $C = 1.25$, Figure 2.5b shows that the power to detect a change of this magnitude is only $1 - \beta = 0.2$. This means the there is only a 20% chance of correctly discerning a change of this magnitude. Ten treatment and ten reference sites would be required to achieve a marginal power of $1 - \beta = 0.5$ for detecting separations of this magnitude in the presence of the inherent scatter in infaunal community structure.

The power of the ordination tests could be improved through the use of a correspondence analysis (CA) rather than PCA. Although it is not evident in Figure 2.4a, the distribution of sample scores was distorted because few species were shared between the treatment and reference sites. This artificially distorted the distribution of PCA sample scores in the shape of a horseshoe (Coats *et al.*, 1999; pg A-2). CA tends to reduce the severity of this horseshoe effect and provides a more suitable measure of separation distance. Nevertheless, separation indices are likely to be similar under most types of ordination analyses. For example, increasing the number of PCA axes is unlikely to increase power unless the additional axes reveal large separation distances between the means. Specifically, a comparison of the sample-size diagrams shown in Figure 2.5 shows that for the same separation size, a larger number of samples (n) are required to achieve the same power when the number of dimensions increases.

The community-composition analyses that are described above show similar results to the sample-size computations performed on the intertidal abundance of individual taxa that were described in the previous section. Namely, a comparison of treatment and reference intertidal populations within any given year is likely to yield low power unless samples are collected at a large number of sites (>10) or unless large effect sizes are being tested.

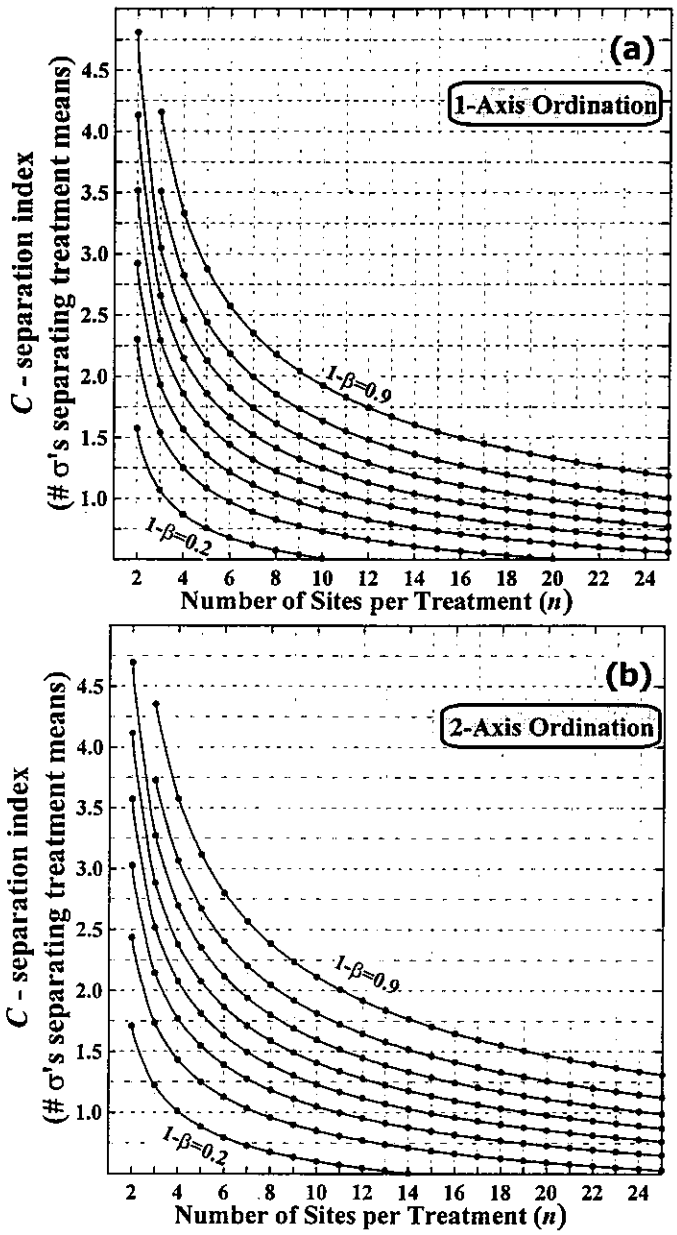

Figure 2.5. Sample-size chart showing the number of sites (n) per treatment that are needed to detect a treatment effect that causes a separation of C in the treatment means in a (a) one-dimensional, (b) two-dimensional, or (c) three-dimensional ordination.

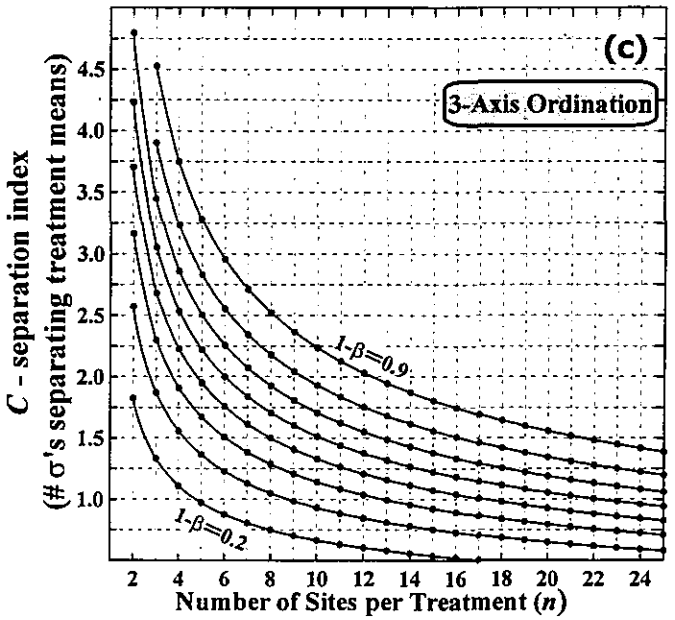

Figure 2.5. Continued.

CHAPTER 3. RECOVERY

Because of the unexpected nature of an oil spill, there is often little opportunity to establish paired plots or collect adequate baseline data prior to impingement of the spill onto the shoreline. In their absence, tests of impact and recovery can be based on a comparison of temporal trends in mean abundance at reference and impact sites. This chapter examines the sample sizes that would be required to detect recovery of intertidal populations from acute impacts caused by hydrocarbon exposure, invasive cleanup methods, or other localized disturbance. The statistical formulation is based on a contrast of two or more years of sampling at a number of sites within and beyond the *Exxon Valdez* spill zone (Skalski and Robson, 1992; Skalski *et al*, 2001). This constitutes a test for parallelism in the time histories of populations at impacted and reference sites.

Impact Size

Under the null hypothesis of no impact, or after impacted sites have recovered, temporal trends at control and treatment sites would track or parallel one another over time. Although abundance may fluctuate widely from year to year due to regional climatic influences, these large-scale influences would probably affect reference and impact sites in a similar manner. As a consequence, the mean abundance at reference and impact sites would tend to fluctuate in unison and the resulting time series of intertidal abundances would be parallel. This formulation does not, however, require that the mean levels be equal at the reference and impact sites. Instead, it allows for inherent differences in the carrying capacity between locations within and beyond the spill zone; differences that were probably present prior to the spill's occurrence.

Under the alternative hypothesis of impact, or ongoing recolonization of damaged habitats, temporal trends in mean abundance at reference and impact sites would not be parallel. The sampling design consists of selecting l_R reference sites beyond the spill and l_I impact sites. These $l_R + l_I$ sites are sampled concurrently over t years. Because regional influences cause the year-to-year fluctuations in intertidal abundance to be correlated, the assumption of independence is violated in univariate F-tests. This precludes the use of sampling year as a factor in an ANOVA. Instead, a sequential test must be applied to test parallelism as described in Appendix E. In the parallelism test, the more transitory the impact or repopulation event, the more likely the test will detect temporal differences in populations at impact and reference sites.

Generalizing sample-size calculations for departures from parallelism is complicated because there are endless ways that the impact and reference time series can differ. In Chapter 2,

differences between intertidal communities at reference and treated sites could be summarized in the form of a measure of the size of an effect (Δ). For individual taxa, effect size was parameterized by the percent increase or decrease in abundance (Equation B.8). For community responses, effect size was represented by a separation index (C) (Equation B.26). These measures of effect size determined the value of the noncentral parameters used in power analyses. In an analogous manner, Appendix E describes a power formulation for sequential tests where the deviation from parallelism is measured by a difference in linear trends.

Power Analyses

Because the power formulation for parallelism has an endless variety of realizations, it cannot be easily parameterized in terms of an effect size representative of a specific difference in trends. Instead, the range of required sample sizes for intertidal monitoring can be discerned from power curves determined for specific assemblages that were monitored for a number of years in PWS after the *Exxon Valdez* oil spill. Six representative assemblages from the PWS database are explored here.

Two types of intertidal impacts were identified in the PWS dataset (Coats *et al.*, 1999). The largest amplitude impact arose from invasive cleanup procedures that involved hot-water washing of intertidal substrate along large portions of beaches at a number of sites. These Category-3 sites experienced eleven-fold reductions in intertidal abundance. Other sites were impacted by oil cover but did not experience the same level of mechanical disturbance that the Category-3 sites experienced. On average, these Category-2 sites experienced three-fold reductions in intertidal abundance. Sample sizes needed to detect impacts (or recovery) from the combined effects of oiling and habitat disturbance can be determined from a comparison of temporal trends in intertidal abundance at the Category-3 sites and trends at reference (Category-1) sites that were not impacted by the oil spill. Sample sizes needed to detect lower-amplitude impacts from hydrocarbon exposure alone can be determined from a comparison of Category-2 and Category-1 sites as described below.

Habitat Disturbance

Population trends within PWS intertidal habitats that were subjected to severe habitat disturbance from invasive cleanup techniques are shown by solid squares (⊠) in Figure 3.1. For comparison, the time series of mean populations at reference sites (◊) is also shown for each of the six assemblages.

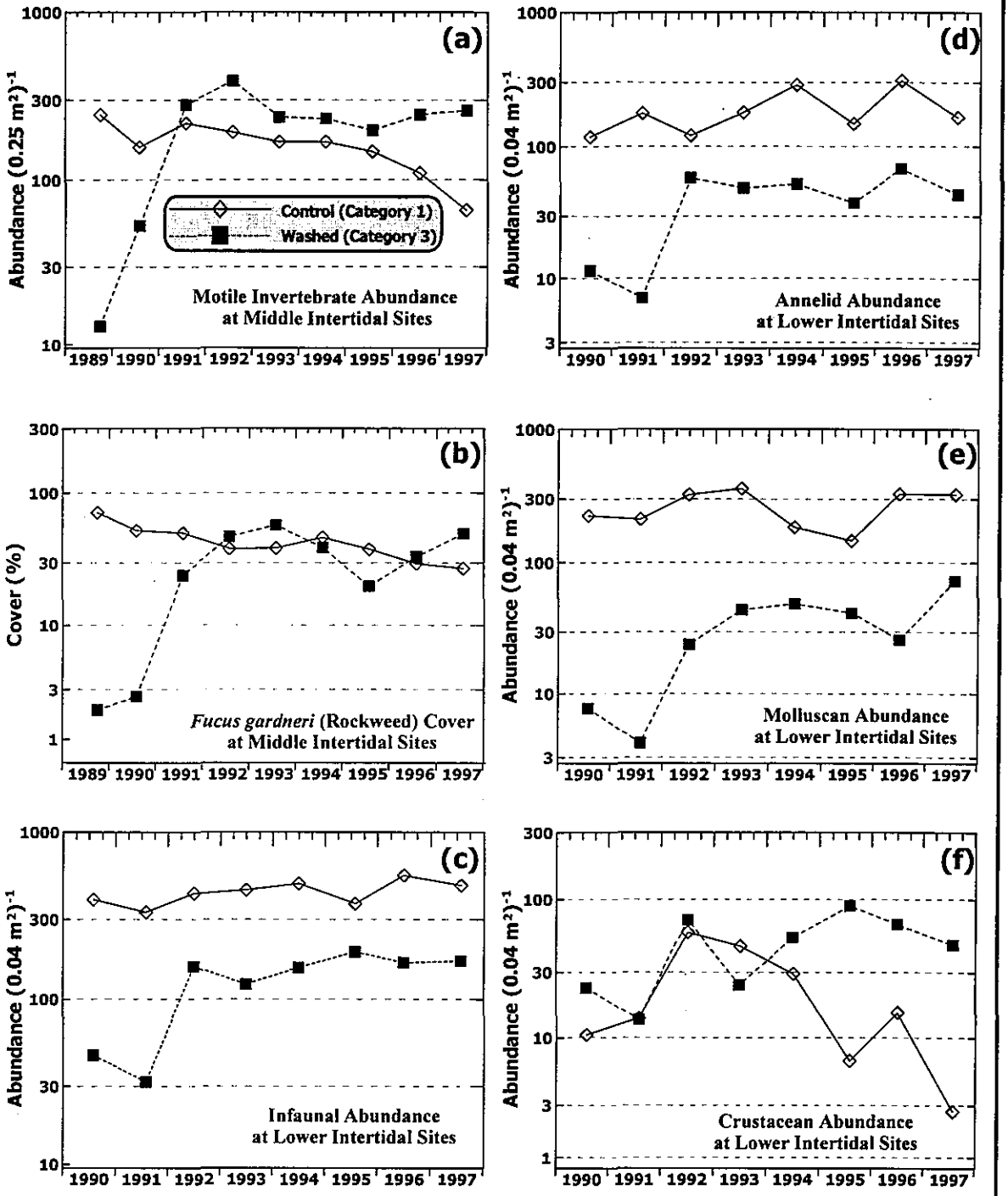

Figure 3.1. Comparison of mean populations at reference and washed sites for six PWS intertidal assemblages: (a) Motile invertebrates, (b) *Fucus*, (c) Total infauna, (d) Annelids, (e) Mollusks, and (f) Crustaceans

The sample-size chart shown in Figure 3.2 was computed using the power formulation presented in Appendix E (Equation E.8). It indicates the number of impact sites (l_I) that would be need to be sampled in order to detect repopulation events of the magnitude shown in Figure 3.1 for a variety of power ($1-\beta$) levels assuming a two-tailed significance level of $\alpha = 0.1$. Sampling at an equal number of reference or control sites (l_R) is also assumed. The first four years of data shown in Figure 3.1 were used in the computation. The three core sites used to compute the population averages shown in Figure 3.1, were used to compute the deviations from parallel linear trends in the numerator of Equation E.8. These same data were used to compute the variability about those trends in the denominator of Equation E.8. These deviation and variance values were used to extrapolate results to other sample sizes (number of sites) using the alternative formulation of Skalski and Robson (1992; Equation 6.44 on Page 204). These six assemblages span a broad range of intertidal populations and capture many of the differences in repopulation and recovery that can occur after severe habitat disturbance.

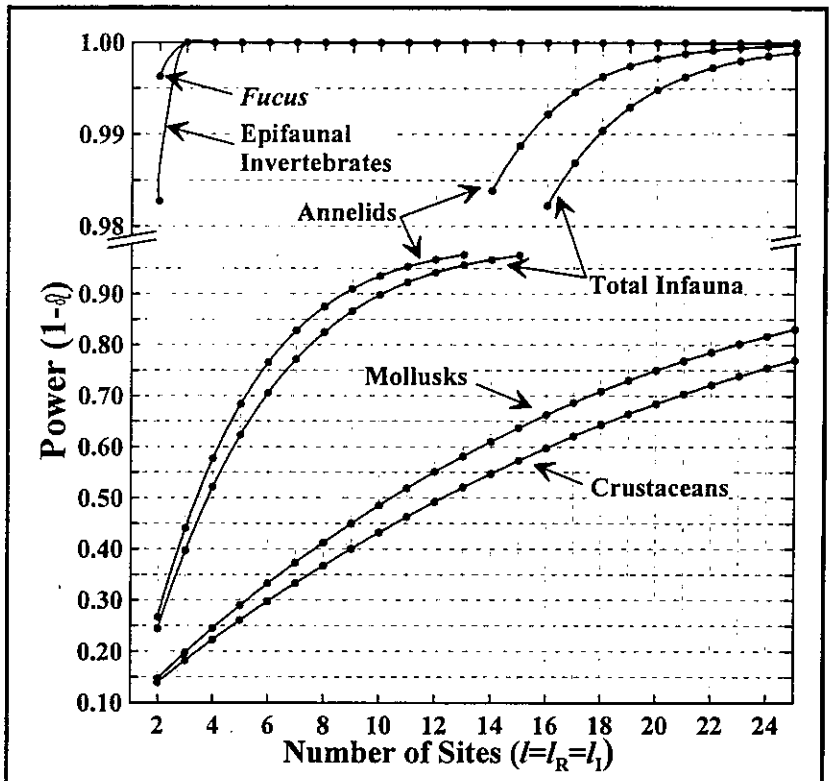

Figure 3.2. Sample sizes needed to detect linear departures from parallelism during the abrupt repopulation event at PWS intertidal sites subjected to hot-water washing.

The two major epibiotic assemblages at the middle-intertidal elevation, motile invertebrates (Figure 3.1a) and *Fucus gardneri* (Figure 3.1b), exhibited marked population increases between 1989 and 1992 at impacted sites relative to the time series of mean populations at reference sites. After 1992, populations stabilized and tended to track the population fluctuations at reference sites. In the invertebrate case, the mean post-recovery population at washed sites consistently exceeded that of the reference sites. This suggests that the inherent carrying capacity of the three sites that were subjected to oiling and intense washing may have actually been higher than the reference sites prior to the spill. Such differences, however, do not affect the determination of impacts and recovery based on parallelism tests.

Similar repopulation events are evident in the lower-intertidal infaunal assemblages shown in Figure 3.1cde. Repopulation of infaunal crustaceans is not as visually evident in Figure 3.1f. Also, in contrast to epifaunal invertebrates, post-recovery infaunal populations at the impacted sites were consistently lower than at the associated reference sites. Although this may be due to pre-spill differences in lower intertidal habitats, it is likely that alteration of the habitat by the hot-water washes had a major effect; namely, removal of fine-grained sediments. Infaunal communities are sensitive to changes in grain-size distribution, and the removal of fine-grained sediments may have affected the ability of certain elements within the community to recolonize the washed sites. In addition to the differences in mean abundance, the post-recovery community structure at washed sites was measurably different than the community structure at reference sites (Coats *et al.*, 1999). Other environmental factors, such as organic content, bacterial populations, food supply, and trophic interactions, covary with grain size and may be more directly responsible for the observed differences in infaunal abundance (Snelgrove and Butman, 1994).

Another difference between the infaunal and epibiotic assemblages was the amount of between-site variability. Although the population time series at individual sites is not shown in Figure 3.1, the epibiotic populations at the individual sites closely tracked their respective mean. This resulted in relatively low estimates of between-site variability of around 1.1 as listed in Table 3.1. In contrast, infaunal populations at individual sites fluctuated widely about the mean from year to year and resulted in variability of 3.0 or greater in the infaunal assemblages. This difference in variability is also partially reflected in the CV_B's listed in Table 2.4. Median CV_B's for middle-intertidal algae and motile invertebrates were 0.8 or less while the infaunal median was 1.0. As discussed below, the increased between-site variability in the infaunal time series increased the number of sites that need to be sampled to achieve a given power to detect non-parallelism.

Table 3.1. Amplitude of the departure of from parallel linear trends and variability about the mean trends

	Washed		Oiled	
Assemblage	$\lvert \mathbf{b}'\bar{\mathbf{x}} \rvert$ [a]	$\sqrt{l \cdot \mathbf{b}'\Sigma\mathbf{b}}$ [b]	$\lvert \mathbf{b}'\bar{\mathbf{x}} \rvert$	$\sqrt{l \cdot \mathbf{b}'\Sigma\mathbf{b}}$
Fucus gardneri	5.5	1.1	1.9	0.6
Motile invertebrates	4.9	1.1	1.1	1.5
Total Infauna	1.7	3.0	1.3	2.9
Mollusks	3.0	32.0	1.7	7.3
Annelids	2.4	5.3	1.1	5.0
Crustaceans	1.4	9.0	1.5	7.4

[a] Amplitude of the deviation from parallel linear trends (See Equation E.8)
[b] Between-site variability about mean population trends

Figure 3.2 shows that the severe habitat disturbance and subsequent epibiotic recovery experienced in PWS should be easily detected with a four-year monitoring program. Very high power $1-\beta > 0.98$ for epifaunal invertebrate assemblages was achieved with sampling at as little as two reference and two impact sites. In contrast, the high between-site variability associated with the infaunal populations makes departures from parallel linear trends difficult to detect without sampling at a larger number of sites. For the three core sites sampled in the PWS monitoring program, the observed departures from parallelism in annelid and total infaunal populations only achieved respective powers of 0.45 and 0.40. Powers above 0.7 would require sampling at six reference and six impacted sites. Such a sampling effort would provide a 70% chance that population trends of the magnitude seen in Figure 3.1 would be detected by the intertidal monitoring program.

Figure 3.2 also suggests that detection of departures from parallel linear trends in crustaceans and mollusks would require far more sites, more than 21 impact and 21 reference sites, to achieve a power of 0.7. This is not surprising for the crustacean populations because they did not visually exhibit marked population increases in the four years after the spill (Figure 3.1f). However, the mollusk population increase in Figure 3.1e is clearly evident. The apparently low power associated with the mollusk test resulted from an anomalously high variability among the populations that were enumerated at Category-3 sites. This resulted in a variability estimate that was an order-of-magnitude greater than for other assemblages (Table 3.1).

Hydrocarbon Exposure

The repopulation events at oiled sites that did not experience severe habitat disturbance were smaller in amplitude. This is evident from a comparison of the time series plotted in Figure 3.1 and Figure 3.3 after noting the expanded y-axis scales in Figure 3.3.

The difference in amplitude suggests that the impacts at oiled sites were less severe than at sites that were subjected to intensive cleaning. It could also result from inherent differences in the carrying capacity of the sites selected in each category; differences that were present before the spill occurred. In any regard, the smaller amplitude of the recovery increases the difficulty in detecting deviations from parallelism.

Despite the reduced size of the repopulation events, they are still visually evident as departures from parallelism in the time series for some of the intertidal assemblages (Figure 3.3). The time series of mean populations at oiled (shaded circles ⊛) and reference sites (◊) converge, and in most cases, cross one another in the first four years after the spill in 1989. As with the time series at washed sites, early convergence (or divergence) in the time series of crustacean mean populations is not visually evident (Figure 3.3f). The convergence in the *Fucus* time series (Figure 3.3b) provides the strongest signature and the trends were more linear than in other assemblages. *Fucus* had the highest magnitude of departure from parallel trends at oiled sites (1.9 in Table 3.1).

These features are to some extent reflected in the sample sizes that are projected for oiled sites in Figure 3.4. As with washed sites, however, the computed amplitude of the between-site variability around the mean time series also determines the power to detect deviations from parallel linear trends. Except for the low *Fucus* variability (0.6), other assemblages had variability that ranged between 1.5 and 7.4. Because *Fucus* cover also had a large linear convergence trend that was consistently reflected at all sites (low variability), only small sample sizes would be needed to detect non-parallelism with high levels of confidence (Figure 3.4). Sampling at three oiled and three reference sites over four years would be capable of detecting a repopulation event similar to Figure 3.3a with a power of more than 0.95. With a power of more than 0.7, monitoring at 9 or more oiled sites would be required to detect the small population changes for motile epifaunal invertebrates and total infauna shown in Figure 3.3bc. There would be little opportunity to confidently $(1-\beta > 0.7)$ discern repopulation events in infaunal mollusks, annelids, or crustaceans without sampling at more than 14 to 20 oiled sites.

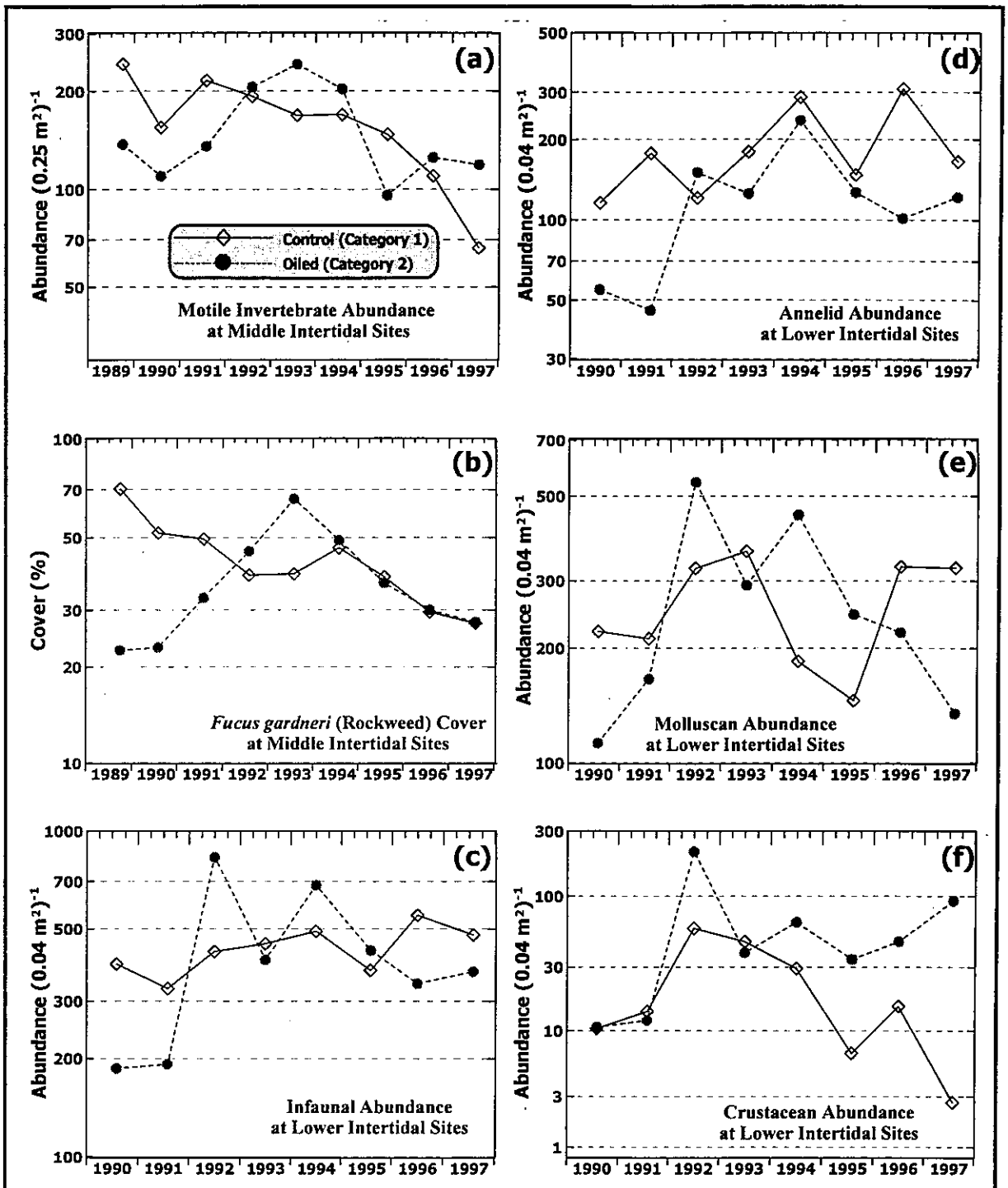

Figure 3.3. Comparison of mean populations at reference and oiled sites for six PWS intertidal assemblages: (a) Motile invertebrates, (b) *Fucus*, (c) Total infauna, (d) Annelids, (e) Mollusks, and (f) Crustaceans

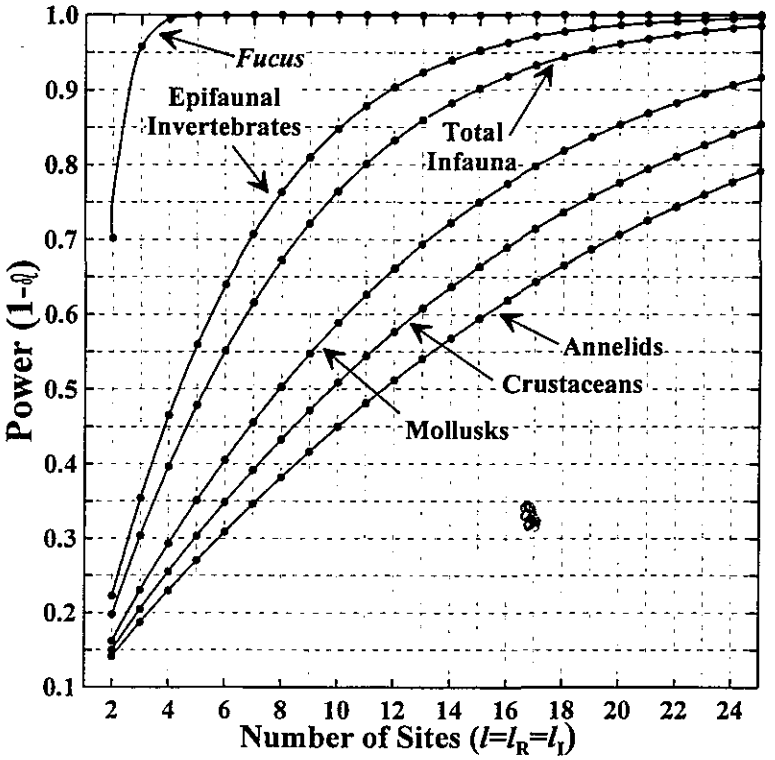

Figure 3.4. Sample sizes needed to detect linear departures from parallelism during the abrupt repopulation event at PWS intertidal sites subjected to oiling but not invasive cleanup techniques.

The striking differences in sample-size recommendations among the various intertidal assemblages emphasize the importance of selecting optimal biological variables to include in the monitoring program. The assemblage of concern must not only be exposed to contamination or habitat disturbance, but it must also have the ability to demonstrate impacts or recovery within the practical constraints of field sampling. Optimal taxa are ubiquitous, not extremely clumped or patchy in distribution, and respond uniformly at all sites to the impact. Taxa with these attributes have the greatest likelihood of demonstrating statistically significant effects.

Chapter 4. Chronic Effects

After acute impacts have dissipated and recolonization has largely occurred, ongoing monitoring could productively investigate whether an intertidal ecosystem had reached a long-term stable state or whether there were lingering effects resulting from the spill and cleanup efforts. Persistent chronic effects such as bioaccumulation of toxicants may go undetected over short periods of time, but may result in long-term changes that only become evident years after the spill event. For example, Fukuyama *et al.* (2000) described chronic effects in PWS intertidal clams that became evident five to six years after the *Exxon Valdez* spill. Resident clams contained lingering hydrocarbon burdens in their tissues and hydrocarbon uptake was demonstrated in transplanted clams. Similarly, Shigenaka *et al* (1999) found slowly increasing populations in resident Littleneck Clams at impacted sites (*Protothaca staminea*). This gradual recolonization contrasts sharply with the abrupt population increase seen in most other intertidal taxa and suggests that these clams were experiencing a long-term recovery from chronic impacts related to the spill. Finally, population reverberations from the abrupt recolonization event are visually apparent in the time series of many PWS taxa (Coats *et al.*, 1999). Houghton *et al.* (1996) ascribed the oscillation in rockweed cover to the senescence of a single cohort (age-class) that colonized during a brief one-year period between 1990 and 1991. These repercussions of the abrupt recolonization event are the subject of ongoing manipulative experiments now being conducted by NOAA within Alaskan intertidal zones.

In the absence of climatological, global, or catastrophic environmental events, a stable biological system might be expected in which population levels fluctuate about some stationary long-term mean. Under those conditions, there would be no regression relationship between intertidal abundance and time. Consequently, one method for determining long-term stability is to test for a significant regression relationship between abundance and time at sites that were previously impacted by the spill or cleanup. Under the null hypothesis of no long-term ecosystem instability, the regression coefficients, other than the intercept, should not be significantly different from zero.

Conversely, under an alternative hypothesis that is applicable when chronic impacts are present, mean population levels would be expected to slowly change over time, and the slope of a regression line would be significantly different from zero. Appendix F formulates a statistical test for long-term linear trends and develops noncentral parameters that can be used to determine the power to detect non-zero slope coefficients. This Chapter applies this power formulation to

the PWS dataset and provides sample-size recommendations that can be used to test for long-term chronic impacts to intertidal taxa.

Chronic Effect Size and Duration

The ability to detect a long-term chronic effect depends on the magnitude of the annual change caused in the intertidal population of concern and on the length of time over which the effect can be measured. For long-term chronic effects, the power to detect change will be limited by the duration of the monitoring program. In other cases, populations may stabilize after a few years and extending the monitoring program will be of little benefit.

Between 1990 and 1993, most PWS intertidal populations at impacted sites increased by a factor of more than two. This recolonization event was too large and too short term to be considered a recovery from chronic effects. Instead, it represented a widespread recovery from the initial acute impacts of hydrocarbon exposure and habitat disturbance. The parallelism tests described in the previous chapter address how to detect these abrupt high-amplitude recolonization events. The magnitude of the recolonization event suggests that the acute impacts from the spill caused an initial reduction in abundance of at least a factor of two, and population impacts from invasive cleanup techniques that were much larger. Detection of these acute initial impacts was the subject of Chapter 2. Based on the observed abrupt changes in PWS intertidal abundance, the dissipation of any lingering chronic effects will be reflected in long-term population increases that are smaller than 200% overall and that occur over periods of more than three years.

These considerations establish approximate limits on the size of chronic effects to be used in the sample-size determinations presented in this chapter. Sample-size determinations were based on this working definition of chronic effects and encompass a range of scenarios that address small overall population changes occurring across a number of years. Table 4.1 lists the scenarios that characterize the ability to detect annual population changes from 3% to 30% over 5 and 10-year periods.

Power Analyses

Appendix G presents sample-size plots needed to detect chronic effects in intertidal taxa for the scenarios presented in Table 4.1. Within-site and between-site variability in mean abundance for a given year was represented by the three CV levels that were computed for abundant taxa from PWS data in Chapter 2. Various combinations of annual increase and duration are reflected in the sample-size plots. As in the detection of acute impacts, the power to detect chronic impacts is

Table 4.1. Total population change as a function of various annual population increases and study durations presented in this Chapter and in Appendix G

Annual Increase	Total Increase	
	5-year	10-year
3%		27%
5%		45%
10%	40%	90%
11.5%	46%	104%
15%	60%	135%
20%	80%	180%
25%	100%	
30%	120%	

determined by both the number of quadrats or infaunal cores sampled at each site (m), and the total number of sites (n) visited each year.

The functional form of the noncentrality parameter used to determine sample sizes for chronic effects (Equation F.11) is similar to that for testing treatment effects (Equation B.8). Consequently, the shapes of the power curves shown in Appendix G are similar to those of Appendix D, and the same recommendations for sampling design apply to the detection of chronic effects. In particular, adding more than eight replicate samples within each site is generally less important for determining chronic impacts than adding additional sites. Most of the power curves in Appendix G approach a vertical asymptote above eight replicate samples (m), so increasing the number of replicate samples above this threshold does little to enhance the detection of chronic effects. Consequently, collecting approximately six replicate samples at each site is optimal in most cases. Similarly, the power curves start to approach a horizontal asymptote below $m = 4$ and are distinctly horizontal below $m = 2$. This indicates that the optimal sampling design for the detection of long-term trends in the dominant intertidal taxa consists of collecting six to eight replicate samples at as many sites as possible.

Further assuming that a nominal 60% chance of detecting a long-term trend is the minimum acceptable power level, the number of impact sites that require sampling can also be determined for various chronic effect sizes. From the Appendix-G power curves, sampling at six sites (with seven replicate samples) would be required to reliably discern a 10% annual trend in assemblages with low variability over a 5-year period (Figure G.1a). A 15% annual increase could be resolved by collecting six samples at half this many sites (Figure G.1b). However, approximately eight replicate samples at thirteen sites would be needed to discern the 15% annual trend in taxa with moderate variability (Figure G.2a). If the 15% annual increase persisted

for ten years, monitoring five replicates at six impact and six reference sites would be capable of marginally resolving the trend in nearly all the dominant intertidal assemblages (Figure G.6a).

Protothaca Application

The long-term trend in populations of littleneck clams (*Protothaca staminea*) observed in the PWS monitoring program provides an important application of power analysis for chronic effects. Shigenaka *et al* (1999) found that at impacted sites, littleneck clams (*Protothaca staminea*) exhibited a different pattern of impact and recovery than other infaunal species. In particular, they did not show the abrupt population increase observed among many recovering species. Instead, Category-3 (washed) populations were gradually approaching reference-site populations, which suggested that the clams were recovering from chronic impacts after the spill. Figure 4.1 shows how the abundance of littleneck clams at the unoiled reference sites and those at sites that were oiled and subsequently washed, has steadily converged since 1992.

In PWS, *P. staminea* is frequently encountered on gravel beaches, and the clam is a regular part of the subsistence diet for native villagers residing in the region. While butter clams (*Saxidomus gigantea*) constitute the majority of the harvest, littlenecks are also popular. Both are members of the Veneridae family. Figure C.7 shows that the pre- and post-recolonization population levels for the Veneridae clams were very similar. This contrasts with the significantly lower

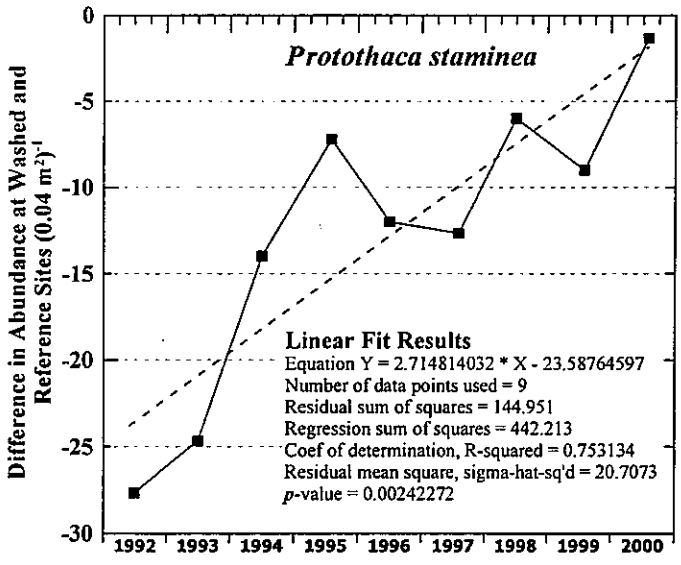

Figure 4.1. Long-term trend in the difference between average littleneck clam (*P. staminea*) populations at Category-3 washed sites and Category-1 reference sites in PWS.

populations observed in most other infaunal taxa during the impact period from 1990 through 1991. The spatial variability in the Veneridae clam distribution within sites, as reflected by a $CV_W = 1.2$, was also lower than most other infaunal taxa where the median variability was reflected by a $CV_W = 2.2$ (Table 2.4). The between-site spatial variability for littleneck clams was lower than for the Veneridae family as a whole. Post-recovery variability for *P. staminea* was $CV_W = 1.12$, $CV_B = 0.65$.

These parameters can be used in Equation F.10 to determine the power to detect linear trends of the magnitude shown in Figure 4.1. The slope in Figure 4.1 constitutes an 11.5% ($\Delta = 0.115$) annual increase in abundance or a 92% increase over nine years. In other words, the clam populations almost doubled in the span of nine years. In the PWS monitoring program, five infaunal cores were collected at each site ($m = 5$). Three Category-3 sites form the core group of sites representing impacts from high-pressure washing ($n = 3$). The approximate power can be estimated from the sample size curves provided in Appendix G. Figure G.4b shows that a 5% annual increase in low variability taxa over 10 years results in a 60% probability ($1 - \beta = 0.6$) of detecting a change of this magnitude or larger. *P. staminea* exhibited a slightly higher annual increase so the true power would be slightly higher than 0.6.

Figure 4.2b displays more precise sample-size curves for determining the power to detect the observed long-term recovery of *P. staminea* within PWS. At $m = 5$ and $n = 3$, the power to detect the observed trend is close to 0.7 over the nine-year trend. Thus, there is a 70% chance that the statistical test correctly discerned a temporal trend in abundance if a non-zero trend actually existed in the littleneck clam populations. The power curves show that doubling the number of samples collected at each site ($m = 10$) would only increase confidence by 0.05% or $1 - \beta = 0.75$. In contrast, if these 15 additional samples were instead collected at three additional sites (i.e. $m = 5$ and $n = 6$), the power would exceed 0.85. This example shows again that for a given total number of samples, increasing the number of sites is more beneficial than increasing the number of samples within each site.

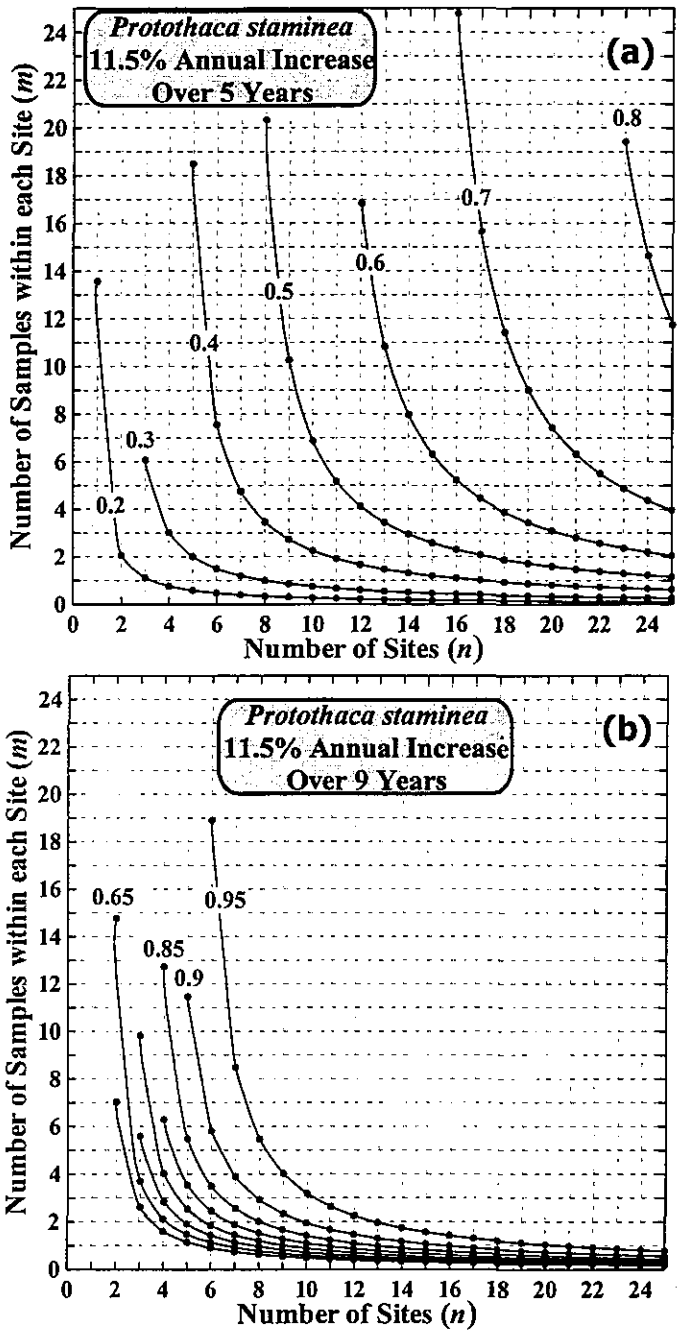

Figure 4.2. Sample-size chart showing the number of sites (n) and the number of replicate samples (m) needed to detect an impact that causes the 11.5% annual increase in intertidal populations over a (a) 5-year and (b) 9-year period with a statistical power ($1-\beta$) at the one-tailed significance level of $\alpha = 0.1$. The curves correspond to different levels of statistical power in an environment with the low natural biological variation associated with littleneck clams (*P. staminea*) ($CV_W = 1.12$, $CV_B = 0.65$).

The duration of sampling also strongly affects the ability to reliably discern long-term trends. Figure 4.2a shows that if monitoring had ceased after five years, then the power to detect the 11.5% trend would have only been 0.30 with $m = 5$ and $n = 3$. This constitutes an unacceptably high probability (70%) of committing a Type II error whereby a meaningful trend in populations would be missed. Under these circumstances, where long-term monitoring was limited to only five years, eleven or more sites would need to be sampled just to reach a break-even power level $(1-\beta = 0.5)$.

Assuming that littleneck clam populations at washed sites continued to increase indefinitely, extending the monitoring program beyond the year 2000 could also markedly improve statistical power up to a point. Figure 4.3 projects the statistical power into future years assuming the same sampling design used in PWS were to continue. The power $(1-\beta = 0.7)$ at nine years corresponds to the observed PWS case where long-term monitoring was discontinued after 2000. Had monitoring continued another year into 2001 and impacted clam populations continued to increase at the same rate, then the statistical power would have been increased by 10% to 0.8. However, as shown in Figure 4.3, monitoring the trend beyond 14 years would be of little incremental advantage with regard to increasing statistical power.

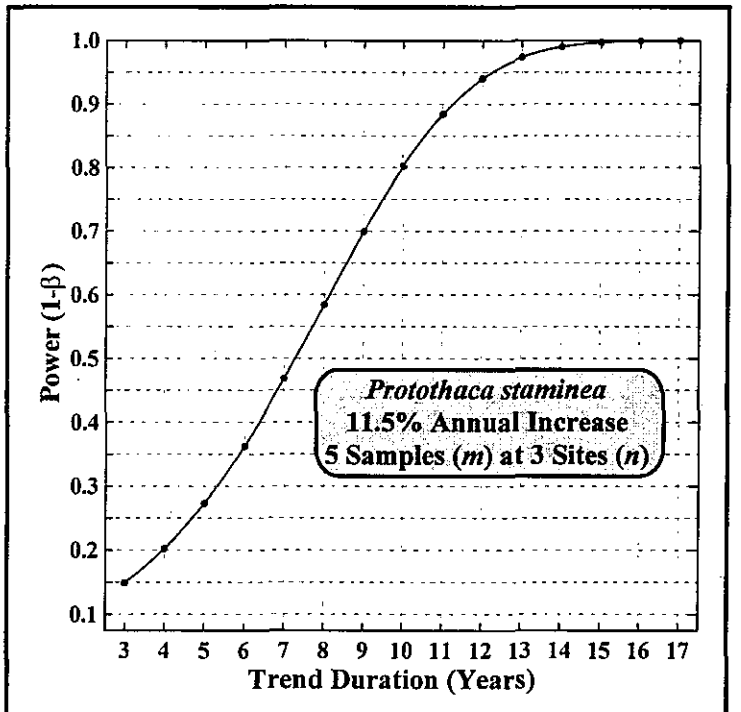

Figure 4.3. Power to detect the observed 11.5% increase in littleneck clam populations with 5 samples collected at each of 3 sites as a function of the duration of the monitoring program.

Chapter 5. Conclusions

This report demonstrates that there is no simple answer to the question of how many intertidal samples to collect. It is difficult to generalize the monitoring-design recommendations and still rigorously quantify the power to detect a given impact to a specific intertidal taxon. Also, much depends on the type of population change that is of interest. In this report, sample-size requirements were specified for three types of changes to intertidal populations: treatment effects, abrupt recolonization, and chronic long-term effects. Some of the sample-size recommendations are summarized in Table 5.1. These were derived from comprehensive power analyses conducted on data acquired from the PWS intertidal monitoring program. Guidance for most monitoring-design requirements is covered by the extensive array of power curves presented in Appendices D and G. However, these power curves may not apply to the design of monitoring programs conducted in regions where the intertidal spatial variability is radically higher than for PWS taxa. For those cases, this report presents the methodology for developing new site-specific power analyses based on pilot studies. The mathematical formulations for determining sample sizes and the associated site-specific variability estimates are provided in Appendices B, E, and F.

Table 5.1. Representative sample-size recommendations[1]

Monitoring Goal	Taxon or Abundance[2]	Spatial Variability[3]	Effect Magnitude[4]	Sampling Duration[5]	Replicate Samples (m)[6]	Sites (n)[7]
Treatment Effects	Sparse	Low	15%	1	5	9
Treatment Effects	Sparse	Low	25%	1	4	4
Treatment Effects	Sparse	Moderate	33%	1	5	22
Treatment Effects	Sparse	Moderate	50%	1	5	8
Treatment Effects	Sparse	High	67%	1	5	15
Treatment Effects	Sparse	High	75%	1	4	11
Treatment Effects	Intermediate	Low	25%	1	5	14
Treatment Effects	Intermediate	Low	33%	1	4	9
Treatment Effects	Intermediate	Moderate	50%	1	5	21
Treatment Effects	Intermediate	Moderate	67%	1	4	10
Treatment Effects	Intermediate	High	75%	1	5	18
Treatment Effects	Intermediate	High	83%	1	4	12
Treatment Effects	Abundant	Low	25%	1	5	13
Treatment Effects	Abundant	Low	50%	1	4	3

[1] Assuming a one-tailed significance level of $\alpha = 0.1$ and a power of at least $\beta = 0.6$

[2] Sparsely populated intertidal taxa have an average count or percent cover that is less than 0.07 in the PWS dataset, while abundant taxa have average densities that exceeded 3. See Appendix A and Table 2.6.

[3] See Table 2.6.

[4] Percent reduction per year.

[5] Number of years sampled (number of annual sampling events).

[6] Number of samples collected within each site.

[7] Number of treatment/impact sites assuming an equivalent number of reference sites are also sampled.

Monitoring Goal	Taxon or Abundance[2]	Spatial Variability[3]	Effect Magnitude[4]	Sampling Duration[5]	Replicate Samples (m)[6]	Sites (n)[7]
Treatment Effects	Abundant	Moderate	50%	1	5	11
Treatment Effects	Abundant	Moderate	67%	1	4	5
Treatment Effects	Abundant	High	67%	1	4	16
Treatment Effects	Abundant	High	75%	1	4	10
Treatment Effects	PWS Infauna	Ordination	2-D C=1.5[8]	1	3	9
Recovery	*Fucus*		Washed[9]	4		2
Recovery	Epifaunal Invertebrates		Washed	4		2
Recovery	Annelids		Washed	4		5
Recovery	Total Infauna		Washed	4		5
Recovery	Mollusks		Washed	4		14
Recovery	Crustaceans		Washed	4		16
Recovery	*Fucus*		Oiled	4		2
Recovery	Epifaunal Invertebrates		Oiled	4		6
Recovery	Annelids		Oiled	4		16
Recovery	Total Infauna		Oiled	4		7
Recovery	Mollusks		Oiled	4		11
Recovery	Crustaceans		Oiled	4		13
Chronic Effects		Low	10%[10]	5	4	8
Chronic Effects		Low	15%	5	4	4
Chronic Effects		Moderate	15%	5	4	16
Chronic Effects		Moderate	20%	5	4	4
Chronic Effects		High	25%	5	4	20
Chronic Effects		High	30%	5	4	14
Chronic Effects		Low	3%	10	4	9
Chronic Effects		Low	5%	10	4	4
Chronic Effects		Moderate	5%	10	4	14
Chronic Effects		Moderate	10%	10	3	4
Chronic Effects		High	21%	10	4	6
Chronic Effects		High	20%	10	4	4

Sample Sizes

Monitoring duration is one major difference in sampling programs designed to detect the three types of population change. Given certain assumptions, treatment effects can be evaluated from a single sampling occasion, although sampling before and after treatment is preferable. By assuming intertidal populations at treatment and reference sites were similar before application of the treatment, 50% reductions in the populations of moderately variable abundant taxa could be resolved by a single survey that collects 7 replicate samples at 10 treatment and 10 reference sites (140 samples total; Figure 2.3). This would provide a marginal power of $1-\beta=0.6$, or a 40% chance (β) of missing a 50% population reduction (Δ). There would also be a 10% risk (α) of incorrectly finding a population reduction this large. Collecting additional replicate

[8] Amplitude of the two-dimensional separation index in ordination hyperspace
[9] Amplitude of the increase in abundance equivalent to that observed during the abrupt repopulation event in PWS
[10] Annual increase

samples at these sites, however, would be of little statistical benefit. Alternatively, if enough sites were available, then the same population reduction could be resolved by collecting 3 replicate samples at 13 treatment and 13 reference sites (78 samples total). Treatment effects on community composition can be determined from multivariate analyses. Sample sizes determined from an "analysis of distance" (ANODIS) indicate that a comparison of treatment and reference intertidal communities within any given year is likely to yield low power unless samples are collected at a large number of sites (>10) or unless sizeable treatment effects are present.

The detection of oil-spill related changes to intertidal populations requires that samples be collected over a number of years. Tests for parallelism at impact and reference sites provide one means of detecting abrupt recolonization events of the sort seen in PWS. These parallelism tests allow for differences in populations at the various sites that may have existed before the spill because of differences in the carrying capacity of specific habitats. Sample sizes vary significantly depending on the intertidal assemblage being tested because the power of the test depends on the consistency in population response among the sites in addition to the magnitude of the change. Recolonization from the damage experienced by *Fucus* and epifaunal invertebrates after severe habitat disturbance can be resolved by sampling over four years at only two reference and two impact sites. In contrast, the higher variability in infaunal populations makes detection of departures from parallelism difficult without sampling at least six reference and six impacted sites. Crustaceans and mollusks would require sampling at more than 21 impact and 21 reference sites, to achieve a power of 0.7.

Detecting subtle population trends due to chronic effects from an oil spill requires sampling over longer periods of five to ten years. Required sample sizes can be estimated from tests for statistically significant slope coefficients in linear regression lines fitted to long-term population trends. Sample sizes depend on the magnitude of the annual trend and the duration of monitoring. For assemblages with low variability, sampling at six sites (with seven replicates) would marginally discern a 10% annual trend over a 5-year period, but a 15% annual increase could be resolved by collecting six samples at only half this many sites. The long-term trend in PWS littleneck clam (*Protothaca staminea*) populations demonstrates the advantages of extending the duration of monitoring. The clam populations exhibited an 11.5% annual increase for nine years at three core sites where five replicate samples were collected. The likelihood of correctly detecting an increase of this magnitude was 70% ($1-\beta = 0.7$). Had sampling only been conducted for five years, the detection power would have been reduced to less than 30%. Assuming populations continued to increase beyond nine years, monitoring beyond 14 years would be of little advantage insofar as increasing the statistical power.

Spatial Variability

In addition to the impact assessment itself, this report lends insight into the inherent variability found within intertidal populations. Accurate measurement of spatial and temporal variability in the biological populations forms the basis for the design of an adequate sampling program. For the most part, variability among intertidal assemblages and tidal elevations was similar. Slight differences in spatial variability were found between sparsely populated and abundant taxa. Sparsely populated taxa tended to have lower between-site variability while abundant taxa tended to have lower within-site variability. Also, seven taxa exhibited anomalously high variability due to their inherent tendency to form dense patches or clumps within the intertidal zone. Because of their inordinately high variability, the general sample-size guidelines presented in this report do not apply to them. The remaining 263 intertidal taxa and assemblages were well represented by the sample-size calculations presented herein.

Recommendations

One insight provided by this report's analyses is that a large number of samples are often required to achieve marginal statistical power (Table 5.1). This requirement may conflict with the constraints imposed by traditional intertidal sampling protocols, which are labor intensive and demand the presence of experienced field biologists. By adhering to traditional sampling techniques, there may not be enough time and trained personnel available to collect samples sufficient for minimal statistical credibility. The following recommendations will help to increase the sample collection rate without unduly sacrificing needed statistical rigor. While not all of the recommendations follow directly from the results presented in the body of this report, they represent the collective experience that the authors have gained after participating in many marine monitoring programs.

Sample Before a Spill Impacts Intertidal Sites

The power to detect impacts and recovery is significantly weakened by the lack of data prior to the spill. If at all possible, intertidal sites should be sampled before they are impacted by an offshore oil spill. This will provide before-spill data that is crucial for rigorously testing impacts to intertidal biota. In addition, these data can act as a pilot study for determining biological variance and establishing sample-sizes for a more-extensive post-spill monitoring program. In the past, sampling immediately before a spill impinged on an intertidal zone was unrealistic. However, current levels of oil-spill preparedness may allow rapid response and mobilization of monitoring personnel to the spill region. Moreover, with the advanced predictive capabilities of available oil-spill trajectory models, shoreline impact areas can now be identified with tangible

skill given the location of an offshore spill and real-time metocean data. However, biological assessments would need to be conducted rapidly, and traditional field sampling techniques may need to be relaxed as described below.

Relax Taxonomic Discrimination

Knowledgeable and experienced biologists are required to accurately identify specimens to the lowest taxonomic level in the field. This is a time-consuming and expensive process. The rationale for low-level identification is that a particular species may be especially sensitive to hydrocarbon exposure. Therefore, its identification and enumeration provides needed discrimination for detecting oil-spill impacts. In reality, taxonomic discrimination to species level is rarely exploited in the subsequent data analyses except for certain target species that may have particular economic or societal importance, such as *P. staminea* in PWS, or that may be environmentally sensitive, such as an endangered or threatened species.

Instead, the statistical analysis of multiyear intertidal data is confounded by differences in taxonomic identifications that arise over time because the names of species change or because different biologists identify taxa to different levels. For oil-spill impact assessment, consistent and accurate determination of counts or percent cover of dominant taxa is far more important than determining what species a rare specimen might represent. Identification of lichens (*Verrucaria*) is a case in point. Their exclusion from the analysis is often warranted given their great variation in appearance and widely differing identification and quantification among observers.

To facilitate taxonomic identification in the field, epibiotic taxa that are not an important or dominant species should be identified at the family or higher taxonomic level. Except for certain target species, determinations of impacts and recovery should be based on major taxonomic or functional-form groups, for example, grazers versus algae versus predators. Impact assessments based on these broad categories are more pertinent to overall determinations of impact and recovery. Reducing taxonomic discrimination at the outset during field sampling will save time and will allow larger numbers of well trained but less experienced biologists to be deployed in the field. This will increase the number of samples that can be collected, which directly improves the statistical power to detect impacts and recovery. The issue of taxonomic sufficiency in oil-spill assessments has been the recent focus of discussion in the marine community (Dauvin *et al.*, 2003; Terlizzi *et al.*, 2003; Gomez Gesteira *et al.*, 2003).

Randomize Quadrat Locations

Fixed quadrats are time consuming to setup and maintain. The advantages realized by reduction in temporal variability afforded by fixed quadrats may not be offset by the time and energy needed to establish permanent markers, particularly in multi-year field programs when additional effort is expended finding and repairing the markers in subsequent years. Instead, randomly placed quadrats along given transects could be established each sampling occasion as long as they follow certain guidelines concerning spacing and consistency of habitat. In many cases, the increased statistical power realized by collecting a larger number of additional samples far outweighs the benefits arising from the reduction in variance that is realized by using fixed sampling locations.

Limit the Focus

When biologists consult this report in the field immediately after an oil spill, many of the sampling-design decisions normally afforded an experimentalist will be foregone conclusions. There may be a limited number (n) of impacted beaches available for survey, or a limited number of trained biologists available to conduct surveys. Decisionmakers and stakeholders will be unwilling to specify acceptable error rates (• and •) and an impact threshold (Δ) because they do not understand the implications of such a decision. Precedent and historical levels then dictate the error rates and thresholds to be used in the sampling design. Because of precedence, significance levels (•) exceeding 0.1 are not well received in peer-reviewed scientific journals. Similarly, allowing the risk of missing an important impact (•) to be 50% or more defeats the purpose of the monitoring. Power ($1-\beta$) will need to be at least 0.6, and preferably higher.

Under these circumstances, the best that field biologists can hope for is to not expend sampling effort unnecessarily collecting too many replicate samples at each site. First, they will need to determine the goals and duration of monitoring. Is it limited to a one-time assessment of treatment effects (Chapter 2), or will multi-year post-spill sampling be conducted to quantify recolonization events (Chapter 3) and long-term chronic effects (Chapter 4)? Once these questions are answered, consulting the appropriate sample-size charts in this report will help identify the optimal number of replicate samples to be collected at each site.

Appendix A. Glossary of Selected Statistical Terms and Acronyms

Alpha (•) The statistical significance level. The probability of committing a Type-I error where the null hypothesis of no impact is incorrectly rejected. The probability of incorrectly finding an important impact when it is in fact, inconsequential. •-levels are set at low levels, typically 0.10, 0.05, or 0.01, to indicate to indicate a high degree of confidence (90%, 95%, or 99%) that the measured impact in fact exists when the null hypothesis is rejected.

Alternative Hypothesis (H_a) The oil spill and cleanup measurably affected the abundance of intertidal organisms.

ANOVA An acronym for analysis of variance that examines the contribution of each parameter to the variation in the outcomes of an experiment. It is a method of statistical analysis broadly applicable to a number of research designs, used to determine differences among the means of two or more groups on a variable.

ANODIS An acronym for analysis of distance that is the multivariate analog to an ANOVA. Distance refers to the separation of mean sample scores in ordination hyperspace.

BACI An optimal sampling design where samples are collected before and after the impact at both control (reference) and impacted sites

Beta (β) The probability of missing a meaningful impact. The probability of committing a Type-II error where the null hypothesis of no impact is incorrectly accepted.

β-Diversity Beta diversity measures the differences in diversity among samples. A group of samples with high β-diversity will have completely different species compositions and some pairs of samples may have no species in common. A group of samples with low β-diversity will be similar in species composition throughout. Principal Component Analysis (PCA) functions best at low β-diversity while Correspondence Analysis (CA) behaves best at high β-diversity (ter Braak, 1983).

Between-Site The number (n) or variability (CV_B) of sites or beaches that are sampled. A balanced design is assumed in this report, so there are actually $2n$ sites sampled, n impacted sites and n reference (unimpacted) sites.

Category Classification of PWS sites in terms of their impact exposure. Category-1 sites were reference sites that were unoiled in 1989. Category-2 sites were oiled in 1989 but were either untreated or only lightly cleaned. Category-3 sites were oiled in 1989 and were subjected to high-pressure, hot-water washes.

Clumping The tendency for organisms to cluster together to form dense patches of closely grouped aggregates surrounded by areas that are relatively devoid of specimens. Clumping is synonymous with a contagious spatial distribution that is best represented by a negative binomial frequency distribution where the variance is greater than the arithmetic mean. In the PWS intertidal dataset, taxa with an excessive natural tendency to clump had dispersion indices larger than 20 (see Dispersion Index below).

Coefficient of Variation (CV) A coefficient used to compare the relative amounts of variation in populations having different means. It is defined as the standard deviation divided by the mean.

Correspondence Analysis (CA) An eigenanalysis-based ordination method also known as reciprocal averaging where sample scores and species scores are calculated simultaneously as weighted average of one another by maximizing the correlation between them. These methods perform best when species have unimodal distributions along environmental gradients (ter Braak and Verdonschot, 1995).

Dispersion Index $\left(\frac{1}{K}\right)$ A measure of the degree to which individuals in an intertidal population clump together or form patches within a site. It is an inverse measure of dispersion.

Effect Size (Δ) The amplitude of the change in biological properties (impact) that is considered important or meaningful. The degree to which the oil spill and cleanup changed intertidal populations. The degree to which the null hypothesis is false where the null hypothesis implies that the effect size is zero.

Horseshoe Effect A distortion of ordination diagrams that is evident as strong curvature in the distribution of sample scores in the first two principal axes. The curvature can be strong enough that scores along the first axis are involuted and form a horseshoe shape. The horseshoe effect is an artifact of ordination techniques, such as Principal Component Analysis, when they are applied to very long gradients where few species are shared between widely separated samples (high β-diversity). Correspondence analyses tend to reduce the severity of the horseshoe effect.

Least-Squares Regression The process of fitting a function (here a polynomial) to data (abundance versus time) such that the sum of the squared residuals is minimized.

Noncentral Distributions A probability distribution (such as an F, t, or χ^2 distribution) that accounts for a non-zero effect size. Null hypotheses are tested with the familiar (central) F, t, or χ^2 distributions. Alternative hypotheses are tested using noncentral distributions.

Multivariate Analysis An analysis that simultaneously examines the abundance of many different species in a set of intertidal samples. Multivariate methods take advantage of correlations in species response to distill pertinent information about community structure and its response to environmental influences.

NOAA National Oceanic and Atmospheric Administration

Null Hypothesis (H_o) The oil spill and cleanup had no measurable effect on the abundance of intertidal organisms.

Ordination A multivariate technique that arranges or "orders" (as in ordination) samples along an axis based on species composition. This ordination can be conducted along a number of dimensions (usually 2 or 3) that approximate some pattern of response of the intertidal community to underlying environmental gradients (such as grain size or hydrocarbon exposure). Thus, ordination condenses the complex species-abundance database into a few factors responsible for observed variability within the intertidal community, while retaining ecologically meaningful biological information.

Parallelism A condition where time profiles of average abundance at control and impact sites track one another through time. Observed temporal excursions act in unison so that a constant difference in (logarithmic) abundance is maintained.

Power ($1-\beta$) The probability of correctly finding an important impact. It is the complement of β, which is the probability of missing a meaningful impact. It measures the desirable likelihood of correctly rejecting the null hypothesis (H_0). The power depends on the significance criterion ($\bullet$), the variability of the sample results, and the size of the impact (Δ).

Principal Components Analysis (PCA) An ordination technique that involves an eigenanalysis of the correlation matrix. Ideally, the first principal component will represent the dominant environmental gradient. The second component will be orthogonal (completely uncorrelated) with the first, and will explain some of the residual variation. This class of ordination techniques works best for monotonic distributions where species abundance steadily increases or decreases along an environmental gradient. In reality, organism abundance tends to have a unimodal distribution (rises and falls along the gradient), but may appear to be monotonic if small portions of the gradient are sampled.

***p*-Value** The measured probability of incorrectly finding an important impact when it is in fact, inconsequential. It is compared to the •-value to indicate the statistical significance of the hypothesis test.

Sample Score The coordinates along ordination axes specifying the location of a sample. They are often related to environmental gradients and represent the specific intertidal community that is best suited to a particular ecological niche.

Separation Index (*C*) Measures the separation between the mean community composition at impact and reference sites on ordination diagrams in terms of the number of standard deviations determined from the scatter of observations around each mean.

Singleton A taxon where only a single organism was collected in a given set of samples. These exceedingly rare, sparsely populated taxa are analytically problematic because their variability is indeterminate.

Sparse Taxa Sparsely populated intertidal taxa have an average count or percent cover that is less than 0.07 in the PWS dataset, while abundant taxa have average densities that exceeded 3. The area sampled in the PWS infaunal cores was 0.009 m^2 while epibiotic quadrats covered an area of 0.25 m^2. Taxa with intermediate abundance lie between these density measures, i.e., 7.8 m^{-2} < infaunal density < 333 m^{-2}, 0.3 m^{-2} < motile epifaunal invertebrate density < 12 m^{-2}, or 0.3 % < sessile epibiotic cover < 12 %. Sparse taxa are not necessarily synonymous with taxa that are rare or infrequent in samples, although for the size of the sampling units in the PWS study, this was largely the case.

Type I Error Incorrectly rejecting the null hypothesis. Finding that the oil spill or cleanup had a tangible effect on the abundance of intertidal organisms, when in fact there was no effect.

Type II Error Incorrectly accepting a false null hypothesis. Finding that the oil spill or cleanup did not have a tangible effect on the abundance of intertidal organisms, when in fact the alternative hypothesis was correct and there was a measurable impact.

Unimodal Distribution A species frequency distribution with one mode indicating that the species has one optimal environmental condition. Any increase or decrease in environmental conditions from this optimum will be less hospitable to the species and result in lower abundance. Ordination techniques based on correspondence analysis perform best when species have unimodal distributions.

Within-Site The number (m) or variability (CVw) of replicate samples collected along a site or beach.

Appendix B. Power Formulation for Treatment Effects

This Appendix provides the basis for the computational procedures used to determine sample sizes for the detection of treatment effects from a comparison of mean populations at reference sites and sites subjected to a specific cleanup method or experimental manipulation. Sample-size determinations are based on power analyses, which require estimates of ambient variability in the intertidal biota. Techniques for estimating variance and coefficients of variation (CVs) are also presented in this Appendix. These formulations were applied to intertidal data collected in Prince William Sound (PWS) in the decade following the *Exxon Valdez* oil spill as described in Chapter 2 of this report. The statistical constructs can also be applied to intertidal data collected in other locales where variability in the distribution of intertidal organisms is thought to differ from those of PWS.

The statistical formulation separates variability into two components. Small-scale or "*within-site*" variability is associated with differences in population measurements among epibiotic quadrats or infaunal sediment cores collected at adjacent sites along a particular beach. Large-scale differences between beach sites are quantified by a "*between-site*" measure of variability. Between-site variability among impacted sites can arise because the severity of oil-spill or cleanup effects differ because of inherent environmental differences among the beaches subjected to the spill. Ambient intertidal populations can also differ among unoiled beaches due to natural differences in the physical character of the beaches. These two scales of variability determine the number of within-site (m) and between-site (n) samples that are required detect effects of size Δ at a statistical significance level α with a statistical power of $1-\beta$.

Following the format of Chapter 2, two separate power formulations are presented in this Appendix. The first applies to the analysis of changes in the abundance of individual species, taxa, or assemblages. The second formulation focuses on the detection of changes in communities as a whole. The latter is based on an analysis of multivariate distances determined from a principal component analysis (PCA), correspondence analysis (CA), or similar orthogonal ordination procedure.

Species Response

Test Statistic

An experiment to test the effect of a single treatment on a species or taxonomic group would compare the mean abundance at reference sites (μ_1) with the mean at the treated sites (μ_2). The statistical test would evaluate the null hypothesis

$$H_o: \quad \mu_1 = \mu_2 \tag{B.1}$$

against the alternative hypothesis

$$H_a: \quad \mu_1 \neq \mu_2. \tag{B.2}$$

In biological systems, environmental changes often have a multiplicative effect on abundance. Under these circumstances, a logarithmic transformation of abundance yields a more stable distribution, and variance is less dependent population size. Tests performed on log-transformed abundance, suggest rewriting the hypotheses in Equations B.1 and B.2 as:

$$H_o: \quad \frac{\mu_2}{\mu_1} = 1 \tag{B.3}$$

$$H_a: \quad \frac{\mu_2}{\mu_1} \neq 1 \tag{B.4}$$

where the means are now computed from log-transformed counts. For most intertidal taxa, deleterious environmental influences will be expected to result in a reduction in the population. In this case, the alternative hypothesis (H_a) can be written as

$$H_a: \quad \frac{\mu_2}{\mu_1} < 1 \tag{B.5}$$

which can be tested against a one-tailed sampling distribution. However, opportunistic taxa may actually increase in abundance after an oil spill or other habitat disturbance due to either their enhanced tolerance of hydrocarbon exposure, or decreased inter-species competition. In the absence of other information, the alternative hypothesis (B.4) should be tested against a two-tailed sampling distribution to allow for increases or decreases in populations. In practice, after an oil spill, population increases are rarely of concern.

Under log-transformation, Hypothesis (B.3) can be evaluated using a two-sample t-test parameter

$$t_{n_1+n_2-2} = \frac{(\bar{x}_2 - \bar{x}_1) - 0}{\sqrt{s^2_{\text{Pool}}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \tag{B.6}$$

where:

$\bar{x}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}$;

x_{ij} = log-abundance for the j^{th} replicate ($j = 1,\ldots,n_i$) in the i^{th} treatment ($i = 1,2$);

n_i = number of replicate sites for i^{th} treatment ($i = 1,2$); and

s^2_{Pool} = variance computed from pooled data at the treatment and reference sites.

The test statistic (B.6) is *t*-distributed with $n_1 + n_2 - 2$ degrees of freedom under H_o. Because the test is based on the assumption that H_0 is true, *i.e.*, that there is no difference in means, it is tested against a central *t*-distribution (the distribution is "central" because the true difference in means is assumed to be zero). Compilations of the theoretical central *t*-distribution are readily available and comparisons with the *t*-test parameter can be made at a variety of significance levels (α).

Power Formulation

In contrast to the test for acceptance of H_o, a test of the alternative hypothesis, namely the power of the statistical test to reject H_o, can be only be calculated by comparing a noncentral parameter Φ, which is related to the size of the difference in means, with noncentral ***F***-distributions. Tables of noncentral distributions are less accessible than commonplace central distributions. Utilizing replicate samples collected from within each site under the alternative hypothesis, the noncentral test statistic is

$$\Phi_{1,n_1+n_2-2} = \frac{1}{\sqrt{2}} \cdot \frac{\left|\ln\left(\frac{\mu_2}{\mu_1}\right)\right|}{\sqrt{\frac{\left(\frac{\sigma^2_{B_1}}{n_1} + \frac{\sigma^2_{W_1}}{n_1 m_1}\right)}{\mu_1^2} + \frac{\left(\frac{\sigma^2_{B_2}}{n_2} + \frac{\sigma^2_{W_2}}{n_2 m_2}\right)}{\mu_2^2}}} \tag{B.7}$$

where:

$\sigma^2_{B_i}$ = between-site variance for the i^{th} treatment ($i = 1,2$);

$\sigma^2_{W_i}$ = within-site variance for the i^{th} treatment ($i = 1,2$); and

m_i = number of samples drawn within a site for the i^{th} treatment ($i = 1,2$).

This noncentrality parameter can be rewritten in terms of the coefficients of variation (CV) and the magnitude of the change (Δ) as

$$\Phi_{1,n_1+n_2-2} = \frac{1}{\sqrt{2}} \cdot \frac{\left|\ln(1+\Delta)\right|}{\sqrt{\left(\frac{CV_{B_1}^2}{n_1} + \frac{CV_{W_1}^2}{n_1 m_1}\right) + \left(\frac{CV_{B_2}^2}{n_2} + \frac{CV_{W_2}^2}{n_2 m_2}\right)}} \qquad \text{(B.8)}$$

where:

$CV_{B_i} = \frac{\sigma_{B_i}}{\mu_i}$;

$CV_{W_i} = \frac{\sigma_{W_i}}{\mu_i}$; and

$\Delta = \frac{\mu_2}{\mu_1} - 1 =$ the fractional difference in abundance at treated sites relative to reference sites.

For example, if the treatment causes a 25% reduction in mean abundance, then $\Delta = -0.25$. In a two-sided test where the abundance of an opportunistic taxon increases in abundance by 33.3% at the treated sites, then $\Delta = +0.33$. All else being equal, the power of these two effect-sizes is identical because the noncentrality parameter in (B.8) is the same in each case by virtue of the absolute value of the $\left|\ln(1+\Delta)\right|$ factor in the numerator.

It is important to note that the noncentrality parameter (B.8) is also a function of both the between-site (*i.e.*, CV_B) and within-site (*i.e.*, CV_W) variability. Empirical values for CV_B and CV_W can be used in conjunction with Equation (B.8) to determine the power of the test for specified numbers of between-site (n) and within-site (m) samples. Ideally, a site-specific preliminary survey would be used to estimate the CVs in Equation B.8. The resulting power analysis would yield optimal sample sizes to be used in the design of a full oil-spill monitoring program. Alternatively, observed values of CVs from the PWS intertidal monitoring study can be used as preliminary estimates for sample size calculations. The projected power of the statistical hypothesis test (Equation B.3) can then be determined by comparing empirical values of Equation B.8 with tables of theoretical values of the noncentral F-distribution (Tiku, 1967, 1972; or Skalski and Robson, 1992).

Variance Estimation

In Chapter 2, sample-size recommendations were based on an evaluation of Equation B.8 using CVs computed from the intertidal data collected as part of the PWS monitoring program. This section describes how the CVs were computed. The techniques for estimating the CVs described in this section can be applied to data from pilot studies in locales where the CVs determined from the PWS intertidal data are deemed inappropriate.

For the PWS data, the between-site and within-site variance components for each taxon or taxonomic group were calculated using data collected during a single year. A standard statistical construct was used to estimate these two variance components, namely, a single-classification ANOVA with unequal sample size (Table B.1).

Table B.1. Single Classification ANOVA used to Estimate Variance Components

Source	df	SS	MS	E(MS)
Total	N			
Mean	1			
$\text{Total}_{\text{Cor}}$	$N-1$	SSTOT		
Between Sites	$n-1$	SST	MST	$\sigma_W^2 + \frac{\left(N - \frac{1}{N}\sum_{i=1}^{n} m_i^2\right)}{(n-1)}\sigma_B^2$
Within Sites	$N-n$	SSE	MSE	σ_W^2

where:

m_i = number of quadrat or core samples collected at the i^{th} site ($i = 1,\ldots,n$), *viz.*, the number of within-site observations;

n = number of sites; and

$N = \sum_{i=1}^{n} m_i$, or the total number of samples collected during the given year that the data was collected

By applying this ANOVA model to the abundance measured in a number of quadrat/core samples collected at several sites or beaches, overall variability (SSTOT) can be partitioned into a component associated with small-scale variability within the sites (SSE) and a component associated with large-scale differences between the various beaches or sites (SST). The within-site variability is measured by the expected mean-square error, $\hat{\sigma}_W^2 = \text{MSE}$, while the between-site variance can be computed from

$$\hat{\sigma}_B^2 = \left\{ \begin{array}{ll} 0, & \text{if MSE>MST} \\ \dfrac{\text{MST-MSE}}{\left(N - \dfrac{1}{N}\sum_{i=1}^{n} m_i^2\right)}, & \text{if MST> MSE} \\ \hline (n-1) \end{array} \right\}. \qquad \text{(B.9)}$$

Two sets of variance estimates were computed from the ANOVAs applied to the PWS data. One set was representative of sites impacted by oil and was computed from intertidal data collected at oiled sites prior 1992. A second set was representative of healthy intertidal populations and was computed from post-recolonization data collected in the years from 1993 through 2000.

Because the variances estimates computed from different years of data did not always include the same number of sites, or even the same number of samples from within those sites, the pooled variance was computed by weighted averaging. The pooled estimate of between-site variance $\left(\hat{\bar{\sigma}}_B^2\right)$ was computed by weighting the between-site variance $\left(\hat{\sigma}_{B_j}^2\right)$ determined for the j^{th} year ($j = 1,\ldots,Y$) by the number of sites $\left(n_j\right)$ used in the ANOVA for that year (Equation B.10).

$$\hat{\bar{\sigma}}_B^2 = \frac{\sum_{j=1}^{Y}\left(n_j - 1\right)\hat{\sigma}_{B_j}^2}{\sum_{j=1}^{Y}\left(n_j - 1\right)} \qquad \text{(B.10)}$$

Similarly, the pooled estimate of within-site variance $\left(\hat{\bar{\sigma}}_W^2\right)$ across years was calculated by weighting the within-site variance $\left(\hat{\sigma}_{W_j}^2\right)$ determined for the j^{th} year by the number of samples within each site used in the ANOVA for that year (Equation B.11).

$$\hat{\bar{\sigma}}_W^2 = \frac{\sum_{j=1}^{Y}\left(N_j - n_j\right)\hat{\sigma}_{W_j}^2}{\sum_{j=1}^{Y}\left(N_j - n_j\right)} \qquad \text{(B.11)}$$

where N_j is the total number of observations collected in the j^{th} year or, $N_j = \sum_{i=1}^{n_j} m_{ij}$.

Coefficient of Variation

With the variance estimates computed using the techniques described in the previous section, the CVs can be determined after scaling by the appropriate estimate of mean abundance. The grand mean $\left(\overline{\overline{x}}_j\right)$ for the j^{th} year is an average of the means $\left(\overline{x}_{ij}\right)$ computed at the n_j sites

$$\overline{\overline{x}}_j = \frac{1}{n_j}\sum_{i=1}^{n_j}\overline{x}_{ij}. \tag{B.12}$$

The grand mean can also be written as a function of population densities $\left(x_{ijk}\right)$ in the individual samples collected at each site within a given year.

$$\overline{\overline{x}}_j = \frac{1}{n_j}\sum_{i=1}^{n_j}\left(\frac{1}{m_{ij}}\sum_{k=1}^{m_{ij}}x_{ijk}\right) \tag{B.13}$$

where x_{ijk} is the population density measured within the k^{th} epibiotic quadrat or infaunal core ($k = 1,\ldots,m_{ij}$) collected at the i^{th} replicate site ($i = 1,\ldots,n_j$) in the j^{th} year ($j = 1,\ldots,Y$).

For the j^{th} year, estimates of the within-site coefficient of variation ($\hat{CV}_{W_j}$) and the between-site coefficient of variation ($\hat{CV}_{B_j}$) are found by normalizing the estimated variance by the grand mean.

$$\hat{CV}_{W_j} = \frac{\sqrt{\hat{\sigma}_{W_j}^{\,2}}}{\overline{\overline{x}}_j} \tag{B.14}$$

$$\hat{CV}_{B_j} = \frac{\sqrt{\hat{\sigma}_{B_j}^{\,2}}}{\overline{\overline{x}}_j} \tag{B.15}$$

As with the variance estimates, weighted averaging across years increased the reliability of the CV estimates.

$$\hat{CV}_W = \frac{\sum_{j=1}^{Y}\left(N_j - n_j\right)\hat{CV}_{W_j}}{\sum_{j=1}^{Y}\left(N_j - n_j\right)} \tag{B.16}$$

$$\bar{e}v_B = \frac{\sum_{j=1}^{Y}(n_j - 1)\bar{e}v_{B_j}}{\sum_{j=1}^{Y}(n_j - 1)} \quad \text{(B.17)}$$

These CVs were computed from the PWS data for individual taxa at the three tidal elevations. The results are presented graphically and numerically in Appendix C. The CVs for data collected immediately after the spill are presented on the left-hand side along with mean abundances. The variability and abundance of post-recovery populations are presented on the right-hand side of the figures.

Community Response

This section provides the formulation for power tests that can be conducted on multivariate parameters such as those determined from a PCA or a CA. The formulation can simultaneously analyze multiple dimensions and is based on an analog to analysis of variance (ANOVA). It is designated "analysis of distance" or ANODIS where distance refers to separations in a multivariate plot.

Test Statistic

The distance data used in the ANODIS come from the locations of the sample points in an ordination diagram, for example, sample scores for each individual sample that arise from a PCA, or are derived from a CA. The ANODIS is based on the fact that principal components (for PCA) and ordination axes (for CA) are orthogonal or, in other words, independent. In addition, the sample scores are linear combinations of the original observations and should be asymptotically normally distributed. The typical result of PCA is a sample-score plot as illustrated in Figure 2.4.

The null hypothesis of the ANODIS is that all the samples come from a single population with a common center located at the multivariate mean of all the samples $[\bar{\mu} = (\mu_x, \mu_y, ...)]$. The alternative hypothesis is that the samples from treated and reference sites come from different populations with different centers in the multivariate hyperspace. A test for differences in two treatments, for example a group of treated samples and a group of reference samples, would evaluated the null hypothesis

$$H_o: \quad \bar{\mu}_1 = \bar{\mu}_2 \quad \text{(B.18)}$$

against the alternative hypothesis

$$H_a: \quad \bar{\mu}_1 \neq \bar{\mu}_2. \tag{B.19}$$

For a two-dimensional ordination, the hypothesis test is based on an ANODIS table setup in the form of a one-way classification (Table B.2).

Table B.2. One-way classification of a bivariate ordination using ANODIS

Source	df	SS	MS	F
$\text{Total}_{\text{Cor}}$	$2\left\{\sum_{i=1}^{K}[n_i]-1\right\}$	$\sum_{i=1}^{K}\sum_{j=1}^{n_i}\left[\left(x_{ij}-\bar{\bar{x}}\right)^2+\left(y_{ij}-\bar{\bar{y}}\right)^2\right]$		
Treatment	$2(K-1)$	$\text{SST}=\sum_{i=1}^{K}n_i\left[\left(\bar{x}_i-\bar{\bar{x}}\right)^2+\left(\bar{y}_i-\bar{\bar{y}}\right)^2\right]$	$\text{MST}=\frac{\text{SST}}{2(K-1)}$	$\text{F}=\frac{\text{MST}}{\text{MSE}}$
Error	$2\sum_{i=1}^{K}(n_i-1)$	$\text{SSE}=\sum_{i=1}^{K}\sum_{j=1}^{n_i}\left[\left(x_{ij}-\bar{x}_i\right)^2+\left(y_{ij}-\bar{y}_i\right)^2\right]$	$\text{MSE}=\frac{\text{SSE}}{2\sum_{i=1}^{K}(n_i-1)}$	

where: K is the number of treatments; n_i is the number of replicate samples for the i[th] treatment; x_{ij} is the sample score along the x-ordination axis (first principal axis) for the j[th] sample in the i[th] treatment; $\bar{x}_i$ is the mean of sample scores along the x-axis for the i[th] treatment; and $\bar{\bar{x}}$ is the grand mean over all the sample scores along the x-axis for all the treatments.

Each term in the sum-of-squares equations addresses a different dimension and these terms can be summed separately. Consequently, the treatment sum of squares can be rewritten as:

$$\text{SST}=\sum_{i=1}^{K}n_i\left(\bar{x}_i-\bar{\bar{x}}\right)^2+\sum_{i=1}^{K}n_i\left(\bar{y}_i-\bar{\bar{y}}\right)^2 \tag{B.20}$$

Each term in Equation B.20 represents a treatment sum-of-squares for a traditional univariate ANOVA performed on a single dimension. The x-axis sample scores are independent of y-axis scores, and assuming the data are normally distributed, each term is chi-squared (χ^2) distributed with $(K-1)$ degrees of freedom. The pair of treatment sum-of-squares for two dimensions has $2(K-1)$ degrees of freedom. The treatment sum-of-squares measures the distances of the means for each treatment from the grand mean of all samples.

Similarly, the error sum-of-squares (SSE) measures the distance of each sample score from its respective treatment mean. If these error distances are small compared to the spread of treatment means, then ***F*** is large and the pattern of sample scores probably resulted from significant differences between treatment and reference sites. As with the treatment sum-of-squares (SST), the error sum-of-squares (SSE) can be partitioned into separate components for each dimension:

$$\mathrm{SSE} = \sum_{i=1}^{K}\sum_{j=1}^{n_i}\left(x_{ij}-\bar{x}_i\right)^2 + \sum_{i=1}^{K}\sum_{j=1}^{n_i}\left(y_{ij}-\bar{y}_i\right)^2 . \qquad \text{(B.21)}$$

With two independent normally-distributed sample scores, the sum of these two χ^2 variables, has a total degrees-of-freedom equal to the sum of the individual terms, or $2\sum_{i=1}^{K}\left(n_i-1\right)$.

The ratio of these two independent χ^2 variables (SST and SSE), divided by their respective degrees-of-freedom, follows an ***F***-distribution with $\mathrm{df}_1 = 2\left(K-1\right)$ and $\mathrm{df}_2 = 2\sum_{i=1}^{K}\left(n_i-1\right)$:

$$\mathrm{F}_{\mathrm{df}_1,\mathrm{df}_2} = \frac{\left[\dfrac{\mathrm{SST}}{2\left(K-1\right)}\right]}{\dfrac{\mathrm{SSE}}{2\sum_{i=1}^{K}\left(n_i-1\right)}}. \qquad \text{(B.22)}$$

This formulation can be extended to more than two dimensions by expanding the number of terms in SST and SSE, and increasing the degrees-of-freedom multiplicatively. The statistical significance of departures from randomly distributed sample scores (the null hypothesis) can be determined from standard (central) ***F***-distribution tables. This empirically determined ***p***-value represents the probability of incorrectly rejecting the null hypothesis of no effects. It can be compared with the preselected α-level, set at 0.1 in this report, to determine whether the departure could have been caused by chance alone. Smaller ***p***-values indicate a greater degree of confidence that the observed difference in community structure was due to the impact or treatment.

Power Formulation

A more rigorous method for evaluating the importance of an observed departure from the null hypothesis is through a power calculation. One advantage of the ANODIS formulation in the ***F***-test described above is that the distributional properties of the test statistic are well described

under both the null and alternative hypothesis. Noncentral ***F***-distributions can be used to evaluate the statistical power of tests and hence optimal sample sizes. No additional statistical theory needs to be developed, and existing noncentral ***F***-tables (Tiku 1967, 1972) can be applied to quantitatively examine perceived differences in community composition among the various treatment groups. The noncentral ***F***-distribution depends on the degrees of freedom in the numerator (df_1) and denominator of the F-test (df_2), and a noncentrality parameter defined by Tiku (1967) for a one-way classification as

$$\phi_{df_1,df_2} = \frac{1}{\sqrt{df_1 + 1}} \cdot \sqrt{\frac{\sum_{i=1}^{K} n_i (\mu_i - \mu)^2}{\sigma^2}} \qquad \text{(B.23)}$$

where n_i is the number of observations in the i^{th} treatment $(i = 1,\ldots,K)$; μ_i is the mean for the i^{th} treatment; μ is the grand mean; and σ^2 is the variance among samples within the individual treatments.

For a two-treatment, completely randomized design that has balanced treatments (*i.e.*, $n_1 = n_2 = n$), the noncentrality parameter for the ANODIS reduces to

$$\phi_{df_1,df_2} = \frac{1}{\sqrt{df_1 + 1}} \cdot \frac{\sqrt{n}}{2} \cdot \frac{|D_{\mu_1-\mu_2}|}{\sigma} \qquad \text{(B.24)}$$

where $D_{\mu_1-\mu_2}$ is the distance between the mean sample scores for treatment and reference samples in the multivariate analysis. In three dimensions, for example, a trivariate PCA, the distance between the two treatment means can be determined from

$$D_{\mu_1-\mu_2} = \sqrt{(\bar{x}_1 - \bar{x}_2)^2 + (\bar{y}_1 - \bar{y}_2)^2 + (\bar{z}_1 - \bar{z}_2)^2}. \qquad \text{(B.25)}$$

In a one dimensional PCA, the distance measure is simply the difference in mean values in an ANOVA. An estimate of σ is obtained from the square-root of the error term (*MSE*) in the ANODIS, which represents the average separation of samples from their treatment mean, for example, as formulated in Table B.2 for two dimensions.

The relative separation of the treatment means can also be characterized in the form of an index (*C*). It is defined in a manner similar to the univariate case of an ANOVA following Kirk (1982: p. 144-145), Bratcher *et al.* (1970), and the "effect size" of Cohen (1988: p 20-27; 274-288):

$$C = \frac{D_{\mu_1 - \mu_2}}{\sigma}. \tag{B.26}$$

Substitution of Equation (B.26) into Equation (B.24) yields:

$$\phi_{df_1, df_2} = \frac{C}{\sqrt{df_1 + 1}} \cdot \frac{\sqrt{n}}{2}. \tag{B.27}$$

Treatment separation parameterized in the form of the index C offers several advantages. Intuitively, the quotient C expresses the signal-to-noise ratio in a comparison of two treatments, for example, when comparing treatment and reference samples. The numerator of C is the Euclidean (straight-line) distance between treatment means in any number of ordination dimensions. Its denominator is the standard deviation of the distances between samples within each treatment. Thus, it is a measure of how many standard deviations the treatment means are separated by. The larger C becomes, the easier it is to detect changes for a given sample size n. As a unitless measurement of the relative size of the difference in treatments, it is helpful for interpreting ordination plots whose axis units cannot be easily related to the physical measurements. C expresses distances among sample scores in units of variability common to impacted and reference populations.

When there are three or more treatments, power can still be formulated in terms of the quotient C. However, the index is expressed as:

$$C' = \frac{D_{max}}{\sigma} \tag{B.28}$$

where the numerator is the greatest expected distance between any of the two treatments in the study. Using C', the minimum statistical power of the F-test is calculated assuming the remaining treatments have centers coincident with the grand centroid of the data. In any other configuration, the power of the F-test will be greater than that specified by C'.

Appendix C. Variance Distributions

The following plots summarize the abundance and variability of intertidal taxa enumerated within Prince William Sound. Taxa are ranked by the mean population during the non-impact period beginning in 1993 for infauna and starting in 1994 for epibiota. Results for populations measured when exposure to oil was greatest are shown on the left side of the plots. Statistics for these "impacted" populations were determined from epibiotic data collected in 1989 and 1990, and for infaunal data collected in 1990 and 1991.

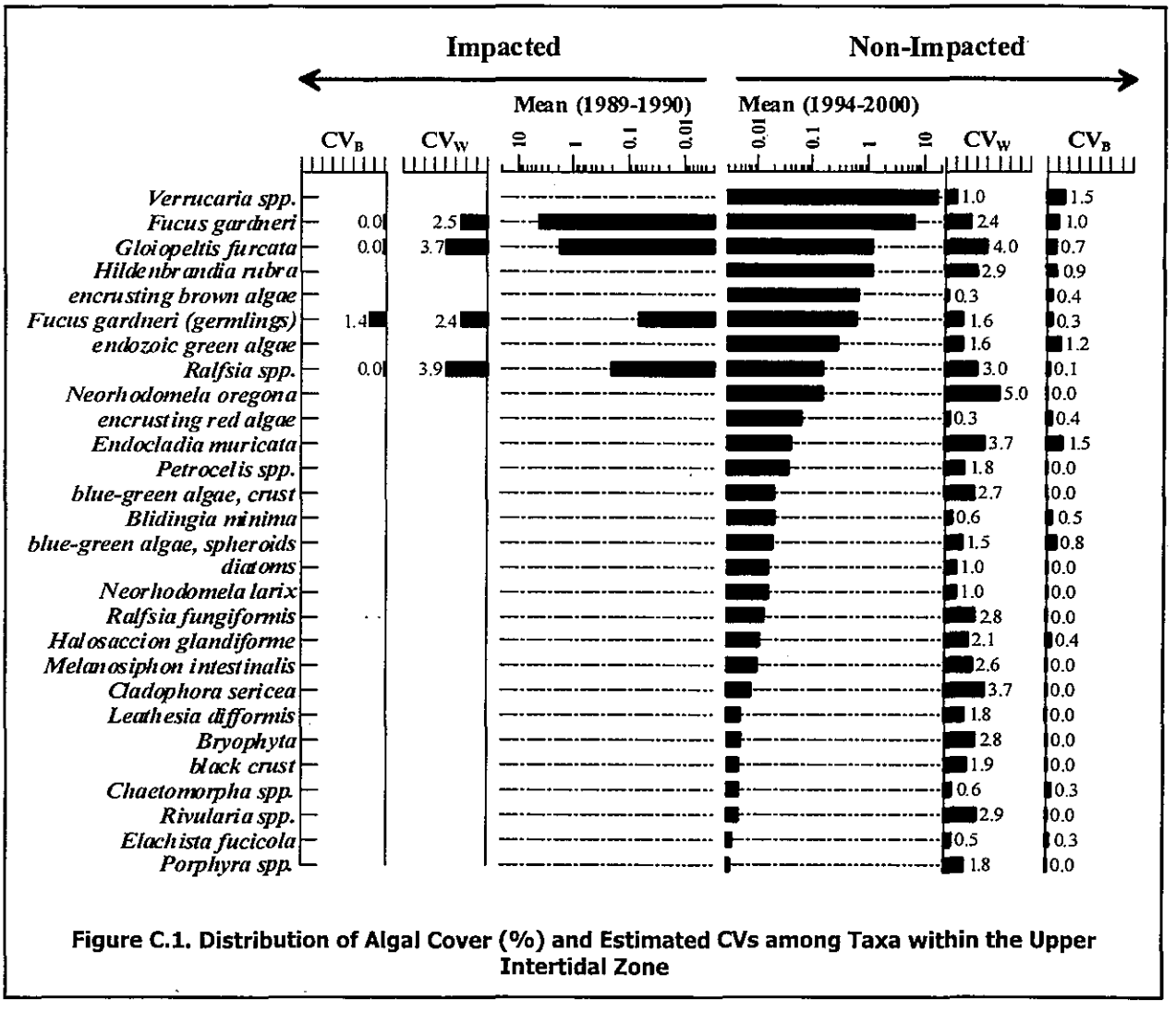

Figure C.1. Distribution of Algal Cover (%) and Estimated CVs among Taxa within the Upper Intertidal Zone

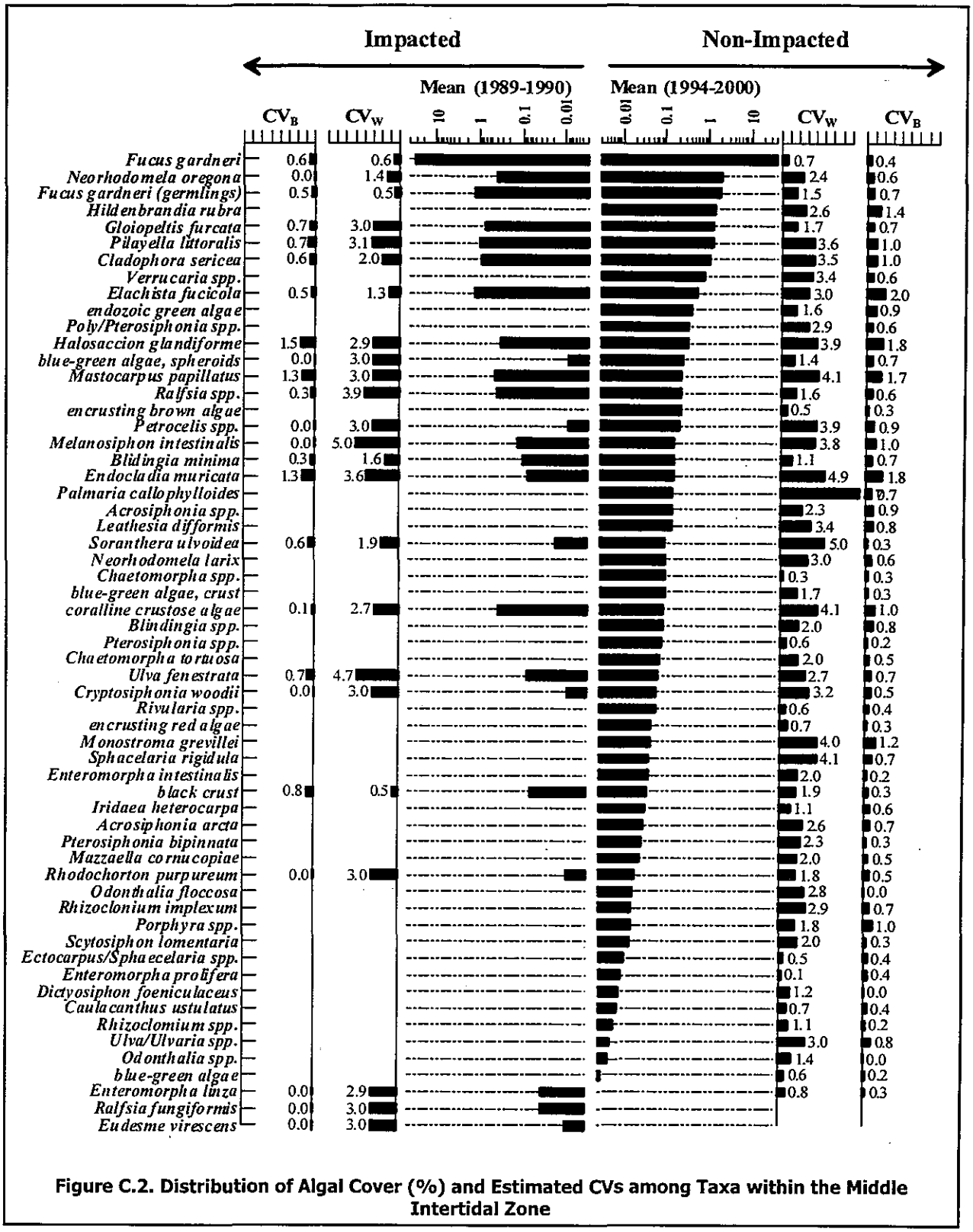

Figure C.2. Distribution of Algal Cover (%) and Estimated CVs among Taxa within the Middle Intertidal Zone

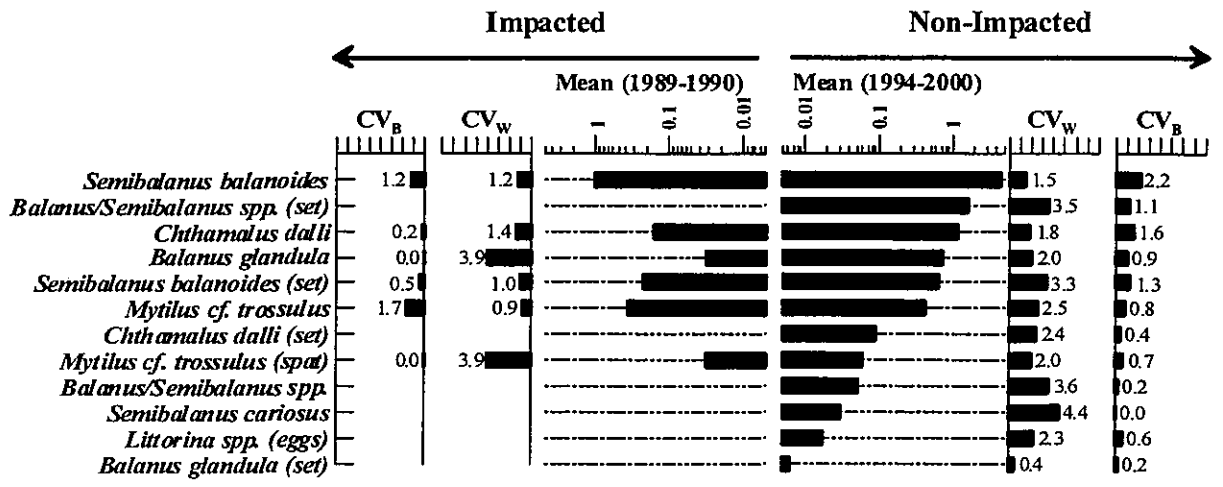

Figure C.3. Distribution of Invertebrate Cover (%) and Estimated CVs among Taxa within the Upper Intertidal Zone

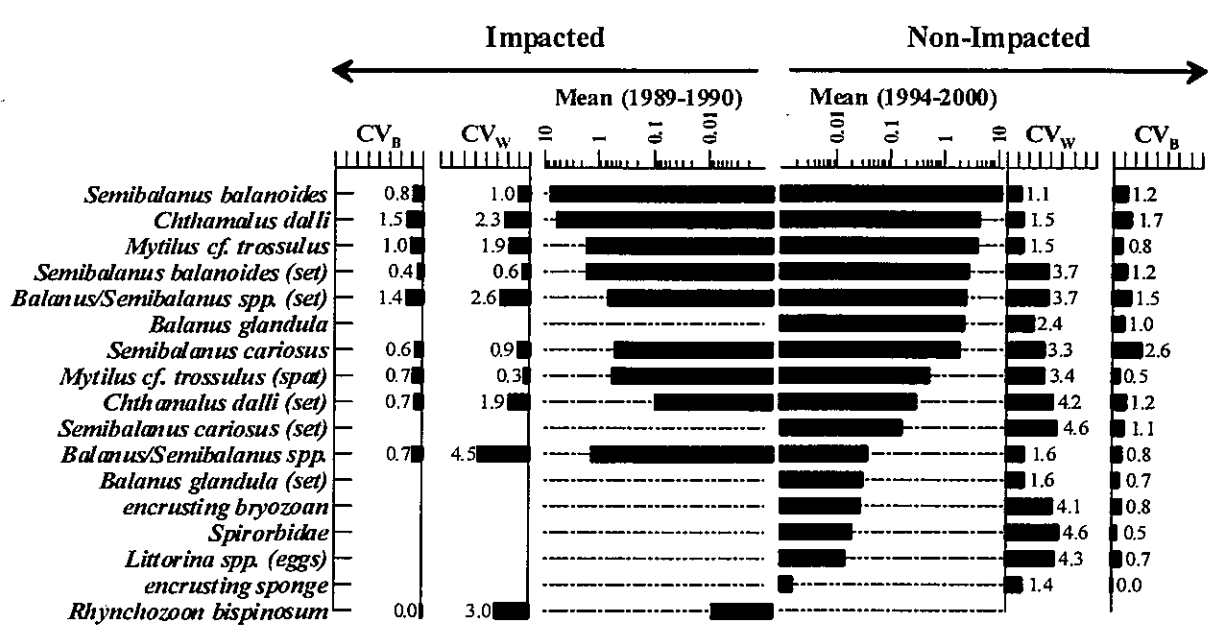

Figure C.4. Distribution of Invertebrate Cover (%) and Estimated CVs among Taxa within the

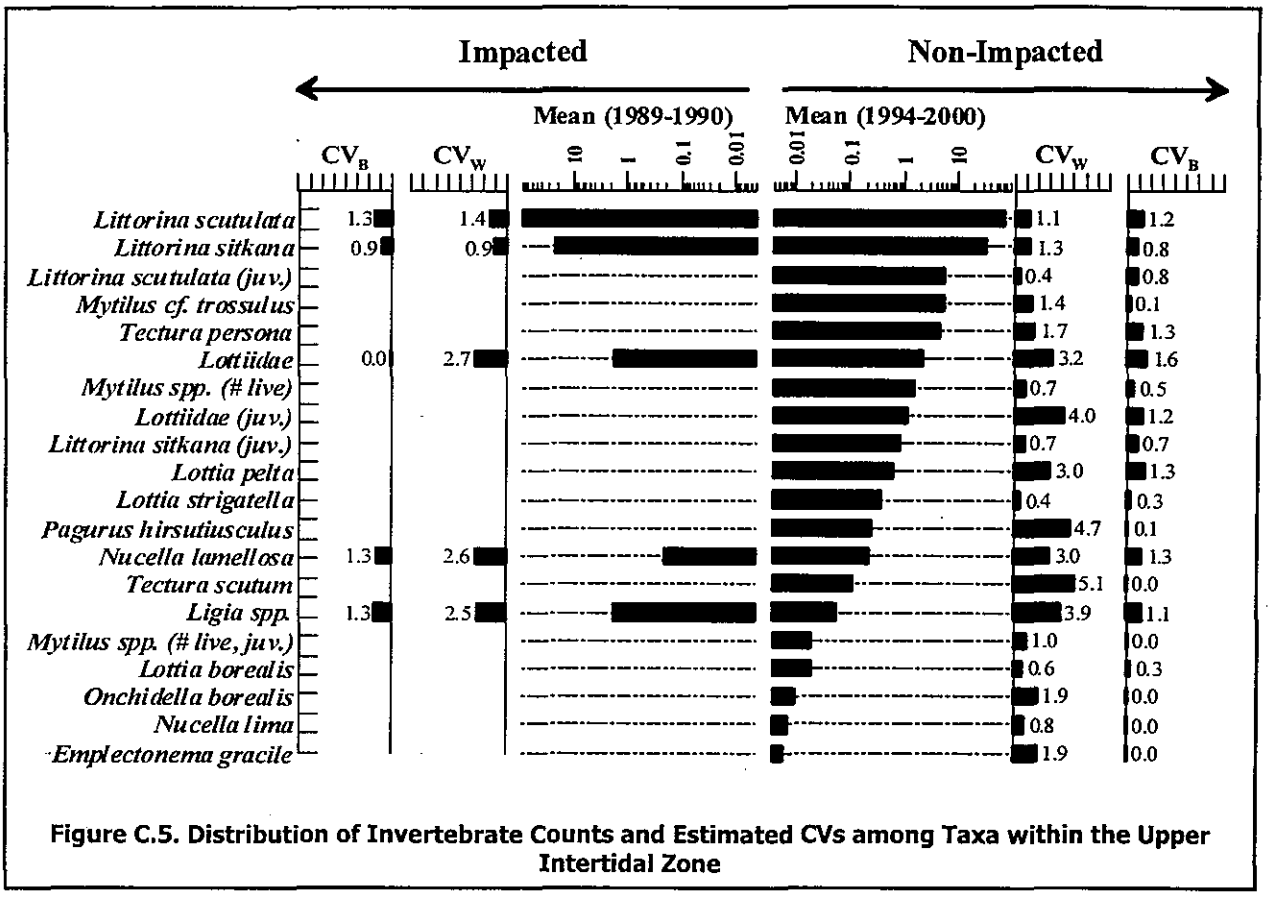

Figure C.5. Distribution of Invertebrate Counts and Estimated CVs among Taxa within the Upper Intertidal Zone

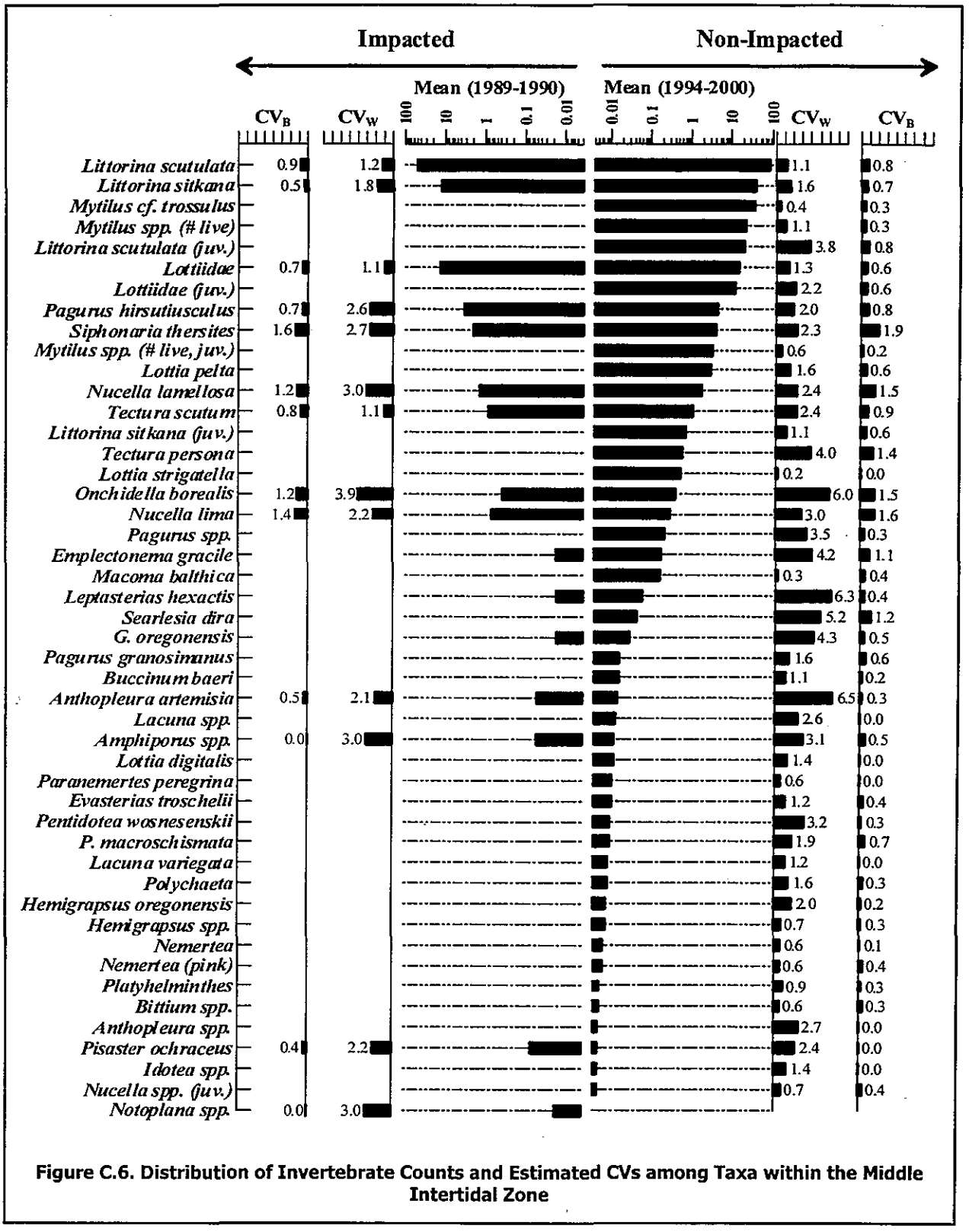

Figure C.6. Distribution of Invertebrate Counts and Estimated CVs among Taxa within the Middle Intertidal Zone

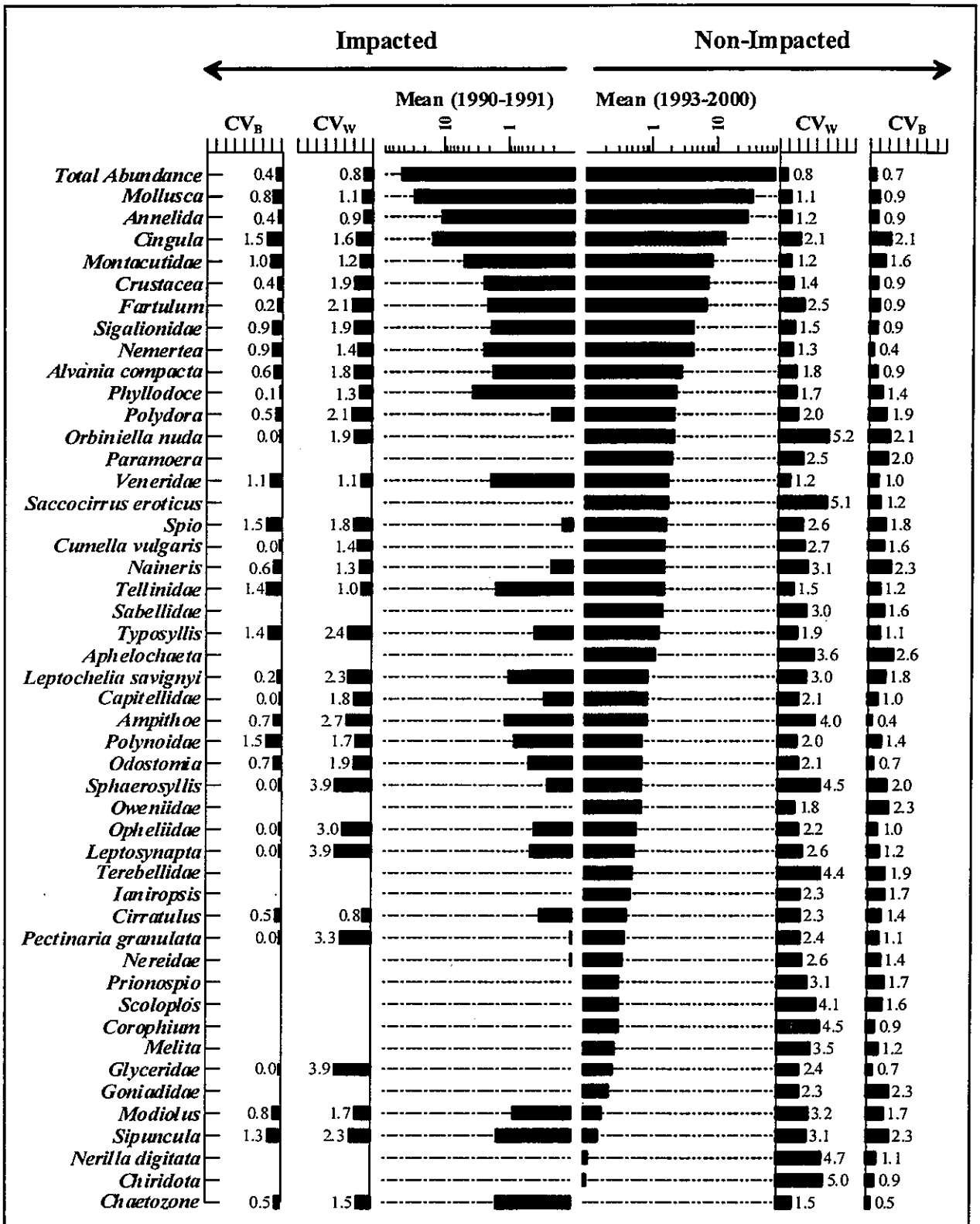

Figure C.7. Distribution of Infaunal Counts and Estimated CVs among Taxa within the Lower Intertidal Zone

Appendix D. Power Curves for Treatment Effects

The following plots provide the number of replicate samples (m) that need to be collected at n reference sites and at n treatment sites to achieve various powers between 0.2 and 0.9. Plots are provided for taxa that are sparsely populated (density < 0.07), have intermediate abundances (0.07 < density < 3), and are abundant (3 < density). Sample sizes are computed for taxa with low, moderate, and high variability within each of the three abundance ranges. As described in Chapter 2, intertidal variability was estimated from the 10^{th}, 50^{th} (median), and 90^{th} percentiles of the intertidal data collected within Prince William Sound.

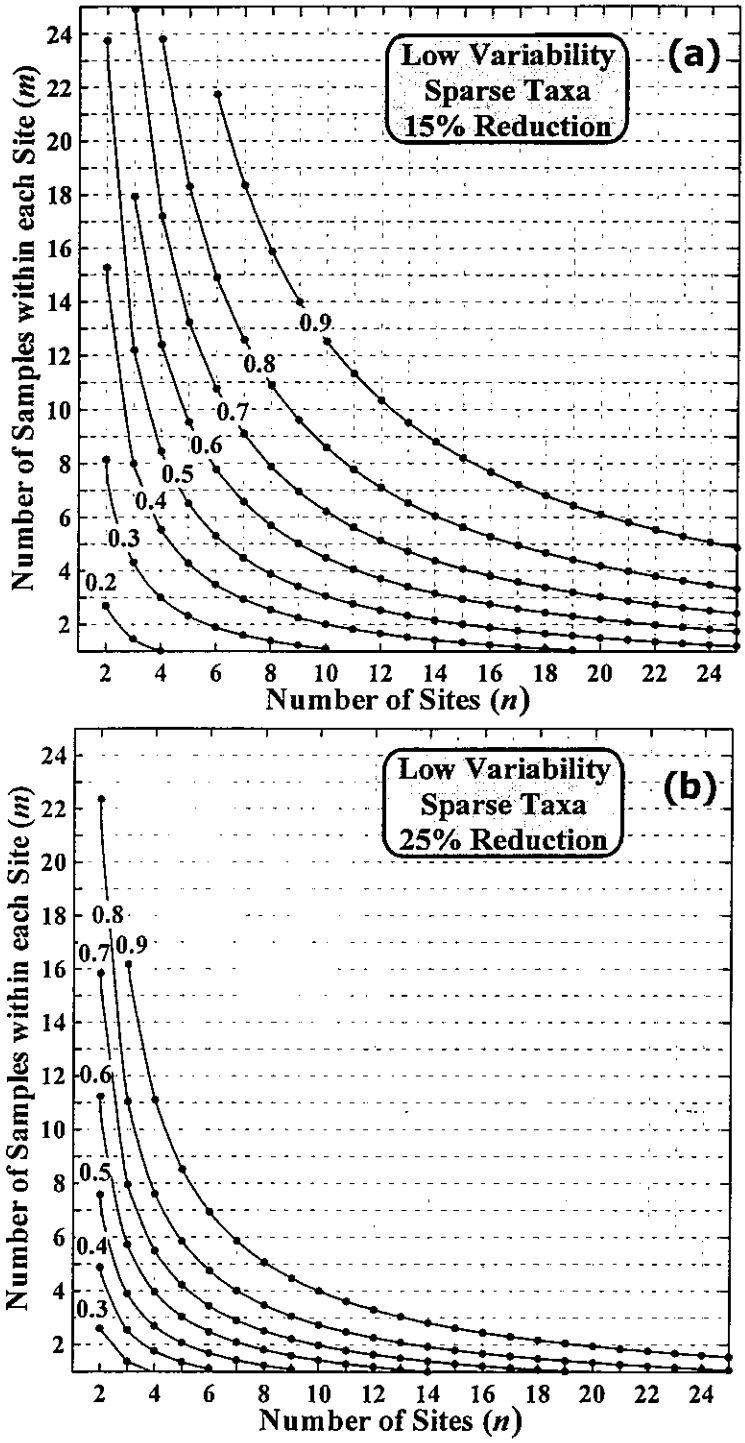

Figure D.1. Sample-size chart showing the number of replicate samples (m) collected at n reference and n treatment sites that are needed to detect a (a) 15% reduction (18% increase) or (b) 25% reduction (33% increase) in sparse intertidal populations with a statistical power ($1-\beta$) at the one-tailed significance level of $\alpha = 0.1$. The curves correspond to different levels of statistical power in an environment with low natural biological variation ($CV_W = 0.49$, $CV_B = 0.00$).

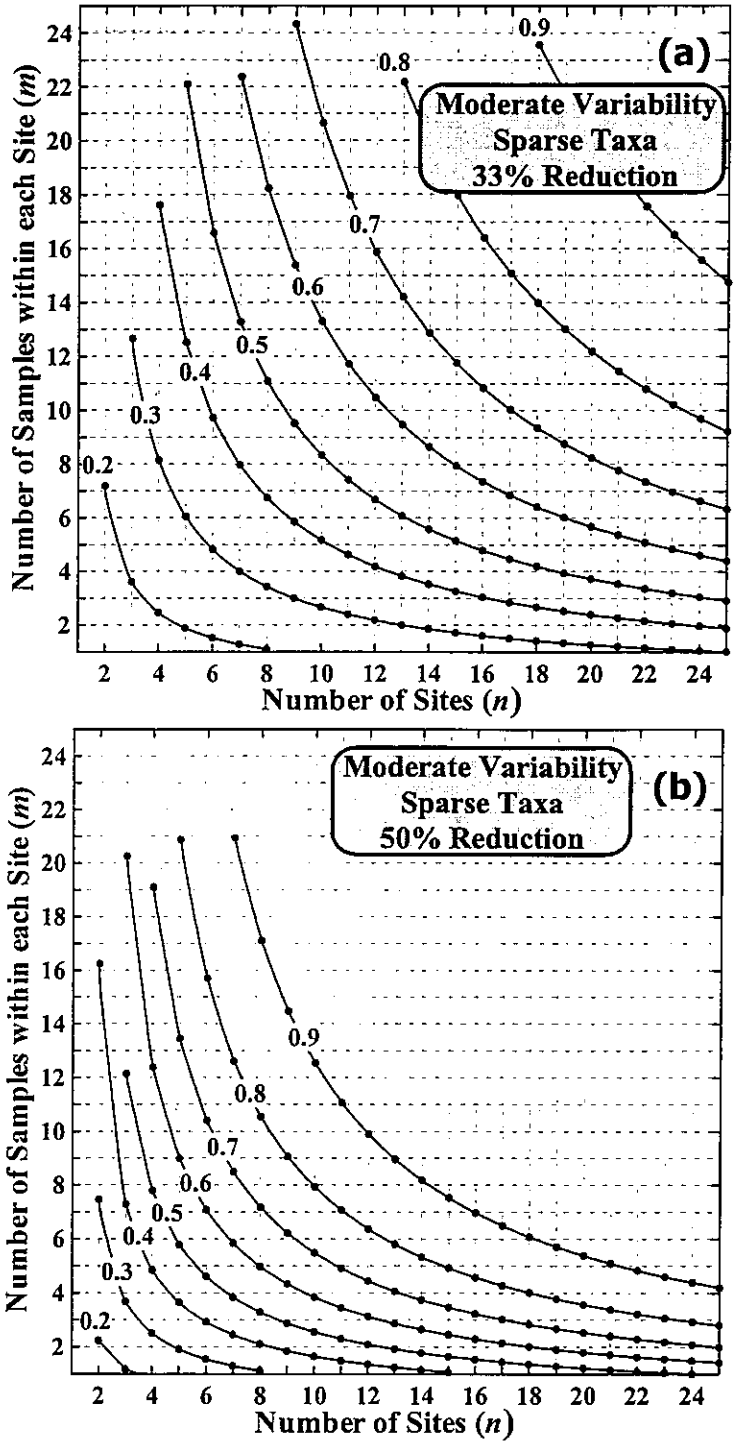

Figure D.2. Sample-size chart showing the number of replicate samples (m) collected at n reference and n treatment sites that are needed to detect a (a) 33% reduction (50% increase) or (b) 50% reduction (100% increase) in sparse intertidal populations with a statistical power ($1-\beta$) at the one-tailed significance level of $\alpha = 0.1$. The curves correspond to different levels of statistical power in an environment with moderate natural biological variation ($CV_W = 1.86$, $CV_B = 0.27$).

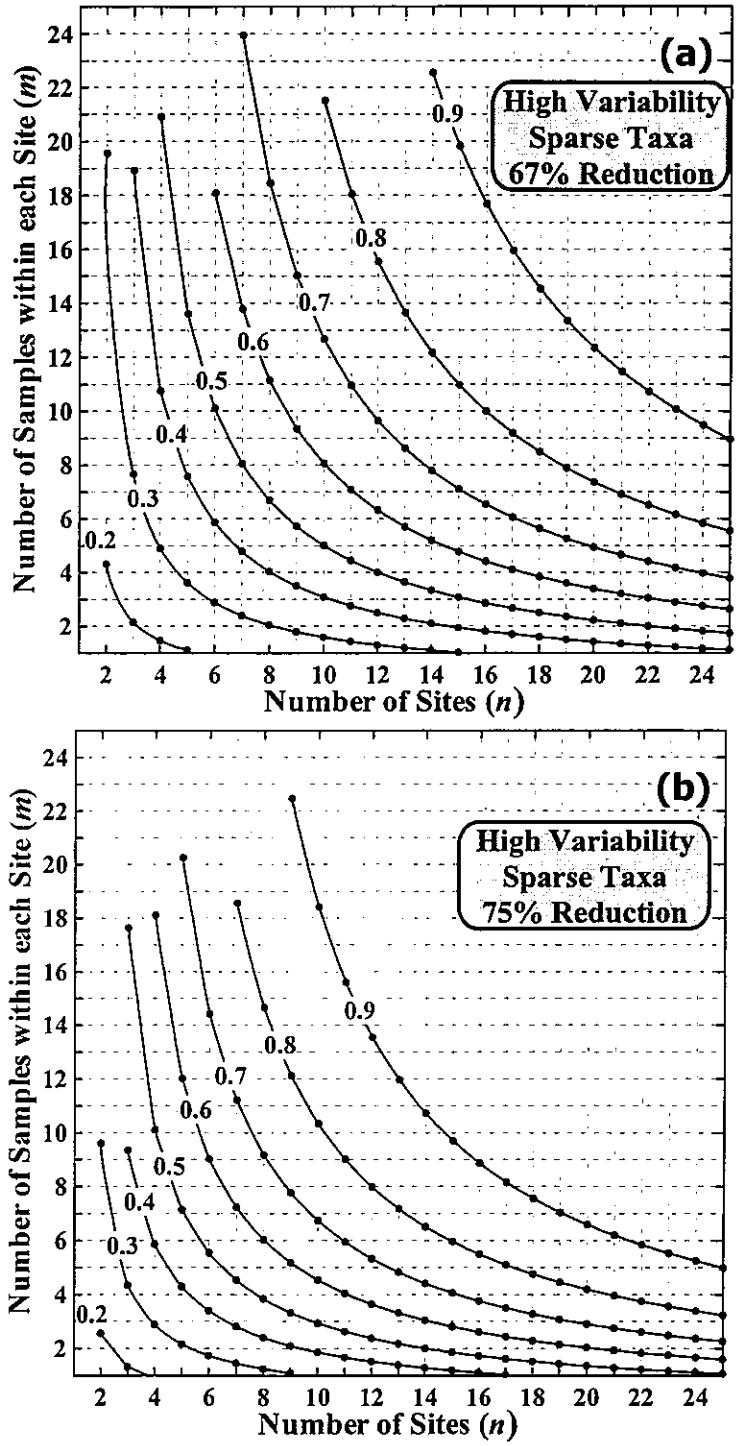

Figure D.3. Sample-size chart showing the number of replicate samples (m) collected at n reference and n treatment sites that are needed to detect a (a) 67% reduction (200% increase) or (b) 75% reduction (300% increase) in sparse intertidal populations with a statistical power ($1-\beta$) at the one-tailed significance level of $\alpha = 0.1$. The curves correspond to different levels of statistical power in an environment with high natural biological variation ($CV_W = 3.88$, $CV_B = 0.76$).

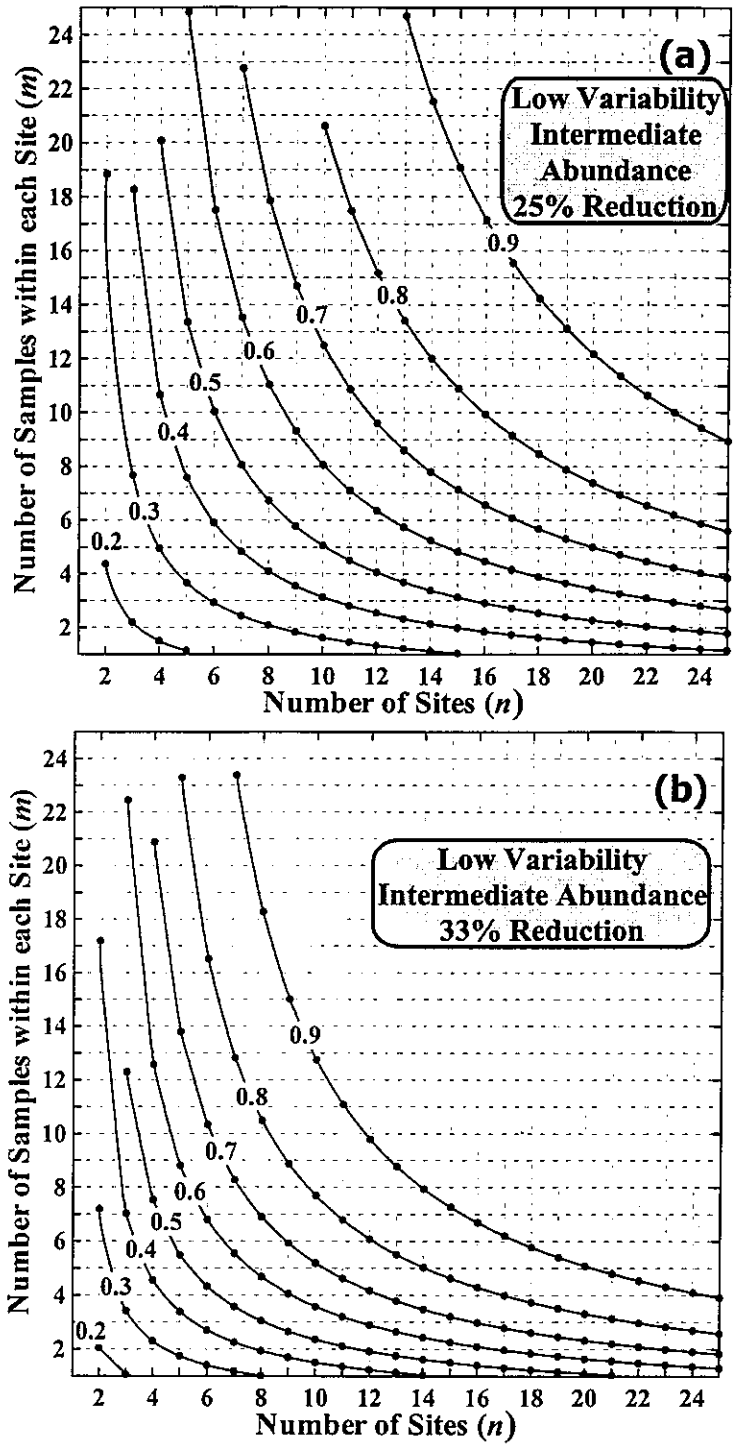

Figure D.4. Sample-size chart showing the number of replicate samples (m) collected at n reference and n treatment sites that are needed to detect a (a) 25% reduction (33% increase) or (b) 33% reduction (50% increase) in moderately dense intertidal populations with a statistical power ($1-\beta$) at the one-tailed significance level of $\alpha = 0.1$. The curves correspond to different levels of statistical power in an environment with low natural biological variation ($CV_W = 1.03$, $CV_B = 0.19$).

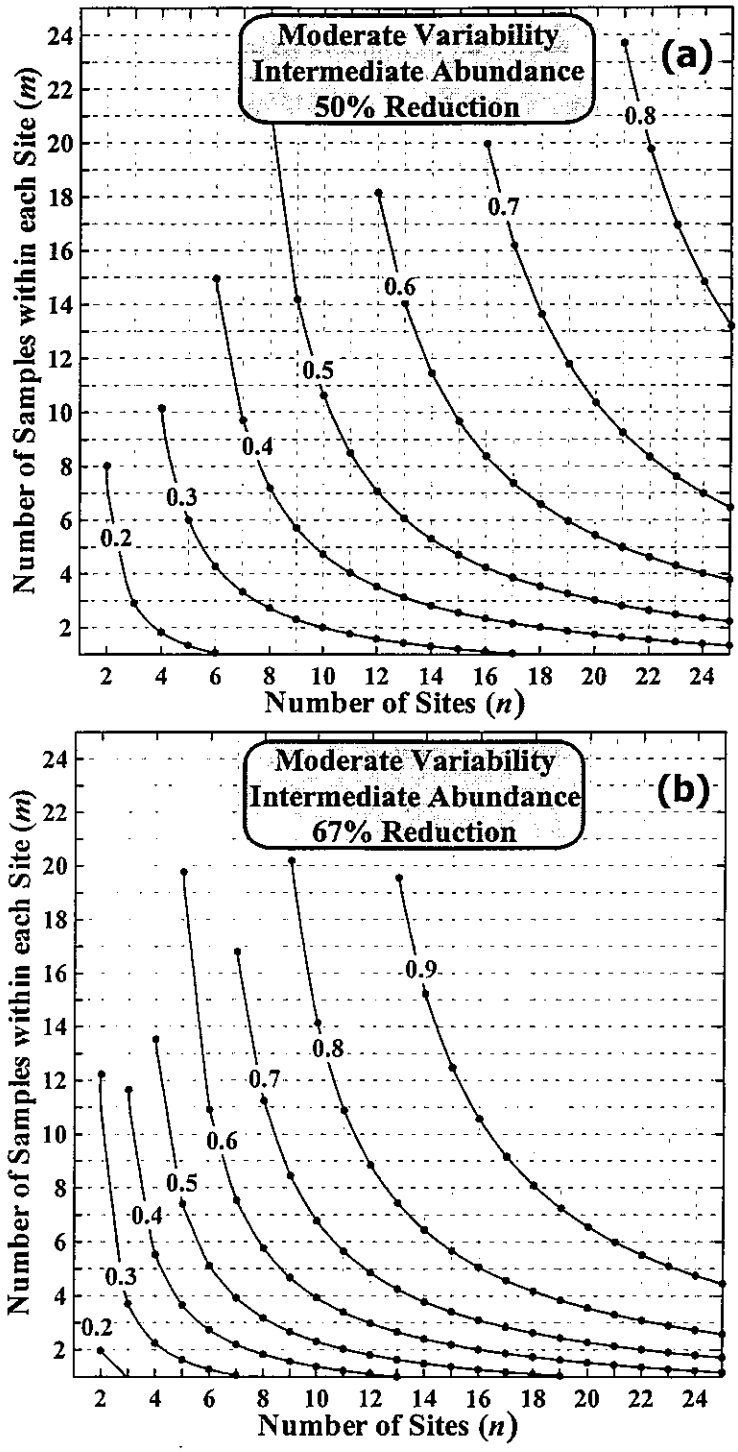

Figure D.5. Sample-size chart showing the number of replicate samples (m) collected at n reference and n treatment sites that are needed to detect a (a) 50% reduction (100% increase) or (b) 67% reduction (200% increase) in moderately dense intertidal populations with a statistical power ($1-\beta$) at the one-tailed significance level of $\alpha = 0.1$. The curves correspond to different levels of statistical power in an environment with moderate natural biological variation ($CV_W = 2.52$, $CV_B = 0.91$).

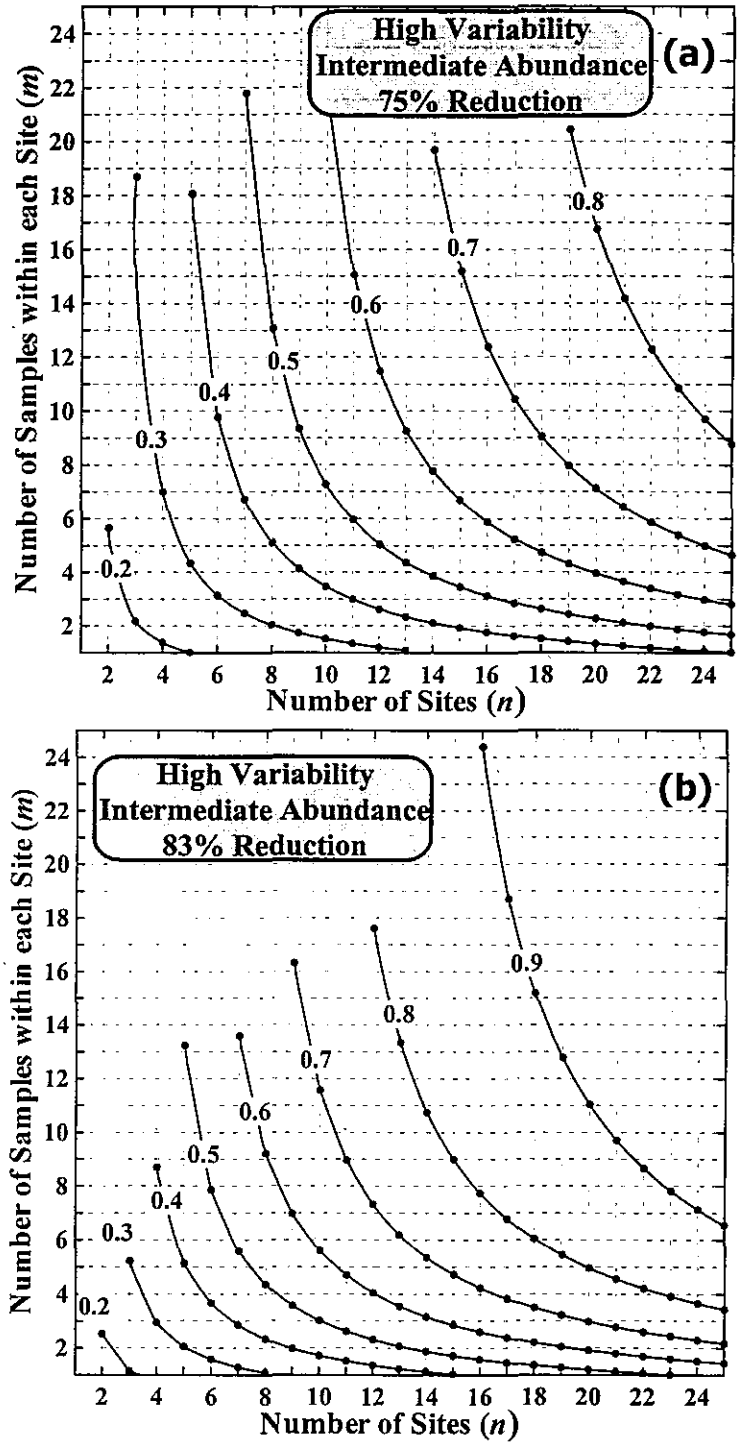

Figure D.6. Sample-size chart showing the number of replicate samples (m) collected at n reference and n treatment sites that are needed to detect a (a) 75% reduction (300% increase) or (b) 83% reduction (500% increase) in moderately dense intertidal populations with a statistical power ($1-\beta$) at the one-tailed significance level of $\alpha = 0.1$. The curves correspond to different levels of statistical power in an environment with high natural biological variation ($CV_W = 4.44$, $CV_B = 1.73$).

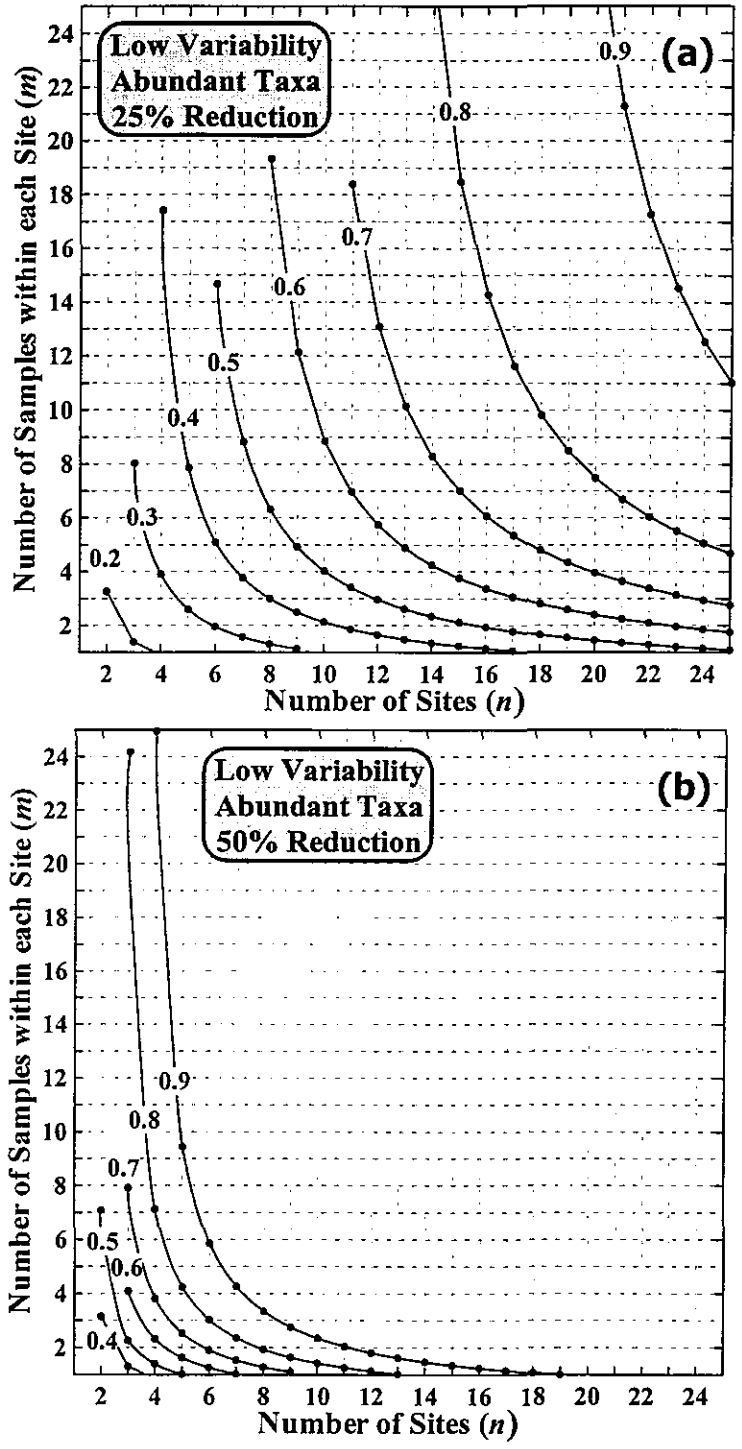

Figure D.7. Sample-size chart showing the number of replicate samples (m) collected at n reference and n treatment sites that are needed to detect a (a) 25% reduction (33% increase) or (b) 50% reduction (100% increase) in abundant intertidal populations with a statistical power ($1-\beta$) at the one-tailed significance level of $\alpha = 0.1$. The curves correspond to different levels of statistical power in an environment with low natural biological variation ($CV_W = 0.76$, $CV_B = 0.32$).

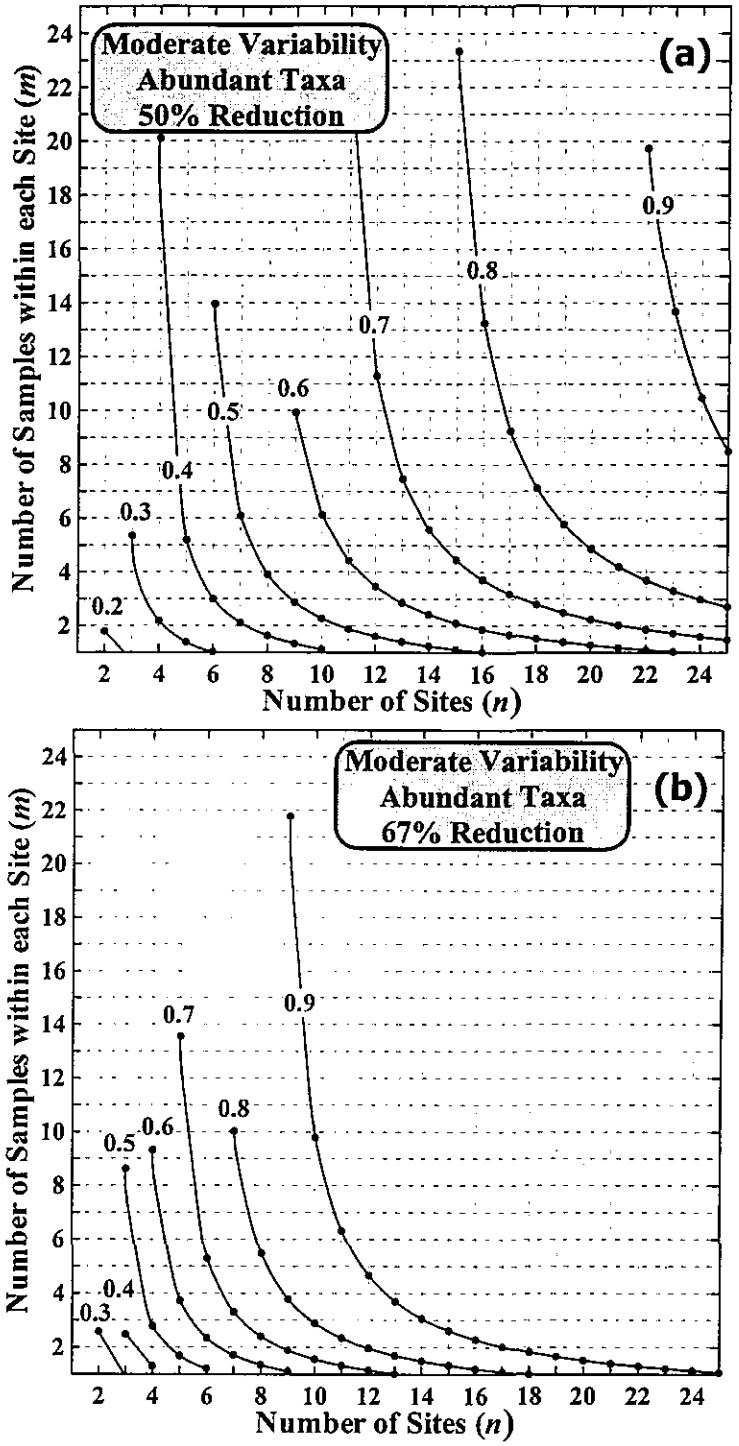

Figure D.8. Sample-size chart showing the number of replicate samples (m) collected at n reference and n treatment sites that are needed to detect a (a) 50% reduction (100% increase) or (b) 67% reduction (200% increase) in abundant intertidal populations with a statistical power ($1-\beta$) at the one-tailed significance level of $\alpha = 0.1$. The curves correspond to different levels of statistical power in an environment with moderate natural biological variation ($CV_W = 1.28$, $CV_B = 0.84$).

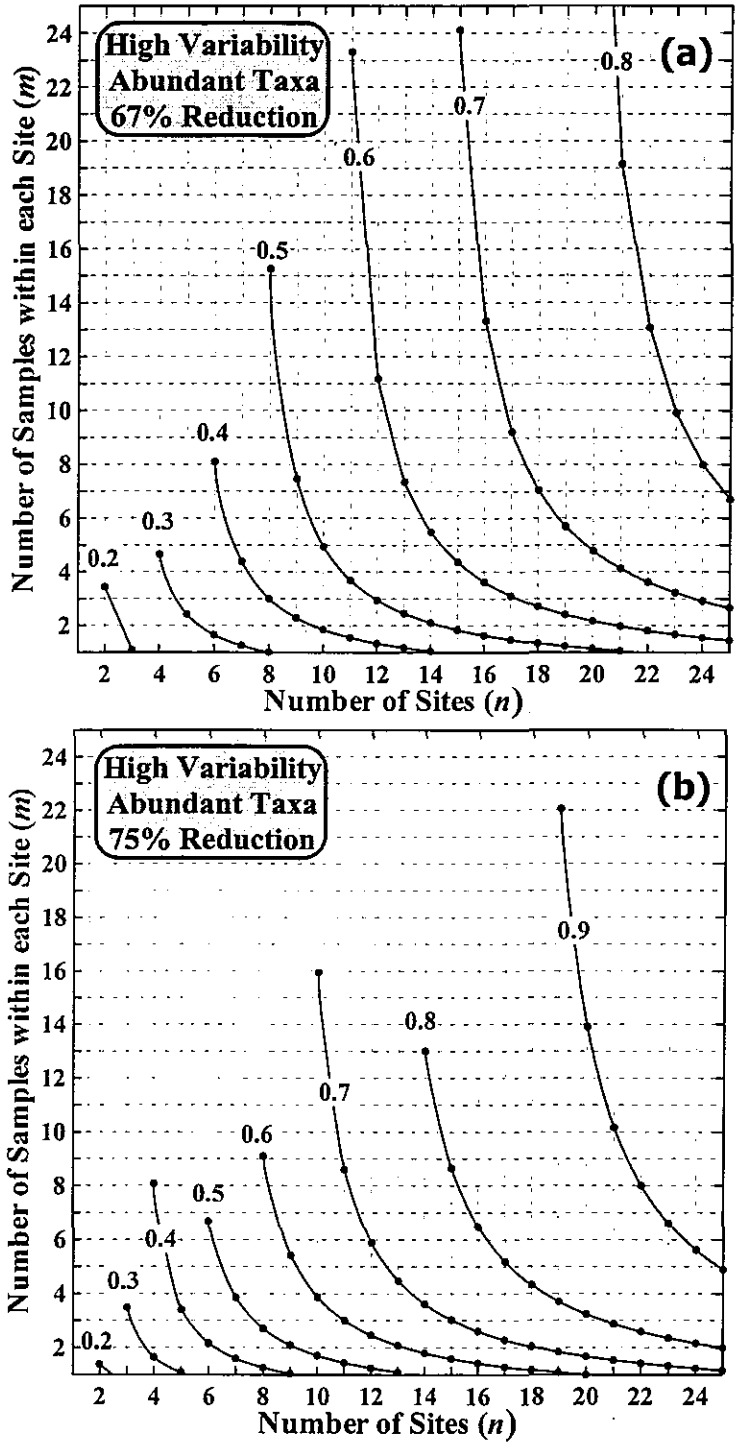

Figure D.9. Sample-size chart showing the number of replicate samples (m) collected at n reference and n treatment sites that are needed to detect a (a) 67% reduction (200% increase) or (b) 75% reduction (300% increase) in abundant intertidal populations with a statistical power ($1-\beta$) at the one-tailed significance level of $\alpha = 0.1$. The curves correspond to different levels of statistical power in an environment with high natural biological variation ($CV_W = 2.35$, $CV_B = 1.57$).

APPENDIX E. POWER FORMULATION FOR TESTING RECOVERY

This Appendix provides the basis for the computational procedures used to determine sample sizes for the detection of abrupt recolonization events that are revealed as a difference in temporal trends in abundance at impact and reference sites. The power formulation for sequential tests presupposes that the deviation from parallelism can be measured by a difference in linear trends. The monitoring design consists of samples collected concurrently at reference and impact sites on an annual basis for two or more years.

Test Statistic

If the abundance data is multivariate normal, with covariance matrices that are equal for impact and reference sites, then a counterpart of the Student's *t*-Statistic, known as the Hotelling's T^2-statistic (1947), can be used to evaluate the sequential data for parallelism. If l_R reference sites are sampled along with l_I impact sites every year for *t* years, then the Hotelling's T^2-statistic is a function of the vector of deviations in mean values between each sequential sample. It can be formulated into an equivalent *F*-test of the null hypothesis of parallelism:

$$\mathrm{H_o}: \begin{bmatrix} \mu_{R1} - \mu_{R2} \\ \mu_{R2} - \mu_{R3} \\ \vdots \\ \mu_{Rt-1} - \mu_{Rt} \end{bmatrix} = \begin{bmatrix} \mu_{I1} - \mu_{I2} \\ \mu_{I2} - \mu_{I3} \\ \vdots \\ \mu_{It-1} - \mu_{It} \end{bmatrix} \tag{E.1}$$

against the alternative of no parallelism:

$$\mathrm{H_a}: \begin{bmatrix} \mu_{R1} - \mu_{R2} \\ \mu_{R2} - \mu_{R3} \\ \vdots \\ \mu_{Rt-1} - \mu_{Rt} \end{bmatrix} \neq \begin{bmatrix} \mu_{I1} - \mu_{I2} \\ \mu_{I2} - \mu_{I3} \\ \vdots \\ \mu_{It-1} - \mu_{It} \end{bmatrix}. \tag{E.2}$$

With *t* years of monitoring, there are $t-1$ paired sets of differences in consecutive means and the significance test for parallelism has $t-1$ degrees of freedom. As discussed in previous chapters, log-transformation of the abundance data provides a multiplicative response model that achieves approximate normality and additivity, and stabilizes variances for intertidal populations. The more transitory the nature of the impact or recovery between annual sampling events, the more

likely the test of parallelism will be able to detect it. The formulation of the hypothesis does not account well for diminishing effects over time. Thus, the analysis best applies to tests for recovery from acute impacts that occur over a period of only a few years. The PWS data set is well-suited to parallelism tests because there was a marked increase in intertidal populations over a one- to two-year time span.

Power Formulation

Convenient sample-size charts cannot be provided for a power analyses based on Equation (E.2) because there are a variety of ways that the time series at impact and reference sites can converge or diverge. Consequently, in order to specify a generally applicable measure of the impact size, namely, the degree of departure from parallelism, linear trends in time are assumed. Thus, the test for acute impacts, or repopulation (recovery) after an impact, consists of examining a single degree-of-freedom contrast that is but one realization of the overall test of parallelism described by Hypotheses (E.1) and (E.2). However, if an acute impact or abrupt recovery is present in the time series of abundance, then at least a portion of the observed changes can invariably be approximated with a linear trend. This may not be true, however, of chronic or secondary impacts on populations that can cause the time series to oscillate from year-to-year as an abrupt recolonization event reverberates over long periods (Coats *et al.*, 1999).

When the parallelism hypothesis (E.1) is applied to data collected over two years, the test automatically reduces to a test of linear trends:

$$H_o: \ (1\mu_{R1} - 1\mu_{R2}) = (1\mu_{I1} - 1\mu_{I2}) \qquad \text{(E.3)}$$

where μ_{R1} and μ_{R2} refer to the mean abundance at reference sites during the first and second years of sampling, and μ_{I1} and μ_{I2} are the means at impact sites in the same consecutive years. For sampling in three consecutive years, the hypothesis of equal linear trends is more restrictive:

$$H_o: \ (1\mu_{R1} + 0\mu_{R2} - 1\mu_{R3}) = (1\mu_{I1} + 0\mu_{I2} - 1\mu_{I3}) \qquad \text{(E.4)}$$

Table E.1 lists all the orthogonal polynomial coefficients for testing for equal linear trends in monitoring studies that span periods of two to ten years.

Table E.1. Orthogonal polynomial coefficients for linear trends in annual sampling over periods of two to ten years

	Years									
Length of Study	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
2	+1	-1								
3	+1	0	-1							
4	+3	+1	-1	-3						
5	+2	+1	0	-1	-2					
6	+5	+3	+1	-1	-3	-5				
7	+3	+2	+1	0	-1	-2	-3			
8	+7	+5	+3	+1	-1	-3	-5	-7		
9	+4	+3	+2	+1	0	-1	-2	-3	-4	
10	+9	+7	+5	+3	+1	-1	-3	-5	-7	-9

Visual inspection of the time series for the PWS monitoring program that are shown in Figure 3.1 suggests that populations had stabilized in 1992, approximately four sampling years after the oil spill in 1989. The test for equal linear trends over a four-year period is obtained from Table E.1:

$$H_o: \quad +3\mu_{R1} + 1\mu_{R2} - 1\mu_{R3} - 3\mu_{R4} - 3\mu_{I1} - 1\mu_{I2} + 1\mu_{I3} + 3\mu_{I4} = 0 \tag{E.5}$$

The contrast evaluated in Equation (E.5) can be written in vector form as:

$$H_o: \quad \mathbf{b}'\boldsymbol{\mu} = 0 \tag{E.6}$$

where:

$$\mathbf{b} = \begin{bmatrix} 3 \\ 1 \\ -1 \\ -3 \\ -3 \\ -1 \\ 1 \\ 3 \end{bmatrix} \text{ and } \boldsymbol{\mu} = \begin{bmatrix} \mu_{R1} \\ \mu_{R2} \\ \mu_{R3} \\ \mu_{R4} \\ \mu_{I1} \\ \mu_{I2} \\ \mu_{I3} \\ \mu_{I4} \end{bmatrix}$$

The null hypothesis can be tested using the *d*-statistic proposed by Skalski and Robson (1992)

$$d = \frac{\mathbf{b}'\bar{\mathbf{x}}}{\sqrt{\mathbf{b}'\boldsymbol{\Sigma}\mathbf{b}}} \tag{E.7}$$

where $\bar{\mathbf{x}}$ is the vector of sample means after log-transformation

$$\bar{\mathbf{x}} = \begin{bmatrix} \bar{x}_{R1} \\ \bar{x}_{R2} \\ \bar{x}_{R3} \\ \bar{x}_{R4} \\ \bar{x}_{I1} \\ \bar{x}_{I2} \\ \bar{x}_{I3} \\ \bar{x}_{I4} \end{bmatrix}$$

and $\boldsymbol{\Sigma}$ is the covariance matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix} Var(\bar{x}_{R1}) & Cov(\bar{x}_{R1},\bar{x}_{R2}) & Cov(\bar{x}_{R1},\bar{x}_{R3}) & Cov(\bar{x}_{R1},\bar{x}_{R4}) & & & & \\ Cov(\bar{x}_{R2},\bar{x}_{R1}) & Var(\bar{x}_{R2}) & Cov(\bar{x}_{R2},\bar{x}_{R3}) & Cov(\bar{x}_{R2},\bar{x}_{R4}) & & & & \\ Cov(\bar{x}_{R3},\bar{x}_{R1}) & Cov(\bar{x}_{R3},\bar{x}_{R2}) & Var(\bar{x}_{R3}) & Cov(\bar{x}_{R3},\bar{x}_{R4}) & & & & \\ Cov(\bar{x}_{R4},\bar{x}_{R1}) & Cov(\bar{x}_{R4},\bar{x}_{R2}) & Cov(\bar{x}_{R4},\bar{x}_{R3}) & Var(\bar{x}_{R4}) & & & & \\ & & & & Var(\bar{x}_{I1}) & Cov(\bar{x}_{I1},\bar{x}_{I2}) & Cov(\bar{x}_{I1},\bar{x}_{I3}) & Cov(\bar{x}_{I1},\bar{x}_{I4}) \\ & & & & Cov(\bar{x}_{I2},\bar{x}_{I1}) & Var(\bar{x}_{I2}) & Cov(\bar{x}_{I2},\bar{x}_{I3}) & Cov(\bar{x}_{I2},\bar{x}_{I4}) \\ & & & & Cov(\bar{x}_{I3},\bar{x}_{I1}) & Cov(\bar{x}_{I3},\bar{x}_{I2}) & Var(\bar{x}_{I3}) & Cov(\bar{x}_{I3},\bar{x}_{I4}) \\ & & & & Cov(\bar{x}_{I4},\bar{x}_{I1}) & Cov(\bar{x}_{I4},\bar{x}_{I2}) & Cov(\bar{x}_{I4},\bar{x}_{I3}) & Var(\bar{x}_{I4}) \end{bmatrix}$$

with $Var(\bar{x}_{R1}) = \dfrac{\sum_{i=1}^{l_R}(x_{R1i} - \bar{x}_{R1})^2}{(l_R - 1)l_R}$ and $Cov(\bar{x}_{R1},\bar{x}_{R2}) = \dfrac{\sum_{i=1}^{l_C}(x_{R1i} - \bar{x}_{R1})(x_{R2i} - \bar{x}_{R2})}{(l_R - 1)l_R}$.

The d-statistic is asymptotically t-distributed with $l_R + l_I - 2$ degrees of freedom. It measures the amount of deviation from parallel linear trends (numerator) relative to the variability about the mean trends (denominator). The power of the null hypothesis test can be calculated from the noncentrality parameter

$$\Phi_{1,l_R+l_I-2} = \frac{1}{\sqrt{2}} \cdot \frac{|\mathbf{b}'\bar{\mathbf{x}}|}{\sqrt{\mathbf{b}'\boldsymbol{\Sigma}\mathbf{b}}} \tag{E.8}$$

In general, the larger the d-statistic, the more the linear trends depart from being parallel. In Chapter 3, the sample sizes needed to detect non-parallelism of the sort seen in the PWS intertidal data are computed for six specific assemblages using Equation (E.8).

APPENDIX F. POWER FORMULATION FOR TESTING LONG-TERM STABILITY

This Appendix provides the basis for the computational procedures used to determine sample sizes for the detection of long-term trends in abundance. In Chapter 4 and Appendix G, these procedures are applied to PWS intertidal data collected at previously impacted sites that were largely repopulated after the initial impact of the *Exxon Valdez* oil spill. The power formulation is based on a test for significant slope in a straight-line regression model applied to annual abundance measurements collected concurrently at several sites for a number of years.

Test Statistic

Long-term trends in intertidal abundance are likely to be reflected in a non-zero slope ($\hat{b}$) in the linear regression model

$$y = \hat{a} + \hat{b} \cdot i \qquad \text{(F.1)}$$

where the coefficients ($\hat{a}$ and $\hat{b}$) are determined from least squares regression of annual abundance measurements (y_i) collected over Y-years ($i = 0,1,...,Y-1$). The null hypothesis of no trend

$$\mathrm{H}_0: \; b = 0 \qquad \text{(F.2)}$$

can be tested against the alternative hypothesis

$$\mathrm{H}_a: \; b \neq 0 \qquad \text{(F.3)}$$

using the statistic

$$t_{Y-2} = \frac{|\hat{b}|}{\sqrt{\dfrac{MSE}{\sum_{i=0}^{Y-1}(i-\bar{i})^2}}} \qquad \text{(F.4)}$$

which is *t*-distributed with $Y-2$ degrees of freedom under H_0. *MSE* is the mean square error left unexplained by the regression. If the trend accurately captures ongoing recovery, where abundance gradually increases over time, then the alternative hypothesis can be written as

$$H_a: \quad b > 0 \tag{F.5}$$

and the test statistic can be evaluated using the one-tailed *t*-distribution.

Power Formulation

Under the alternative hypothesis (F.3), the test statistic (F.4) has a noncentral *F*-distribution with a noncentrality parameter that can be formulated using within-site (σ_W^2) and between-site (σ_B^2) variances

$$\Phi_{1,Y-2} = \frac{1}{\sqrt{2}} \cdot \frac{|b|}{\sqrt{\frac{\left(\frac{\sigma_B^2}{n} + \frac{\sigma_W^2}{nm}\right)}{\sum_{i=0}^{Y-1}(i-\bar{i})^2}}} \tag{F.6}$$

where n is the number of sites surveyed annually and m is the number of epibiotic quadrats or infaunal core samples collected at each site. The expected value of the regression coefficient $(\hat{b})$ depends on the annual rate of population change and the duration of the study. The least-squares estimate of the rate of population change is

$$\hat{b} = \frac{\sum_{i=0}^{Y-1}(i-\bar{i})\,y_i}{\sum_{i=0}^{Y-1}(i-\bar{i})^2}. \tag{F.7}$$

For a linear trend, the annual intertidal abundance can be represented in terms of the annual fractional increase (Δ) where the expected abundance in each year is given by

$$y_i = y_0(1+i\Delta) \tag{F.8}$$

and the slope coefficient is reduced to

$$\hat{b} = y_0 \Delta . \tag{F.9}$$

After substitution by (F.9), the noncentrality parameter (F.6) can be written in terms of coefficients of variation (CV)

$$\Phi_{1,Y-2} = \frac{1}{\sqrt{2}} \cdot \frac{|\Delta|}{\sqrt{\frac{\left(\frac{CV_B^2}{n} + \frac{CV_W^2}{nm}\right)}{\sum_{i=0}^{Y-1}(i-\bar{i})^2}}}. \tag{F.10}$$

The sum of squares for deviations about the mean year $\left(\sum_{i=0}^{Y-1}(i-\bar{i})^2\right)$ is purely a function of the duration of the monitoring program. Table F.1 lists the value of this sum of squares term for a variety of time spans.

Table F.1. Value of the sum of squares term for deviations about the mean year for monitoring programs lasting from two to fifteen years

Study Duration (years)	$\left(\sum_{i=0}^{Y-1}(i-\bar{i})^2\right)$
2	0.5
3	2
4	5
5	10
6	17.5
7	28
8	42
9	60
10	82.5
11	110
12	143
13	182
14	227.5
15	280

For a 5-year monitoring study, the noncentrality parameter is

$$\Phi_{1,3} = \frac{|\Delta| \cdot \sqrt{5}}{\sqrt{\frac{CV_B^2}{n} + \frac{CV_W^2}{nm}}} \tag{F.11}$$

For a 10-year test of linear trends, the noncentrality parameter is

$$\Phi_{1,8} = \frac{|\Delta| \cdot \sqrt{41.25}}{\sqrt{\frac{CV_B^2}{n} + \frac{CV_W^2}{nm}}} \qquad \text{(F.12)}$$

Variability estimates computed for PWS data in Chapter 2 were used in Chapter 4 to evaluate the noncentrality parameters for 5-year and 10-year monitoring programs.

Appendix G. Power Curves for Detecting Long-Term Trends

The following plots provide the number of replicate samples (m) that need to be collected at n impact and n reference sites to achieve various powers capable of detecting five and ten-year trends in intertidal populations that have two levels of annual increase. Sample sizes are computed for abundant taxa (density > 3) with low, moderate, and high variability within two levels of annual increase. As described in Chapter 2, intertidal variability was estimated from the 10^{th}, 50^{th} (median), and 90^{th} percentiles of the intertidal data collected within PWS. Power was determined at a one-tailed statistical significance of $\alpha = 0.1$.

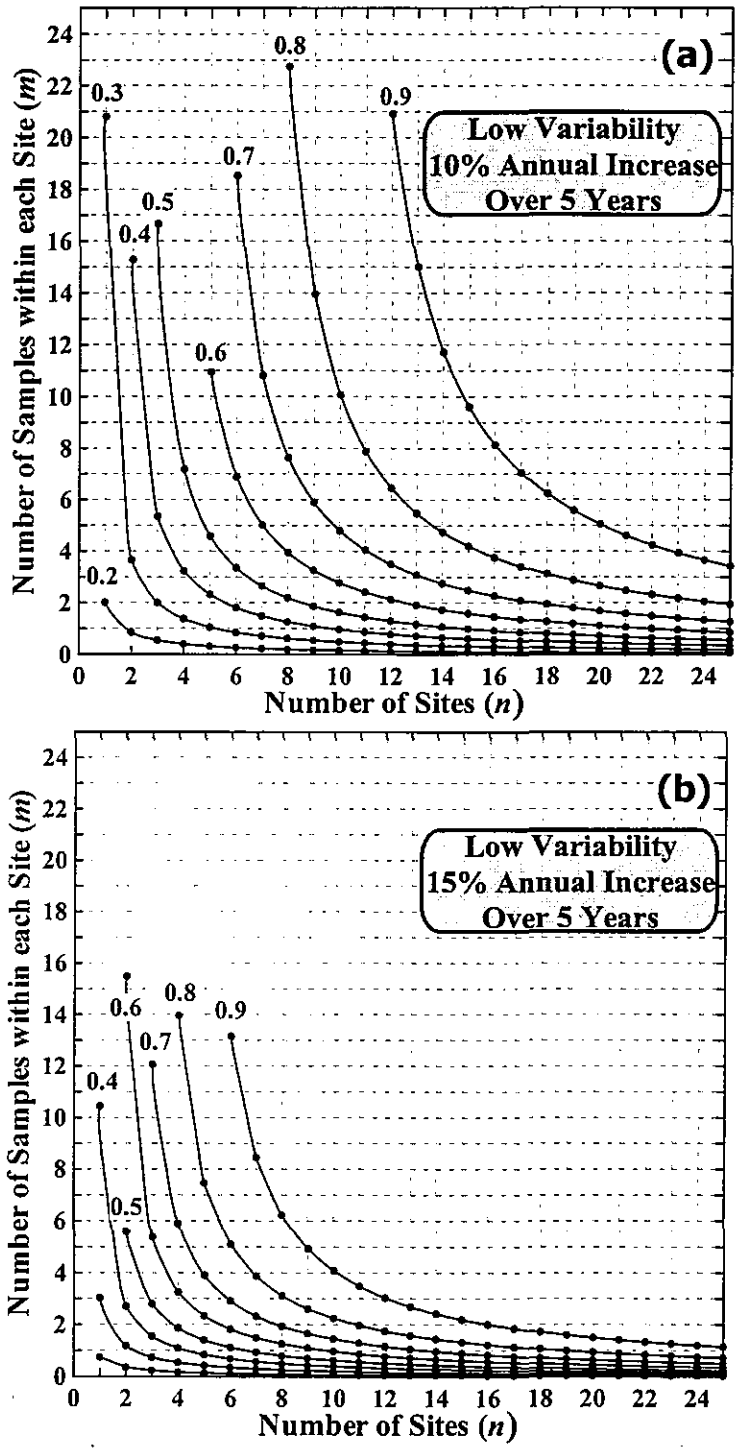

Figure G.1. Sample-size chart showing the number of sites (n) and the number of replicate samples (m) needed to detect an impact that causes a (a) 10% or (b) 15% annual increase in intertidal populations over a 5-year period with a statistical power ($1-\beta$) at the one-tailed significance level of $\alpha = 0.1$. The curves correspond to different levels of statistical power in an environment with low natural biological variation ($CV_W = 0.76,\ CV_B = 0.32$).

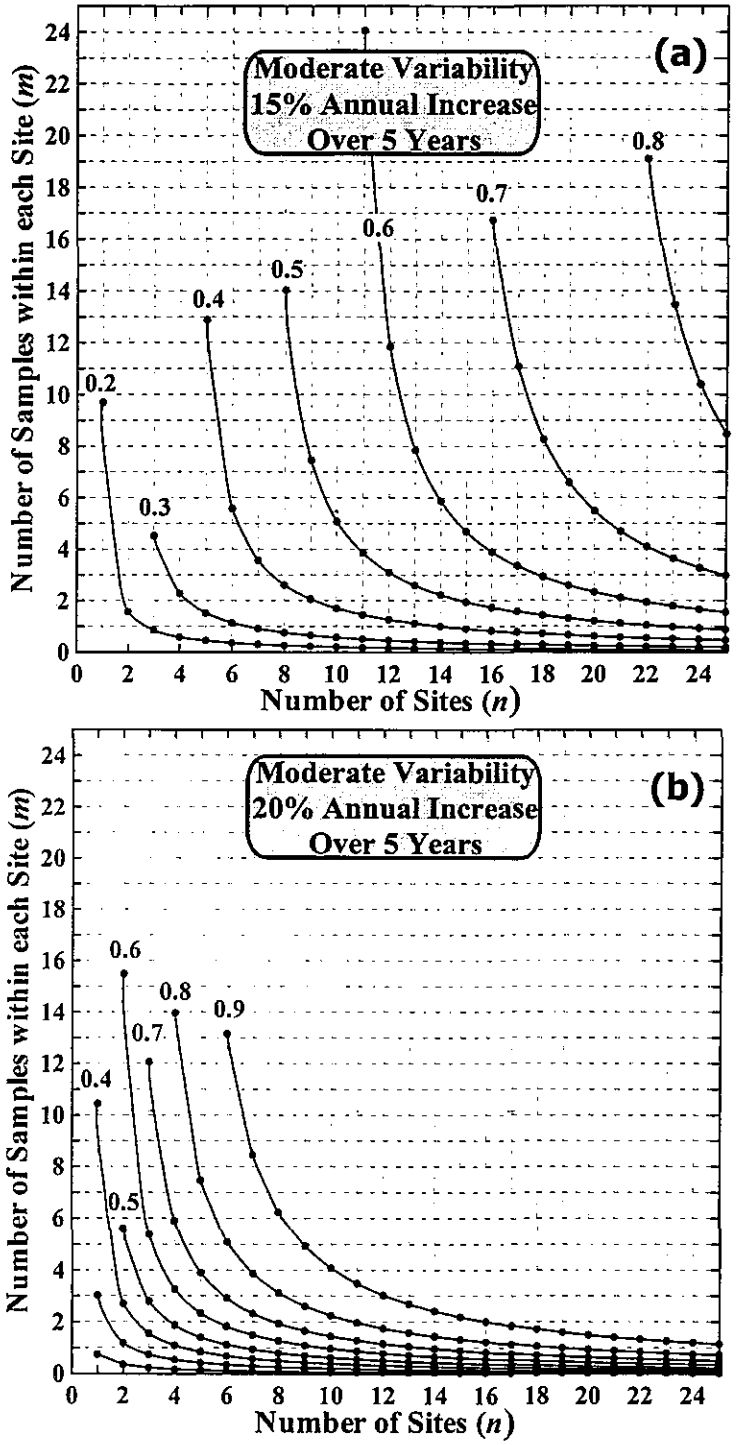

Figure G.2. Sample-size chart showing the number of sites (n) and the number of replicate samples (m) needed to detect an impact that causes a (a) 15% or (b) 20% annual increase in intertidal populations over a 5-year period with a statistical power ($1-\beta$) at the one-tailed significance level of $\alpha = 0.1$. The curves correspond to different levels of statistical power in an environment with moderate natural biological variation ($CV_W = 1.28,\ CV_B = 0.84$).

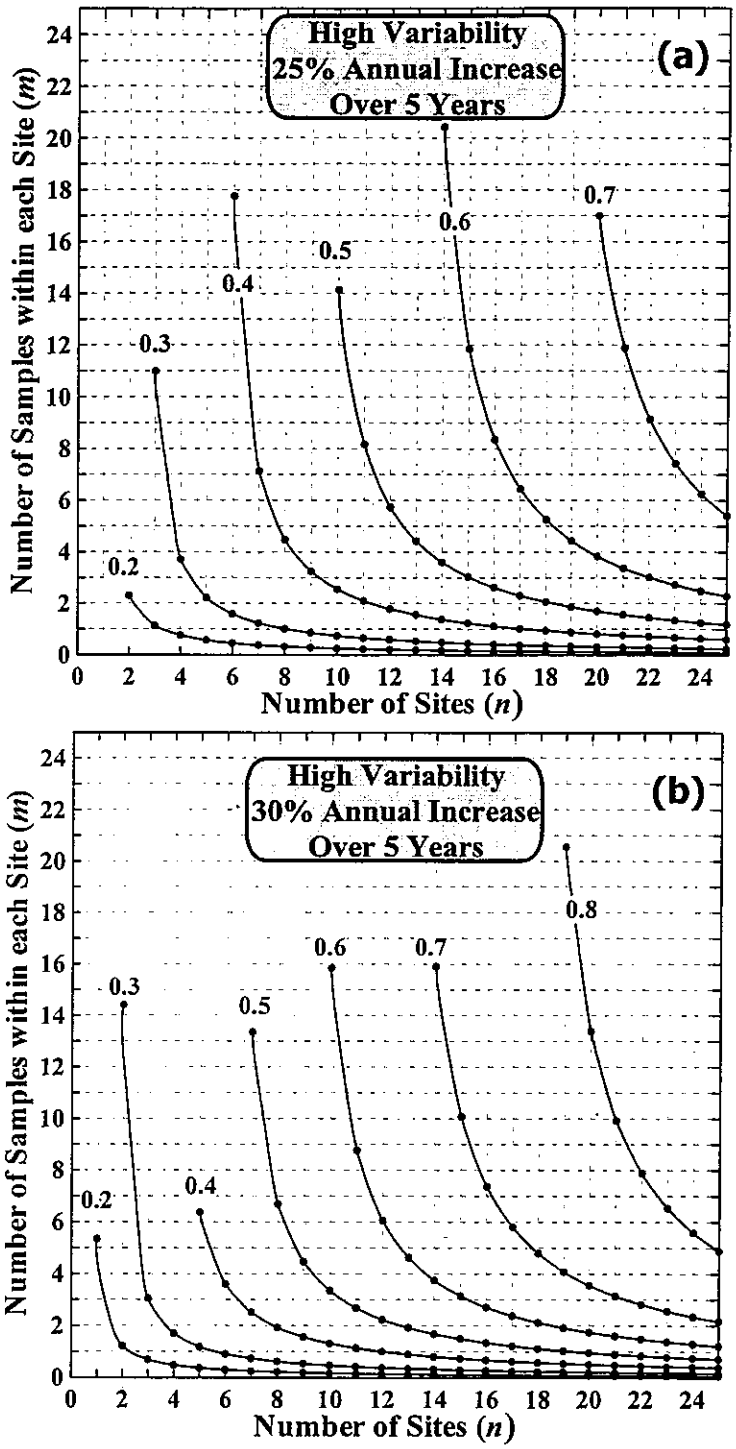

Figure G.3. Sample-size chart showing the number of sites (n) and the number of replicate samples (m) needed to detect an impact that causes a (a) 25% or (b) 30% annual increase in intertidal populations over a 5-year period with a statistical power ($1-\beta$) at the one-tailed significance level of $\alpha = 0.1$. The curves correspond to different levels of statistical power in an environment with high natural biological variation ($CV_W = 2.35$, $CV_B = 1.57$).

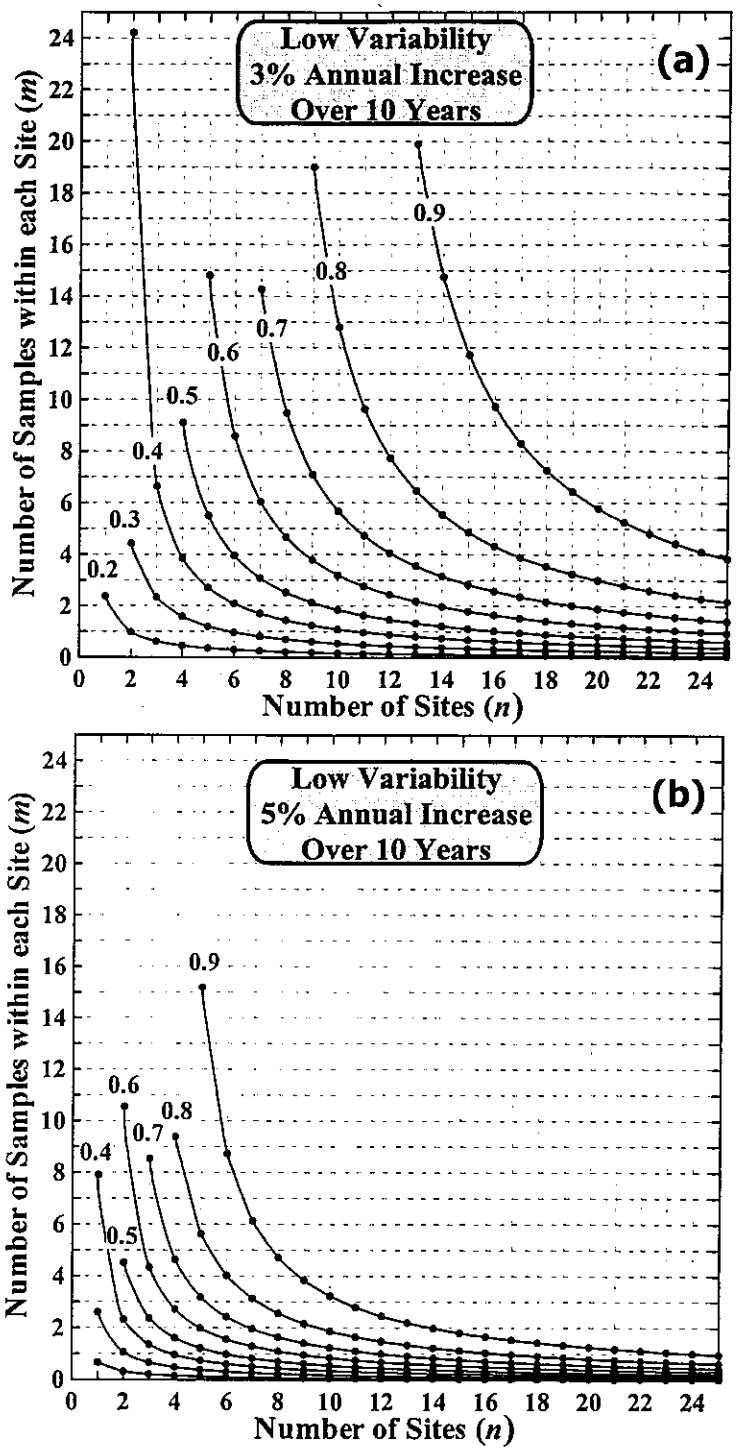

Figure G.4. Sample-size chart showing the number of sites (n) and the number of replicate samples (m) needed to detect an impact that causes a (a) 3% or (b) 5% annual increase in intertidal populations over a 10-year period with a statistical power ($1-\beta$) at the one-tailed significance level of $\alpha = 0.1$. The curves correspond to different levels of statistical power in an environment with low natural biological variation ($CV_W = 0.76$, $CV_B = 0.32$).

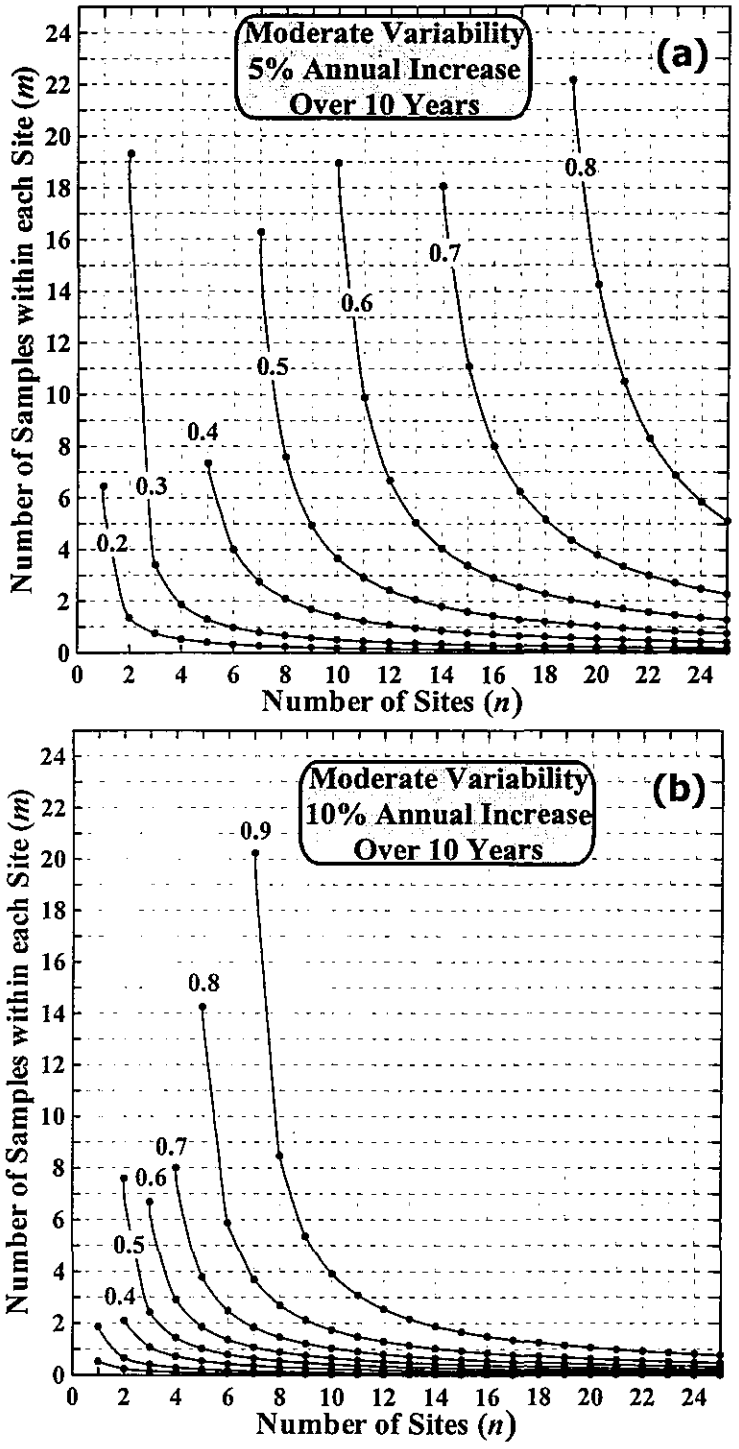

Figure G.5. Sample-size chart showing the number of sites (n) and the number of replicate samples (m) needed to detect an impact that causes a (a) 5% or (b) 10% annual increase in intertidal populations over a 10-year period with a statistical power ($1-\beta$) at the one-tailed significance level of $\alpha = 0.1$. The curves correspond to different levels of statistical power in an environment with moderate natural biological variation ($CV_W = 1.28$, $CV_B = 0.84$).

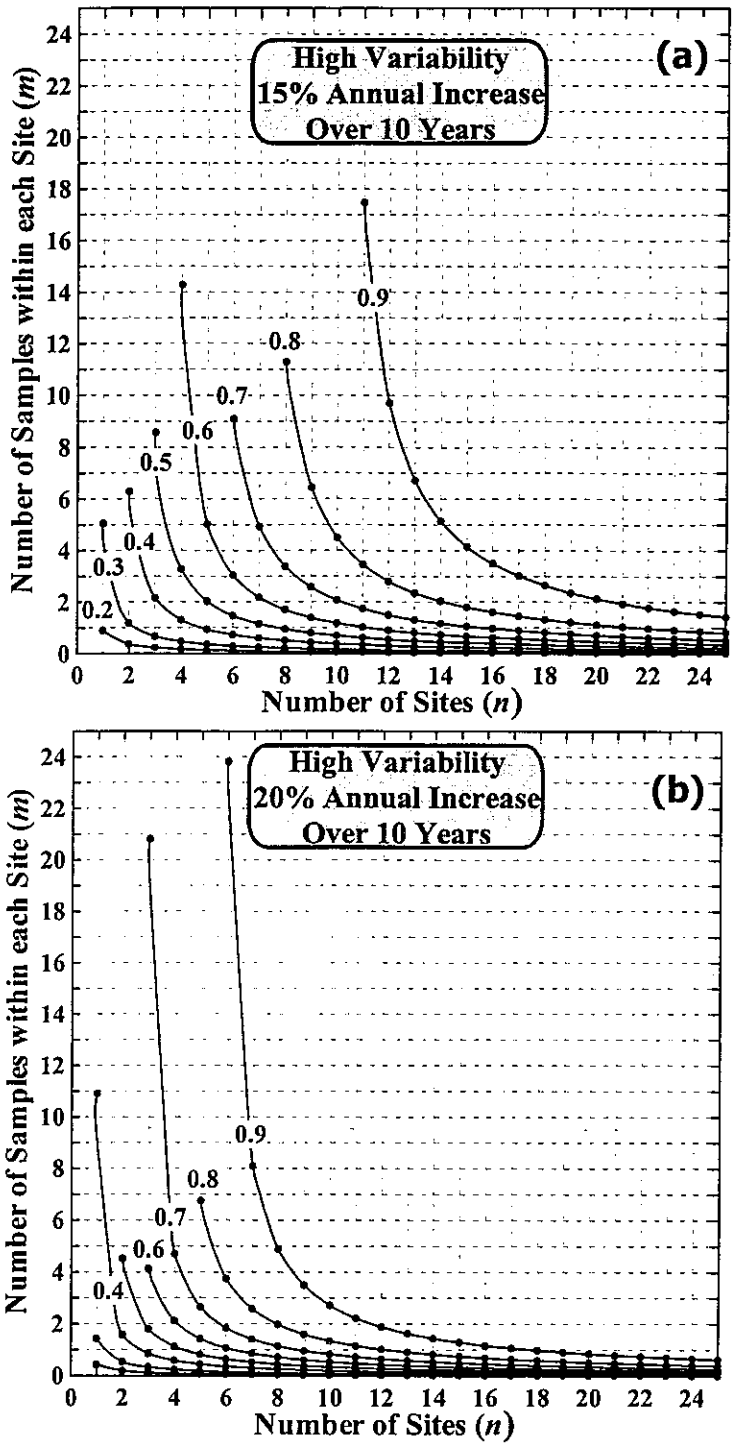

Figure G.6. Sample-size chart showing the number of sites (n) and the number of replicate samples (m) needed to detect an impact that causes a (a) 15% or (b) 20% annual increase in intertidal populations over a 10-year period with a statistical power ($1-\beta$) at the one-tailed significance level of $\alpha = 0.1$. The curves correspond to different levels of statistical power in an environment with high natural biological variation ($CV_W = 2.35$, $CV_B = 1.57$).