JNCC Report No: 598

Monitoring guidance for marine benthic habitats

Noble-James, T., Jesus, A.
& McBreen, F.

July 2017

ISSN 0963-8901

For further information, please contact:

Joint Nature Conservation Committee
Monkstone House
City Road
Peterborough, PE1 1JY
www.jncc.defra.gov.uk

This report should be cited as:

Noble-James, T., Jesus, A. & McBreen, F. 2017. Monitoring guidance for marine benthic habitats. *JNCC Report No. 598.* JNCC, Peterborough.

This report is compliant with the JNCC **Evidence Quality Assurance Policy** http://jncc.defra.gov.uk/default.aspx?page=6675. The draft document was extensively reviewed within JNCC, by monitoring specialists within Natural England, Scottish Natural Heritage (Dualchas Nàdair na h-Alba), Natural Resources Wales (Cyfoeth Naturiol Cymru), and the Department of Agriculture, Environment and Rural Affairs (Northern Ireland). Selected sections were reviewed by a statistician at the Centre for Environment, Fisheries and Aquaculture Science (Cefas).

Summary

Marine benthic monitoring programmes produce evidence against which to evaluate the cause and direction of change in the marine environment. They can also inform which management measures are appropriate, and determine whether they have been successful.

It is crucial that monitoring programmes are well-designed and statistically robust to allow conclusions to be drawn from the acquired data. This ‘best-practice’ guidance aims to provide the information necessary to develop robust monitoring programmes that accurately identify change in the benthic environment. The guidance combines established ecological theory and protocols with JNCC advice and recommendations on benthic monitoring, by means of a step-wise framework which details key stages in the development of a monitoring programme.

Whilst topics such as sample processing and equipment selection have been amply covered elsewhere, this guidance focuses on sampling design, drawing on frequentist theory. The basis of the framework is the development of monitoring objectives, following which the guidance addresses indicator selection, use of existing data, and temporal factors. The importance of statistical power and significance is explored, with guidance on the appropriate levels and ratios for different types of monitoring and the use of power analysis to determine the appropriate sample size. Dependency issues and sampling units are discussed, before guidance on sampling designs is provided. Finally, a statistical analysis section outlines various tests and analyses which can be performed to fulfil a range of monitoring objectives.

Contents

Figures

Tables

1 Introduction

Monitoring programmes produce evidence for effective management of marine benthic habitats and communities, and allow conclusions to be drawn about the cause and direction of natural and anthropogenic change. Management measures are generally based on the results of these conclusions, therefore it is critical that monitoring programmes are well-designed and statistically robust, to avoid damage to the benthic environment or unnecessary exclusion of stakeholders from activities. Clearly defined objectives, careful planning, appropriate sampling designs and judicious application of statistical analyses are all crucial to ensure that the data acquired are representative and the conclusions drawn are accurate.

1.1 Aim

This guidance aims to supply the reader with the information necessary to develop robust monitoring programmes for marine benthic habitats (substrates inclusive of associated communities and species), and answer monitoring questions with a high level of confidence. The information presented here represents monitoring 'best practice', as informed by peer-reviewed and grey literature, and the concepts can be broadly applied to marine benthic habitats in any system or geographical location. This guidance may therefore be used by any organisation or individual undertaking benthic habitats monitoring, although it should be noted that the guidance is designed for application at a relatively small scale (e.g. within Marine Protected Areas (MPAs) or other discrete survey areas), as opposed to large-scale monitoring of regions (e.g. Charting Progress 2 Regions or OSPAR Regional Seas).

1.2 Background and context

The Joint Nature Conservation Committee (JNCC) is leading the UK Marine Biodiversity Monitoring Research and Development Programme (UKMBMP) on behalf of the Statutory Nature Conservation Bodies (SNCBs) and other partners in the UK Marine Monitoring and Assessment Strategy (UKMMAS). The focus of this process is the design of a UK-scale monitoring programme which will collect the evidence required to fulfil all UK marine biodiversity obligations in the most cost-efficient manner.

An overarching UK Marine Biodiversity Monitoring Strategy (Kröger & Johnston 2016) underlies the development of the UKMBMP and identifies two high-level monitoring functions which are driving it:

1) to identify the state of ecological components of biodiversity, and identify whether any changes are due to natural change or anthropogenic activities, to determine whether management measures are required.

2) to identify whether management measures are effective in meeting their objectives.

The Strategy also defines three different 'monitoring types' (described further in Section 2.1) which will be applied to achieve the two high-level monitoring objectives:

1) Sentinel Monitoring of long-term trends (Type 1 monitoring).

2) Operational Monitoring of pressure-state relationships (Type 2 monitoring).

3) Investigative Monitoring to determine management needs and effectiveness (Type 3 monitoring).

To support the development of the UKMBMP, this guidance is structured around the three monitoring types identified in the UK Marine Biodiversity Monitoring Strategy (Kröger & Johnston 2016). However, as previously mentioned the guidance is also applicable outside of UK waters.

1.3 How to use this document

This document presents a stepwise framework which can be used to plan and design a monitoring programme for marine benthic habitats, from setting objectives to statistical analysis. Each section provides background information and best practice guidance for each stage of the design process, with specific advice for sentinel, operational and investigative monitoring. Key points and recommendations are summarised at the end of each section, with flowcharts to visualise key processes. An overview of the document's structure and stepwise framework is presented below in Figure 1.

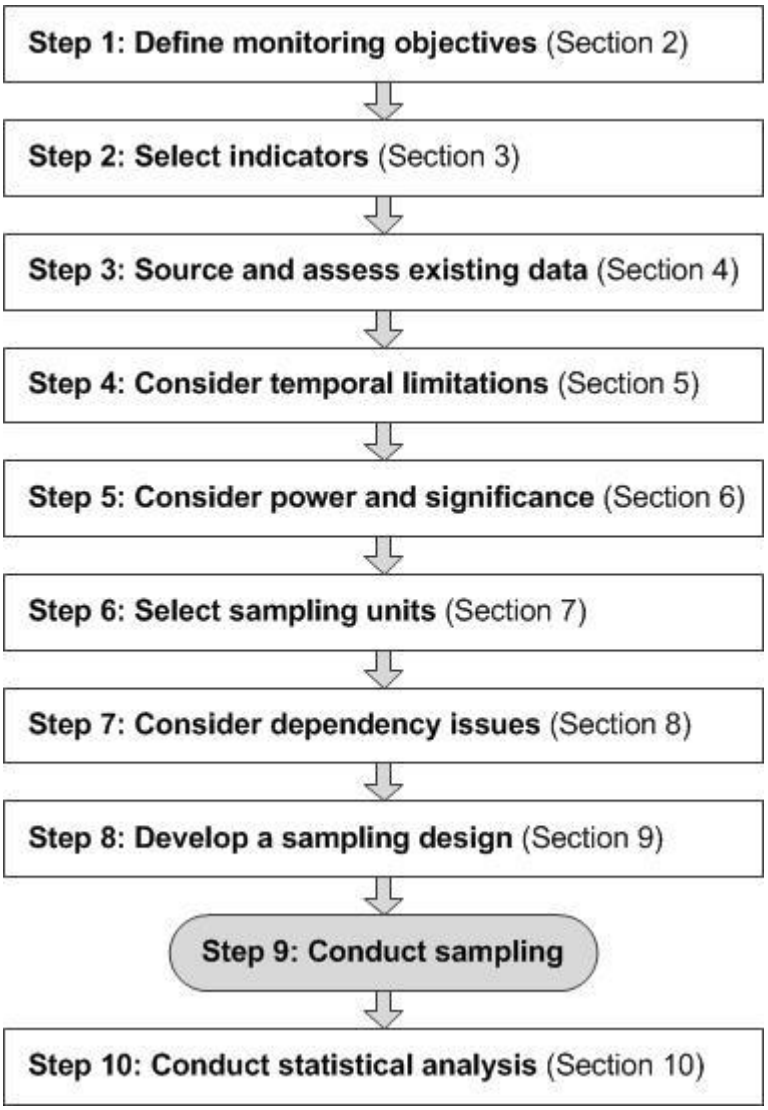


Figure 1. Overview of document structure: a stepwise process for designing a monitoring programme for marine benthic habitats.

1.4 Notes on terminology

Various definitions have been applied to the term 'monitoring' in the context of the marine environment. For example, UKMMAS has adopted the definition by Portmann (2000): 'The taking, on a reasonably regular basis, of any form of observations relative to the (long-term) status of the marine environment, regardless of the frequency of, or purpose for which, the observations are made.' This definition focuses on marine observations 'undertaken more consistently, albeit with varying frequency, over longer periods of time' and it excludes one-off or intermittent field observations. Whilst the Common Standards Monitoring Guidance (JNCC 2004a) applies the definition by Brown (2000) where 'monitoring is an intermittent (regular or irregular) series of observations in time, carried out to show the extent of compliance with a formulated standard or degree of deviation from an expected norm'.

In the context of the UKMBMP, and consequently this document, the term 'monitoring' is used in a very generic sense to mean 'an activity by which evidence necessary to meet the aims of the monitoring programme is collected', broadening the definition to include activities which do not form part of a time-series (e.g. operational monitoring activities), but which contribute to the achievement of monitoring objectives.

The term 'monitoring programme' is used throughout this document to describe a programme of monitoring activities undertaken at a small or local scale to investigate a specific area.

1.5 Outside of document scope

Many guidance documents have already been produced on operational and analytical aspects of benthic habitats monitoring. Therefore, this document does not cover:

- sampling techniques and operations,
- habitat mapping and remote sensing methods,
- laboratory processing and analytical standards.

The JNCC Marine Monitoring Method Finder[1], a web-based information hub, has been developed to provide a single point of access to the numerous guidance documents and tools generated both within and outside the UK, and can be used in conjunction with this document to assure a consistent approach to data collection and analysis.

[1] http://jncc.defra.gov.uk/page-7171

2 Defining monitoring objectives

The foundation of a successful monitoring programme is the early establishment of clear and achievable monitoring objectives. As stated by Underwood and Chapman (2013) '…if the aims and objectives of any study are not clearly defined at the outset, the least damaging outcome will be wastage of time, money and resources. The worst outcome will be a complete lack of valid information on which to build understanding, predictive capability, and managerial/conservatory decision-making.'

The generic objective of all monitoring programmes is to detect the occurrence and degree of change through time and space, and assess this against known impacts or management actions (Parry *et al* 2012), with the aim of preserving or enhancing biodiversity, ecosystem services and natural capital assets.

If monitoring is conducted to assess the condition of habitats within Marine Protected Areas (MPAs) a high-level conservation objective will have been defined for each designated feature, for example; maintain at or recover to favourable condition. Specific monitoring objectives must be developed for each MPA, to allow assessment of whether conservation objectives have been achieved. These objectives are likely to be determined by the vulnerability of the designated features to pressures, the need to assess the effectiveness of management measures, and the level of confidence in habitat extent and condition.

If monitoring is to be conducted outside of MPAs, conservation objectives may not have been explicitly stated, and the aim of the monitoring may be to demonstrate whether habitats meet a specified threshold above which they are considered in good condition.

Monitoring objectives will generally correspond to one or more of the three direct monitoring types mentioned in Section 1.1. These monitoring types are discussed in the following sub-section. Under certain circumstances monitoring objectives may also be partially or fully achieved via indirect monitoring methods, as discussed further in Section 2.2.

2.1 Direct monitoring

Direct monitoring (hereafter simply referred to as 'monitoring') involves acquisition of data from habitats of interest using methods such as photographic and/or physical sampling, and remote sensing techniques (e.g. acoustic survey). The three types of direct monitoring detailed in the UK Marine Biodiversity Strategy are described in greater detail in Box 1.

Box 1: The three monitoring types (Kröger & Johnston 2016)

Sentinel monitoring of long-term trends (Type 1 monitoring)
Objective: to measure rate and direction of long-term change.

This type of monitoring provides the context to distinguish directional trends from short-scale variability in space and time. To achieve this objective efficiently, a long-term commitment to regular and consistent data collection is necessary; this means time-series must be established as their power in identifying trends is far superior to any combination of independent studies.

Operational monitoring of pressure-state relationships (Type 2 monitoring)
Objective: to measure state and relate observed change to possible causes.

This objective complements monitoring long-term trends and is best suited to explore the likely impacts of anthropogenic pressures on habitats and species and identify emerging problems. It leads to setting of hypotheses about processes underlying observed patterns, and is generally best applied in areas where a gradient of pressure is present (e.g. no pressure increasing gradually to 'high' pressure).

It relies on finding relationships between observed changes in biodiversity and observed variability in pressures and environmental factors. It provides inference but it is not proof of cause and effect. The spatial and temporal scale for this type of monitoring will require careful consideration of the reality on the ground to ensure inference will be reliable; for example, inference will be poor in situations where the presence of a pressure is consistently correlated to the presence of an environmental driver (e.g., a specific depth stratum).

Investigative monitoring to determine management needs and effectiveness (Type 3 monitoring)
Objective: to investigate the cause of change.

This monitoring type provides evidence of causality. It complements the above types by testing specific hypotheses through targeted manipulative studies (i.e. excluding an impact or causing an impact for experimental purposes). The design and statistical approach that can be used in these cases gives confidence in identifying cause and effect. It is best suited to test state/pressure relationships and the efficacy of management measures.

2.1.1 Which types of monitoring should be conducted?

Sentinel monitoring to identify long-term trends in habitat condition (and responses to environmentally and anthropogenically driven change) can be conducted within any survey area, assuming there are no logistical or resourcing constraints (e.g. insufficient budget, no availability of a suitable sampling platform, equipment, or personnel). Operational and/or investigative monitoring, however, may not be relevant or feasible depending on factors such as the distribution of anthropogenic pressures and habitats, understanding of pressure-state relationships, and the status of existing or proposed management measures. The following section (Section 2.1.2) can be used to guide selection of relevant monitoring types, whilst Section 2.1.3 addresses whether it is feasible to conduct them. A flow process is provided in Figure 2 to aid assessment of feasibility.

2.1.2 Which monitoring types are relevant?

The following questions will help to determine which monitoring types are relevant.

Are pressure-state relationships already understood?

Our understanding of the relationships between anthropogenic pressures and habitat condition is variable. Some relationships are clearly established, with a large body of supporting evidence (e.g. the impact of abrasion on slow-growing biogenic habitats, such as cold water corals, or sponge beds), whilst others require further investigation (e.g. the impact of abrasion on sedimentary habitats).

Further studies are unlikely to be an efficient use of resources for combinations of habitats and pressures where the relationship is well established, unless there is a specific need for understanding of local conditions, or a requirement for such monitoring under activity licensing conditions. For habitats where pressure-state relationships are less clearly defined, operational or investigative monitoring may be appropriate to inform management measures and improve assessments of condition.

Do established or proposed management measures result in complete or zoned restrictions?

The nature of planned or existing management measures may influence whether investigative monitoring is required. If management measures reduce the risk of impacts to the lowest possible level (i.e. the complete exclusion of pressures to which habitats are sensitive), investigative monitoring may not be necessary, unless evidence of habitat or species recovery is required at the local scale (e.g. where the socio-economic impacts of the closure are high). For example, where fishers have been excluded from a popular fishing ground it may be necessary to provide evidence of improved habitat condition within the closure.

Where zoned management is proposed or in place (e.g. only closing specific areas within an MPA, or excluding certain gear types) investigative monitoring studies may be conducted to enable adaptive management of the site.

2.1.3 Which monitoring types are feasible?

Operational and/or investigative monitoring are not always feasible, even if they are relevant, as they involve more complex designs than sentinel monitoring (discussed further in Sections 9.3 and 9.4). These types of monitoring will also benefit from a higher level of confidence in habitat distribution than required by sentinel monitoring, and some knowledge of the spatial and temporal distribution of pressures is also needed.

The following questions will help the reader follow a suggested assessment process to determine whether operational or investigative monitoring are feasible (illustrated in Figure 2).

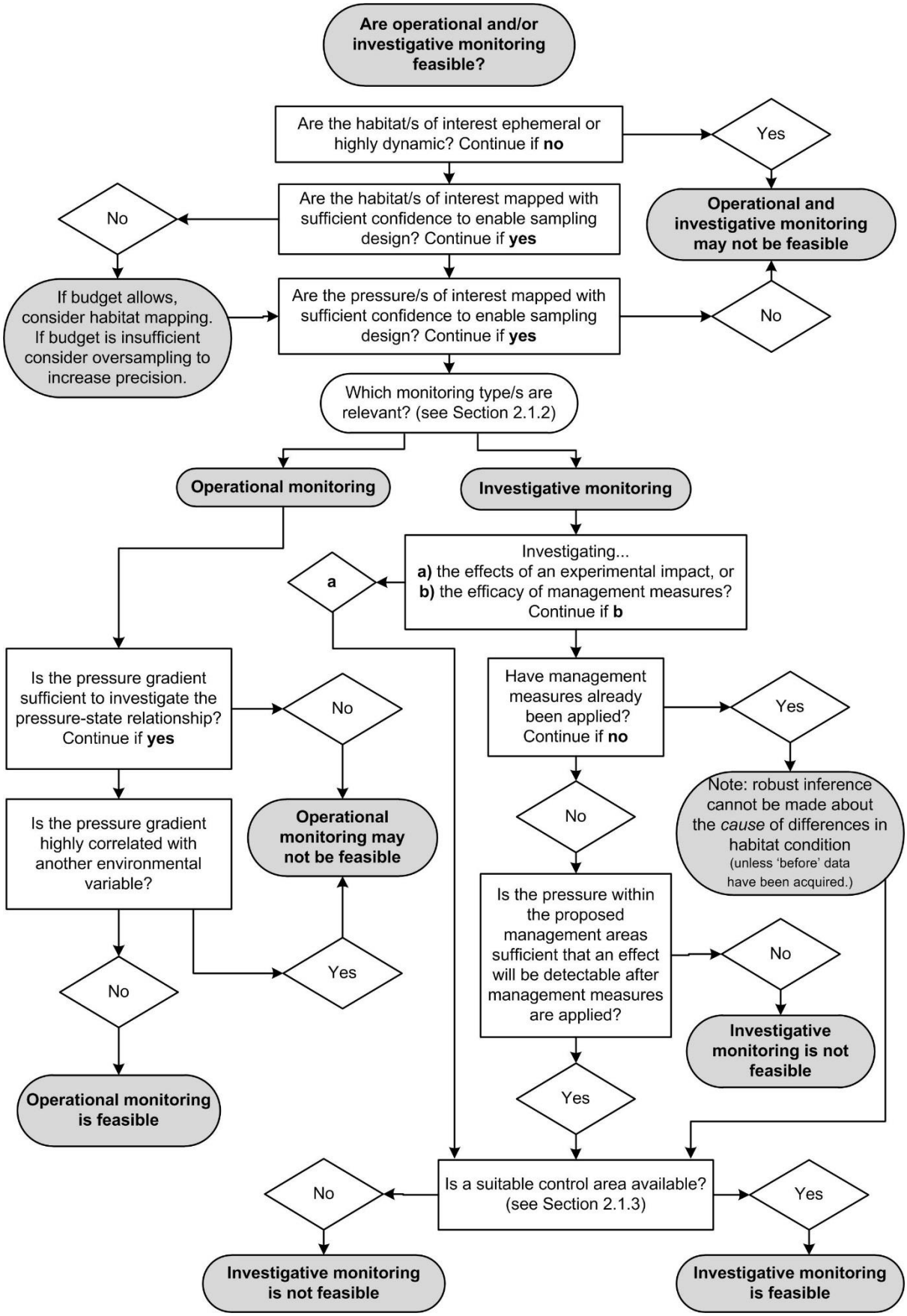


Figure 2. Defining monitoring objectives: a process to determine whether operational and/or investigative monitoring are feasible.

Is habitat distribution sufficiently understood to enable sampling design, or is acquisition of new acoustic data feasible?

Operational and investigative sampling designs are optimised where the distribution of habitats is known with reasonably high confidence, and sampling effort can be distributed or 'stratified' using this information (discussed further in Section 9), however the resolution and level of confidence in habitat maps can vary widely. For example, in the UK a number of areas (including some MPAs) have been mapped using remote sensing data sources (e.g. bathymetry, backscatter, side scan sonar, LIDAR[2]) validated by ground-truthing data (e.g. grab sampling and photographic data). Subsequently confidence in habitat distribution can be relatively high for some areas, whilst no such data are available for others. In these areas modelled products will indicate predicted habitat distribution (e.g. EMODnet[3] seabed habitats and bathymetry layers), but the distribution cannot be assumed to be accurate.

At this point acquisition of new remote sensing data should be considered, to allow characterisation of the site and aid sampling design. The strength of the case for and feasibility of acquiring such data will depend on the habitat type, the size of the survey area, the resources required, and the stated monitoring objectives.

It is recommended that new remote sensing data are acquired and a habitat map created if the area does not have an existing high resolution habitat map, or the existing one is likely to be obsolete. However, acquisition of remote sensing data and production of habitat maps can be extremely resource-intensive, particularly for large survey areas, and it is unlikely that mapping will be always be an efficient use of resources. For example, where sediments are mobile (e.g. sandbanks) or features are ephemeral (e.g. *Sabellaria spinulosa* reefs) maps can become obsolete within a relatively short period of time.

For some habitats, it is possible to conduct operational or investigative monitoring activities in the absence of a high confidence habitat map. This approach is likely to be most effective where a habitat of interest is thought to be homogeneous and its distribution is predicted to be consistent across the survey area. Where habitats are highly patchy (e.g. cobble mosaics), variable (e.g. a range of habitats within the site) or discrete (e.g. areas of Methane-Derived Authigenic Carbonate (MDAC)) the lack of a high confidence map is likely to reduce the robustness of the design. If acquisition of remote-sensing data is not possible for areas with patchy or variable habitats, modelled products should be used with caution and with awareness that the design is likely to be less statistically robust. In some situations, where habitat maps are not available, the robustness and precision of the sampling (see Section 6.2) can be improved by increasing the sample size ('oversampling') and applying post-hoc stratification (see Section 9.1.2 and 10.2).

In some cases, it will not be possible to design operational or investigative monitoring activities in the absence of remote sensing data; for example, investigative monitoring of locally discrete habitats (e.g. isolated patches of rocky reef) cannot be conducted where a suitable control site cannot be identified.

Are habitats extremely variable, mobile or ephemeral?

Some habitats, such as sandbanks, display very high levels of natural variability, with substantial small-scale variation in substrates and associated taxa. In such habitats it can be difficult to detect anthropogenic impacts against a background of natural variation (e.g. Collie *et al* 2000; Hiddink *et al* 2006; van Denderen *et al* 2015). A large sample size may

[2] Light Detection and Ranging; a technique that uses light sensors to measure distance.
[3] The European Marine Observation and Data Network.

compensate for the effect of such variation; however, collection of an adequate sample may not always be achievable given financial and logistical constraints.

Where habitats are highly mobile (e.g. some sandbanks), or ephemeral (e.g. *Sabellaria* reefs), investigative monitoring studies should only be undertaken with extreme caution, if at all. In such habitats it is likely that substrate and community composition will change naturally within both the control and impact areas, and therefore it will not be possible to categorically attribute an improvement or decline in condition to the implementation of management measures.

Are the spatial and temporal distributions of pressures understood?

Operational and investigative designs require knowledge of the spatial distribution and intensity of relevant anthropogenic pressures across the habitat/s of interest. This information may also be required for interpretation of patterns observed from sentinel monitoring data. The availability and reliability of pressures information should therefore be considered in the development of monitoring objectives.

Pressure mapping can be relatively straightforward in the intertidal zone, but is increasingly difficult in the inshore and offshore areas. Some non-dispersive pressures can be mapped with high accuracy in subtidal zones (e.g. physical change directly below a newly placed piece of infrastructure), but the vast majority of pressures cannot be accurately mapped in these areas. Subtidal pressure mapping therefore generally entails a degree of modelling or inference (e.g. modelling of contaminant dispersion, aggregation of Vessel Monitoring System (VMS) transmission 'pings' or sightings-per-unit-effort (SPUE) to grid cell format). Some marine activities and pressures cannot be mapped with any confidence (e.g. marine litter distribution). In cases where the pressure distribution is unknown, qualitative information or expert judgement is required to determine whether a pressure-state study is feasible.

The level of confidence in pressure mapping products may also be reduced where pressure intensity is low. For example, fisheries VMS 'pings' can be aggregated to grid cells to map estimated 'swept area' of demersal abrasion within offshore waters (where the operational fleet is primarily >12m length); if the ratio of the area 'swept' to the area of the cell is very low (i.e. there are very few pings), it is unlikely that sampling points will coincide with areas within the cell which have been exposed to the pressure. It is essential that the method used to map pressures is fully understood.

In addition to mapped pressures, the likelihood that additional unmapped pressures exist within the survey area should be considered. This is particularly relevant for the inshore region, where anthropogenic activity is substantially higher and generally more diverse due to land proximity. It may be possible to account for additional unmapped pressures in analysis (e.g. addition of hydrocarbon concentration as a covariate, or calculation of contamination indices); if this is not possible operational or investigative monitoring may not be appropriate.

The temporal limitations of pressures data should also be appraised in the context of the habitat/s being monitored. If the data are not recent it is possible that the habitat has recovered from any impact, and that a relationship will not be detected. This is particularly relevant for habitats such as sandbanks which are subject to high levels of natural hydrodynamic disturbance, and are likely to recover more quickly than low-energy habitats (e.g. Dernie *et al* 2003; Kaiser *et al* 2006).

Further considerations: operational monitoring

If confidence in pressure distribution is satisfactory in terms of both time and space, the data should be assessed to determine whether pressure gradients are sufficient for further investigation;

Is the pressure of interest sufficient to investigate the pressure-state relationship?

Identification of a biological response to an anthropogenic pressure requires sampling across a large range of pressure intensity. If the pressure appears to be evenly distributed, or shows little variation across the habitat of interest, it will be difficult or impossible to determine how habitat condition changes in response to increasing pressure intensity (see Section 3). The level of pressure variation needed to detect a clear response in biological variables will vary between habitats and systems. For example, a higher range of anthropogenic disturbance is required to detect indicator response in dynamic systems, where benthic communities are adapted to natural disturbance (e.g. sandbanks, see Jenkins *et al* 2015). A lower level of pressure would be sufficient for habitats which are not naturally subject to disturbance and recover more slowly from anthropogenic impacts (e.g. corals or deep-sea sponge aggregations).

Is the pressure gradient highly correlated with another environmental or physical parameter?

Where a pressure gradient is highly correlated with another environmental or physical parameter (e.g. sediment type or seabed depth) it will be difficult or impossible to detect a biological response to pressure against natural variation. In some cases, it may be possible to distinguish responses to anthropogenic pressures by modelling correlated parameters, but where the pressure & environmental parameter/s are collinear (very highly correlated) it will be difficult to identify causation in the relationships observed. In such cases, investigative monitoring studies are likely to be more appropriate for exploring relationships between pressures and habitat state. Further detailed guidance on survey designs for operational monitoring is provided in Section 9.3.

Further considerations: investigative monitoring

Where investigative monitoring is relevant to identify whether management measures have been effective, the feasibility of this type of monitoring is dependent on several conditions. These can be explored using the questions below:

Are management measures in place, or are future management areas known?

The ability to attribute change to management measures will be severely limited if sufficient data have not been collected prior to their implementation. If management measures have already been implemented and ‘before’ data have not been acquired it will be possible to compare managed areas and control sites, however it cannot be assumed that they were in the same condition before application of the measures.

Management measures can result in exclusion of activities from an entire area of habitat (e.g. of all rocky reef within an MPA), or from a series of areas within a wider habitat (zoned management). If the planned management is zoned, the locations of the management areas must be known with confidence to enable a balanced and statistically robust investigative design. If the locations of the management areas are not known with confidence, it may be possible to use sentinel monitoring samples for a before and after comparison, depending on the number of sentinel sampling points which have fallen within the management areas on the initial survey. The probability of being able to use sentinel monitoring data for investigative monitoring studies are improved if a systematic sampling strategy is used (e.g.

a triangular grid pattern, see Section 9.1.3), however a robust dataset cannot be guaranteed without knowing the size and location of management areas.

Are pressure/s sufficiently high that a clear change in habitat condition could be detected after implementation of management measures?

The likelihood of detecting a biological response after management implementation may be reduced where the pressure of interest has historically been low. The probability of detecting a change will vary depending on the sensitivity of the habitat/s to the excluded pressure/s. In cases where the likely effects of management measures are thought to be too subtle to detect against a background of natural variation, or confidence in pressure mapping products is low, investigative monitoring is unlikely to be feasible.

Is a suitable control site available?

A robust sampling design to detect the effects of management measures will generally feature sampling before and after the measures have been implemented, at both the 'impact' site(s) (where management measures are implemented, or where an experimental impact has been applied), and at 'control' site(s) (where the status quo is maintained) as a minimum. This type of design, termed Before-After-Control-Impact (BACI), is particularly powerful because it controls for both temporal and spatial variation, improving the robustness of conclusions on management effects. For a BACI study to be a viable option, a suitable control site must be available (see Section 9.4.6 for further details).

Control sites should:

- ideally be in relatively close proximity to the 'impact' site (area/s where management measures are to be introduced),
- not be directly adjacent to the impact site to avoid biological 'overspill' or edge effects (e.g. concentrated fishing pressure at the boundaries of a managed area),
- ideally be located where there is high confidence in habitat distribution; particularly where comparable substrates are likely to be isolated or limited in extent (e.g. a rocky outcrop), although high resolution maps are less critical where substrates are likely to be homogeneous,
- have comparable environmental conditions (e.g. hydrodynamic regime and organic inputs) to those of the impact site,
- have a sufficient level of the same pressure that a difference between the control and impact sites may be detected using an appropriate indicator following management.

Finding suitable control sites can be difficult when habitats are locally rare or isolated. For example, where an MPA has been designated to protect an isolated or rare habitat (e.g. MDAC concretions), identification of a control site is likely to involve considerable extra acoustic survey time and resources, which can render BACI-type studies unfeasible. Where habitats are likely to be more broadly distributed (e.g. sedimentary habitats) control sites can be identified by using modelled habitats data (such as EMODnet seabed habitats[4]) in combination with pressures information, and variation can be limited post-hoc by only comparing samples with a similar sediment composition to the impact site.

[4] http://www.emodnet.eu/seabed-habitats

Further detailed guidance on investigative monitoring designs is available in Section 9.4.

2.2 Indirect monitoring

Indirect monitoring involves monitoring of pressures to infer habitat condition, as an alternative to or in combination with direct monitoring. Indirect monitoring can be used as an early warning tool to flag up the need for urgent direct monitoring where pressure levels have changed, or it can provide an effective and economical alternative to frequent direct monitoring where certain conditions are met. It is expected that indirect monitoring will play a role in most monitoring programmes, however it should be stressed that this approach is not always suitable as a replacement for direct monitoring.

The extent to which a monitoring programme relies on indirect monitoring should be careful considered; it is recommended that monitoring programmes should only rely heavily on indirect monitoring or use it as a replacement for direct monitoring when;

1. **The distribution and intensity of pressures is well understood.**
   Availability of information on pressure distribution and intensity of pressures varies widely. As mentioned previously, relatively accurate information on pressure footprint is available for some pressures in specific areas, whilst others are poorly understood (e.g. indirect or dispersive pressures such as hydrocarbon contamination or sediment dispersal). For example, reliable VMS data are currently only available for vessels >12m length, which primarily operate in the offshore region. Sighting and logbook information can provide an indication of fishing activity inshore, for example an inshore SPUE map has been produced by Cefas (Breen *et al* 2014), however the inshore distribution of the pressure is unlikely to be understood as well as offshore.

2. **The number of pressures is low**
   Pressures are likely to be more numerous in some areas, such as the inshore, and shallower or less remote offshore regions. At present understanding of the cumulative effects of pressures is limited, therefore indirect monitoring is most suitable where a single pressure or low number of pressures are likely to be present.

3. **Pressure-state relationships are firmly established**
   Indirect monitoring should only be considered where the pressure-state relationship is well understood, and the change in habitat condition in response to a particular pressure is consistent and predictable.

2.3 Duration and frequency of monitoring

A long-term commitment to ongoing regular and consistent data collection is needed to achieve sentinel monitoring objectives in an efficient way. The appropriate frequency of sentinel sampling will vary across habitats based on the relative risk of the habitat to human pressures, as well as the life cycles of biota and natural variability in parameters selected for monitoring. These sampling cycles will be further influenced by reporting schedules and availability of funding and resources.

Operational monitoring may consist of a one-off case study (e.g. a research and development study designed to develop an indicator of habitat condition; see Section 3), or occur multiple times (e.g. regular pressure gradient monitoring required as a condition of development consent). Investigative monitoring studies will require repeated sampling, with at least two sampling events; before and after an impact or the introduction of management measures. Similar to operational monitoring the frequency and number of sampling events is

flexible, and will depend heavily on the resources available, the purpose of the monitoring, the evidence requirement and the characteristics of the habitat in question.

Where indirect monitoring is used in a monitoring programme it should be conducted at a frequency which is relevant to the pressure and habitat combination, dependent on the availability of data.

2.4 Summary of key points and recommendations

Section 2: Defining monitoring objectives

Key Points:

- The foundation of a successful monitoring programme is the early establishment of clearly defined and achievable objectives.
- Monitoring objectives will generally correspond to one or more of the three monitoring types described in the UK Marine Biodiversity Monitoring Strategy (Kröger & Johnston 2016):
    - Sentinel Monitoring of long-term trends (Type 1 monitoring)
    - Operational Monitoring of pressure-state relationships (Type 2 monitoring)
    - Investigative Monitoring to determine management needs and effectiveness (Type 3 monitoring)
- Sentinel monitoring can be conducted wherever there are no logistical constraints, but operational and investigative monitoring will not always be relevant or feasible.
- Monitoring can be direct (e.g. collecting physical and/or remote sensing data), or indirect (e.g. monitoring of pressures to infer habitat condition).

Recommendations:

- If operational and/or investigative monitoring are required to achieve monitoring objectives, the following questions should be used to determine whether they are feasible:
    - *Is habitat distribution sufficiently understood to enable sampling design, or is acquisition of new acoustic data feasible?*
    - *Are habitats extremely variable, mobile or ephemeral?*
    - *Are the spatial and temporal distributions of pressures understood?*
    - *Is the pressure of interest sufficient to investigate the pressure-state relationship? (Operational)*
    - *Is the pressure gradient highly correlated with another environmental or physical parameter? (Operational)*
    - *Are management measures in place or are future management areas known? (Investigative)*
    - *Are pressure/s sufficiently high that a clear change in habitat condition could be detected after implementation of management measures? (Investigative)*
    - *Is a suitable control site available? (Investigative)*
- The flowchart in Figure 2 can be used to determine which monitoring types are feasible.
- Indirect monitoring should only replace direct monitoring when:
    - *The distribution and intensity of pressures is well understood.*
    - *The number of pressures is low.*
    - *Pressure-state relationships are firmly established.*

3 Selecting indicators

Having defined monitoring objectives, the next step is to determine which ecological parameters should be measured. Marine benthic habitats are highly complex, therefore it is common practice to limit the number of monitored parameters by using one or more indicators to represent key functional and structural aspects of the ecosystem (OSPAR 2012).

According to OSPAR (2012) advice on the selection of indicators (for assessment under the Marine Strategy Framework Directive, MSFD), a marine biodiversity indicator is defined as:

> *'any measurable feature or condition of the marine environment that is relevant to the stability and integrity of habitats and communities, the sustainability of ecosystem goods and services (e.g. primary productivity, maintenance of food chains, nutrient cycling, biodiversity), the quality and safety of seafood, and the status of amenities of socio-economic importance.'*

OSPAR has identified two different types of indicator which can be used to assess differences between actual and desired environmental condition, or 'state':

- **State indicators** reflect the actual environmental condition within a given geographical area, and include selected species, assemblage characteristics, and biotic functional groups, in addition to habitat characteristics.

  Physical or chemical properties, such as hydrodynamic parameters, shear stress, light attenuation or nutrient levels may also be used as indicators, where they are very closely linked to the condition of habitat/s (Alexander *et al* 2014).

- **Pressure indicators** indicate the prevailing anthropogenic pressures (e.g. VMS data, contaminant measurements or dispersion models), and may be used indirectly to infer the environmental condition where pressure and state are closely linked.

State and pressure indicators, either singularly or in combination, can be used to assess whether the desired environmental status has been achieved. For example, whether MPAs have achieved conservation objectives, or whether specific benthic habitats have achieved targets as defined by wider marine policy drivers (e.g. attainment of Good Environmental Status (GES) under the MSFD).

Broad-scale indicators have been and continue to be developed to fulfil requirements of policy drivers or directives, for example; MSFD, the OSPAR convention, Birds & Habitats Directives, Marine and Coastal Access Act, Marine (Scotland) Act and Water Framework Directive (WFD). Such indicators should be incorporated into monitoring programmes to maximise consistency and efficiency; however, it may also be necessary to develop additional indicators at the habitat level, or for specific geographical areas.

The complexity of marine ecosystem processes, functions and interactions, and the relative paucity of information (particularly in the offshore region), make the selection of state indicators an extremely difficult task. Indicators which have been selected subjectively (i.e. based on assumptions or observations) may undermine the monitoring objectives by not accurately reflecting true environmental condition, or impacts of anthropogenic pressures. The robustness of state indicators is crucial for the success of a monitoring programme, and indicators should be developed in a logical and objective way.

The following sections, 3.1 and 3.3, summarise key steps in the state indicator development process, including building and using CEMs, assessing attributes of indicators, and testing and validation.

3.1 Developing and using Conceptual Ecological Models (CEMs)

Conceptual Ecological Models (CEMs) provide frameworks by which ecological parameters can be systematically selected for development as indicators, taking the complex processes which drive marine benthic ecosystems into account (e.g. Maddox *et al* 1999; Manley *et al* 2000; Gross 2003).

CEMs are diagrammatic representations of the influences and processes which occur within an ecosystem, with important aspects of habitats and their biological communities being represented by discrete model components (e.g. sediment type, recruitment, infauna). They can be used to identify critical aspects of an ecosystem which may serve as a basis for indicator development. The strength and direction of ecosystem processes can be displayed, to identify which model components are likely to show the strongest response to natural variability and anthropogenic pressures. Using this information, ecological parameters (e.g. number of taxa, abundance per taxon, or biomass) can then be identified to either directly or indirectly measure these important model components, and to be taken forward for development as indicators.

JNCC has commissioned a series of CEMs for subtidal temperate habitats which will be used in the UK to identify appropriate ecological components for state indicator development; for example, shallow sublittoral coarse sediments (Alexander *et al* 2014), shallow sublittoral rock (Alexander *et al* 2015), shallow sublittoral muds (Coates *et al* 2015). It should be noted that CEMs are unlikely to be developed for the full range of subtidal habitats for which the UK has a monitoring or reporting obligation.

In the absence of CEMs for specific habitats, existing datasets, pilot studies, scientific and grey literature should be critically reviewed to determine which parameters may be suitable for development as state indicators.

3.2 Developing state indicators

Not all ecological parameters identified from CEMs will make effective state indicators; they must possess certain attributes which allow change to be detected in an efficient and logistically sustainable way, against a background of natural variation. Once ecological parameters have been selected they must be assessed for suitability as potential indicators.

3.2.1 Attributes of effective state indicators

The International Council for the Exploration of the Seas (ICES) Advisory Committee on Ecosystems[5] defines a ‘good’ indicator as one that specialists and non-specialists can both easily comprehend, that is sensitive to (and tightly linked in space and time to) human activity, is accurately measurable, has a low responsiveness to natural changes in the environment, is based on currently available data, and is widely applicable over large areas. In advice on the selection of indicators for descriptors of marine biodiversity, OSPAR (2012) set out selection criteria. Although OSPAR have specified these criteria in reference to MSFD indicators, they can be broadly applied outside of this context and are shown in Table 1. Ecological parameters should be assessed alongside this list with consideration of

[5] www.ices.dk/community/groups/Pages/ACOM.aspx.

whether it will be feasible to measure the indicator in the long-term, given financial and equipment resources.

Table 1. OSPAR (2012) state indicator selection criteria (adapted from ICES and UK scientific indicator evaluation).

Criterion	Specification
Sensitivity	Does the indicator allow detection of any type of change against background variation or noise?
Accuracy	Is the indicator measured with a low error rate?
Specificity	Does the indicator respond primarily to a particular human pressure, with low responsiveness to other causes of change?
Simplicity	Is the indicator easily measured?
Responsiveness	Is the indicator able to act as an early warning signal?
Spatial applicability	Is the indicator measurable over a large proportion of the geographical area to which it is to apply e.g. if the indicator is used at a UK level, is it possible to measure the required parameter(s) across this entire range or is it localised to one small scale area?
Management link	Is the indicator tightly linked to an activity which can be managed to reduce its negative effects on the indicator (i.e. are the quantitative trends in cause and effect of change well known?)
Validity	Is the indicator based on an existing body or time-series of data (either continuous or interrupted) to allow a realistic setting of objectives?
Communication	Is the indicator relatively easy to understand by non-scientists and those who will decide on their use?

3.3 Testing and validating state indicators

Ecological parameters selected as potential indicators should be validated and tested before they are made operational. Ideally this process should use data specifically acquired for this purpose (e.g. an operational monitoring design where the indicator is measured along a pressure gradient). Existing data could be used if this is not possible, provided the spatial and temporal distribution of the data is adequate for testing the indicator. Validation requirements will vary depending on the nature of the indicator, however it must be shown to reliably and consistently respond to anthropogenic disturbance.

3.4 Summary of key points and recommendations

Section 3: Selecting indicators

Key Points:

- An indicator is defined by OSPAR (2012) as:

  'any measurable feature or condition of the marine environment that is relevant to the stability and integrity of habitats and communities, the sustainability of ecosystem goods and services (e.g. primary productivity, maintenance of food chains, nutrient cycling, biodiversity), the quality and safety of seafood, and the status of amenities of socio-economic importance.'

- State indicators reflect the actual environmental condition within a given geographical area, whilst pressure indicators indicate the prevailing anthropogenic pressures.
- Broad-scale indicators have been and continue to be developed to fulfil requirements of policy drivers or directives, however it may be necessary to develop additional indicators at the habitat level or for specified geographical areas.
- Conceptual Ecological Models (CEMs) provide frameworks by which ecological parameters can be systematically selected for development as indicators, taking into account the complex processes which drive marine benthic ecosystems.
- Not all ecological parameters are suitable for development as state indicators. OSPAR (2012) have published a list of criteria to aid state indicator selection.

Recommendations:

- JNCC has commissioned a series of Conceptual Ecological Models (CEMs) for a limited number of subtidal temperate habitats (available on the JNCC Report Series webpage). These CEMs can be used to select ecological parameters for development as state indicators.
- Where CEMs are not available existing datasets, pilot studies, scientific and grey literature should be critically reviewed to determine which parameters should be developed as state indicators.
- Ecological parameters selected for indicator development should be assessed against the OSPAR (2012) list of state indicator selection criteria (Table 1).
- Ecological parameters which have been selected as potential indicators should undergo a testing and validation phase before they are made operational.

4 Sourcing, assessing and using existing data

Designing a benthic monitoring programme to accurately detect change requires careful consideration of spatial and temporal variation within the habitat/s to be monitored. The use of existing data is a valuable and highly cost-effective means of better understanding marine benthic habitats. Existing data typically have two main functions within monitoring programmes:

1) Providing information to facilitate sampling design (e.g. sample data for power analysis (discussed further in Section 6.6), species presence/absence, habitat maps that allow comparison of ephemeral feature distribution),

2) Forming the initial observation/s in a monitoring time-series.

This section provides information on identifying, sourcing and validating existing data, with emphasis on assessing the quality and relevance for benthic habitats monitoring.

4.1 Sourcing existing data

The quality and quantity of available data will vary substantially. Some areas may have been comprehensively surveyed; for example, Dogger Bank has been the subject of a series of academic research studies dating back to the 1920s, in addition to industry and site designation surveys, whilst deep sea MPAs may have limited available data. Sources of existing data include a range of public data-sharing initiatives, regional monitoring and mapping projects (e.g. Strategic Environmental Assessments, SEA), industry data, peer-reviewed publications and grey literature (e.g. reports by SNCBs and NGOs). In the UK various information centres and databases facilitate access to sources of publicly-available data collected by governmental organisations, academic institutions, NGOs and citizen science programmes. Short descriptions and links to these data sources are provided in Annex I. Academic literature should also be reviewed for the area and/or habitat in question, to identify any datasets that are not publicly available and create opportunities for collaborative working.

4.2 Assessing the suitability of existing data

Existing data can be used to improve sampling designs and provide initial data points in monitoring time-series, however they must be carefully evaluated to ensure they are suitable for these purposes.

Various factors can limit the quality of existing datasets; for example, the time of collection, acquisition techniques and equipment, sample processing protocols, and many others. Failure to appraise data quality can result in problems such as the generation of an inappropriate sample size (e.g. too small to accurately detect change, or too large, causing an unnecessary waste of resources) or inaccurate conclusions about change when used as the initial event in a monitoring time-series (e.g. detecting change where there has been none, or failing to detect change). Some basic considerations to aid data evaluation are presented in Table 2; it should be noted that these are intended as a guide only, and may not cover all issues in specific datasets.

Table 2. Considerations and questions to aid review of existing data for use in benthic habitat monitoring programmes.

Data type	Considerations	Questions
All data types	Omission of basic information and provision of data in inappropriate formats can substantially limit the use of existing datasets.	*Is basic information such as seabed depth and sample co-ordinates supplied?*
		Are geodetic parameters specified?
		Are the data available in the required format?
	Existing data may not be suitable for fulfilling monitoring objectives, especially if 'non-standard' measurements are required, such as bivalve size.	*Have the required measurements been taken, and have environmental variables which are likely to covary with the indicator been measured?*
	Inaccurate conclusions may be drawn if natural variance and change are measured using data which have been influenced by an unusual disturbance (e.g. a significant storm event, sea surface temperature anomaly or a contamination incident).	*Were the data collected during a period and from an area which, as far as possible, represents 'undisturbed' conditions in relation to the site, so that any change observed may be attributed to management measures or changes in parameters of interest (e.g. abrasion pressure)?*
	Existing data may have been acquired prior to infrastructure development within or near the original sampling locations. Access to original sampling areas should be considered if this is the case, and the data are to form the initial point in a data-series.	*Do any developments or infrastructure exist within the site, or are they planned, potentially restricting future access and limiting re-sampling opportunities?*
Grab or core sample data	The size, shape and mechanism of the sampler used can substantially affect resulting biological and physico-chemical measurements.	*Is the sampler type, sample volume and sample depth known?*
	The size of the sieve mesh aperture, sieving and sorting techniques used to process infauna can significantly affect the recorded community structure (see MESH ROG: Guerra & Freitas 2012; Bishop & Hartley 1986; Phillips *et al* 2014).	*Is the sieve mesh aperture size known? Which sieving technique has been used?*
	Physico-chemical analysis procedures may vary, and some procedures have been rendered largely obsolete due to technological advances (e.g. infrared detection of hydrocarbons).	*Are the analytical procedures accurately recorded, and do they correspond to recognised standards (e.g. ISO)? For instance, the method by which organic carbon has been isolated, or the digestion technique used to extract metals.*
		Are the analytical procedures repeatable, and are they considered to be sufficiently accurate in comparison to modern techniques?

Data type	Considerations	Questions
Grab or core sample data	Macrofaunal datasets are typically subject to a process of rationalisation, involving aggregation and exclusion of specific taxa, and removal of juveniles to reduce the effect of newly-settled ephemeral components of the assemblage (OSPAR 2004).	*Are the raw data available? If not, is the rationalisation procedure clear, and is it known whether the dataset includes juveniles? Are species names consistent with current nomenclature?*
	The experience of laboratory personnel, and the quality control procedures followed can significantly influence the accuracy of the data.	*Is the laboratory used reputable, and does it participate in a quality control scheme such as the National Marine Biological Analytical Quality Control scheme (NMBAQC)? If not, caution should be exercised in data interpretation.*
Photographic and video data	Data quality can be compromised by operational and environmental factors, such as vessel positioning capability, wave height, image quality, light and turbidity (see MESH ROG: Coggan *et al* 2007; NMBAQC guidance: Hitchin *et al* 2015).	*Are the data of sufficient quality to extract the required information? Was the positioning equipment on the camera or was position taken from the vessel?*
	Substantial variation may exist in the taxonomic resolution to which biota are identified. Analytical errors may include incorrect identification of organisms which are difficult or impossible to identify to generic or species level from photographic data (e.g. sponges), and loss of information where taxonomic resolution is too low for full identification.	*Have the biota been identified to an appropriate taxonomic resolution, and are they comparable between datasets (further data rationalisation may be necessary if not)?*
	The true distribution of rare or patchy habitats, species or communities may be obscured if video transect data are not logged at appropriate intervals (e.g. 10 / 50 / 100m segments).	*Have transect data been logged at sufficient intervals to accurately reflect the distribution of the indicator in question?*
Scientific trawl or dredge data	The type of trawl used, trawl length and speed, and the aperture of the net mesh will influence which types of organisms are retained within the trawl or dredge (see MESH ROG: Curtis & Coggan 2007).	*Was the equipment used appropriate to adequately sample the communities of interest?*
Acoustic data	The positioning, resolution and quality of acoustic data (e.g. multibeam bathymetry, backscatter or side scan sonar) are extremely important if these data are to be used to create habitat maps, or to inform sampling design (see MESH ROGs: Long 2005; Hopkins 2007).	*Are acoustic data of sufficient quality to determine the distribution and/or character of habitats within the site?*
		Is the acoustic coverage sufficient for monitoring objectives to be met?

4.3 Ensuring comparability of existing and new data

Differences in survey timing, operational methods, equipment, processing and analysis techniques can all have implications for the quality and comparability of data. If existing data will provide the initial data point/s in a monitoring time-series, future monitoring practices should be aligned with the existing data set/s as far as possible, to ensure that any change detected is authentic.

Seasonal variations in benthic ecosystems can introduce additional 'noise' into time-series data, which can obscure or magnify trends. Ideally the effects of seasonality should be minimised by conducting each sampling event within the same season. If this is not logistically possible it is important to be aware of temporal differences in the datasets and handle them accordingly; e.g. if juveniles have been excluded from an infaunal dataset the same protocol should be followed for comparing current data to minimise seasonal fluctuations (Reiss & Kröncke 2005). Existing data should not be treated as an initial monitoring event if the seasonal differences are substantial and cannot be corrected.

Benthic samples can vary considerably due to the different capacities, mechanical actions, and bite profiles of sampling devices, with performance also influenced by sediment type (Blomqvist 1991; Barrio Froján & Mason 2010). Differences in grab sampler volume are likely to create disparities in faunal characterisation, as larger samplers are more likely to capture widely dispersed or rare taxa (Boyd *et al* 2006). A similar bias occurs for benthic dredges and trawls, in relation to different designs, modifications and tow speeds (Eleftheriou & Moore 2013). Where current habitat condition will be assessed against existing data, best practice dictates that the sampler type, capacity and method of use should correspond to that originally used, unless a gear comparison study indicates that a different sampler is comparable. If sampling equipment cannot be aligned due to logistical constraints or other overriding factors (e.g. the desire to maintain inter-agency or international consistency), potential differences in the datasets should be acknowledged throughout the analysis and reporting process.

Processing and analysis specifications can also limit comparison of existing and current data. Comparison of samples processed using sieves with different mesh apertures (i.e. 0.5mm or 1.0mm for macrofauna, or 0.30mm or 0.25mm for meiofauna) will be incomparable due to selectivity bias (Reish 1959; Lewis & Stoner 1981). Samples which have been processed using different sieving techniques (e.g. autosiever versus manual sieving) should also be compared with caution.

Analytical techniques and reporting standards can introduce variation in physico-chemical datasets (e.g. particle size classification system, organic content analysis methods, heavy- and trace metal digestion techniques), rendering time-series data incomparable. This is particularly an issue for datasets spanning decades where methods, equipment and protocols may have become obsolete or less commonly used.

4.4 Summary of key points and recommendations

Section 4: Sourcing, assessing and using existing data

Key Points:

- Existing data are a valuable and highly cost-effective source of information on the variation expected in specific areas or habitats. These data have two main functions within monitoring programmes:
  - Providing information to aid sampling design.
  - Constituting the initial event/s in a monitoring time-series.
- Existing data collected by governmental organisations, academic institutions, NGOs and citizen science programmes are publicly available from a variety of information centres and databases.

Recommendations:

- Existing data should be carefully evaluated to ensure that they are suitable for use in monitoring programmes.
- The data sources listed in Annex I can be used as a starting point to obtain existing UK data.
- The considerations provided in Table 2 can be used to help evaluate whether the existing data are suitable for use.
- When using existing data as the first point in a monitoring time-series, current monitoring practices should be aligned wherever possible (e.g. in terms of survey timing, operational methods, equipment, processing and analysis techniques), and addressed in analysis and reporting.

5 Considering temporal limitations

Logistical factors such as financial resources, vessel availability, and reporting cycles will often limit the timing of monitoring surveys. It may also be necessary to time surveys to avoid adverse weather or environmental conditions (e.g. visibility can be reduced by organic detritus in the water column, or by sediment suspended by increased wave action).

As mentioned in Section 4.3, sampling will ideally occur in the same season throughout the lifespan of a monitoring programme, unless there is evidence of low biological and environmental variation in the habitats of interest (JNCC 2004c). Where not dictated by logistical factors or the need to align with existing data, efforts should be made to sample at the optimum time for measuring the selected indicators.

Macroalgal communities display tangible seasonal trends, and habitats in the photic zone may support dense ephemeral assemblages during the summer months (e.g. Maggs 1983; Howson & Davison 2000). Seasonal effects are also observed in seagrass communities, with die-back of seagrass blades and epiphytic assemblages present in the autumn and winter months (Short *et al* 1988). It is generally accepted that macroalgal and seagrass communities should be surveyed in the late summer months (JNCC 2004c), unless evidence suggests that this timing is inappropriate for particular species of interest.

Epifaunal communities may also exhibit ephemeral cycles, with seasonal patterns often coinciding with those of algal assemblages (Jensen *et al* 1994). Bryozoans (e.g. *Bugula* spp: Hayward & Ryland 1998) and hydroids (e.g. *Tubularia indivisa*: Fish & Fish 1996) demonstrate seasonal cycles of growth in spring and summer and die-back in late autumn and winter, entering a phase of dormancy (see also Ryland 1976; Gili & Hughes 1995). Conversely some taxa, such as the soft coral *Alcyonium digitatum*, spawn in winter with larvae settling before the spring plankton bloom (Hartnoll 1975).

Biogenic habitats can also display significant temporal variation. *Sabellaria spinulosa* reefs vary seasonally in quality and extent, with settlement of *S. spinulosa* juveniles recorded during March and April in south-western English waters (George & Warwick 1985; Wilson 1970), and rapid annual growth of 2-3cm thick sheets occurring during a single spring/ summer growing season (Holt *et al* 1998). The location of reefs is also likely to change over time, as reefs are subject to 5-7 year cycles of aggregation and degeneration (Wilson 1971), and reefs are destroyed or eroded by winter storms (Holt *et al* 1998). *Mytilus edulis* reefs are also susceptible to partial or total destruction by erratic winter storms in exposed areas (Nehls & Thiel 1993), while recruitment is thought to be favoured by cold preceding winters (Holt *et al* 1998).

Many infaunal taxa have seasonal reproductive patterns which dramatically alter the number of individuals present at different times of the year. Some polychaete worms have semelparous or ‘boom and bust’ life history strategies where the mature adults spawn synchronously and then die. The number of adults present will depend on the stage in their life cycle, whilst larval settlement and recruitment of juveniles can result in a massive increase in population size at certain times of the year (JNCC 2004b).

Although the life cycles of benthic taxa should always be considered, practical constraints may dictate that sampling is conducted within a sub-optimal timeframe. Where this is necessary, the temporal disparity should be acknowledged in analysis and reporting, in addition to stochastic events which may have impacted the benthic environment, such as anomalously cold or stormy winters (Davies *et al* 2001; JNCC 2004b).

When sampling cannot occur at the optimum time it may be possible to rationalise data to reflect species or communities at the preferred period. For example, Reiss and Kröncke (2005) observed an increase in benthic macrofaunal abundance during spring and summer due to recruitment, which decreased in the winter in response to food limitation and predation pressure. Sampling the adult population in a state of equilibrium would therefore require monitoring at a time of year when substantial survey time would probably be lost to adverse weather conditions. Given the scale of resources required for marine survey, the benefit of sampling in winter is unlikely to offset the cost in terms of sample numbers. It would therefore be justifiable to conduct the survey in spring or summer, and to remove the juvenile fraction from the main dataset.

5.1 Summary of key points and recommendations

Section 5: Considering temporal limitations

Key Points:

- The appropriate monitoring season will depend on the ecology and life history of the relevant indicator taxa.
- Seasonal variations in benthic ecosystems introduce variation into time-series data, which can obscure, or artificially elevate or decrease, the effect the monitoring aims to detect.
- Seasonal variation is generally caused by reproductive patterns and ephemeral cycles (e.g. algal and epifaunal die-back).

Recommendations:

- The timing of sample collection should be planned in relation to the known biology of the organism or community of interest, and temporal variation of the ecosystem.
- The effect of stochastic events such as anomalously cold or stormy winters should also be taken into account in analysis and reporting.
- Seasonal variation should be reduced as far as possible by undertaking repeated monitoring surveys in the same season, wherever logistical limitations allow.
- When it is not possible to temporally align repeated monitoring surveys, the potential impact on the time-series should be acknowledged and explored.

6 Considering inference, power and significance

Statistical inference, the process of deducing properties of a population from quantitative sample data, is a highly effective tool for detecting change, when data are representative and acquired by means of well-designed sampling strategies (Steele 2001). This section describes how to acquire a statistically robust sample with sufficient power to enable statistical inference and answer monitoring questions with confidence.

6.1 Statistical inference

There are various methods of drawing inference from quantitative sample data. Frequentist (or classical) inference is the most common, using an objective framework of hypothesis testing to objectively calculate probability (p-values). Although this method has become ubiquitous across many different fields it is not without limitations, given its inherent vulnerability to errors (discussed further in Section 6.5), and the potential for misinterpretation and/or misuse of the resulting p-values (e.g. Anderson *et al* 2000; Nuzzo 2014). Some statisticians advocate supplementing or even replacing p-values with other approaches. These include methods that emphasize estimation over testing, such as confidence or prediction intervals, decision-theoretic modelling and false discovery rates, and other approaches such as likelihood ratios or Bayesian methods (Wasserstein & Lazar 2016).

Bayesian methods calculate probability in respect to knowledge of the 'prior distribution' of a parameter, and are becoming increasingly popular in ecology. The frequentist and Bayesian paradigms are conflicted; although the advantages of Bayesian inference are well-documented, this branch of statistics is considered controversial by frequentists due to its perceived subjectivity and bias (Gelman 2008).

The following sections of the guidance focus on frequentist principles, and where 'statistical inference' is referred to throughout the remainder of this document, the term will refer to frequentist inference. A core emphasis on frequentist principles is considered appropriate due the need to reduce subjectivity, and to ensure that the principles of sampling design and the results of monitoring are accessible to those using monitoring products. It is also likely that the 'prior distribution' of a parameter required for Bayesian statistics may not be known for the initial monitoring survey.

It is acknowledged that there are various issues surrounding the use of p-values from frequentist analyses, but rigorous interpretation and transparent presentation of p-values and associated evidence should mitigate the common pitfalls. For instance, proper inference should be accompanied by full disclosure of the method used to determine sample size, the exclusions made, and the manipulations performed (Simonsohn *et al* 2013; Wasserstein & Lazar 2016). A p-value does not measure the size of an effect or, crucially, the importance of a result; therefore, effect sizes and confidence intervals should therefore always be reported to convey the magnitude of the effect (Nuzzo 2014). It is particularly important to emphasise that scientific conclusions, and resulting management or policy decisions should not be based solely on whether a p-value passes a specific threshold (Wasserstein & Lazar 2016).

Further guidance on the proper interpretation and use of p-values is presented in a statement issued by the American Statistical Association (Wasserstein & Lazar 2016).

6.2 Sampling design terminology

The robustness of statistical inference relies on accurate measurement of the selected indicator/s or parameters, through the acquisition of a well-designed sample (Eleftheriou 2013).

Table 3 outlines fundamental terms and concepts related to sampling design which will be used throughout the following sections.

Table 3. Sampling terminology.

Term	Definition
Population	In a statistical sense, a population is a collection of elements, objects or organisms of interest, to which the findings of a study are extrapolated (Steele 2001). *e.g. all sea pens within an MPA, or all epifauna on a rocky outcrop*
Sampling unit	A sampling unit is one of the units into which an aggregate (i.e. a population) is divided for the purpose of sampling, each unit being regarded as individual and indivisible when the selection is made. (Dodge 2003). Sampling units can be considered as individual 'items' which provide measurements of a particular variable, attribute or characteristic (Steele 2001). A sampling unit can be defined arbitrarily, such as a quadrat or transect, or naturally, such as an individual organism, depending on the monitoring objectives (Dytham 2011). *e.g. arbitrary units are often used for monitoring benthic habitats, such as a 100m camera transect, or a 0.1m^2 grab sample. An example of a naturally defined unit would be a single fish sampled to measure mercury concentration.*
Sample (N)	A part of a population, or subset from a set of sampling units (Dodge 2003), about which generalised conclusions can be drawn about the population by inference. *e.g. N = 146 x 0.1m^2 grab samples, or 54 x 100m camera transects*
Observation	The value of a variable taken from a specific sampling unit. *e.g. 32 sea pens observed from a single 100m camera transect*
Inference	The process of deducing properties of an underlying population by analysis of sample data. *i.e. the assumption that the patterns observed from sample data apply to the entire population.*
Variance (σ)	The distribution of data around their mean value.
Bias	The difference between a measured (sample) population mean and an accepted true population value (Bainbridge 1985). Bias is a systematic deviation of an estimate from the true value and is

	caused by artefacts of the method used to obtain the estimate (Andrew & Mapstone 1987), leading to an under- or overestimate of the true population value (Walther & Moore 2005). Measurement bias is mainly caused by faulty measuring devices or procedures, whilst sampling bias is due to unrepresentative sampling of the target population (Walther & Moore 2005).
Precision	The degree of concordance among a number of measurements or estimates for the same population (Cochran & Cox 1957; Sokal & Rohlf 1981; Lincoln *et al* 1982), precision is reflected by the variability of an estimate (Andrew & Mapstone 1987). Precision arises from the variance produced by the measurement device or procedure, in addition to sample variation (Walther & Moore 2005). The precision of a sample can be influenced by a wide range of factors, including measurement error, sample size, sampling unit size (e.g. a small vs large quadrat), sampling design and population variance.
Accuracy	The closeness of a measurement or estimate to the true value of the population (Cochran & Cox 1957; Sokal & Rohlf 1981; Lincoln *et al* 1982), as related to the bias and precision of the measurement.

6.3 Precision and accuracy

The probability that inference made about a population is correct and unbiased depends on the precision and accuracy of the sample. Precision and accuracy are interrelated, but they can vary independently when bias is present, leading to consistent over- or underrepresentation of the true parameter mean (as illustrated in Figure 3). For example, consider a survey that aims to measure infaunal abundance across a single habitat within a large MPA. If a large amount of sampling effort was concentrated in one corner of the site the variance in abundance could be low, resulting in high precision, but the accuracy might also be low, as the true population mean for the site would be biased if abundance varied across the MPA. In this case accuracy could be improved by distributing the sampling effort more evenly across the site.

Acquiring a precise and accurate sample that reflects the population is essential for robust inference (Lindenmayer & Likens 2010; Addison 2011), but this can often prove challenging in the benthic environment where indicator response (e.g. to pressures) must be identified against a background of natural variation or ‘noise’.

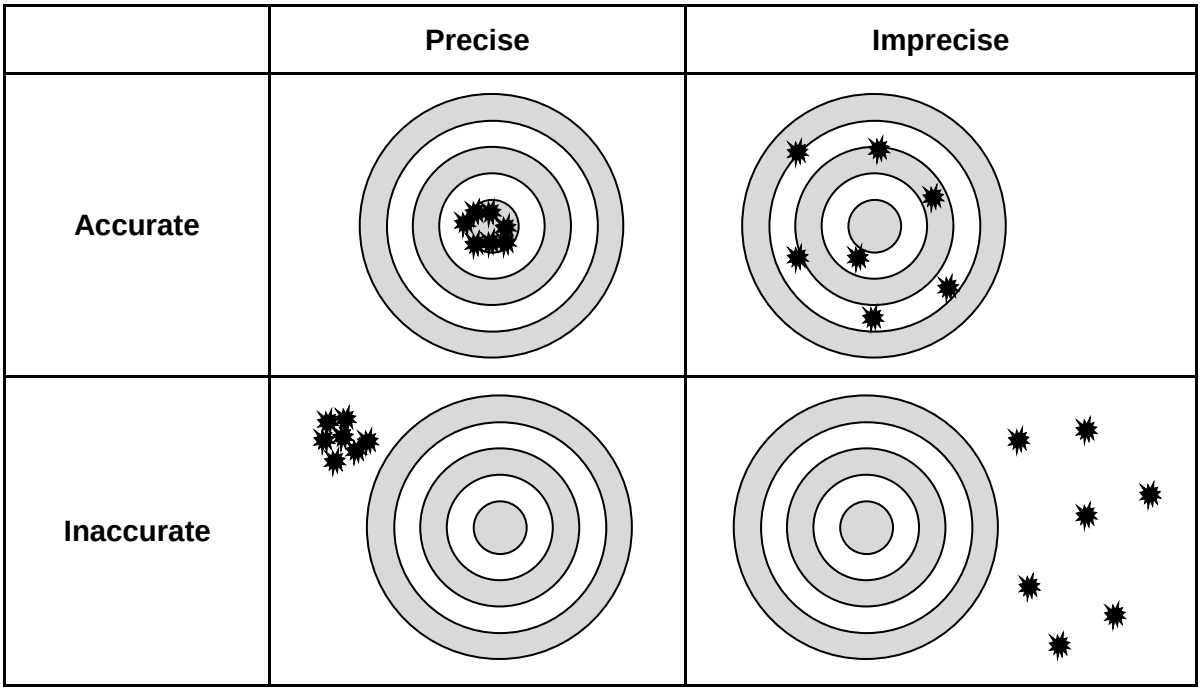


Figure 3. Theoretical illustration of sampling accuracy versus precision.
Note: The bullseye represents the true population mean

Marine flora and fauna can be extremely patchy in abundance and distribution, being influenced by many interacting processes at a variety of scales (Underwood & Chapman 2013). Environmental, biological and anthropogenic factors which cause noise in data include, but are not limited to:

- geographical location
- sediment composition
- habitat type
- hydrodynamic regime
- weather and temperature events
- nutrient availability
- pollution
- life cycles
- recruitment
- competition and predation
- anthropogenic disturbance

An additional layer of noise can be introduced through variation in sampling procedures; for example, variation in grab sample volume, or differences in sieving techniques or identification skills between video analysts.

Environmental parameters should be measured as covariates to reduce noise within datasets. The most commonly measured parameters for benthic habitats are depth and associated particle size distribution (PSD). Other parameters measured or quantified may include:

- light penetration
- organic matter
- nitrates & phosphates
- current speed
- temperature

- turbidity
- dissolved oxygen
- redox potential
- hydrocarbons and metals concentrations

The appropriate parameters to measure will vary according to the monitoring objectives, the selected indicators, and the location of the survey area. A well-researched sampling design and analysis specification will reduce noise within sample data, however it should be recognised that unexplained variation is likely to exist due to factors such as unmapped pressures or poorly understood faunal distributions.

6.4 Hypothesis testing

The formulation and testing of hypotheses about indicator response is central to evaluating whether change has occurred over time (Addison 2011; Eleftheriou 2013), and determining whether management measures are needed or have been effective. Hypotheses should be defined before sampling design to avoid ambiguity about what is actually being measured (Addison 2011).

Hypothesis testing is a method of statistical inference, which generally involves the comparison of two datasets, or the comparison of a dataset obtained by sampling against a synthesised data set from an idealised model. An alternative hypothesis (H_1) is proposed for the statistical relationship between the two data sets, and this is tested against a null hypothesis (H_0) that proposes no relationship between two data sets. Testing of a null hypothesis is based in the empirical falsification theory of Karl Popper (1935), which states that a theory can never be proven to be true, but it can be falsified. The essence of this theory is that it is not possible to prove the alternative or null hypotheses, only to reject the null hypothesis based on the probability that it is false.

Alternative hypotheses can be directional or non-directional. A non-directional alternative hypothesis simply states that the null hypothesis is incorrect, whereas a directional hypothesis states that the null hypothesis is incorrect and also specifies whether the true value of the parameter is greater than or less than the reference value specified by the null hypothesis. For example:

Box 2.

Non-directional alternative hypothesis:

H_1 : Density of sponges will ***change*** within an area closed to trawling for 6 years
H_0 : Density of sponges will remain the same within an area closed to trawling for 6 years

Directional alternative hypothesis:

H_1 : Density of sponges will ***increase*** within an area closed to trawling for 6 years
H_0 : Density of sponges will remain the same within an area closed to trawling for 6 years

The use of a directional hypothesis enables a one-tailed statistical test to be performed, which provides more power to reject the null hypothesis when it is false, by not testing the effect in both directions (Ruxton & Neuhäuser 2010). It is therefore beneficial to use a directional hypothesis wherever confidence in the predicted direction of change is high.

Where there is ambiguity in the predicted direction of change, a non-directional hypothesis should be used.

If the purpose of the monitoring is to identify long-term trends (sentinel monitoring), the rate and direction of change is likely to be unknown, therefore it is appropriate to use non-directional hypotheses, which have power to detect either an increase or a decrease in the chosen indicator over a selected time period (e.g. between two sampling events, or a longer time-series).

Where the purpose of the monitoring is to investigate pressure-state relationships (operational monitoring), or determine management needs and effectiveness (investigative monitoring), confidence in the direction of change should be higher, and use of a directional hypothesis may be justified.

Once the hypotheses have been defined, calculation of a test statistic and p-value will inform acceptance or rejection of the null hypothesis. The p-value is defined as 'the probability of observing results as extreme (or more) as observed, if the null hypothesis is true' (Dytham 2011). In plain terms the lower the p-value, the lower the probability that the null hypothesis is true. The threshold at which the null hypothesis is rejected (e.g. $p \leq 0.05$ or $p \leq 0.01$), known as the significance level (α), is selected based on the strength of the evidence required to conclude an effect.

At this point, consideration should be given to which statistical analyses will be most suitable to test the null hypothesis, as this will determine the type of power analysis and sampling design required. Further guidance on statistical analyses is presented in Section 10.

6.5 Type I and Type II errors

Hypothesis testing provides a powerful form of inference when used correctly, however the process is intrinsically prone to error. The two main forms of error in statistical testing are referred to as Type I and II errors (see Table 4), resulting in either incorrect acceptance or rejection of the null hypothesis. These errors are likely to lead to spurious conclusions about habitat condition, management effectiveness and pressure-state relationships, which could result in serious consequences for the marine environment and/or stakeholders (Green 1979; Fairweather 1991; Mapstone 1995; Underwood 1997b; Quinn & Keough 2002).

Table 4. Type I & II error: the four alternative outcomes of hypothesis testing.

		Truth	
		No significant effect actually occurring (H_0)	**Significant effect actually occurring (H_1)**
Decision made	**Reject H_0 (significant effect detected)**	Type I error	✓
	Don't reject H_0 (no significant effect detected)	✓	Type II error

A Type I (false positive) error occurs when a significant effect is detected, where in reality one has not occurred, resulting in erroneous rejection of the null hypothesis. The probability of Type I error is expressed as the significance level (α or p-value), which is conventionally but arbitrarily set at 0.05. This probability equates to a 5% (1 in 20) chance of falsely rejecting the null hypothesis (i.e. concluding that a change has occurred when it has not). The arbitrary significance level of $p \leq 0.05$ was initially suggested by Fisher (1925) and is

ubiquitous throughout many disciplines as a reasonable significance threshold. The traditional p≤0.05 threshold is, however, skewed towards reduction of Type I errors. This threshold is appropriate in certain situations; a useful parallel is that of the judicial system, where the need for proof 'beyond any reasonable doubt' in criminal prosecutions makes it less likely that an innocent person will be convicted (Type I error), but also more likely that a guilty person will go free (Type II error) (Peterman & M'Gonigle 1992). Statisticians have criticised the application of this arbitrary threshold for environmental management, as the level of proof required will reduce the likelihood of detecting subtler changes (Mapstone 1995; Buhl-Mortensen 1996).

A Type II (false negative) error occurs where no significant effect has been detected, when in fact one has occurred, resulting in erroneous acceptance of the null hypothesis. The issues surrounding Type II error (β) are more complex than Type I error, since Type II error is inversely related to statistical power (Addison 2011), which is determined by several different factors (discussed further in Section 6.6).

Type I and Type II error are inversely related to one another, so that by increasing α, β is reduced. When working with a set amount of resource which restricts maximum sample size, setting the level for α will determine β, therefore using a more lenient level of significance will result in increased power (and vice versa).

An example of the potential outcomes of Type I and II errors is presented in Box 3.

Box 3.

H_1 : Density of sponges will increase within an area closed to trawling for 6 years
H_0 : Density of sponges will remain the same within an area closed to trawling for 6 years

If a Type I error occurs (H_0 is erroneously rejected), the researcher would incorrectly conclude that sponge density had increased when in fact it had not. This could result in an overprotective management approach.

If a Type II error occurs (H_0 is erroneously accepted), the researcher would incorrectly conclude that sponge density had remained the same, when in fact it had increased. This could result in an underprotective management approach.

6.6 Conducting power analysis

Power analysis is a means of optimising the precision of a sample, giving the researcher the ability to select a sample size to detect an effect of a given magnitude, whilst controlling the degree of Type I and II error considered acceptable.

An increase in sample size (and concurrently cost) reduces variance in the sample, thereby increasing its precision and power. Although an increased sample size will always result in increased precision, the relationship between power and sample size is curvilinear, being analogous to a species-area curve (Figure 6). As the sample size increases, there are diminishing returns beyond a certain point on the power continuum. Power analysis allows researchers to determine where this point occurs, and simultaneously maximise statistical robustness and cost-effectiveness.

In addition to being conducted prior to sampling (*a priori*), power analysis can also be applied after a study has been completed (*post-hoc*) to reveal whether the sample was

sufficiently large for the desired level of power, thus informing future sampling designs and the degree of confidence with which inference can be made.

The statistical power (1-*β*) of a test can be described by the equation below (Di Stefano 2003):

$$\text{Power} \propto (ES \times \alpha \times \sqrt{N}) / \sigma$$

where α is the Type I error rate, ES is the effect size, N is the sample size and σ is the population standard deviation (Green 1989; Fairweather 1991; Osenberg *et al* 1994; Mapstone 1995).

These four elements are highly related, such that each is a function of the other three (see Table 5), and they can all be manipulated to varying degrees.

Table 5. The four elements of statistical power.

Element	**Set according to…**	**Power is increased where… (Underwood, 2013)**
Significance (α)	Socio-economic and environmental consequences of a Type I error.	α is less strict (the probability of Type I error is increased).
Variance (σ)	Estimated variance, using previous data for the same sampling area, or a proxy area of similar habitat.	Variance in data is small (resulting in high precision / low standard deviation).
Effect size (ES)	The magnitude of change to be detected in the selected indicator/s.	The size of the effect to be detected is large.
Sample size (N)	The resources available and the required level of power (1-β) and significance (α). N is fixed in *post hoc* analysis.	The sample size is large.

The following sections (Sections 6.6.1 to 6.6.4) discuss the process of defining the elements required to conduct power analysis.

6.6.1 Defining ratios and levels of power and significance

The need to balance conservation objectives with socio-political considerations will influence the relative importance of committing Type I and II errors in the context of each monitoring programme (Di Stefano 2001). This is particularly relevant within a system of adaptive management where stakeholder access to specific areas is reviewed on a periodic basis (for example, closures within an MPA). From a conservation perspective, the failure to detect an impact (Type II error; false negative) is more serious than incorrectly concluding an effect has occurred (Type I error; false positive) (Peterman 1990; Taylor & Gerodette 1993; Di Stefano 2003). This concept aligns with the precautionary principle of environmental management, which advocates measures to reduce the probability of Type II errors by adopting a less conservative approach to hypothesis testing, thus improving ability to detect more subtle changes in the marine environment (Gray 1990, 1996; Peterman & M'Gonigle 1992; Underwood 1997a).

Despite the more serious environmental implications of Type II errors, the level of 'adequate' statistical power (1-β; describes the ability of a test to detect an effect if the effect actually exists) is commonly defined as 0.80, whilst statistical significance (α; the probability of not detecting an effect when in fact it exists) is conventionally set at 0.05. Adherence to this 'five-eighty convention' (Di Stefano 2003) results in a ratio of α to 1-β which equates to a 5% to 20% ratio of error probability, meaning that you are four times less likely to detect an effect when it exists (Type II error) than to falsely detect an effect when it does not exist (Type I error). This ratio may be appropriate in some situations (e.g. where the burden of proof is high), however it is less sensitive to more subtle changes in habitat condition and should not automatically be applied when monitoring for conservation purposes (Buhl-Mortensen 1996; Di Stefano 2003). The ratio of α and 1-β should be defined on a case-by-case basis according to perceived costs of committing Type I and Type II errors to both stakeholders and the environment, taking into account the trade-off between the resources required and the need to provide robust evidence.

Once an acceptable ratio of α and 1-β has been determined, minimum levels should also be defined and adhered to, as monitoring an indicator with low power is potentially a waste of limited resources.

6.6.2 UKBMBP approach to defining ratios and levels of power and significance

As part of the UK Marine Biodiversity Monitoring Programme (UKMBMP) JNCC has developed an approach to defining appropriate ratios and levels of α and 1-β for sentinel, operational and investigative monitoring, according to the relative risks and costs to the environment and stakeholders of committing Type I and Type II errors.

Ideally levels of α and 1-β will be as low and as high, respectively, as possible, and the ratio balanced (e.g. α = 0.05 / 1-β = 0.95), so that the risks of committing Type I and Type II errors are equal. This can, however, result in an unfeasibly large sample size, and ratios and levels may need adjustment if the budget and/or length of sampling period are not sufficient. The UKMBMP approach advocates defining minimum ratios and levels of α and 1-β on a case-by-case basis, within the framework presented in Table 6. This framework provides guideline recommendations for minimum and optimum levels and ratios of α and 1-β, giving the user the flexibility to select levels and ratios which will produce a robust dataset for the specific survey area and monitoring objectives.

It should be noted that the minimum values presented in the framework will not be universally sufficient, and the optimal values will not always be achievable. The requirements for each monitoring event should be assessed on a case-by-case basis to determine the minimum acceptable values prior to power analysis. If these values are not achievable the user should consider other methods of increasing power, for example; increasing the detectable effect size (if this is ecologically valid), considering a different indicator if multiple indicators are available, or diverting resources from lower priority monitoring objectives.

When deciding on the minimum ratios and levels of power and significance, the associated costs and risks to the environment and stakeholders should be well understood and acknowledged. These are discussed below with respect to the three monitoring types.

Sentinel monitoring

The primary function of sentinel monitoring is to detect change before irreversible damage occurs. If a Type I error were committed it would result in the failure to detect an actual decline in habitat condition, and continued damage to benthic habitats by unmanaged

activities could occur. It can therefore be argued that the costs to the environment of reaching an erroneous conclusion would be greater when efforts are made to minimise the likelihood of committing a Type I error (e.g. setting a low significance threshold) because this automatically reduces power. Conversely, committing a Type II error could result in detecting a decline in habitat condition when it had not actually occurred, thus unnecessarily triggering costly operational or investigative monitoring activities.

Given the 'early warning' function provided by sentinel monitoring it is recommended that the ratio of α to 1-β is not skewed towards minimising either the risk of Type I or II errors, and that they should be consistently set as equal, with a balanced ratio.

Levels of α and 1-β should be set as low and high, respectively, as possible, taking into account logistical and financial constraints and the desired level of change to be detected. It is critical that the initial monitoring event provides a robust first dataset which will improve understanding of local variance, against which changes can be identified. It is recommended that the levels of α and 1-β should be set to a minimum of 0.20 and 0.80, however, where resources allow, the robustness of the initial dataset should be maximised by the application of more stringent levels (e.g. 0.05 and 0.95). If post-hoc power analyses and species accumulation curves suggest that the initial sampling effort exceeded the amount required to characterise the variance, the levels could then be relaxed for subsequent events.

Operational monitoring

Operational monitoring studies are conducted to improve understanding of the relationships between intensity of human pressures and habitat condition. The results of these studies will provide a basis for assessing observed changes in habitat state, inform the development of targets for acceptable pressure levels, and provide evidence for the development of indicators.

Using the example of demersal fishing abrasion, committing a Type I error could result in falsely concluding that there are differences caused by abrasion between one or more pressure levels, or between reference areas and areas subject to abrasion, when in fact there are none. This could lead to the introduction of inappropriate management measures or the setting of unnecessarily low levels of acceptable pressure, potentially resulting in higher economic costs for the fishing industry. On the other hand, if falsely concluding that there are no pressure-related differences in indicator values (i.e. committing a Type II error), this could result in setting the levels of acceptable abrasion pressure too high or implementing insufficient management measures, leading to a high risk of adverse changes to the environment. Similarly, if these false conclusions were used to develop pressure-based indicators, habitats may be assessed to be in a better state than they actually are. It is therefore recommended that the risk of committing a Type I error is kept low (by setting α low) to retain sufficient confidence in conclusions on the direction and shape of the pressure-state relationship.

As the risks associated with committing both Type I and Type II errors are high for both biodiversity and stakeholders, the ratio of α to 1-β should be equal where possible. It is recommended that α is set at 0.05 or less due to the requirement for robust scientific evidence for or against a relationship between a pressure and habitat condition. The probability of committing a Type II error should, where feasible, also be set at 0.05 or less (equating to 1-β ≥0.95) to reduce the risk of underestimating or failing to detect a pressure-related difference in habitat condition along pressure gradients. If a balanced ratio of α and 1-β is not achievable due to resource limitations, and effect size cannot be altered, the ratios of α and 1-β can be adjusted to maintain a minimum α of 0.05. As a minimum, 1-β should be set at 0.80 (α to1-β = 1 to 4), to preserve confidence in the conclusions of the study.

Investigative monitoring

Since the aim of investigative monitoring is to test the effects of management measures, or to test a hypothesis in the form of an experiment, a close link is expected between the results of the statistical analyses from such an experiment and action taken in terms of management. Again, using the example of demersal fisheries abrasion, falsely concluding that a closure had improved habitat condition (a Type I error) could result in an unnecessary continuation of the closure despite no negative effects of fishing, causing economic burden for the fishing industry and potential damage to the credibility of the monitoring programme. On the other hand, if concluding that the closure had not improved habitat condition when in fact it had (a Type II error) resulted in removing the closure or not taking management action in other areas, the risk to the environment would be high and resulting damage could be irreversible.

As with operational monitoring, the risks of committing both Type I and II errors are potentially high for biodiversity and stakeholders, therefore the ratio of α to 1-β should be equal wherever possible. However, α should always be set to 0.05 or less, due to the close link of results to management measures and requirement for strong evidence (i.e. rigorous hypothesis testing). Where resources allow, 1-β should be set at ≥0.95 to maintain the balance of α and 1-β. If the ratio must be adjusted to maintain a low likelihood of Type I errors in the context of available resources, it is recommended that the minimum 1-β is set at 0.80 (α to 1-β ratio of 1 to 4). Ratios and levels of α and 1-β from the initial monitoring event should be retained for subsequent events to ensure comparable levels of precision and power.

Table 6. UKMBMP proposed optimum and minimum ratios and levels of statistical significance (α) and statistical power (1-β) for sentinel, operational and investigative monitoring.

Monitoring type	Level / ratio	Optimum		Minimum	
		α	1-β	α	1-β
Sentinel monitoring (Type 1)	**Level**	≤0.05	≥0.95	0.20	0.80
	Ratio	1	1	1	1
Operational monitoring (Type 2)	**Level**	≤0.05	≥0.95	0.05	0.80
	Ratio	1	1	1	4
Investigative monitoring (Type 3)	**Level**	≤0.05	≥0.95	0.05	0.80
	Ratio	1	1	1	4

6.6.3 Estimating variance

In power analysis variance is estimated using previous data from the area of interest, or if this is not available, using proxy data acquired from a similar habitat (preferably from the same geographical region). Ideally the parameter from which variance is estimated should be a fully validated indicator or series of indicators, however where indicators are not yet fully developed it may be appropriate to use a broad proxy indicator (e.g. a univariate metric). If the responsiveness of the proxy indicator is unclear, a precautionary approach could involve testing several univariate parameters (e.g. total abundance, richness, diversity), and selecting the parameter with the greatest level of variability, thus requiring the highest number of stations (i.e. that which is most variable in space and time).

The accuracy of a power analysis result is wholly dependent on whether the data used accurately reflect the true variance of the habitats being monitored. With this in mind it is important that data are assessed for suitability prior to analysis; for instance, is the number of stations sufficient to describe variance in the indicator/s of interest, is the spatial coverage sufficient in comparison to the area of interest, and are the levels of pressure similar? To help evaluate the degree to which existing data have described benthic communities, it may be useful to generate species accumulation curves for infaunal data.

6.6.4 Selecting an effect size

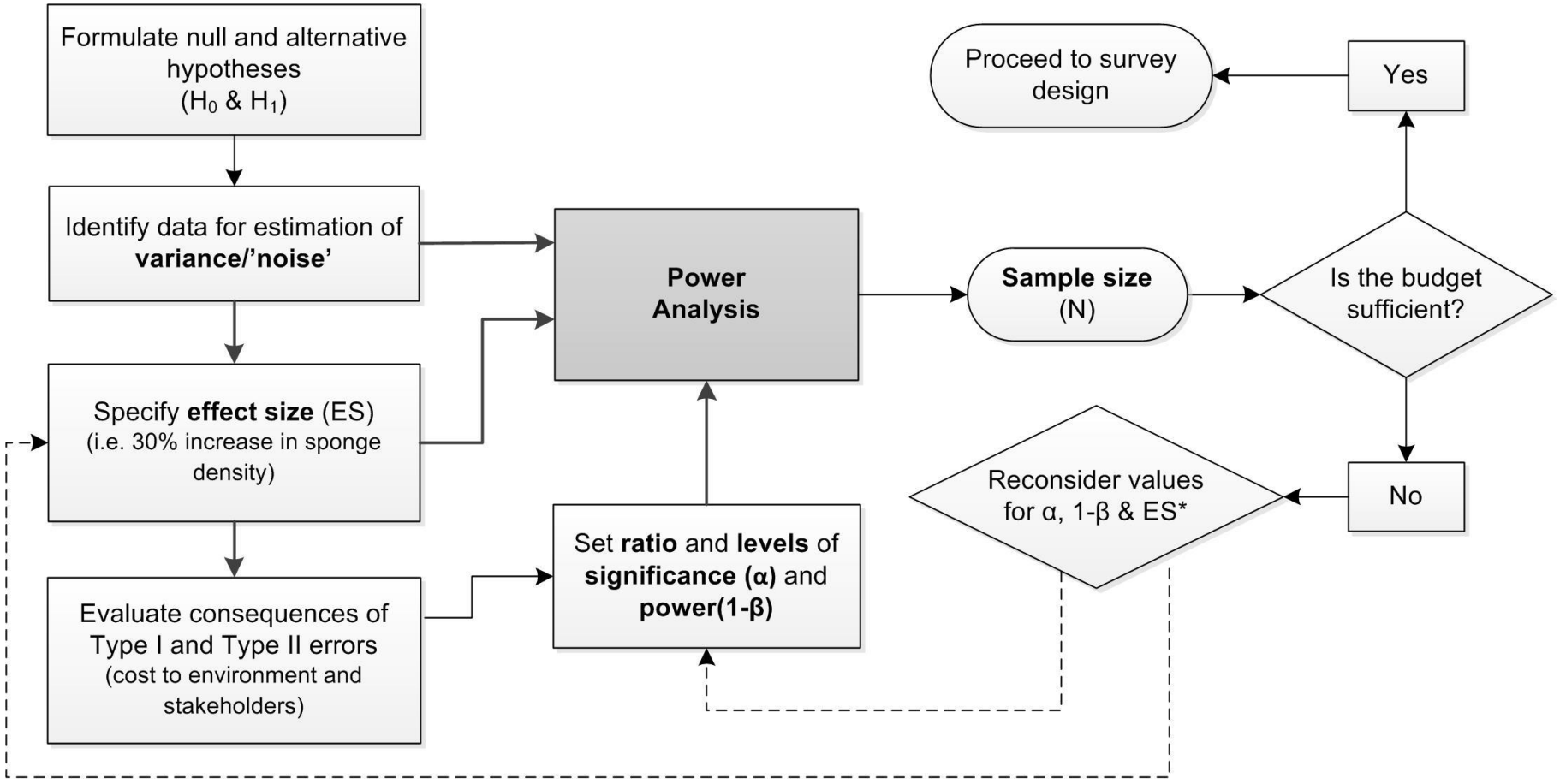
The effect size is the magnitude of change to be detected in the selected indicator/s, and can be extremely difficult to define. The selected effect size should ideally be based on a good ecological understanding of the habitat and associated communities of interest, and should involve judgements about the level of change that is likely to occur, and that is considered unacceptable (Mapstone 1995). In reality, many benthic systems and pressure-state relationships are poorly understood and alternative methods may be necessary. Munkittrick *et al* (2009) reviewed alternative methods for determining effect sizes, and recommended that where understanding is poor effect sizes should be selected through discussion with stakeholders, or by adopting effect sizes from comparable studies that used similar parameters.

The detectable effect size may also be constrained by resources, and can be altered (within pre-defined acceptable limits). For example, if an original desired effect size of 20% results in an unfeasibly large sample, it may be necessary to reduce the detectable effect size to 30% to maintain the required ratio and levels of power and significance.

6.6.5 Conducting *a priori* power analysis

A priori power analysis can be conducted once data have been selected from which to estimate variance, and power, significance, and the effect size has been defined (as illustrated in Figure 4).

Power analyses are available in many standard platforms and packages, for example R (emon package; Barry & Maxwell 2017), SPSS, Minitab, and specialist software programmes (e.g. GPower, PASS), all of which support a range of analytical designs. The analysis will output the required sample size for a specified effect size, and ratio and level of power and significance. If this sample size is unachievable given the available resources, the researcher must reconsider the ratio and levels of power, significance, and/or effect size. It is, however, imperative that minimum values for all three of these elements are specified and upheld to ensure that data are sufficient to achieve environmental and socio-political objectives.



* It should be noted that although values for α, 1-β and ES may be adjusted, the values must be sufficient to achieve the socio-political and ecological objectives of the monitoring programme. Minimum values should be set on a case by case basis.

Figure 4. Flowchart illustrating the *a priori* application of power analysis.

Where the monitored habitats display distinct variation (e.g. different habitat types or bathymetrically distinct areas), the sampling effort can be partitioned into separate 'strata' (discussed in Section 9.1.2). To increase precision, the number of sampling points can be determined via power analyses performed for each stratum, however this approach could potentially result in an unaffordable sample size. Alternatively (but requiring more computation), the power analysis could be based on the stratified mean of the area from which the strata were taken (i.e. the sum of the stratum means weighted by the number of sample units within each stratum). This would allow selection of sample sizes for individual strata proportional to their variance and the area of the stratum (Cochran 1977).

Before conducting power analysis, it is important to understand the distribution of the underlying data, to enable the appropriate distribution to be fitted (e.g. Poisson, Negative Binomial, Gaussian, Lognormal). The most straightforward way of determining the correct distribution is to plot a density estimate of the actual data (essentially a smoothed histogram) against a selection of fitted distributions. The resulting plot will allow comparison of the actual and fitted distributions to select which one is most similar (see Figure 5, created in R). An applied example of *a priori* power analysis is presented below.

Example: *a priori* power analysis

The following example details *a priori* power analysis conducted prior to sampling at an offshore MPA. The purpose of the monitoring was to provide the 'before' dataset in a BACI study (see Section 9.4) to determine the effectiveness of management measures (investigative monitoring), where specific zones of the MPA were to be closed to demersal fishing.

Previously acquired macrofaunal taxon richness data were available from >200 sampling stations within the MPA, and were stratified into coarse sediment, mixed sediment and sand strata. A density estimate curve was generated for each stratum, and plotted against fitted Poisson, Negative Binomial and Gaussian distributions. The appropriate distribution for the power analysis was then selected (e.g. in Figure 5, the distribution best fitted to the data is the Negative Binomial).

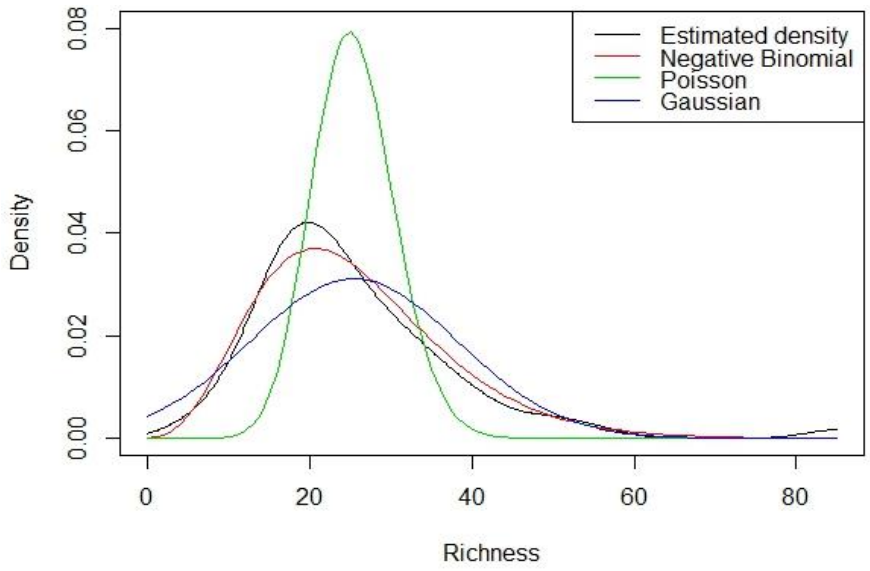


Figure 5. Plotting fitted distributions against a density estimate for actual data within a coarse sediment habitat stratum.

A one-tailed test (power.BACI in R emon package (Barry & Maxwell 2017)) with a Negative Binomial distribution was used to derive power values for 10% incremental increases in effect size, using the following directional hypotheses:

H_0 : Mean taxon richness will increase by (*10% increment*) in an area closed to demersal fishing in comparison to an area which remains open (assuming identical environmental characteristics).

H_1 : Mean taxon richness will not change in an area closed to demersal fishing in comparison to an area which remains open (assuming identical environmental characteristics).

The power curves displayed in Figure 6 were generated to determine the sampling effort required to detect incremental increases in taxon richness (10% - 50%) derived from grab samples of course sediment. Taxon richness was used as a proxy metric, in place of a fully developed indicator, and the expected effect size was unknown, therefore it was determined that a precautionary approach would be to use a low effect size of ≤30%. An adaptive management approach had been specified for the MPA, and the result of the study was likely result in actions which could have consequences for stakeholders. Therefore, the significance level (α) was set to 0.05 (5% probability of a Type I error), with the desired power level set at 0.90 to increase the robustness of the design (10% probability of committing a Type II error).

The point at which the curves intersect the horizontal dotted line (1-β = 0.90) indicate the sample size (N) required to achieve 90% power at 0.05 significance. In this case, the sample size required to detect an effect size of 10% was unfeasibly large, however a sample size of 44 was required to detect a 20% increase in taxon richness, which was achievable within the monitoring budget.

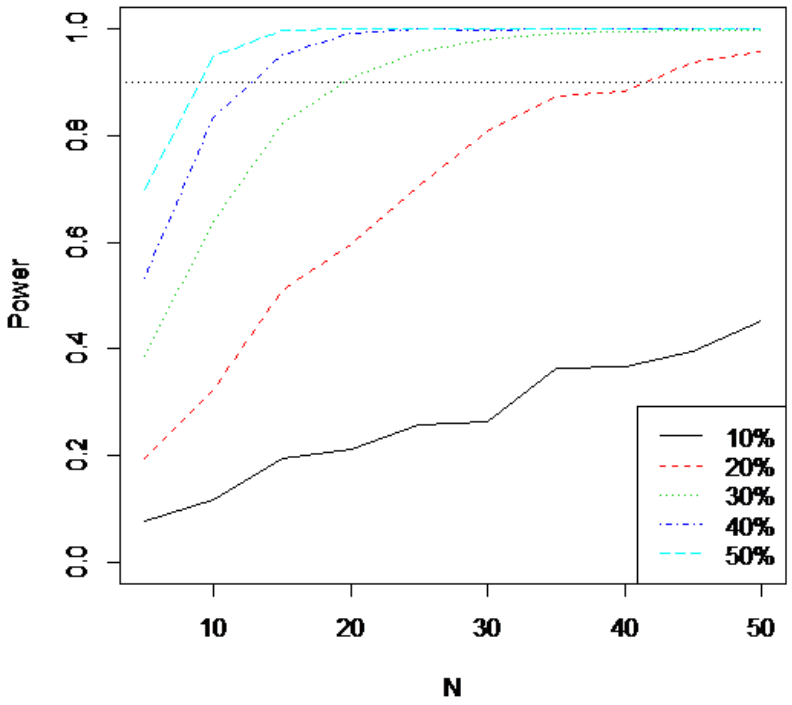


Figure 6. Power curves generated to determine the sampling effort required to detect 10% incremental increases in taxon richness in grab samples of coarse sediment, with significance (α) set at 0.05 and power set at 0.90.

6.6.6 Conducting *post hoc* power analysis

Power analyses conducted retrospectively (*post hoc*) allow the researcher to determine whether the number of samples taken has generated the desired level of power, given the inherent uncertainty in *a priori* estimation of effect sizes. Post hoc analysis will increase confidence in acceptance or rejection of the null hypothesis (i.e. is a non-significant effect really non-significant?), and will also provide information on whether the sample size should be increased or reduced for subsequent monitoring events.

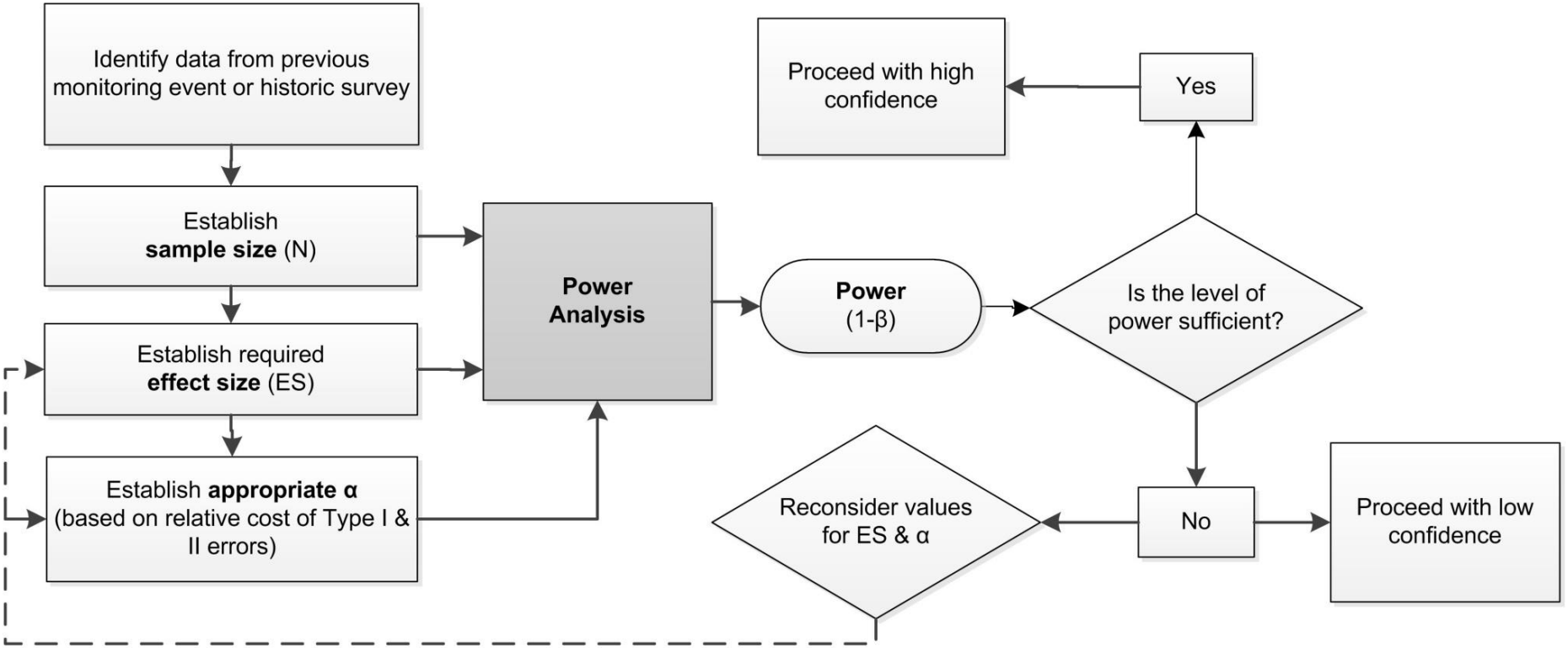
Post hoc power analyses are similar to those conducted *a priori*, using the fixed sample size acquired, the variance of the sample, and the specified α to calculate post hoc power values for various different effect sizes. These values are used to determine whether the sample size acquired has sufficient power to detect the required effect size. If the level of power is not sufficient, effect size and/or α may be altered within the limits set prior to analysis to improve power, or the user must proceed acknowledging the limited power of the design and caveating the results of statistical analysis.

A flow process for *a priori* power analysis is displayed in Figure 7.

6.7 Conducting precision analysis

The precision of a sample can be calculated using the sample mean and confidence intervals using the precision function in the R emon package (Barry & Maxwell 2017), if the data distribution is approximately normal (J. Barry, Cefas, pers. comm. 2017). This analysis will supplement the results of power analyses and determine whether the desired level of precision (which will vary dependent on the variable being measured) is likely to be achieved given the number of samples calculated for a given power and significance.

If the available resources do not allow attainment of the high precision in addition to the required level of power, the level of precision achieved should be noted and the results of analysis caveated with this information (J. Barry, Cefas, pers. comm. 2017).



* It should be noted that although values for α and ES may be adjusted, the values must be sufficient to achieve the socio-political and ecological objectives of the monitoring programme. Minimum values should be set on a case by case basis.

Figure 7. Flowchart illustrating the *post hoc* application of power analysis.

6.8 Summary of key points and recommendations

Section 6: Considering inference, power and significance

Key Points:

- Statistical inference (e.g. hypothesis testing) is central to evaluating whether change has occurred, or whether a relationship exists.
- In hypothesis testing, a Type I error occurs when a significant effect is detected, where in reality none has occurred. A Type II error occurs where no significant effect has been detected, when in fact one has occurred.
- Statistical power (1-β) is the probability that a test correctly rejects the null hypothesis when it is false, and is a product of the statistical significance level (α), the magnitude of the effect size (ES), sample size (N) and parameter variability (σ).
- The commonly used α to β ratio of 5% to 20% error probability results in a test which is skewed towards reduction of Type I errors. This ratio may result in the failure to detect change when it exists.
- Power analysis can be used to determine how large the sample (N) must be to detect change against a background of natural variation.

Recommendations:

- Non-directional hypotheses should be used for sentinel monitoring where the direction of change is unknown, or for operational or investigative monitoring if confidence in the direction of the effect is low. Directional hypotheses should be used for operational and investigative monitoring where confidence in the direction of the effect is high.
- Levels of significance and power should be selected according to the relative costs to biodiversity and stakeholders of committing Type I and Type II errors.
- JNCC have developed a flexible framework which can be used to help define appropriate ratios and levels of power and significance (Table 5).
- Power analysis should be conducted *a priori* (for each stratum) to determine how large the sample (N) must be to detect change of a given magnitude at a given level of significance. *Post hoc* power analysis should be conducted retrospectively to determine whether the sample was sufficiently large.
- Precision analysis should be conducted to supplement the results of power analysis.
- Environmental parameters which are thought to strongly influence variation in the distribution of indicators or add noise to the data should be measured.

7 Selecting sampling units

After generating a statistically robust sample size (N) through power analysis, it is important to ensure that the sampling units provide accurate observations of the indicator/s in question. A number of factors which can determine the effectiveness of sampling units must be considered as part of the design process, the most influential of which are the size and type of the sampling unit (see Eleftheriou 2013), and the amount of replication required within each sampling unit.

7.1 Sampling unit size

The distribution of benthic taxa varies at a range of spatial scales in response to natural and anthropogenic factors. Furthermore, flora and fauna are likely to be found in patchy aggregations as opposed to an even distribution (Underwood & Chapman 2013). Selecting the correct size and type of sampling unit is therefore extremely important for effective sampling, and identification of distribution patterns.

The issues that might arise from using a sampling unit which is too large are demonstrated in Figure 8, using the example of a single polychaete species. It is clear that with too large a sampling unit, the population would be described as evenly distributed with a low variance (Figure 9B). This is obviously not representative of the target population, and using a smaller sampling unit allows the patchy distribution to be observed with a larger variance, because of the greater potential to sample the areas between clusters.

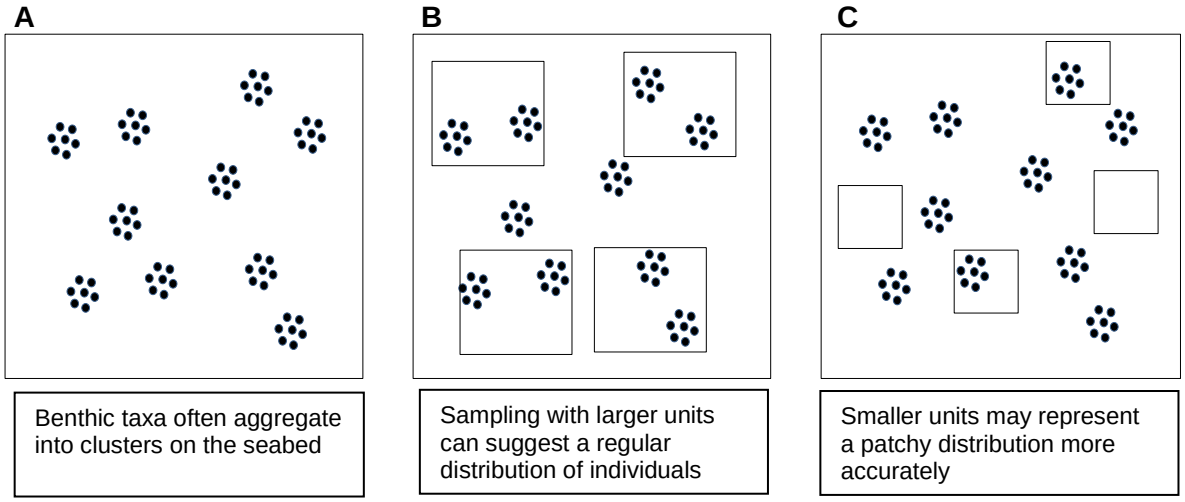


Figure 8. The effect of different sized sampling units on representation of a clustered faunal distribution (adapted from Underwood & Chapman 2013).

In another example (Figure 9), the indicator is the density of sea pens, which are sparsely distributed. In this case it is likely that a large sampling unit would be required to accurately detect the true sea pen distribution. In this case video transects would be more appropriate as sampling units to describe the distribution, as opposed to grab sampling at discrete points. It will generally be appropriate to split long video transects into smaller segments to record variation along the transect. The segment interval will vary on a case-by-case basis, but should be ecologically meaningful and related to the expected distribution of the habitat or organism(s).

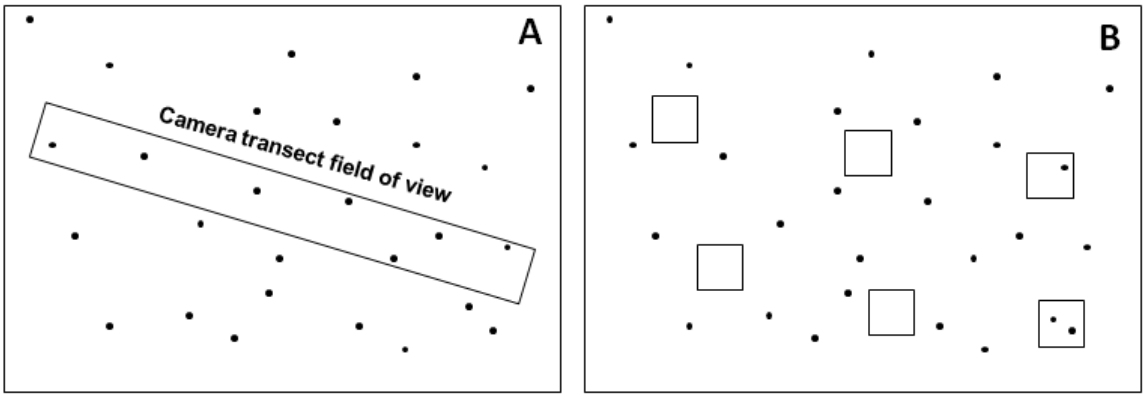


Figure 9. Comparative estimates of sea pen distribution using a camera transect (A), and grab samples (B).

Ideally, the precision and accuracy of different sampling units should be investigated in a pilot study assessing retention of the species or community of interest, generating species accumulation curves and considering spatial autocorrelation (e.g. between grab samples or segments of a video transect, see Section 8.1). If resources do not allow this, the life history and ecology of the species, or a similar proxy species should be researched to inform decisions on the type of sampling unit used (Sutherland 1996; Underwood & Chapman 2013).

7.2 Replication within sampling units

Replication within sampling units (e.g. a grab station) is commonly practised to reduce the effects of random variation, and to improve understanding of small-scale variation, particularly in systems where assemblages are likely to display a patchy or heterogeneous distribution (Hurlbert 1984). Replication also increases the likelihood that rare or sparsely distributed taxa will be captured within the sample. The replicates can be analysed to evaluate within-station variance, then aggregated and averaged for comparison with other sampling units across the survey area.

MESH Recommended Operational Guidelines (ROGs) for grab sampling state that each a minimum of three successful replicate grab samples should be taken at each station (Guerra & Frietas 2012). However, the optimum amount of replication at a single station is likely to vary depending on the habitats and indicators in question. For instance, high energy sediments are intrinsically more variable than those found in more depositional environments, and the amount of within-station replication needed may be higher.

Inevitably the need to understand small-scale variability through within-station replication must be balanced with the requirement to collect data at a wide range of separate stations. In advice to JNCC, Holtrop and Brewer (2013) recommended that when resources are limited collecting samples from a wider range of sampling locations should be prioritised over within-station replicates, however this approach is likely to lead to a reduced understanding of localised variation.

Decisions on whether or not to replicate within stations, and the amount of replication, needed should be taken on a case-by-case basis depending on the monitoring objectives. It is recommended that within-station replication is conducted where possible, however this may not be achievable if resources are limited.

7.3 Summary of key points and recommendations

Section 7: Selecting sampling units

Key Points:

- It is important to ensure that the sampling units provide accurate observations of the indicator/s in question.
- The size and type of the sampling unit determine the effectiveness of the unit for drawing inference about a population.
- Replication within sampling units reduces the effects of environmental 'noise' or random variation and provides a more accurate and precise estimate, particularly where biota are likely to display a patchy distribution.

Recommendations:

- The size and type of the sampling unit should be tailored to the size and expected distribution of the indicator; a unit which is too large or too small may result in the inability to detect spatial patterns.
- If possible, the precision and accuracy of different sampling units should be investigated in a pilot study.
- Replication within sampling units should be conducted where resources allow (e.g. three grab samples to provide a mean value per sampling station).

8 Considering dependency issues

A key assumption of many statistical analyses is independence of observations. For two observations (i.e. the data acquired from two sampling units) to be considered independent, the occurrence of one must not affect the probability of the other occurring. Achieving independence of observations may prove difficult in the marine environment, where abiotic and biotic parameters vary and interact over space and time, and processes may be stochastic. The implications of non-independent units are potentially less serious for some multivariate analyses in comparison to univariate tests (Clarke 1993), however it is important that dependence within or between sampling units is minimised wherever possible (Underwood & Chapman 2013).

Biological and environmental interactions result in correlations within response variables in space and/or time; spatial autocorrelation and serial correlation. These forms of correlation are discussed in the following sections.

8.1 Spatial autocorrelation

Spatial autocorrelation refers to the pattern in which observations from nearby locations are likely to have values more similar than would be expected due to chance alone (Fortin *et al* 2002), and can be positive or negative. Negative autocorrelations may occur when individuals engage in resource competition, territoriality or avoidance behaviour (Legendre & Fortin 1989), creating a 'checkerboard' distribution, as described by Diamond (1975). Positive autocorrelation occurs when taxa are distributed in clumps or patches, or form aggregations. For example, *Sabellaria spinulosa* reefs are colonised by gregarious settlement, with existing aggregations of *S. spinulosa* encouraging settlement of larvae (Wilson 1970), therefore two sampling units taken in close proximity are likely to be highly spatially autocorrelated.

Traditional random selection of sampling locations can result in samples being taken in close proximity to each other. This problem can be avoided by taking a 'pseudo-randomisation' approach to sampling, in which sampling locations are selected at random, but a minimum distance (a buffer) is maintained between them. Spatial analysis can be used to calculate an appropriate buffer around sampling locations, using existing data (from the site in question, or the nearest possible analogue). These calculations assess the similarity of a metric (e.g. taxon richness) at pairs of sampling points, as a function of the distance between them (Wilding *et al* 2015), and provides the minimum distance between sample locations required to ensure independence.

Commonly used methods of identifying spatial autocorrelation for sampling designs include;

- **Production of semivariograms**; these plot semivariance (half the variance in the differences between the values of a variable at two locations) against sampling distance. The point at which the semivariance levels out, the sill, indicates the distance within which sampling points are spatially autocorrelated, taking into account residual random variation which includes measurement error (nugget variance) that is not spatially correlated (Bourgeron *et al* 2001). A theoretical example of a semivariogram is presented in Figure 10. The reliability of the semivariance estimate increases with sample size for any given pair-wise distance, and is dependent on the spatial distribution of the samples. Crawley (2013) recommends at least 30 data points within each sampling stratum to determine semivariance. Semivariograms can be plotted in the R environmental monitoring package 'emon' (Barry & Maxwell 2017), using the 'svariog' function.

- **Moran's I or Geary's C coefficients**; these coefficients provide a measure of spatial autocorrelation for a single variable. Moran's I computes the degree of correlation between the values of a variable as a function of distance, whilst Geary's C measures the difference among values of a variable at nearby locations (Fortin *et al* 2002).

- **Mantel test;** this test provides a linear estimate of the relationship between two distance matrices based on data sets (i.e. environmental variables and biological metrics) obtained at the same sampling locations (Bourgeron *et al* 2001).

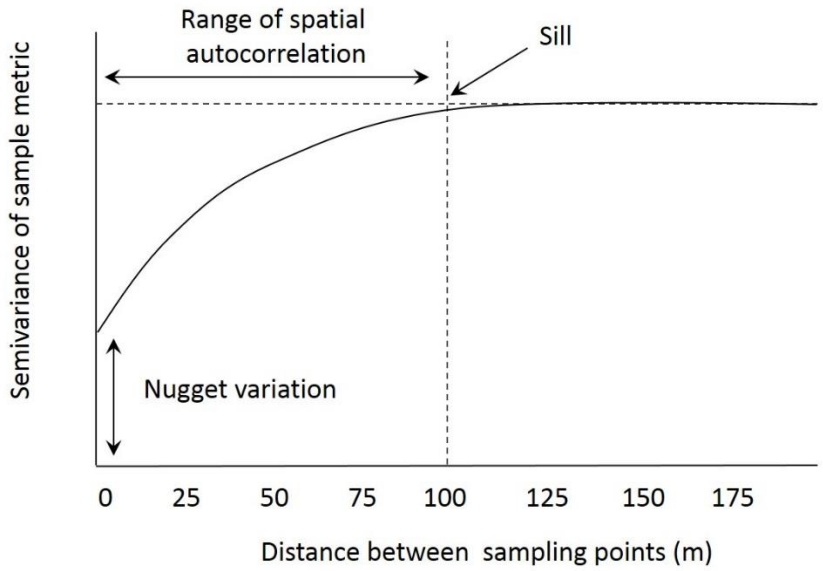


Figure 10. Example of a semivariogram. The sill (point beyond which samples are spatially independent) is reached at 100m distance. Nugget variation may be attributed to measurement error, or variation at scales smaller than the sampling distance.

8.2 Serial correlation

Serial correlation (or temporal autocorrelation), is the correlation of a variable with itself across different points in a time series. Serial correlation violates the assumption of independence between observations, the result of which can be erroneous rejection of the null hypothesis due to artificially exaggerated 'goodness of fit' within regression models, or identification of a trend which does not exist. This error can be common in time-series and monitoring data, particularly when sampling periods are close together, sampling points are fixed, and when indicator organisms are sessile, slow-growing or long-lived.

Serial correlation is less likely to be a significant issue for offshore monitoring, as the frequency of data points in the time-series is generally limited by logistical and financial constraints. It is likely that serial correlation will arise more frequently in nearshore or intertidal monitoring, particularly when permanent plots or transects are used. Monitoring surveys should be planned to minimise the strength of serial correlations where possible, by establishing a sufficient interval between sampling events based on prior knowledge of the indicator (Underwood & Chapman 2013). For example, organisms which are slow-growing and long-lived will require a longer interval than those which are short-lived and have a high population turnover (unless the monitoring is specifically focused on determining recovery rates).

Re-randomisation of sampling locations will reduce temporal autocorrelation, however, this will result in the loss of information about specific features (e.g. recovery of a particular coral mound). If the objective of the monitoring is to infer the overall condition of a habitat, then within-site sampling locations should not remain fixed. Rather, they should be re-randomised for each sampling event to further minimise the potential for serial correlation (Jon Barry, Cefas, pers. comm. 2015). If the aim of the monitoring is to make inference about discrete areas, such as measuring growth-rates or cover at specific locations, then sampling locations should remain fixed, and a repeated measures analysis should be used to model the serial correlation.

Replacement, or double-counting, of individuals can also confound the assumption of independence, particularly where species are highly mobile. The likelihood of double-counting is negligible for extractive survey activities, as all organisms are typically retained from grab samples or trawls. The majority of epifaunal organisms are unlikely to be double-counted in the sample during camera operations, however video analysts should be aware of this possibility for highly-motile fauna such as fish.

8.3 Pseudoreplication

Sampling units that are highly spatially or temporally correlated can be described as ‘pseudoreplicated’ if they are treated as if they were independent in analysis. Hurlbert (1984) brought attention to the issue of pseudoreplication in experimental designs as ‘…probably the single most common fault in the design and analysis of ecological field experiments’.

Hurlbert specifically addressed the issue of pseudoreplication in experimental designs, such as might be employed for operational or investigative monitoring activities (discussed further in Sections 9.3 and 9.4), stating that ‘…pseudoreplication most commonly results from use of inferential statistics to test for treatment effects with data from experiments where either treatments are not replicated (though samples may be) or replicates are not statistically independent’. Pseudoreplication thus refers not to a problem with experimental design per se but to a combination of experimental design and statistical analysis that is inappropriate for testing the hypotheses of interest. The most common examples of pseudoreplication are wrongly treating multiple samples from one ‘experimental unit’ (e.g. a ‘control’ or a ‘impact’ site in a BACI study, or an individual VMS abrasion cell in an operational monitoring study) as multiple experimental units, and using experimental units that are not statistically independent (Heffner 1996).

To avoid spatial pseudoreplication it is important to intersperse experimental units. A basic example of interspersion is displayed in Figure 11; in this example an operational monitoring study is conducted to investigative infaunal response to abrasion pressure, using experimental units assigned to a ‘low’, ‘moderate’ and ‘high’ pressure categories, located in an area with a homogeneous substrate. Within Box A, all experimental units from each pressure category are grouped in similar geographical areas, and the geographical distance between experimental units within the same category is smaller than between different categories. This design is likely to be affected by pseudoreplication, as the assumption of spatial independence has been violated. Box B shows the same study designed to account for spatial autocorrelation, with low, moderate and high category experimental units interspersed. The results of statistical analyses conducted on data resulting from this study are likely to be more reliable than those from the previous design.

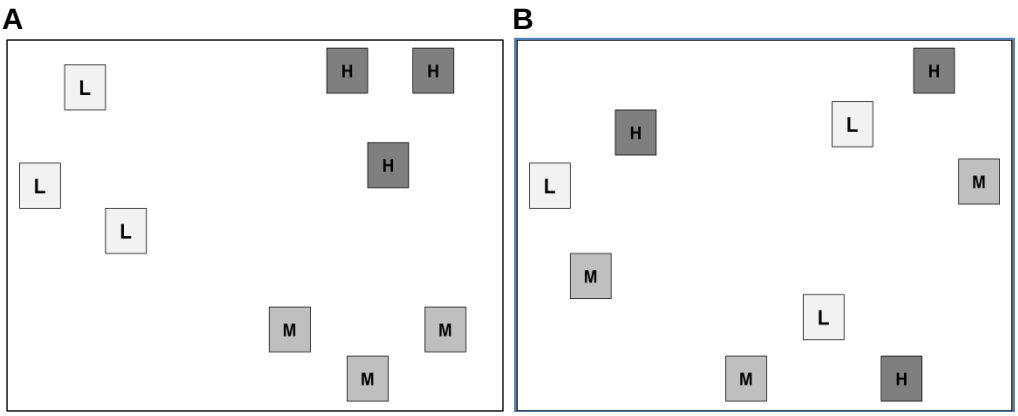


Figure 11. A) Grouping of experimental units (**L**ow, **M**edium, **H**igh) in the same areas may lead to spatial autocorrelation, B) Interspersion of experimental units controls for avoidance of similarities caused by spatial autocorrelation.

The concept of spatial pseudoreplication can also be applied to replicates within sampling units; if multiple replicates have been acquired from a single sampling point (as discussed in Section 7.2) they should not be treated as separate sampling units. To avoid dependency issues replicates should be pooled and a mean value calculated to provide a single value for each sampling unit. The same aggregation technique can be applied where several closely spaced sampling units are spatially autocorrelated.

Temporal pseudoreplication occurs when differences in sampling timing between experimental units mean that the design is confounded, or when repeated measures data are analysed inappropriately (Davies & Gray 2015). For example, in a BACI study, if 'control' and 'impact' data were acquired a month apart, where the 'before' and 'after' monitoring events were to be spaced by three years, it is possible that an unrelated environmental change or disturbance event could have occurred between collection of 'control' and 'impact' data. This would confound the assumption that environmental conditions within both the 'control' and 'impact' areas were the same. Statistical robustness will therefore be maximised where efforts are made to collect data from all experimental units during the same sampling period.

In practice, it can be extremely difficult to achieve a design which is entirely unaffected by pseudoreplication. For example, areas of different pressure intensity can be very spatially distinct, and/or associated with specific sediment types and benthic communities. The selection of experimental units can also be limited by benthic infrastructure, and it may be necessary to develop strategies using low confidence habitat and pressure maps. Logistical, operational and budgetary limitations may also prevent acquisition of all data within the same sampling period. Where pseudoreplication is unavoidable, it is possible to reduce or remove its influence by using the correct statistical models (e.g. mixed-effects models or generalized linear mixed models; see Millar & Anderson 2004).

8.4 Summary of key points and recommendations

Section 8: Considering dependency issues

Key Points:

- A key assumption of many statistical tests is independence of observations. Correlations within response variables in space (spatial autocorrelation) and/or time (serial correlation), are common in the marine environment, violating the assumption of independence.
- Spatial autocorrelation refers to the pattern in which observations from nearby locations are likely to have a similar value than expected due to chance alone. This can result in erroneous inference if not accounted for in analysis.
- Serial correlation, or temporal autocorrelation, is the correlation of a variable with itself across different points in a time-series. This type of dependency is particularly common within time-series monitoring data, and can result in the detection of a trend which does not exist.
- Pseudoreplication refers to a particular combination of experimental design (or sampling) and statistical analysis which is inappropriate for testing the hypothesis of interest.
- In practice, it may be difficult to attain a design which is entirely free of spatial dependency issues.

Recommendations:

- Spatial autocorrelation can be reduced by application of a minimum distance, or buffer, between sampling locations. This can be achieved using existing data from the MPA or a proxy area to produce semi-variograms, in addition to calculating Moran's I or Geary's C coefficients.
- Where the monitoring objective is to infer characteristics of a population (as opposed to measuring growth-rates or cover at specific locations) the potential for serial correlations can be minimised by re-randomising sampling locations. The strength of serial correlations should be minimised where possible, by establishing a suitable sampling event interval for indicator/s in question (e.g. slow-growing biogenic habitats will require a longer monitoring interval than dynamic habitats).
- If autocorrelation cannot be avoided the appropriate analyses, such as a generalized linear mixed model, should be used.
- Where the monitoring objective is to investigate change at fixed locations (or repeated observations on the same individuals), a repeated measures analysis should be used to account for the inevitable serial correlation.

9 Developing a sampling design

To recap the various aspects of monitoring design discussed in the previous sections, at this point the following questions should have been addressed;

- What are the monitoring objectives?
- Which type/s of monitoring will be used?
- Are new acoustic data and/or a habitat map required?
- Which existing data are suitable for inclusion in the monitoring time-series or as proxy data for power analysis / pilot investigations?
- Which indicator/s will be investigated?
- Which time of year is optimal for investigating these indicators?
- Which hypotheses will be tested, and which tests will be used?
- Which levels of power (1-β) and significance (α) are required?
- What effect size (ES) will the monitoring detect?
- What is the required sample size (N)?
- Which type and size of sampling unit will be used to sample the population?
- How should sampling units be arranged to avoid spatial autocorrelation?
- How long should the interval between sampling events be to avoid serial correlation?
- Is dependency unavoidable, and if so which statistical techniques can be used to account for this?

When these questions have been answered, the next step is development of a sampling design. This section provides general information on the advantages and limitations of commonly used sampling designs, whilst specific guidance on sentinel, operational & investigative monitoring designs is provided in Sections 9.2, 9.3 and 9.4.

9.1 Sampling designs

Sampling designs provide frameworks by which to draw sampling units from the population, and are either probabilistic or non-probabilistic. In a probabilistic design each sampling unit has the same theoretical probability of being selected, and therefore this type of design is generally considered more statistically robust. For example, within a specified sampling area, a 0.1m^2 grab sampler has the same probability of being placed on any 0.1m^2 area of seabed.

In a non-probabilistic (or judgement) design, some sampling units have no chance of being selected. The design is reliant on the subjective judgement of the researcher and therefore inference cannot automatically be made about the wider population (Albert *et al* 2010). Non-probabilistic sampling designs are generally considered less rigorous; however, situations may arise when they can be advantageous. For example, a researcher studying cockle density and size might decide to sample only where they had previously encountered cockles due to limited time or resources. In this situation the researcher can answer questions about cockle populations in specific areas, but should not assume that they are characteristic of wider populations.

The advantages and disadvantages of the most commonly used probabilistic sampling designs; simple random sampling, stratified random sampling and systematic sampling, are discussed in Sections 9.1.1 to 9.1.3, and are represented diagrammatically in Figure 12. Non-probabilistic (judgement) designs are discussed in Section 9.1.4.

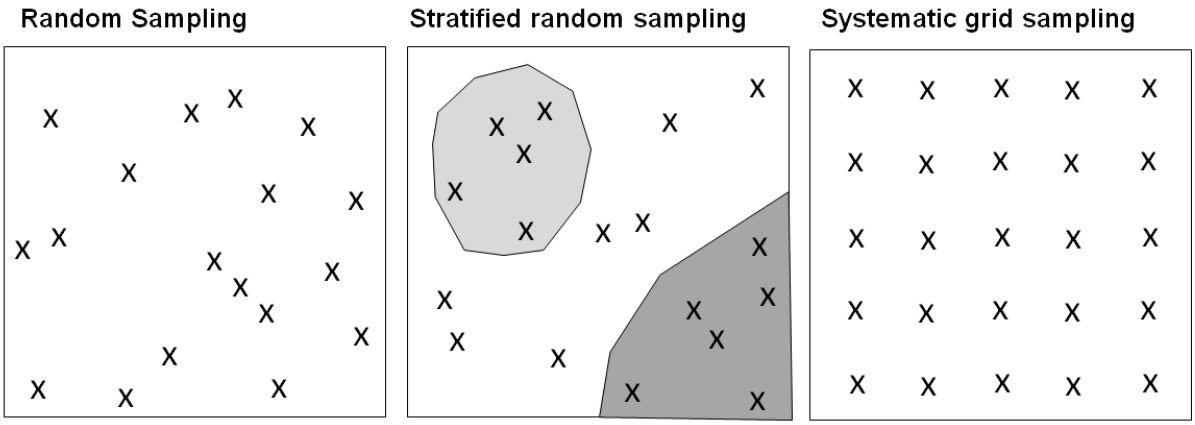


Figure 12. Examples of the three most common probabilistic sampling designs

9.1.1 Simple random sampling

Simple random sampling is the most basic form of probabilistic sampling, and is considered the most effective method for sampling populations where the substrate and environmental conditions are known to be reasonably homogeneous. The method results in an unbiased and non-subjective sample with no input from the researcher, however the resulting data may not be entirely representative of the population within the sampling area. This issue is particularly relevant when substrates are heterogeneous, or environmental conditions and pressures vary across the site. Under these conditions a random sample may not adequately sample the variance within the survey area, particularly when the site is large, and numbers of samples taken are relatively small. Another issue with this type of sampling is the potential for spatial autocorrelation between sampling units; in theory sampling units could be drawn adjacent to each other or in close proximity, violating the assumption of independence. To avoid spatial dependence a minimum distance, or buffer, can be applied between randomly selected sampling points, as previously described in Section 8.1. This approach will, however, not provide a truly random sample and the researcher will need to weigh the importance of this against the more complicated modelling required if sampling units are spatially autocorrelated.

The most straightforward method of producing a random sampling strategy is to generate random sampling points in a GIS package, within the boundary of a specified feature or layer. The same method may be used to generate pairs of coordinates which may be linked to create transects or trawl paths, although this may result in unfeasibly lengthy lines if the area in question is large.

9.1.2 Stratified random sampling

In comparison to the simple random sampling design, stratified random sampling can considerably increase accuracy and precision by ensuring that all the main habitat types, or otherwise-defined areas of different environmental character, are adequately represented in the sampling strategy (Brown 2000; Davies *et al* 2001). This approach is a deviation from simple random sampling, whereby the population is initially divided into distinct strata, so that sampling units within each stratum are more similar to one another than to those between strata. The aim of a stratified random design is to achieve a higher level of precision than that expected from a simple random design. A further advantage of this strategy is increased flexibility, particularly the potential for analysing data from each stratum separately, or aggregating them to a higher level.

Strata are likely to be defined with reference to seabed maps, including but not limited to substrate, topography and habitat maps. Seabed maps, particularly habitat maps, are often

created through extrapolation and modelling of limited data, and associated confidence levels vary greatly. Even where the entire area of interest has been mapped using acoustic data, it should be recognised that natural systems are dynamic, and that the distribution of strata may move and change over different temporal scales (i.e. shifting of sandwave features, biogenic reefs and sediment mosaics). When defining strata on which to base a stratified random design, the researcher should endeavour to use maps with high associated confidence, and acknowledge any limitations. It is recommended that a full acoustic seabed characterisation survey (e.g. multibeam bathymetry and backscatter) is conducted prior to sampling design, however as discussed in Section 2.1.3 this will not always be feasible.

If confidence in habitat maps is low researchers could use a system of field verification, such as selecting random grab sampling points following verification of the substrate by camera or video. This method may require oversampling to adequately sample each stratum (for example, sand and mud cannot always be distinguished from photographic data). An efficient alternative which will ensure uniform coverage of habitats within a sampling area is the systematic sampling design (see Section 9.1.3).

9.1.3 Systematic sampling

Systematic sampling involves placement of sampling points at regular intervals, usually in a grid pattern, and preferably starting from a randomly generated point to remove subjectivity from the design. Systematic sampling is not reliant on high confidence habitat maps, and provides more uniform coverage of a survey area than simple random sampling. The systematic grid design can be used to increase the probability that samples represent the whole sampling area when it cannot be reliably stratified, or where confidence in maps is low. It provides an efficient means of mapping distribution, and is the most effective design when an estimation of spatial pattern or extent is required (Davies *et al* 2001). The use of a systematic sampling design will reduce the probability of spatial autocorrelation by maintaining a uniform distance between sampling points (Olea 1984), in addition to providing the option of stratifying and sub-sampling at a later date if strata are defined subsequent to the sampling period. Systematic sampling grids can also be applied to different strata within the same survey area, resulting in a systematic stratified design. This approach is particularly useful when substrates are known to be highly heterogeneous, and the standard systematic grid is likely to result in an unbalanced design (i.e. insufficient coverage of all habitats). An additional advantage of systematic sampling is that ‘before’ data can be acquired within a survey area (e.g. within an MPA) if management measures are to be implemented but the location of the management areas is not yet known with confidence.

The sampling interval in a systematic grid will generally be determined by the number of sampling points required, however care should be taken to ensure that the interval is not correlated with a periodic seabed feature, e.g. peaks or troughs of sandwaves. If the sampling points are not correlated with a regular feature, and if the samples are sufficiently far apart to be independent, then a systematic sample may effectively be treated as a simple random sample in analysis (Manly & Navarro 2015). To support this assertion, Cabral & Murta (2004) reported that random, stratified random, and systematic sampling designs resulted in similar mean variance ratios in the density of benthic infauna.

Triangular grid patterns are typically preferable to square grids, as this reduces the chance of bias towards a regularly spaced feature (Byrnes 2000). The pattern used to space the systematic sampling points can also affect the ability of the sample to detect certain seabed features; for instance, Barry and Nicholson (1993) determined that a triangular grid was the most efficient pattern for detection of circular patches, in comparison to square grids or random sampling.

Ephemeral habitats (e.g. *Sabellaria spinulosa*) can be difficult to monitor, as the habitat distribution can change throughout time. Seabed maps should be used with caution where habitats are ephemeral (Limpenny *et al* 2010), and should not be used for stratification unless they are recent enough that confidence in distribution is high. Such habitats are likely to be best monitored using a systematic sampling design (e.g. using video tows or drop camera), where effort is evenly distributed throughout the sampling area. If time-series data are available, areas of ephemeral habitat which have been shown to persist through time may also be targeted.

9.1.4 Judgement sampling

Judgement sampling is a non-probabilistic method of drawing a sample from a population, and involves the researcher subjectively selecting the sampling units without any form of randomisation. The risk of bias associated with this method is high, although it can offer an efficient alternative to probability sampling, particularly when the populations in question are well studied and resources are limited, where rare species or habitats are known to occur, or in areas considered to be representative of a certain condition (Davies *et al* 2001). Judgement sampling also provides a method of targeting multiple gradients simultaneously when the sample size is less than optimal due to resource limitations (e.g. targeting a range of sediment types and seabed depths). When sample sizes are small this approach can reduce the probability of recording a 'truncated gradient', whereby the full ranges of environmental and biological gradients are not captured (Albert *et al* 2010).

Results from data acquired using a judgement sampling design should not generally be extrapolated to the entire population, unless the researcher is highly confident that the wider population shares the same characteristics and presents the findings with caveats. Whilst judgement sampling data may be used descriptively to identify broad trends and ranges, hypothesis testing and inference of causality is not appropriate where empirical evidence is required to justify management measures (Steele 2001).

Gradient-directed transect or 'gradsect' is a low-input, high-return judgement method, which is targeted to investigate indicator response along a specified environmental gradient. If well-designed, with adequate knowledge of the system and indicator/s in question, the gradsect strategy can improve precision and efficiency, by capturing data within the full variation range of the specified parameter (Wessels *et al* 1998). The theory behind the gradsect method is that distribution of biota is generally non-random, and therefore sampling designs which employ random or systematic models may fail to detect underlying non-random patterns (Gillison 1984). The gradsect method can be applied to investigate gradients where a random or systematic approach is unlikely to capture the full range of environmental variation (e.g. if the full variance range was restricted to a certain area within a large MPA), or if the number of stations required would be too high using more traditional sampling methods. Although the gradsect method can improve efficiency and precision in some cases it should be acknowledged that natural variation in communities is unlikely to be fully captured, due to the smaller number of stations sampled.

9.1.5 Choosing fixed or re-randomised locations

Fixed monitoring locations (typically comprising plots, smaller quadrats or transects) can provide a very precise measure of change by reducing random variability in parameters such as substrate composition, and physico-chemical conditions. This strategy is typically used for monitoring the growth, density, cover or condition of biota such as biogenic reefs or solitary corals, marine flora and sessile fauna, or organisms that are only known from specific locations (Davies *et al* 2001). For such biotic parameters re-sampling of the same areas of habitat is likely to be a more effective method of monitoring change than random allocation of sampling across larger spatial strata (Van der Meer 1997; Kingsford & Battershill 1998;

Hill & Wilkinson 2004). Fixed locations are also appropriate to measure responses in localised areas of persistent anthropogenic impact.

Fixed monitoring locations are generally most feasible in the intertidal and nearshore environments where locations may be marked and revisited easily, whilst it can be extremely difficult to revisit exact fixed locations in deeper areas.

Despite the advantages of using fixed locations they may be unrepresentative of the survey area as a whole, and only allow inference to be made about discrete locations. Repeated monitoring can cause localised damage and there may be financial overheads associated with marking and maintenance. Care must also be taken to ensure that the act of sampling does not confound the experiment, e.g. trampling or disturbance by surveyors can make it impossible to detect whether there has been a true change. Repeated observations from fixed locations are also highly likely to display serial correlation, confounding the assumption of independence through time and requiring repeated measures analyses.

As reported by Davies *et al* (2001), repeated monitoring should only be conducted at fixed locations when:

- minimising sampling variation is of prime importance (e.g., where subtle changes must be detected at sites which are highly heterogeneous) or information is needed on turnover and species dynamics,
- sample locations are representative of the site and sufficient samples are taken to minimise the risk of chance events reducing their representativeness,
- provision is made for the unexpected loss of sample locations,
- the feature and the surrounding environment will not be significantly altered or damaged by repeat visit.

9.1.6 Summary of key points and recommendations

Section 9.1: Sampling designs

Key Points:

- Sampling designs provide a framework by which to select sampling units from the population, and may be either probabilistic (random) or non-probabilistic (non-random).
- Probabilistic sampling designs typically minimise systematic error and are considered to be more statistically rigorous. They include simple random sampling, stratified random sampling, and systematic sampling.
- Non-probabilistic (judgement sampling) designs involve the researcher subjectively selecting the sampling units without any form of randomisation. They should not be used for inference about a wider population.

Recommendations:

- It is recommended that a full acoustic seabed characterisation survey (multibeam bathymetry and backscatter) is conducted prior to sampling design, if resources allow.
- It is recommended that probabilistic sampling designs are used, so that inference can be drawn about the wider population.
- Simple random sampling (with a buffer) should be conducted where sediments are homogeneous, and pressures are reasonably consistent across the site.
- Stratified random sampling should be conducted where sediments or pressures are clearly stratified across the site, and confidence in habitat maps is moderately high.
- Systematic sampling should be conducted where the seabed cannot be reliably stratified, and where full coverage of the survey area is required. It should also be used when management areas or closures are likely to be established after the sampling has been completed, to ensure that these areas are covered to some extent.
- For systematic sampling a triangular grid pattern should be used in preference to a square grid, to reduce the probability of bias towards regularly spaced features. The grid should start from a randomly generated point to remove subjectivity from the design.
- Judgement sampling should only be used when the researcher has a well-developed knowledge of the indicator/s and system in question, and where resources do not allow a probabilistic design. Judgement sampling is not suitable where empirical evidence is required to justify management measures.
- Fixed sampling locations should generally only be used for monitoring the growth, density or cover of biota, such as biogenic reefs, marine flora, and sessile fauna, or those that are only known from specific locations.

9.2 Sentinel monitoring sampling designs

The sampling strategy employed for sentinel monitoring is likely to depend on the level of confidence in the distribution of habitats.

In areas where substrate and depth range are known to be reasonably homogeneous, a **simple random sampling** design may be appropriate. However, if confidence in habitat maps is low, or where habitats are heterogeneous, the simple random design may not provide sufficient coverage of all habitats. When employing a simple random design for sentinel monitoring, the design should incorporate adequate spatial coverage of the entire area of interest to capture the full range of environmental and biological variation throughout both time and space. If the feasible sample size is not sufficient for adequate geographical coverage of the survey area or the habitats within it when distributed randomly, a **systematic sampling** approach should be considered. A systematic approach may also be preferable when substrates are likely or known to be highly heterogeneous, precluding determination of distinct strata, and when confidence in seabed maps is low (e.g. Figure 13). In cases where future monitoring is likely to be required to determine whether management measures have been successful, sentinel monitoring data can serve as 'before' data in an investigative Before-After-Control-Impact (BACI) design (assuming it meets power and significance requirements); this is generally the case where management areas have not been determined prior to the initial monitoring event. In this instance a systematic design covering the entire site is likely to be the most appropriate, to ensure coverage of future management areas. The potential for such data to be used in quantitative analysis will, however, depend on the levels of natural variation and the number of sampling points which have fallen within the undefined management areas.

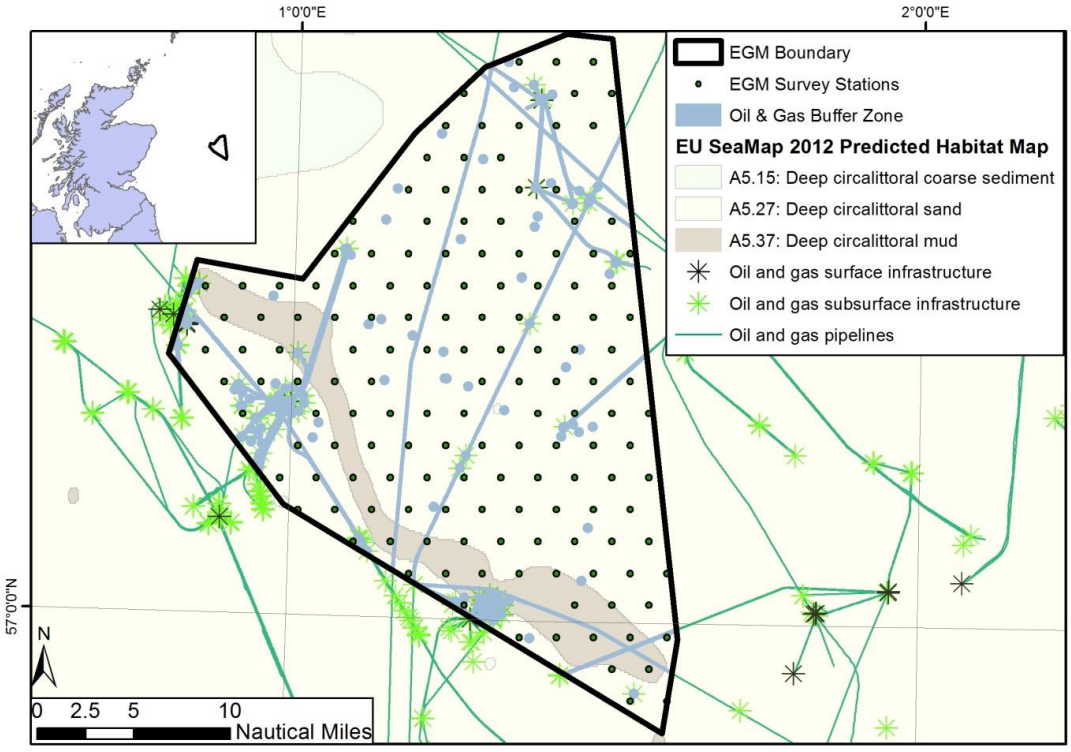


Figure 13: A systematic sampling design for the initial sentinel monitoring survey at East of Gannet and Montrose Fields Nature Conservation MPA (NCMPA).

Stratified random sampling is considered the optimum strategy for sentinel monitoring where confidence in seabed maps is high, and strata are distinct (e.g. circalittoral rock in Figure 14). Any stratification should be based on major ecosystem drivers of variance (e.g.

depth, biogeographic region, water currents, sediment type) and not exclusively on human pressures, whose spatial scale in the long-term is likely to change.

A stratified random sampling design is recommended if high confidence habitat maps have already been produced, or if remote sensing data for the entire survey area can be acquired and processed in the field within the constraints of the budget (assuming strata can be resolved).

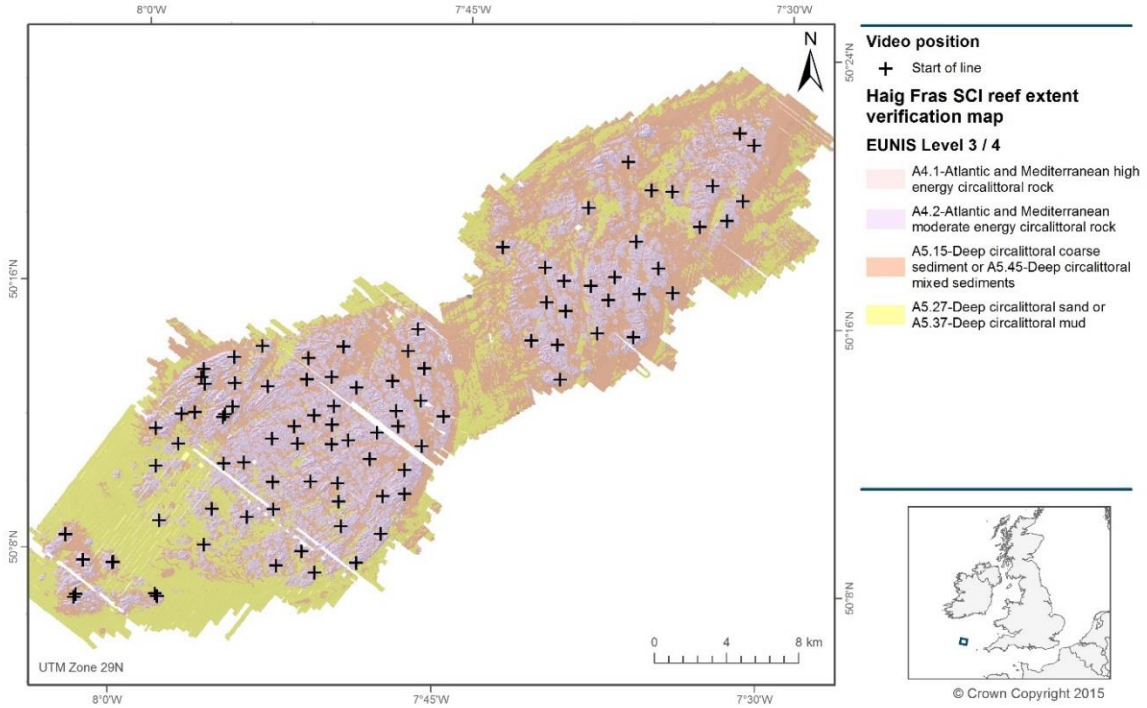


Figure 14: A stratified random sampling design for the initial sentinel monitoring survey at Haig Fras Special Area of Conservation (SAC) (one stratum; moderate energy circalittoral rock)

The flowchart presented in Figure 15 provides a guide to aid selection of an appropriate design for sentinel monitoring.

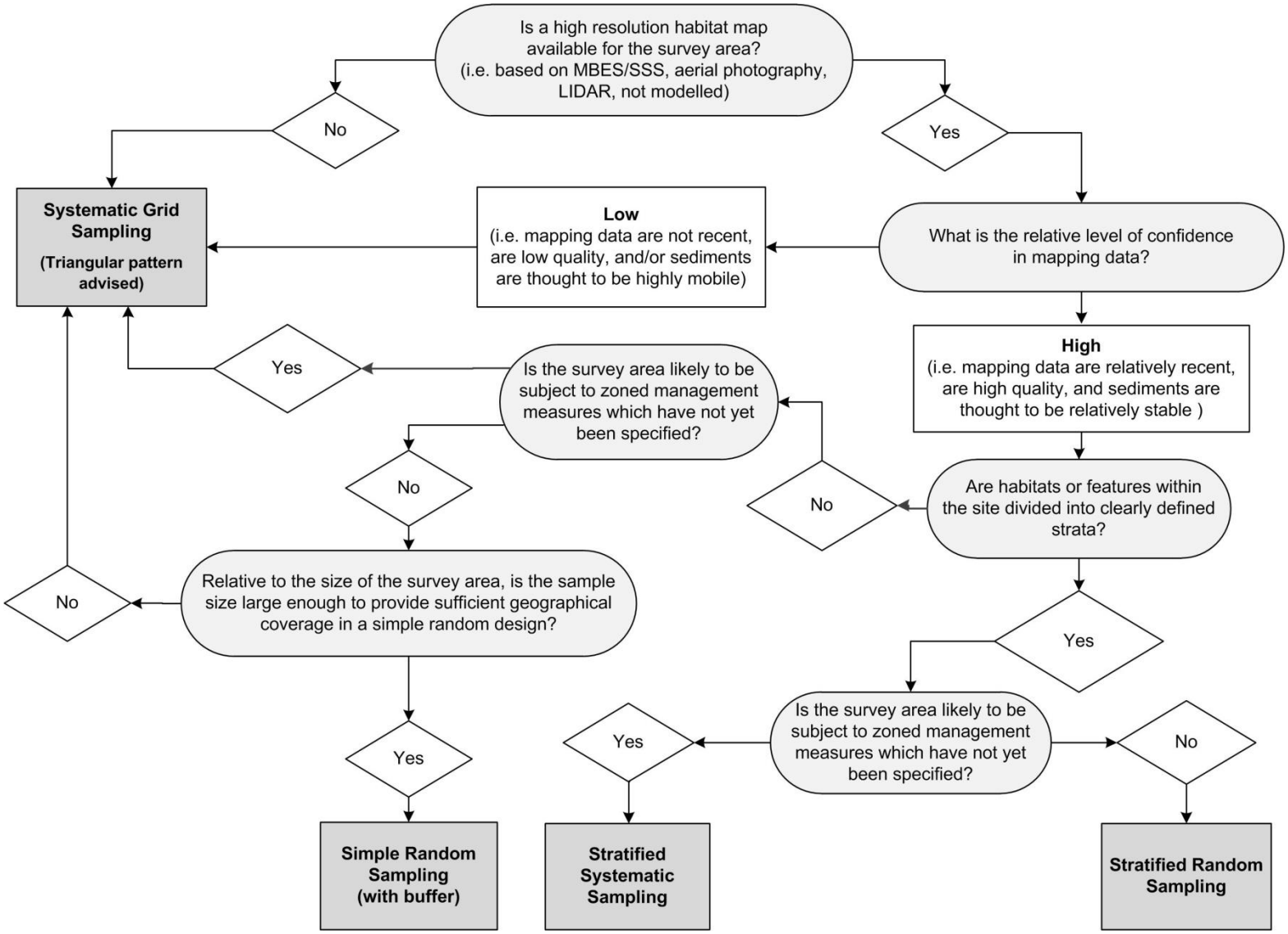


Figure 15. A systematic approach to determining appropriate sampling designs for sentinel monitoring.

9.2.1 Summary of key points and recommendations

Section 9.2: Sentinel monitoring sampling designs

Key Points:

- The sampling strategy employed for sentinel monitoring within and around MPAs is likely to depend on the availability of acoustic seabed data for the site, and the level of confidence in the mapped distribution of habitats.

Recommendations:

- Sentinel monitoring will use simple designs, and are therefore covered by the Section 9.1 recommendations.
- The flow diagram presented in Figure 15 can be used to aid selection of an appropriate sentinel monitoring sampling design.

9.3 Operational monitoring sampling designs

Operational monitoring designs are more complex than those explored in the previous section. They require knowledge of pressure intensity and distribution, and are optimised where there is a reasonably high level of confidence in habitat distribution (as discussed in Section 2.1.3).

Sampling designs for operational monitoring will vary depending on the nature of the pressure under investigation (e.g. dispersive or non-dispersive), the resolution to which the pressure may be mapped, and the confidence with which pressures can be modelled.

A robust design to investigate pressure-state relationships will usually consist of a number of discrete 'pressure units', defined as standardised areas within which the intensity of a specific pressure is known and may be categorised (e.g. a gridded VMS cell, see Figure 16). Ideally pressure units will be replicated within each category of the pressure gradient (e.g. low, medium, high, very high) to increase statistical power. Replicates should be taken within each pressure unit, the number of which should be determined by power analysis once the number of strata (i.e. the pressure categories) and pressure unit replicates within each pressure category have been determined. Replication within pressure units should be conducted using **simple random** or **systematic** designs where substrates are homogeneous. Where the substrate varies within pressure units it may be appropriate to use a **stratified random** approach, however efforts should be made to reduce such variation wherever possible (e.g. by selecting pressure units in areas of homogeneous sediment).

The sampling strategy presented in Figure 16 was designed to investigate the pressure-state relationship between subsurface abrasion and infaunal metrics at the Dogger Bank candidate Special Area of Conservation/Site of Community Importance (cSAC/SCI). Pressure units (0.05 decimal degree abrasion cells) were defined using a standardised method of VMS data aggregation developed by Church *et al* (2016). Based on the resources available, and prior experience of similar studies on comparable habitats, it was decided that the gradient would consist of four pressure categories (and a zero-pressure category). Two replicate pressure units ('a' and 'b' cells) were selected within each pressure category, resulting in ten pressure units overall, each of which contained ten replicate sampling stations.

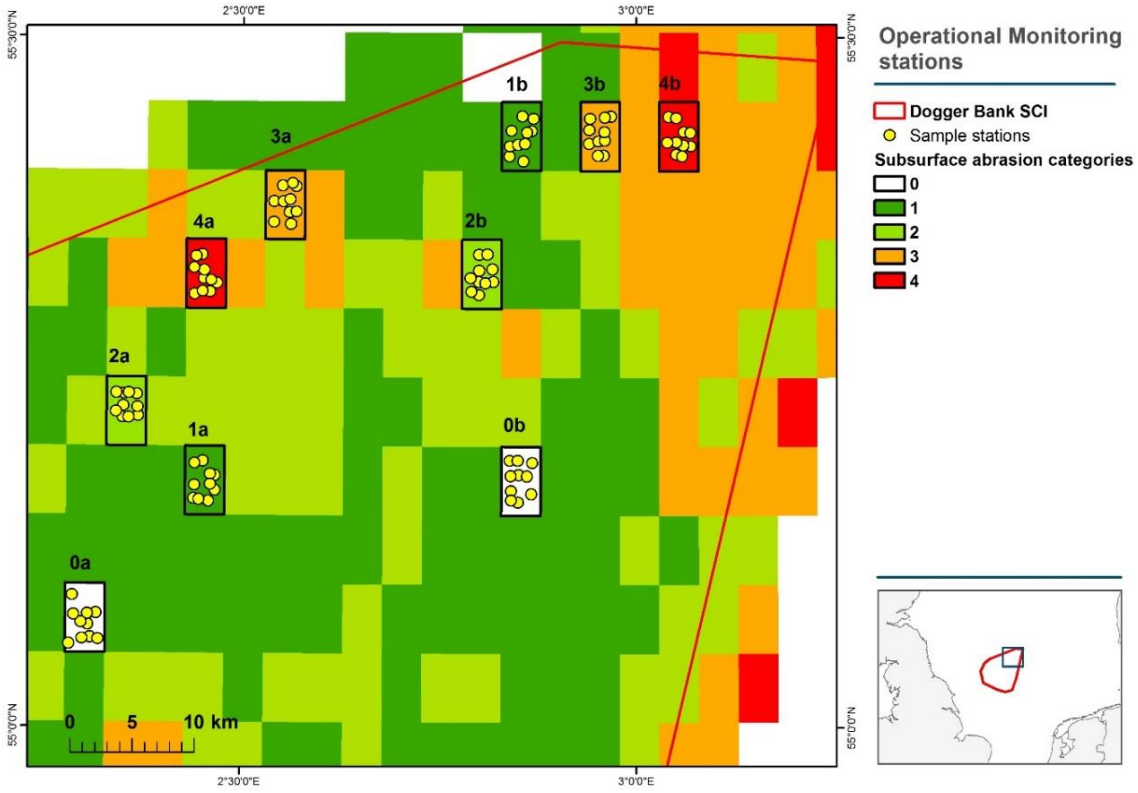


Figure 16: An operational monitoring design for investigating the relationship between infaunal communities and subsurface abrasion pressure (simple random sampling within replicated pressure units (grid cells) along a gradient).

9.3.1 General principles for operational monitoring

Although sampling designs for operational monitoring will vary according to the pressure/s under investigation, some general principles should be universally applied to all designs:

1) Pressure units should be appropriate for the pressure and should be of an ecologically meaningful size.

The choice of pressure unit will depend on the resolution to which a pressure can be mapped, or the confidence with which its distribution can be predicted. It is important that the pressure unit is ecologically meaningful, and neither too large nor too small for the impacts of the pressure to be detected.

For example, the design presented in Figure 16 uses a standardised method to grid VMS pings to a cell format. Assigning a pressure value to a VMS cell requires interpolation, therefore there is a risk of over or under-estimating fishing pressure spatially, depending on the scale of the grid selected. When using VMS data and other types of data requiring interpolation it is important that the scale of the unit is considered in the context of the survey area and habitats present. For more information, see Jenkins *et al* (2015).

Selecting pressure units within which to replicate can be difficult for dispersive pressures, where the distribution and intensity of the pressure is not generally known. In this case, expert judgement must be used, and the likely direction and range of the dispersal must be gauged using hydrodynamic information and modelled products where necessary. Based on the information available, a consistent and ecologically meaningful pressure unit should be selected, which will be comparable along the entire gradient.

An example theoretical design for a point source of contamination is displayed in Figure 17. The design incorporates regularly spaced pressure units which extend out from the contamination source in the direction of the likely prevailing current, beyond a distance thought to be the likely limit of contaminant effects. The size of these units must be decided by the researcher based on the likely range of dispersion and the parameters being investigated, however the size should be standardised for comparability.

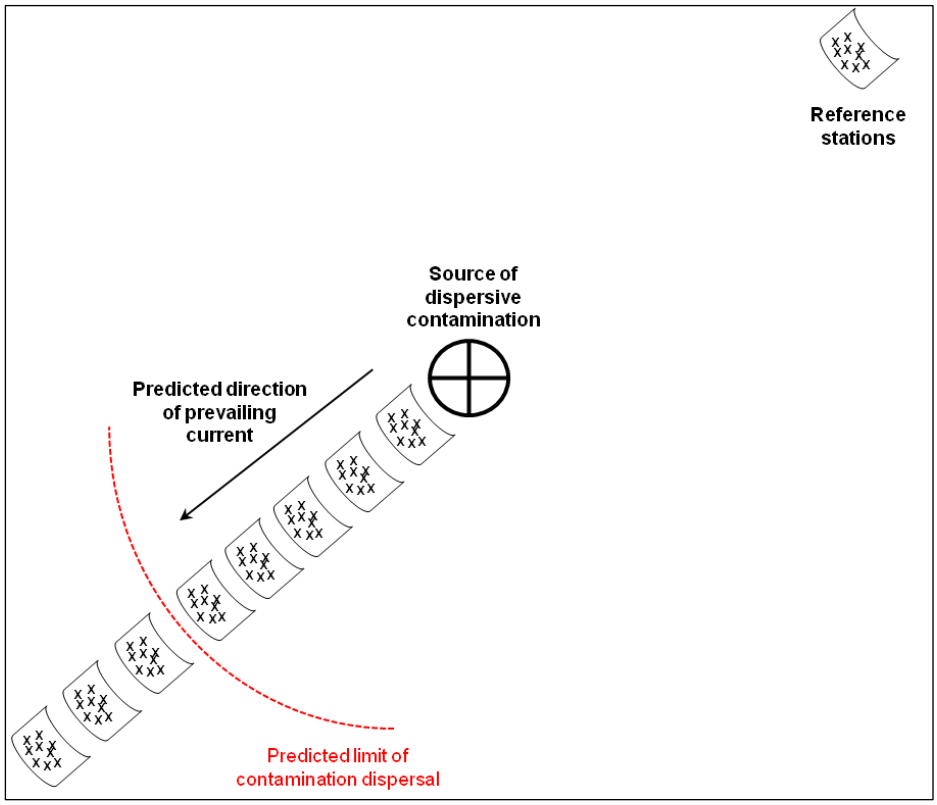


Figure 17. A theoretical operational monitoring design for a point source of dispersive contamination (gradsect design).

2) The distribution of pressures within pressure units should be considered

Where pressure units have been derived using interpolated data or mapping products it may be useful to evaluate the likely distribution of pressure within the unit using expert judgement and supplementary information (if available). For example, the gridded format of VMS cells may imply that pressure is uniform throughout the cell, however this is not generally the case. Review of VMS pings overlain onto the gridded cell, along with qualitative evidence (e.g. information gathered through interviewing fishers), may allow patterns to be identified which may not be evident from interpolated mapping products.

If a pressure appears to be highly skewed towards one area of a pressure unit it may be necessary to exclude it from the design to avoid biasing the dataset, or to sub-divide the cell and sample only within the high-pressure area.

3) Where possible the design should be balanced and should sample the entire pressure gradient at appropriate intervals

Truncated gradients occur when a sampling design fails to record indicator response to the full range of pressure intensity (Thullier *et al* 2004). It is important that the full range (or likely range) of the pressure gradient is identified, and that pressure units are allocated to sample

the gradient along its entire range (Underwood & Chapman 2013). Pressure units outside of the gradient (i.e. non-impacted) should also be sampled to provide controls for comparison.

The number of categories in the gradient will generally depend on the resources available, and the number of pressure units and within-unit replicates required for a robust design. A larger number of categories may allow the pressure-state relationship to be modelled more accurately, however in order to maintain a balanced design there must be a trade-off between the number of categories and the number of replicates within each pressure unit. Where a balanced design is not achievable (e.g. if some samples are found to consist of a different habitat or substrate type which is likely to introduce excessive variance), it may be possible to weight the data in analysis.

Where spatial data are available, various methods can be used to systematically classify data into categories (e.g. user defined, equal intervals, quantile and natural breaks). The optimal method will depend on the distribution of the data; for example, the quantile method places equal numbers of observations into each category, and is best used for data which are evenly distributed across the range, whilst the natural breaks method uses natural groupings to maximise between-category differences, and is best used for data that are unevenly distributed across the range. Whichever classification method is used, it is important that the design is balanced, with an equal number of pressure units assigned to each category of the pressure gradient.

4) Temporal pressure datasets should be combined (or not), based on the resilience and resistance of indicator/s to the pressure

Disturbance caused by anthropogenic pressures may be temporary (e.g. a single contamination incident) or persistent (e.g. sustained trawling over a number of years). Where persistent disturbance is present, and a pressure data time-series is available (e.g. cumulative annual VMS data or aggregate extraction data) a decision must be made about how or whether to combine this data in an ecologically meaningful way. It is important that the design (and subsequent analysis) is based on pressure data which is most likely to reflect the true response of the indicator. The decision on whether to combine datasets should be based on the likelihood of cumulative disturbance effects, and a review of the resistance (tolerance) and resilience (recoverability) of the indicator/s to the pressure/s under investigation.

Where habitats are subject to a high level of natural disturbance (e.g. high energy systems such as sandbanks) biota may be naturally resistant and/or resilient to anthropogenic disturbance, and as such may only show a response to recent human disturbance events. In such cases it may be appropriate to use only the most recent dataset or to combine a small number of datasets (e.g. combining data from two or three annual datasets) to reflect a short-term cumulative effect. Where habitats are subject to a low level of natural disturbance or are slow-growing, they are display lower resistance and resilience to anthropogenic pressures, and may therefore be impacted for a substantial period after the impact. If so, it is appropriate to combine a higher number of datasets to cover the expected longer period of impact.

The Marine Evidence based Sensitivity Assessment (MarESA), available on the Marine Life Information Network (MarLIN)[6], is a key resource for information on the sensitivity of different species and habitats in British Isles waters, and can be used to help determine the likely rate of recovery from pressures.

[6] http://www.marlin.ac.uk/species/sensitivity_rationale

5) Variance should be minimised in the design wherever possible

It is important that the response of the indicator to the pressure is clearly identifiable against background variation. Therefore, all environmental and anthropogenic factors likely to cause such variance must be addressed in the design. This can be achieved by ensuring uniformity of conditions across sampling units wherever possible; e.g. substrates and depth ranges should be comparable between pressure units. The pressure units should also not be distributed across a large geographical area unless necessary. Furthermore, sampling should not occur where other anthropogenic pressures are present which are likely to confound the results, unless they can be measured and accounted for in analysis.

Where it is not possible to limit sources of variance within the design (e.g. the number of pressure units in different pressure categories is locally limited within the habitat/s under investigation), the variables likely to introduce variance should be quantified and used as covariates in analysis.

6) Pressure units should be spatially independent where possible

As discussed in Section 8.1, spatial dependency of pressure units (i.e. spatial autocorrelation) can occur when observations from nearby units have values more similar than those from units that are further apart. In an operational monitoring study this situation is likely to occur when areas of similar pressure intensity are geographically distinct within the survey area. Where this pressure distribution occurs, inferential capability is reduced, and it may not be possible to conclude that any relationship or lack thereof is the result of a pressure-state interaction, as opposed to influenced by natural spatial variation.

Wherever possible, spatial independence should be optimised by:

- interspersing pressure units from different pressure categories (as illustrated in Figure 16),
- maximising distance between pressure units within reason (e.g. try not to locate units directly adjacent to each other),
- ensuring that replicates within pressure units are closer to each other than they are to replicates in a different unit (using a buffer if necessary).

As mentioned in Section 8.3, if spatial independence cannot be attained statistical analyses which explicitly model the spatial dependence should be used (e.g. mixed effects or generalized linear mixed models).

9.3.2 Summary of key points and recommendations

Section 9.3: Operational monitoring sampling designs

Key Points:

- Operational monitoring designs are more complex than those explored in the previous section, and will be optimised where confidence in the distribution of pressures is reasonably high.
- Sampling designs for operational monitoring will vary depending on the nature of the pressure under investigation (i.e. dispersive or non-dispersive), the resolution to which the pressure can be mapped, or the degree of confidence in pressures modelling.

Recommendations:

- A robust design to investigate pressure-state relationships will usually consist of a number of sampling units within 'pressure units' (e.g. a VMS cell or other standardised area of pressure) which are classified into pressure categories to cover the entire gradient of the specified pressure.
- Replicate samples should be taken within each pressure unit, and each category of the pressure gradient should be replicated to increase power.
- It is recommended that the following principles are taken into account for operational sampling designs (see explanations in Section 9.3.1):
    1. Pressure units should be appropriate for the pressure, and should be of an ecologically meaningful size.
    2. The distribution of pressures within pressure units should be considered.
    3. Where possible the design should be balanced, and should sample the entire pressure gradient at appropriate intervals.
    4. Temporal pressure datasets should be aggregated (or not), based on the resilience and resistance of the indicator/s to the pressure.
    5. Variance should be minimised in the design wherever possible.
    6. Pressure units should be spatially independent wherever possible.

9.4 Investigative monitoring sampling designs

Investigative monitoring involves designing an experiment to find evidence of cause and effect within a given area or areas (i.e. the effect of an 'impact' on an indicator). In this context, 'impact' refers to a change to the status quo, either experimentally (e.g. creating disturbance by trawling the seabed) or through management measures (i.e. the removal or addition of a pressure to a specific area), in contrast to 'control' conditions where the status quo is maintained.

In an experiment, a factor is an explanatory variable which has two or more levels. Two factors are typically used in investigative monitoring designs; the manipulation (e.g. exclusion of a pressure by management or experimental disturbance), and sampling period (e.g. before and after the manipulation). Investigative designs use combinations of these factors, or treatments, to test for differences between groups of samples and determine whether the manipulation has resulted in a change to the selected indicator/s (Table 7).

Table 7: A 2x2 factorial design for an experimental manipulation.

		Management measures / experimental disturbance	
		Control	**Impact**
Sampling Period	**Before**	Treatment 1	Treatment 2
	After	Treatment 3	Treatment 4

According to Green (1979), an optimal experiment of this kind has several basic features:

- the type of impact, time of impact, and place of occurrence should be known in advance,
- the impact should not have occurred yet,
- control areas should be available.

Where the features listed above are not present the design will be very limited in its ability to infer the cause of change; for example, a single factor Control-Impact study may easily be confounded by temporal variation (see Section 9.4.1), whilst a double factor design with multiple replicates in time and space ('Beyond BACI', see Section 9.4.4) is substantially more robust. As described in Section 6, available resources must be balanced against the need to provide statistically robust evidence for the effectiveness of management measures, particularly where measures are adaptive or have substantial impacts on stakeholders. Sections 9.4.1 to 9.4.4 describe the limitations and advantages of designs which are commonly used for monitoring the effectiveness of management measures or detecting change in a manipulative experiment. The designs are compared graphically in Figure 19.

9.4.1 Control-Impact and Before-After designs (CI & BA)

The basic two treatment Control-Impact (CI) design has been widely used in MPA monitoring (as summarised in Halpern 2003 and Osenberg *et al* 2006), yet this design has severe limitations for dealing with natural variability. In a CI design the 'impact' site refers to the area where management measures were implemented or an experimental impact was applied (e.g. the exclusion or introduction of a pressure), whilst the 'control' site consists of a comparable unmanipulated area. The control and impact sites are assumed to be identical in the absence of the experimental manipulation, and under this assumption the difference between the control and impact sites provides an estimate of the impact effect (Osenberg *et al* 2011). In reality spatial variation between the control and impact sites is likely to render this assumption invalid (Osenberg *et al* 2006).

A Before-After (BA) design compares conditions within the impact site prior to and following the manipulation without use of a control. This design assumes that differences between survey periods are caused by the manipulation; an assumption that is equally spurious to that of the CI design, due to the likelihood of change over time regardless of an impact. Confidence in the results of a CI or BA experiment will be low, and these designs will not produce robust evidence for justification of management measures. It is highly recommended that a more complex design is used for investigative monitoring surveys.

9.4.2 Before-After-Control-Impact designs (BACI)

In the 2x2 factorial Before-After-Control-Impact (BACI) design (Green 1979) the control and impact sites are sampled once before and once after the manipulation, allowing influence of background spatial and temporal variance to be accounted for. This design allows the

researcher to test for an interaction between time (Before/After) and manipulation (Control/Impact) to determine whether a manipulation has resulted in an effect.

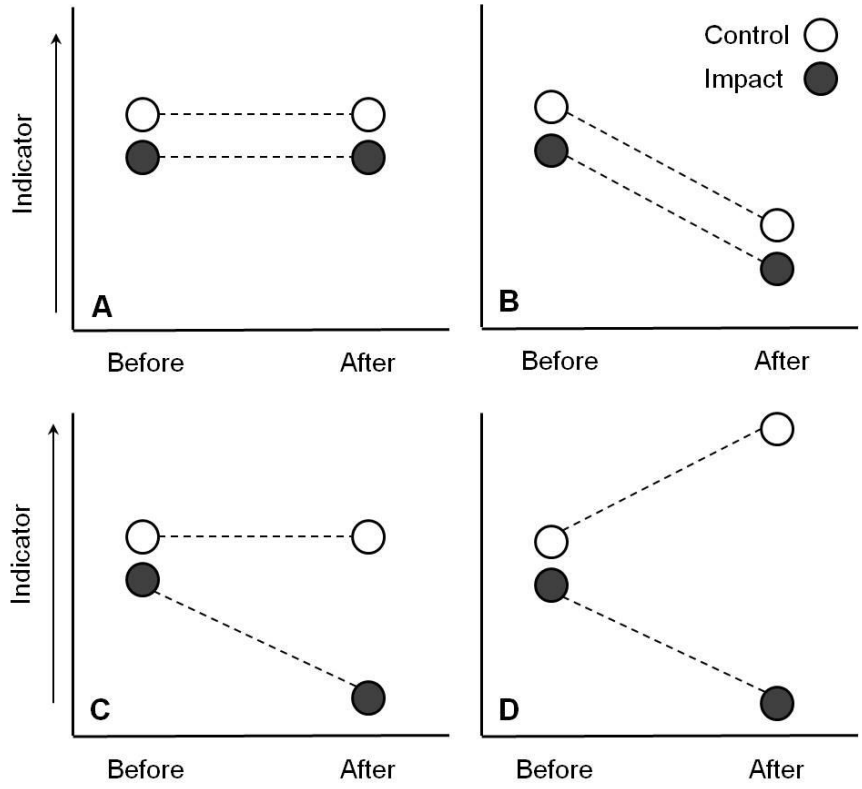


Figure 18: Interaction between control and impact sites in a BACI design.

Figure 18 illustrates the concept of the interaction, with four possible outcomes of a BACI study; A) no change in either the control or impact sites, B) a decrease in both the control and impact sites, it is unclear whether the decrease is due to the same cause, C) a decrease in the impact site with none observed in the control site, which could be interpreted as a manipulation effect, D) an increase in the control site and a decrease in the impact site, indicating that the control site is unsuitable for comparison.

Whilst the BACI design is more robust than Control-Impact or Before-After designs, it is nevertheless generally considered to be flawed (Hurlbert 1984; Bernstein & Zalinkski 1983; Stewart-Oaten *et al* 1986; Underwood 1990, 1992). Despite efforts to select sites with similar physical and ecological characteristics there may be spatial and temporal differences between the impact and control sites which are unrelated to the manipulation (as illustrated in Figure 18), and BACI results should therefore be interpreted with a degree of caution.

Whilst the BACI design is preferable to CI and BA designs, it is recommended that a more complex design is used, involving the addition of more time-series data points (Section 9.4.3), and/or extra control sites (Section 9.4.4).

9.4.3 Before-After-Control-Impact Paired Series designs (BACIPS)

The basic BACI design can be modified to reduce the likelihood of the experiment being confounded. BACI Paired Series (BACIPS) designs involve repeated sampling of the control and impact sites at the same times (or as close together as is feasible), so that shared temporal effects can be identified (Stewart-Oaten *et al* 1986).

BACIPS designs provide more powerful estimates of impact or manipulation effects by accounting for extraneous sources of noise which limit other designs. They require a sustained commitment to monitoring effort and resources, and careful planning. Ideally, multiple time-series measurements would be taken prior to the manipulation being implemented. In practice, it is unlikely to be feasible to take multiple 'before' measurements as part of a monitoring programme, although efforts should be made to include any suitable existing data in BACIPS designs.

9.4.4 Beyond BACI designs

The BACIPS design can be further developed to include multiple control and/or impact sites, sampled at multiple times before and after the impact or manipulation (MBACI) (Keough & Mapstone 1995). Although it is statistically desirable to investigate equal numbers of impact and control locations (a symmetrical design) it is expected that impact sites will be limited in many cases (e.g. management is restricted to a single area in a local context), and therefore an asymmetrical design should be employed, weighted in favour of multiple control sites (Underwood 1990, 1992).

The 'Beyond BACI' design advocated by Underwood (1992) suggests that designs should include a series of spatially independent control sites which have been randomly selected from a set of possible sites with similar characteristics to the impact site. If the manipulation results in authentic changes, the difference between the impact and control sites would be expected to be greater than the differences between control sites; it is therefore clear that as many control sites as feasible should be surveyed to improve the precision of variability estimates. Underwood states that sampling should be conducted at all sites at the same time (or as close as possible); however, the sampling times should be selected randomly within the confines of a specific ecologically justified time period (i.e. a season).

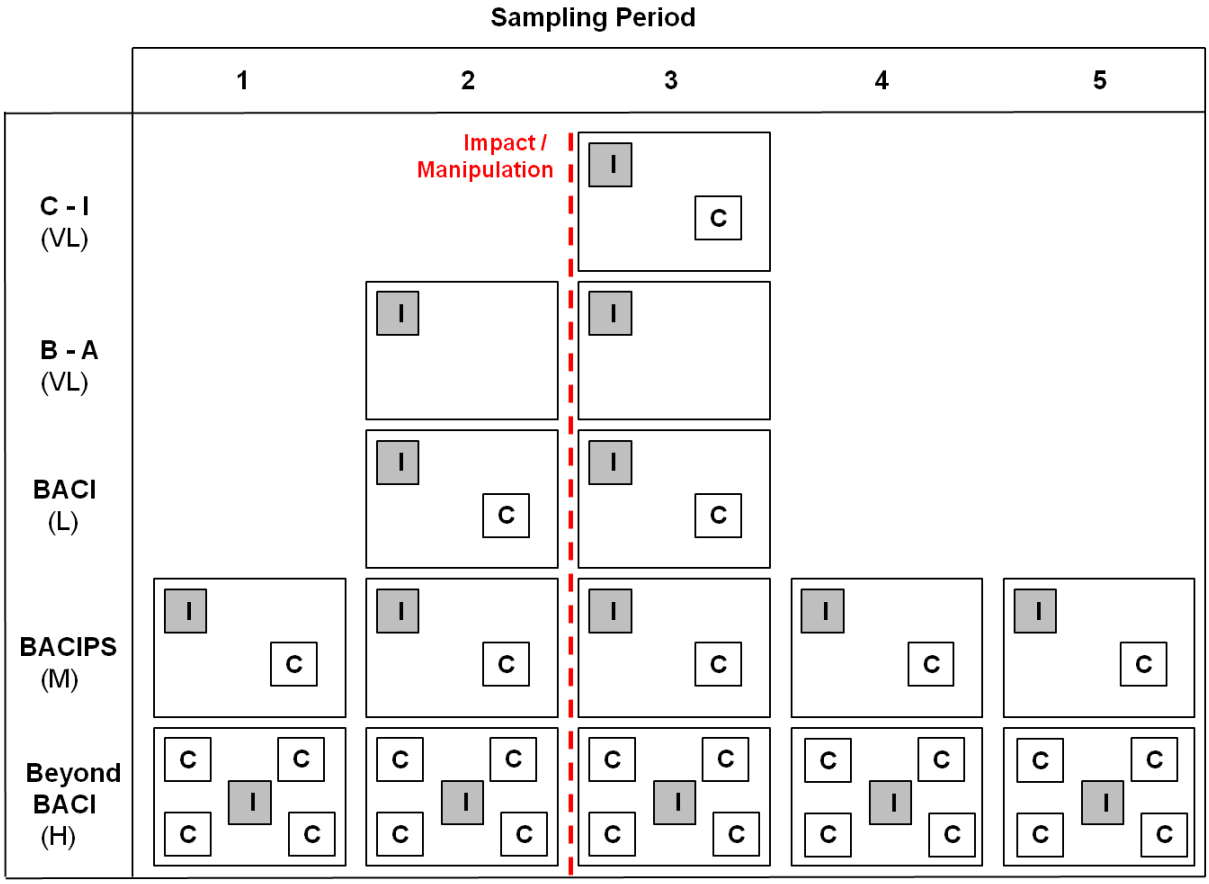


Figure 19. Investigative monitoring design comparison. B = before; A = after; C = control; I = impact; PS = paired series. The degree of confidence in each design method for detection of impact or manipulation effects above natural variation or 'noise' is provided in brackets; VL = very low confidence; L = low confidence; M = moderate confidence; H = high confidence. Numbers of sampling periods and control locations in BACIPS and Beyond BACI designs are not limited to those presented here.

9.4.5 Controlling for variation where 'Before' data are not available

Acquisition of a 'before' dataset is very highly recommended; however, situations may arise where it is not possible to collect pre-disturbance data, or historical datasets do not have sufficient power for a meaningful comparison. In this situation Osenberg *et al* (2011) recommends using habitat availability as a covariate to adjust indicator values, if the distribution of the indicator is known to be affected by this parameter.

Adjusting data by covariance requires a solid understanding of the indicator, system and pressure in question, and the relationships between them. For example; how will management measures affect biota? Will the effect of the closure be on the biota (e.g. an increased number of epifaunal taxa), the habitat (e.g. increased coral density), or both? If the habitat is unaffected by the closure of the area, then the habitat-adjusted data will correctly measure the effect of the closure. However, if the closure also increases habitat availability, then the adjustment of indicator values by habitat will eliminate some of the effect of the closure, thereby underestimating it. This approach offers an opportunity to refine estimates of effect when 'Before' data are not available, but must be applied with caution and an understanding of its limitations.

9.4.6 General principles for investigative monitoring

Where habitats are reasonably homogeneous within the control and impact sites (or where habitat distribution is not known with confidence) a **random** or **systematic** sampling approach should be applied. **Stratified random** sampling may be appropriate if habitats are variable but reasonably balanced between sites (i.e. equal amounts of each habitat in control and impact sites); in this case habitat type may be used as a factor in analysis. A theoretical example of a random stratified BACI design for investigative monitoring is displayed in Figure 20.

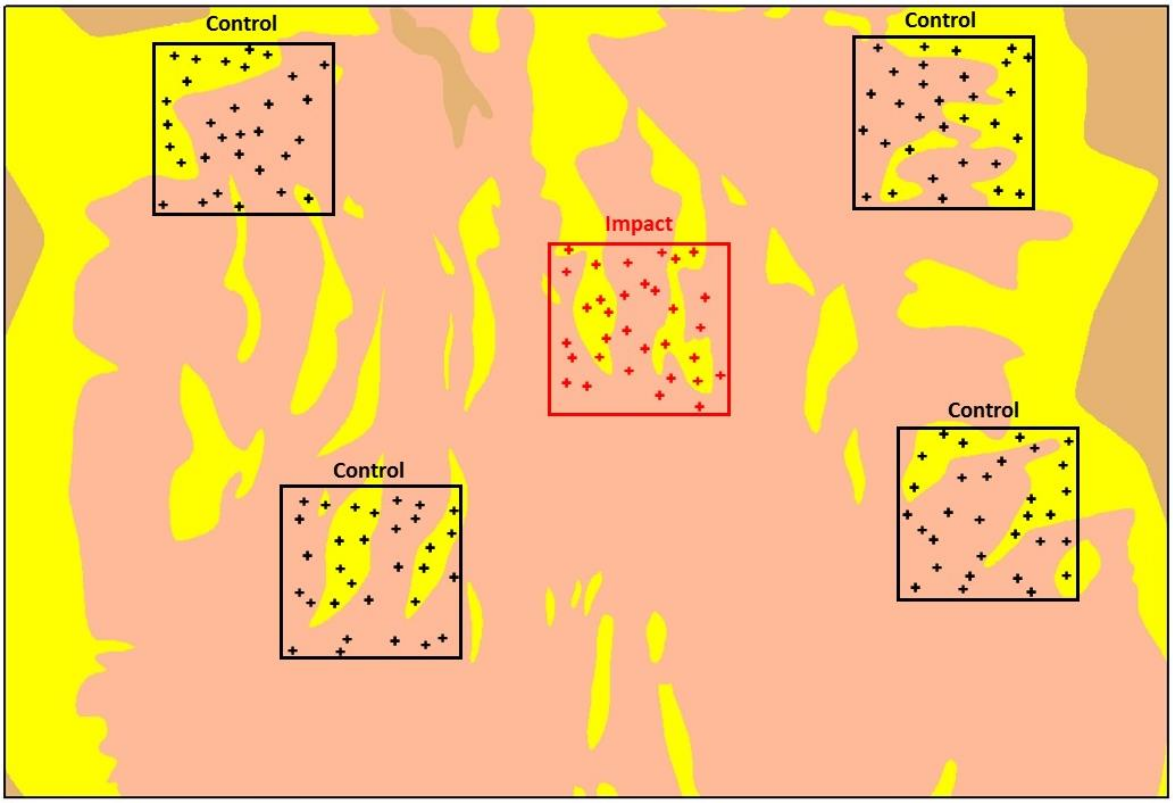


Figure 20. A theoretical random stratified 'Beyond BACI' design with a single impact site (e.g. where an impact has occurred, or management measures been applied) and replicated control sites. Following power analysis, sampling has been stratified by sublittoral coarse sediment (pink - 18 samples per site) and sublittoral sand (yellow - 13 samples per site). The survey area is relatively homogeneous in terms of environmental influences and anthropogenic pressures.

Whichever sampling design is used the following general principles should be applied when designing investigative monitoring studies:

1) Control site placement should be carefully considered

Careful consideration must be given to control site placement to minimise the likelihood that the experiment is confounded by natural variation or changes arising from the experiment itself. Principles for optimal placement of control sites are presented in Table 8. In practice, it may be difficult to fully adhere to these principles due to limitations in control site availability; therefore mitigative measures are suggested, to be applied if necessary. Where the mitigative measures are insufficient to avoid confounding the experiment, an investigative monitoring design may not be appropriate.

Table 8. Principles for optimal placement of control sites.

Principles for optimal control site placement	**Mitigative measures (if principles cannot be met)**
Control sites should be positioned in areas where environmental characteristics (e.g. sediment type, depth and organic inputs) are as similar as possible to those of the impact site. *Choosing homogeneous sites will decrease the effect of noise on detection of indicator response.*	Environmental variables likely to result in between-site variation should be quantified and used as covariates in analysis. If it is not possible to locate a control site with similar environmental conditions and sediment distribution, an investigative design should not be attempted.
Control sites should be located within the same broad vicinity as the impact site. *The appropriate maximum distance will depend on the scale of local variation.*	Where control sites cannot be placed in reasonably close proximity, the comparability of the sites should be thoroughly explored and expert judgement applied to determine whether an investigative design is suitable.
Control sites should be a sufficient distance from the impact site to be spatially independent from it *e.g. not directly adjacent to it*	Where it is necessary to place control sites in very close proximity to the impact site, a buffer should be applied to ensure that sampling points within the control site are closer to each other than to those within the impact site.
Control sites should not be positively affected by the impact or manipulation; *e.g. a control site positioned in close proximity to an area closed to pressures may be subject to biological 'overspill' (e.g. a higher level of larval recruitment than would be expected if management measures were not in place).*	Where control sites must be positioned inside a potential area of positive influence, the likely benefit should be quantified where possible (e.g. larval sampling), and used as covariate in analysis. Where this is not possible expert knowledge should be applied to determine whether the design will be affected to an unacceptable degree.
Control sites should not be negatively affected by the impact or manipulation; *e.g. the control site should not be located in an area where pressures are expected to be displaced to following application of management measures (e.g. MPA 'edge effects').*	Where control sites cannot be placed outside a potential area of negative influence, the pressure should be quantified (if possible) and used as a covariate in analysis. Where this is not possible expert knowledge should be applied to determine whether the design will be affected to an unacceptable degree.
Control sites should display a similar level and distribution of anthropogenic pressure to the impact site before management measures are/were applied.	Anthropogenic pressure should be quantified and used as a covariate or factor in analysis.
Control sites should be located in areas where levels of pressure are not likely to change substantially from those present at the impact site before management measures were applied.	This is difficult to mitigate (particularly for long-term BACIPS and Beyond BACI designs), however areas with reasonably stable historic pressure intensity can be selected, and pressures can be quantified and used as covariates where necessary.

2) Sampling designs should consider pressure distribution within and between control and impact sites

It should not be assumed that the intensity of pressures within the control and impact sites is evenly distributed.

If areas within the impact and control sites have not been subject to the pressure for which management has been applied, then sampling and re-sampling at these locations may diminish the apparent effect of the management within the wider sample. It may therefore be appropriate to exclude sampling from areas within the impact site where the pressure has not occurred, or is thought to be negligible.

For BACI-type designs, power is maximised by selecting control sites which closely mirror pressure intensity and distribution at the impact site *before* the management measures or manipulation. Where pressure varies substantially between or within control and impact sites, the level of pressure should be used as a covariate or factor in analysis.

3) The time-series sampling interval should be optimised to reduce serial correlation

Serial, or temporal, correlation is an unavoidable feature of BACI-type designs, as successive samples taken from the same sites are likely to be correlated with each other (Hurlbert 1984). Sampling events should have sufficient temporal spacing that serial correlation is reduced as far as possible (Stewart-Oaten *et al* 1986). The appropriate timescale will vary based on the indicator/s in question; for example, species or communities with a high turnover of individuals are likely to require a shorter sampling interval than those that are slow-growing or long-lived. As discussed in Section 5, the sampling events in the investigative monitoring time series should be conducted within the same season wherever possible.

4) Sites should remain fixed, whilst within-site sampling points may be re-randomised or fixed

Control and impact sites should remain fixed throughout the time-series, whilst within these sites sampling points may be re-randomised or remain fixed, depending on the monitoring objectives.

As mentioned in Section 9.1.5, fixed locations can provide a precise measure of change when monitoring the growth, density or cover of sessile or slow-growing biota, or those that are only known from specific locations. The primary disadvantage of monitoring fixed locations is that inference can only be made about specific areas, and not the rest of the site or area of habitat; however, in certain situations it will be appropriate to return to the same locations. For example; if specific areas of a coral reef are known to have been damaged by trawling, it will be most effective to return to these locations following a closure if the monitoring objective is to monitor the growth rate and recovery of these areas.

Where monitoring objectives require inference to be made about the entire site, sampling locations should be re-randomised for each sampling event, either by using a random sampling approach, or by re-setting the starting point of a systematic grid.

It is recommended that sampling points are re-randomised wherever possible.

9.4.7 Summary of key points and recommendations

Section 9.5: Investigative monitoring sampling designs

Key Points:

- Determination of cause and effect in an investigative monitoring design is achieved through comparison of 'treatments', or groups of samples subject to different combinations of controlled conditions (known as factors).
- Two main factors are typically used in investigative designs:
    - Management measures or experimental manipulation: i.e. 'Control' and 'Impact' sites.
    - Sampling period: i.e. 'Before' and 'After' management measures or an experimental manipulation.
- As stated by Green (1979), an optimal investigative monitoring experiment has several basic features, without which it will be difficult to infer the cause of change:
    - The type of impact, time of impact, and place of occurrence should be known in advance.
    - The impact should not have occurred yet.
    - Control areas should be available.
- Commonly used investigative monitoring designs are:
    - Control-Impact (CI) and Before-After (BA)
    - Before-After-Control-Impact (BACI)
    - Before-After-Control-Impact Paired Series (BACIPS)
    - Beyond BACI (multiple control sites, multiple sampling events)

Recommendations:

- It is highly recommended that a 'Before' dataset is acquired wherever possible.
- Whilst the BACI design is preferable to CI and BA designs, it is recommended that a more complex design is used (i.e. BACIPS or Beyond BACI)
- The following principles should be taken into account for investigative sampling designs (see explanations in Section 9.4.6):
    1. Control site placement should be carefully considered.
    2. Sampling designs should consider pressure distribution within and between control and impact sites.
    3. The time-series sampling interval should be optimised to reduce serial correlation.
    4. Sites should remain fixed, whilst within-site sampling points may be re-randomised or fixed. It is recommended that sampling points are re-randomised where possible.

9.5 Nesting monitoring types in sampling designs

When either two or three monitoring types are required in the same study, the sampling designs should be 'nested' to prevent unnecessary repetition of sampling effort and maximise resources. An example of a nested design featuring combined sentinel, operational and investigative monitoring stations is presented in Figure 21.

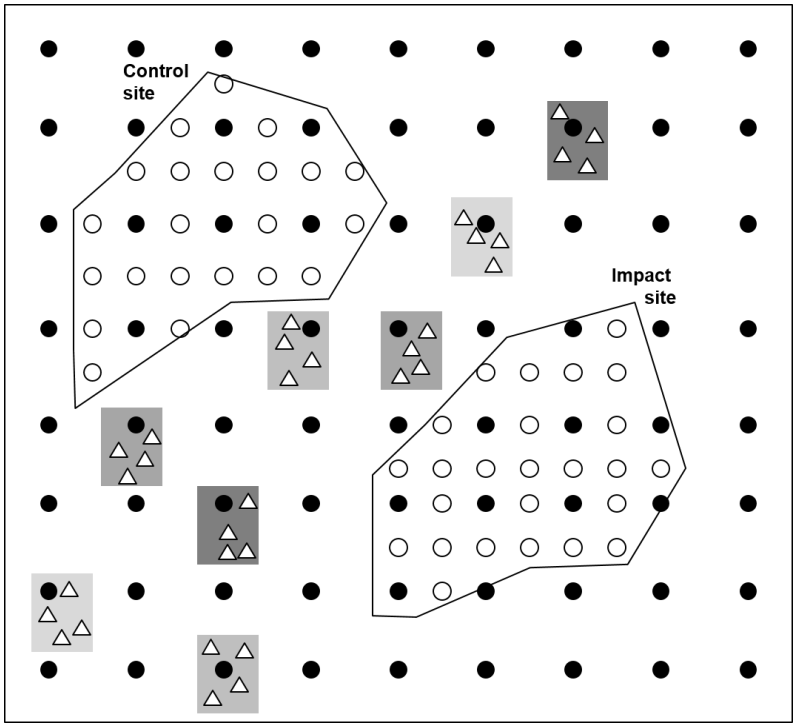


Figure 21. A nested design featuring sentinel, operational and investigative monitoring stations. Black circles = sentinel monitoring points; white triangles = additional points for operational monitoring (eight pressure units under four pressure categories); white circles = additional points for investigative monitoring.

9.6 Sampling designs for large and/or diverse areas

Where inference must be drawn about a large and/or environmentally diverse survey area (e.g. with a large depth range), it may be necessary to modify the standard sampling designs to increase precision.

For example, the Swallow Sand Marine Conservation Zone (MCZ), has an area of 4,746m^2 (see Figure 22). Even with substantial resources, it would be extremely difficult to sample across the entire site with sufficient replication to detect an effect against the inevitable background variation.

In such cases, a more achievable and robust design could involve sampling intensively within nested boxes at various intervals within the survey area. Having been selected to minimise natural variation (e.g. each box covering a small depth range), sampling within these boxes would increase the power of the design and the likelihood of detecting an effect within the context of each box. There are, however, limitations to this design. The unsampled areas can only be assumed to be in the same condition and to support the same communities, therefore strong inference can only be drawn about the boxes sampled. It is possible to improve confidence in this inference by acquiring verification data from

unsampled areas (e.g. qualitative video transects or grab samples to confirm habitat or substrate type outside of the boxes).

An example of a nested box systematic sampling design with wider verification stations, for sentinel monitoring at Swallow Sand MCZ is displayed in Figure 22.

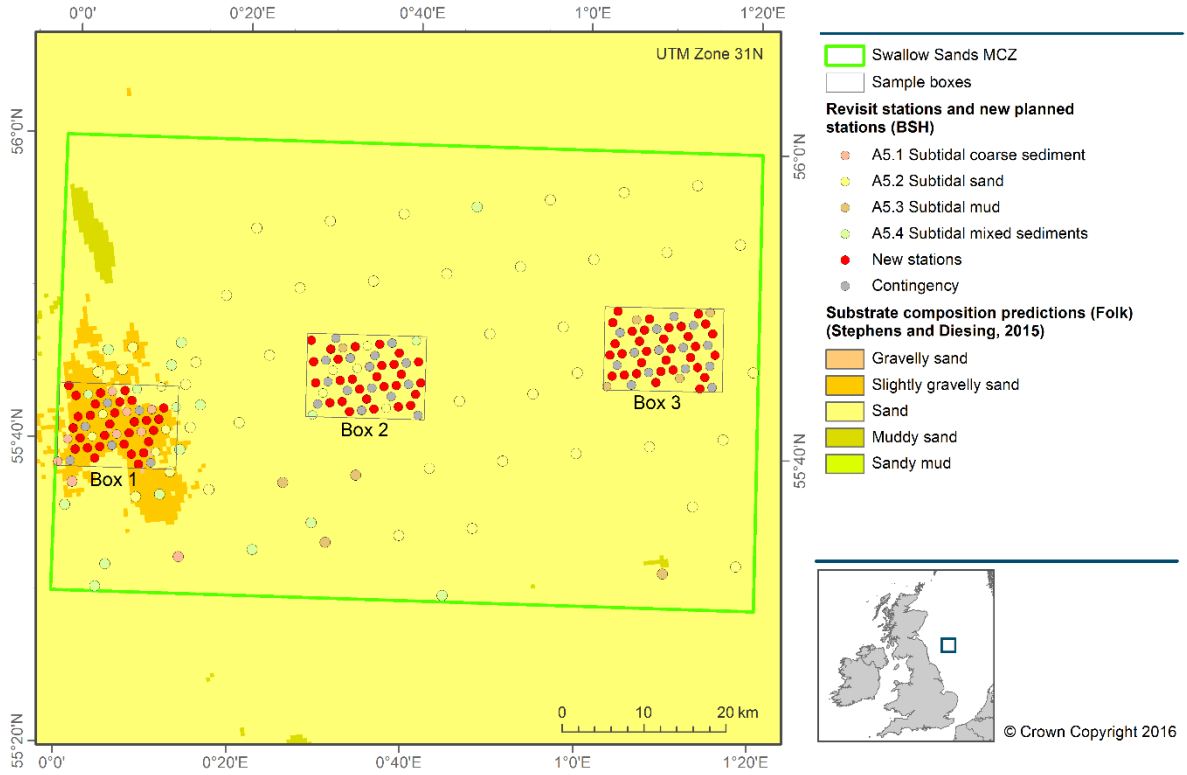


Figure 22. Example of a nested box systematic sampling design (with wider habitat verification stations, for sentinel monitoring at the Swallow Sand MCZ.

10 Conducting statistical analyses

Having successfully acquired monitoring data using a statistically robust design, the next stages are data exploration and analysis. This section provides basic guidance on the types of statistical analyses which can be used for sentinel, operational and investigative monitoring data. It does not attempt to limit the user to particular techniques, or to describe the full range of those available, but supplies basic information on variables and data types, data exploration and statistical analyses (Sections 10.1 to 10.3).

The majority of analyses discussed here can be used for all three monitoring types, therefore statistical advice is presented in the context of investigating patterns in multivariate community data (Section 10.3.1), investigating relationships and trends (Section 10.3.2), and investigating differences between groups (Section 10.3.3). Each section contains a brief summary of different analyses and guidance on when to use them, and a range of tests summarised within a table. Recommended reading boxes direct the user towards appropriate texts and papers for each specific group of analyses, whilst examples of general references are supplied below:

General recommended reading:

DYTHAM, C. 2011. Choosing and using statistics: A biologists guide. 3rd Edition. Wiley-Blackwell

FOWLER, J., COHEN. L. & JARVIS, P. 1998. Practical statistics for field biology. 2nd Edition. Wiley & Sons

QUINN, G.P. & KEOUGH, M.J. 2002. Experimental Design and Data Analysis for Biologists, Cambridge, U.K.537pp

ZUUR, A., IENO, E.N. & SMITH, G. 2007. Analysing ecological data. Springer-Verlag, New York.

10.1 Types of variables and data

The following terminology is used throughout the following sections to describe the different types of variable;

Response variable/s - This refers to the metric/s used to measure the indicator/s (also referred to as dependent variable/s).

Predictor variable/s - This refers to variable/s expected to cause or explain variation in the response variable. They are also commonly referred to as independent or explanatory variables.

To select the appropriate analysis, it is important that the characteristics of the response and predictor variables are recognised. Most data fall within one of three groups:

Numerical data
- Numerical data have a meaning as a measurement. They can be broken down into two further sub-types;
- **Discrete data** are counted (e.g. number of sea pens).
- **Continuous data** are measured (e.g. total hydrocarbon concentration).

	- A discrete or continuous predictor variable may be referred to as a **covariate**.
Categorical data	- Categorical data (including binary data) represent characteristics (i.e. sand / mud / mixed sediment or live / dead coral). - These data may take on numerical values but they do not have mathematical meaning and are not ordered. - A categorical predictor variable may be referred to as a **factor.**
Ordinal data	- Ordinal data are similar to categorical data, but the data follows a clear order, and the order of the scores has mathematical meaning (e.g. low, medium, high as 1, 2, 3).

10.2 Data exploration

Data exploration is the crucial first step of statistical analysis, where the data are 'eyeballed' (visually inspected) and/or plotted to allow to identification of broad trends and patterns, to identify any recording mistakes, and to determine which types of analysis are most suitable (e.g. parametric or non-parametric). Thorough and rigorous data exploration will increase the probability that the correct analytical approach is taken, and ultimately reduces the risk of drawing incorrect conclusions.

In their protocol for data exploration, Zuur *et al* (2010) suggest asking a series of questions which will allow the most common statistical problems to be avoided. This protocol is presented in Table 9, and is recommended as a framework on which to base data exploration. Some of these questions are only required for univariate analysis (e.g. are the data normally distributed?), whilst others should also be applied for multivariate analyses (e.g. are there outliers? Is there collinearity between the covariates?) Further information and advice on remedial actions for assumption violation is available in Zuur *et al* (2007, 2010), Dytham (2011) and Fowler *et al* (1998).

At this point the data should be also reviewed in the context of any original stratification, and decisions made about how to group the data for analysis. For example, if the sediments did not correspond to those predicted from habitat maps or modelled products, *post-hoc* stratification may be required to reduce background variance. It may also be appropriate to exclude data points, at the discretion of the researcher; for example, if a few replicates from one pressure unit of an operational monitoring study were found to comprise a very different substrate to that observed within that unit and other units.

Table 9. A protocol for data exploration (adapted from Zuur *et al* 2010). * Y = response variable, X = predictor variable/s.

Step	Question	Variable/s*	Brief description of the issue	Technique
1	Are there outliers?	Y, X	An outlier is an observation (value) which appears to deviate markedly from other observations. Outlying observations may indicate an input error (e.g. a typo), in which case it must be amended, or random variation. Outliers can introduce bias to statistical models by skewing variance, therefore if the outlier is genuine a judgement must be made as to whether it should be retained in the dataset.	Boxplot Cleveland dotplot
2	Is there homogeneity of variance?	Y	Homogeneity of variance (i.e. each 'population' (group) displays equal variance) is an important assumption of analysis of variance (ANOVA) and other regression models. Violation of this assumption will inflate the Type I error rate.	Conditional boxplot
3	Are the data normally distributed?	Y	Normally distributed data is an assumption of many statistical techniques (i.e. parametric tests), however in reality data are often not normally distributed. Exploring the data distribution allows the correct model to be fitted (see Table 11).	Histogram QQ-plot
4	Are there lots of zeros in the data?	Y	Zero-inflation is a particular problem for count data, as pairs of species/variables consistently recording zero may show correlations where none exist. This can lead to biased parameter estimates and standard errors if the incorrect model is applied.	Frequency plot Correlogram
5	Is there collinearity among the covariates?	X	Collinearity is the existence of correlation between covariates (i.e. % mud and organic content). Highly collinear variables should not be included in the same models, or the likelihood of Type II errors will be increased.	VIF & scatterplot Correlations & PCA
6	What are the relationships between Y and X?	Y, X	It is important to visualise the relationships between response and predictor variables in order to interpret the results of subsequent analyses, and detect observations that do not comply to the general pattern between two variables.	Multipanel scatterplots Conditional boxplots
7	Should interactions be considered?	Y, X	An interaction may arise when the effect of one factor (e.g. two different management measures) is different depending on the levels of another factor (e.g. two different substrate types). Interactions must be identified and modelled as such.	Coplots
8	Are observations of the response variable independent?	Y	As discussed in Section 8, independence of observations in time and space is an important assumption of most statistical techniques, violation of which may result in inflated Type I errors. Where present, dependence must be modelled, or the means of closely spaced samples analysed rather than individual observations.	Auto-correlation functions & variograms Plot Y vs time/space

10.3 Statistical analyses

The main analyses will depend on the monitoring type and specific objectives of the monitoring (e.g. analyses to identify relationships or differences between groups, which should have been defined before sampling design), and a wide range of additional analyses will also need to be carried out for general data exploration. Some techniques will be more appropriate for specific monitoring types, whilst some are more interchangeable. The following sections describe analyses for investigating patterns in multivariate community data (Section 10.3.1), investigating relationships and trends (Section 10.3.2), and investigating differences between groups (Section 10.3.3).

10.3.1 Identifying patterns in multivariate community data

Multivariate community analyses allow exploration of the full biotic community structure and offer multiple visualisation methods and tests. By using multivariate techniques, it is possible to retain as much information as possible from biological and environmental datasets, and to identify patterns which are not apparent when the data are reduced to a single metric. These analyses can be conducted in a number of statistical software packages, the most commonly used of which are PRIMER and R (vegan package).

Multivariate community analyses are particularly appropriate for sentinel monitoring, where analysis may be more descriptive and exploratory than hypothesis-driven, especially for the first sampling event in a time-series. However, the range of multivariate analyses is extremely broad and it's likely that each monitoring type will benefit from using these techniques to varying degrees. Broad groups of multivariate analyses are discussed briefly under the following headings, with a summary table supplied in Table 10.

It should be noted that multivariate community data may require transformation (e.g. square root or fourth root) before some analyses are undertaken. This is to reduce the relative contributions of common and rare species to the overall analysis. It is also recommended that where multiple abiotic variables are used they should be normalised prior to analysis (Clarke & Warwick 2001).

Metric generation

Univariate techniques are used to condense the full benthic community dataset into a single metric for use in univariate analyses (see Sections 10.3.2 and 10.3.3). Commonly used univariate metrics include:

- abundance of individuals,
- richness (e.g. Margalef's species richness),
- evenness (e.g. Pielou's evenness),
- diversity (e.g. Simpson's index, Shannon-Wiener index),
- taxonomic distinctness,
- biological traits metrics
- multimetric indices (e.g. AMBI-IQI).

Distributional techniques

Distributional techniques can be used to graphically display information on patterns of relative species abundance without reducing that information to a single summary statistic (e.g. a diversity index). These techniques typically generate a curve or histogram and include, but are not limited to, the following methods;

– **Ranked species abundance** (dominance) curves provide a means of visually representing species richness and evenness within a sample or series of pooled samples.

– **Species accumulation curves** plot the increasing number of different species observed as samples are successively pooled, providing an indication of whether the sampling effort has captured the full range of species within a community. Observed species curves may be plotted alongside S estimators which generate estimated curves of the number of species accumulated with an infinite amount of effort (see Chao 2005).

– **Abundance-Biomass Comparison (ABC) curves** provide a means of assessing the status of populations which have been subjected to disturbance without the need for reference to control samples (Warwick 1986), allowing community equilibrium to be assessed in terms of the abundance of smaller r-selected opportunist species and larger K-selected species (MacArthur & Wilson 1967), the latter indicating a less disturbed environment.

Classification techniques

Cluster analysis is the most common classification technique used to identify groupings in community or abiotic data. Cluster analysis aims to locate groupings of samples which are similar to each other within a wider group of samples, through analysis of the 'distance' between sample pairs. This distance is obtained using original data or via the production of a similarity matrix; a matrix of scores that represent the similarity between pairs of samples (Clarke & Warwick 2001).

Cluster analysis can be used as an exploratory tool for any monitoring type, some common applications are:

- Identification of differences in community or substrate composition between sites or times.
- Identification of groupings of biota to discern different communities (i.e. habitat or biotope analysis).
- Exploration of whether control or impact sites differ in community or substrate composition, and whether they are therefore comparable (for investigative monitoring).

There are many different clustering techniques, which fall into hierarchical and non-hierarchical categories. **Hierarchical techniques** build a hierarchy of clusters by grouping samples, and then forming further groups at lower levels of similarity (either using top-down or bottom-up method). They do not require the number of clusters to be specified *a priori*, and instead split the data into natural groupings, generating a dendrogram which reveals the relationships between clusters (see Figure 23). It should be noted that various different algorithms can be used to generate the hierarchy (e.g. single-linkage, complete linkage, average linkage), and advantages and disadvantages of the different methods should be evaluated in the context of the specific dataset (see Duda *et al* 2000). Hierarchical cluster analysis can be simultaneously run with a **Similarity Profile (SIMPROF)** test; a permutation test of the null hypothesis that a set of samples do not differ from each other in multivariate structure. The test examines whether the similarities observed in the data are smaller and/or larger than those expected by chance, and allows statistically significant clusters to be identified and displayed on the dendrogram.

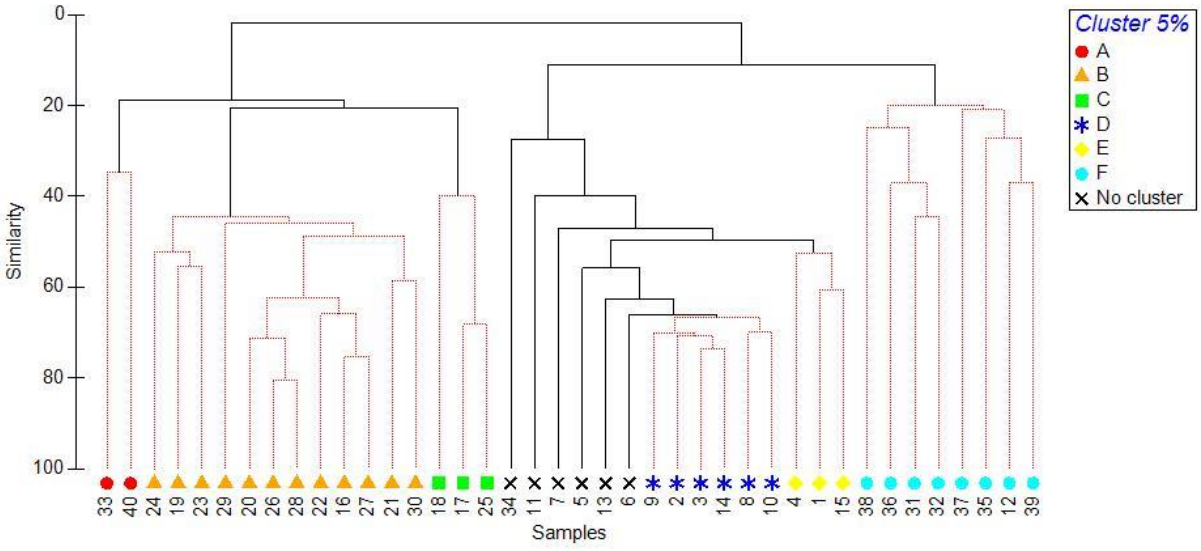


Figure 23: Example of a dendrogram produced by hierarchical clustering of macrofaunal abundance data from sandbank habitat, with a SIMPROF test applied at 5% significance (red lines denote statistically significant clusters).

Non-hierarchical clustering techniques (e.g. k-means clustering, composite clustering) assign and reassign samples to a pre-specified number of groups to achieve maximum within-cluster homogeneity. The main advantage of non-hierarchical over hierarchical techniques is the ability to reassign data which have been incorrectly classified early in the hierarchy; however non-hierarchical techniques provide no information on the relationships between data points.

Further information on cluster analysis and other classification techniques may be found in Duda *et al* (2000).

Ordination techniques

Ordination techniques create a 'map' of samples, usually in either two or three dimensions, in which the placement of the samples reflects the similarity of their biological communities. Distances between the samples correspond to dissimilarities in community structure; i.e. nearby points represent sampling points with similar communities, and points which are far apart have few species in common or the same species at very different levels of abundance (Clarke & Warwick 2001).

Ordination techniques are commonly used to:

- Visualise similarities and differences in;
    - community composition between samples
    - community composition between sites and times (i.e. between control vs impact sites, or before vs after sampling events)
    - community composition between areas of varying pressure
- Explore which environmental variables best explain patterns in community composition.

Multi-dimensional scaling (MDS) is a group of ordination techniques allowing two-dimensional 'mapping' of between-sample similarity, with an associated 2D stress coefficient which indicates how accurately the multidimensional community structure is represented in two dimensions (see Clarke 1993). In addition to visualisation of relationships between samples, MDS plots allow factorial and continuous variables to be superimposed onto the ordination to identify variables which have influenced patterns of distribution and cluster groupings. In addition to the groupings generated through cluster analysis, additional factors can be created for categorical variables, such as site, habitat type, sediment content (e.g. see Figure 24), sampling period, sampling equipment, and abrasion pressure category. Continuous variables such as sediment components, organic matter, contaminants or individual taxa can be displayed as 2-D bubbles of varying size.

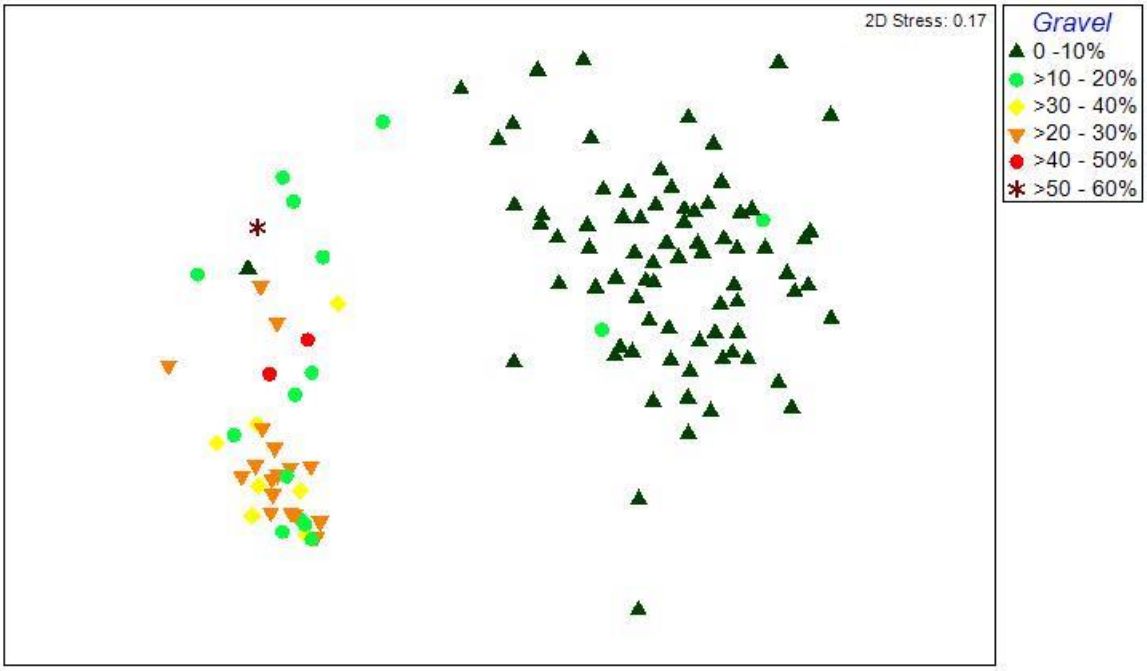


Figure 24: MDS ordination of macrofaunal community abundance data from sandbank habitat, overlain with gravel content classes.

Principal components analysis (PCA) (Chatfield & Collins 1980) is an ordination technique which can be used for various purposes, but is primarily used to explore variance within datasets based on sample dissimilarity, to highlight relationships between groups of variables, and to reduce large numbers of variables into a smaller number (principal components) by combining those that are highly correlated. This ordination method uses Euclidean distance, and is more suited to analysis of normalised environmental data than to biological community data. PCA ordinations generate a two-dimensional plot, displaying relative sample dissimilarity along the primary and secondary principal component axes, and eigenvectors which indicate the direction and strength of correlations between variables.

Exploratory techniques

Various exploratory analyses can aid interpretation of the groupings and patterns identified by classification and ordination techniques, including;

Similarity Percentages (SIMPER) analysis, which calculates within-group similarity, and identifies the most influential taxa within each by ranking average abundances and similarity contributions. The routine also allows pairwise comparison of clusters and other factors (e.g. broadscale habitat, year), by calculating the average group dissimilarity, and identifying the taxa which contribute the most to inter-group dissimilarity. In some cases, this routine may be effective for the identification of potential indicator species.

BIO-ENV analysis (also referred to as BEST analysis when combined with the BV-STEP stepwise selection procedure) finds the 'best' match between patterns in biological communities and associated environmental variables by exploring different variable combinations and ranking the best combinations according to their correlation coefficients.

Linkage trees (e.g. LINKTREE) can be used to explore how the 'best' variables identified through BIOENV analysis relate to groupings identified through cluster analysis, generating a dendrogram which shows which variables best explain splits.

Hypothesis testing

Hypothesis testing of multivariate data uses permutational methods, allowing the user to:

- Identify statistically significant differences in community composition between two or more groups, based on one or more factors (categorical predictor variables),
- Identify statistically significant linear relationships between community structure and continuous environmental predictor variables.

The **Analysis of Similarity (ANOSIM)** routine tests the null hypothesis that there are no differences between groups of samples, specified by levels of a single factor by ranking dissimilarity.

PERMANOVA (permutation-based MANOVA) has a similar function to the ANOSIM test, however, PERMANOVA uses distance measures (Bray-Curtis coefficients or Euclidean distance) rather than ranking to preserve information. This versatile test can handle complex, unbalanced designs including those with multiple factors, fixed factors (where all categories of the factor have been sampled) and random factors (where the levels of the factor have been randomly sampled from a wider 'population'), interaction terms and covariates. When used with multivariate data, the test uses permutations to make it distribution-free. However, when used with univariate data the test gives the same value as a traditional parametric F statistic, provided a Euclidean distance matrix has been calculated and the data are normally distributed (Anderson 2001).

Distance-based linear modelling (DistLM) offers a non-parametric approach to standard linear models. This analysis models the relationship between multivariate community data and one or more predictor variables, with various options for model selection. As with many other multivariate methods, it uses permutations and is based on a resemblance matrix.

Recommended reading:

ANDERSON, M.J. 2001. A new method for non-parametric multivariate analysis of variance. Austral Ecology, **24**, 32-46.

ANDERSON, T.W. 2003. An introduction to multivariate statistical analysis, 3rd Edition. Wiley.

CLARKE, K.R. 1993. Non-parametric multivariate analyses of changes in community structure. Australian Journal of Ecology, **18**, 117-143

CLARKE, K.R. & WARWICK, R.M. 2001. Change in marine communities: an approach to statistical analysis and interpretation. 2nd Edition. PRIMER-E, Plymouth.

DUDA, R.O., HART, P.E. & STORK, D.G. 2000. Pattern Classification, 2nd Edition. Wiley Interscience.

TABACHNICK, B.G. & FIDELL. L.S. 2013. Using multivariate statistics. Pearson Education Limited.

Table 10. Statistical analyses for investigating patterns in multivariate community data.

Analytical Objective	Type	Analysis	Outputs	Analysis Description
Summarise community characteristics in a single metric.	Univariate measures		Univariate metrics	Compute a variety of biodiversity metrics for univariate analysis, e.g. Simpson's diversity index, Margalef's richness, Pielou's evenness)
Explore distributions within community data.	Distributional (examples)	Dominance plots	Graph	Species are ranked in decreasing order of a specified metric, e.g. abundance, biomass, % cover, or other biotic measure.
		Species accumulation curves	Graph	Plot the increasing number of different species observed as samples are successively pooled against *S* estimators, indicating whether sampling effort has been sufficient to capture the full range of species within a community.
		Abundance-Biomass Comparison (ABC) curves	Graph	This allows community equilibrium to be assessed in terms of the abundance of smaller r-selected opportunist species, which may indicate a disturbed environment.
Visualise similarities and differences in community composition between samples.	Ordination	Multi-dimensional Scaling (MDS)	Plot	MDS ordination allows two-dimensional 'mapping' of inter-sample similarity, with an associated stress coefficient indicating how accurately the multidimensional community structure is represented in two dimensions. The resultant 'map' can be overlain with factor symbols (e.g. clusters, sediment types, years) or with 2-D bubbles for continuous variables.
Understand relationships within a set of variables and convert a set of observations into a set of values of uncorrelated variables (principal components).		Principal Components Analysis (PCA)	Plot & principal components	PCA is primarily used to explore variance within datasets based on sample dissimilarity, to highlight relationships between groups of variables, and to reduce large numbers of variables into a smaller number (principal components) by combining those that are highly correlated. It is most suited to analysis of environmental as opposed to biological data.
Identify groups within a dataset, based on similarities in community composition.	Classification	Cluster analysis (e.g. hierarchical cluster analysis, k-means clustering)	Cluster groups and plots/dendrograms	Cluster analyses identify 'natural groupings' of samples which are similar to each other within a wider group of samples through analysis of the similarity coefficients of sample pairs.

Analytical Objective	Type	Analysis	Outputs	Analysis Description
Identify whether differences between hierarchical clusters are statistically significant.	Hypothesis test	Similarity Profile analysis *SIMPROF: PRIMER* *clustsig (vegan): R*	Pi statistic & p-value	SIMPROF is a permutation test of the null hypothesis that a set of samples do not differ from each other in multivariate structure. The test allows statistically significant hierarchical clusters to be identified and displayed on a dendrogram.
Identify which species are responsible for sample groupings (i.e. clusters), and how much they contribute to cluster dissimilarity.	Exploration	Similarity Percentages *SIMPER: PRIMER* *simper (vegan): R*	Ranked species list and % contribution for each cluster	SIMPER analysis calculates within-cluster similarity to identify the most influential taxa within each cluster and calculate the percentage contribution to within-cluster dissimilarity. The output also provides percentage dissimilarity between clusters.
Identify which predictor variable/s best explain assemblage structure.	Exploration	*BEST analysis (BIO-ENV & BV-STEP): PRIMER* *bioenv (vegan): R*	Rho statistic & histogram	These routines find the 'best' match between multivariate assemblage patterns and all associated environmental variables. Correlation coefficients are ranked for each combination of variables.
Identify how 'best' predictor variables affect community group splits (e.g. clusters).	Exploration	Linkage tree *LINKTREE: PRIMER*	Tree diagram & R statistics	Linkage trees use take the combination of variables identified as 'best' at explaining assemblage patterns, and use them to describe how assemblage samples are split into groups.
Model the relationship between community data and one or more predictor variables.	Hypothesis test	Distance-based linear modelling *DistLM: PRIMER* *dbglm: R*	Pseudo-F statistic & p-value	Models the relationship between a multivariate data 'cloud' and one or more predictor variables with various options for model selection. The model is based on a resemblance matrix and uses permutations.
Identify whether differences exist between pre-defined groups of assemblage samples.	Hypothesis test	Analysis of Similarity *ANOSIM: PRIMER* *anosim (vegan): R*	R-statistic & p-value	ANOSIM is broadly analogous to a univariate one or two-way ANOVA, and tests the null hypothesis that there are no community differences between groups of samples, based on ranked dissimilarity. It is essential that groups are pre-defined and not generated by cluster analysis (Clarke & Gorley 2006).
		Permutational MANOVA *PERMANOVA: PRIMER* *R: adonis (vegan)*	Pseudo-F & p-value	Permutational analogue to multivariate ANOVA (but can also be used for univariate data as a more robust alternative to ANOSIM. Allows for more complex designs, inclusion of multiple factors, fixed or random factors, interaction terms and covariates, and unbalanced designs.

10.3.2 Identifying relationships and trends

Analyses to identify relationships and trends are most relevant to operational monitoring, where the relationship between a pressure and indicator state is under investigation, and sentinel monitoring to identify long-term trends where a substantial time-series exists. Relationship and trend analyses will also be used for single-event sentinel and investigative monitoring datasets to explore relationships between indicators and environmental variables (e.g. sediment composition, organic matter content). Various options for modelling linear and non-linear relationships are provided below, and are summarised in Table 11.

Parametric and non-parametric **correlation coefficients** (e.g. **Pearson's correlation, Spearman's rank and Kendall's Tau**) indicate the degree to which two variables are correlated. Causation cannot be implied from these correlations, as the relationship observed may be driven by either or neither of the variables included. Correlations are generally best applied in the exploration phase, to identify relationships between covariates.

Regression models identify whether a relationship exists between a response variable and one or more predictor variables, where the relationship is caused by the predictor/s. **Simple linear regression** (one response, one predictor) and **multiple regression** models (one response, >one predictor) make the assumption that errors follow a normal distribution. In reality this assumption is frequently violated by ecological data; for example, count data typically follow a Poisson distribution. **Generalized linear models (GLMs)**, are a more flexible alternative and allow modelling of a variety of distributions by introduction of link and variance functions. GLMs are used to model various different types of ecological data with different distributions, including random count data (Poisson distribution), clustered count data (Negative binomial distribution), and binary distributions such as presence / absence (Binomial).

Generalized linear models can be limited in their ability to deal with ecological data. If the assumptions of generalized linear models (e.g. a linear relationship between the response variable and linear predictor, independence between response variables) are violated it may be necessary to use a different and more complex approach. **Generalized Additive Models (GAMs)** (Hastie & Tibshirani 1986) are extensions of GLMs which allow the covariates to vary smoothly rather than linearly or in factor groups. GAMs apply smoothing functions to capture patterns in non-linear relationships and show them using smoothed curves. **Generalized Linear Mixed Modelling (GLMM)** can be used to model data where spatial and/or temporal autocorrelation is present.

Recommended reading:

DOBSON, A.J. & BARNETT, A. 2008. An introduction to Generalized Linear Models. Chapman & Hall/CRC

FARAWAY, J.J. 2009. Linear models with R. Chapman & Hall/CRC

FARAWAY, J.J. 2006. Extending the linear model with R. Chapman & Hall/CRC

WOOD, S. 2006. Generalized Additive Models: An introduction with R. Chapman & Hall/CRC Texts in Statistical Science.

ZUUR, A.F., IENO, E.N., WALKER, N.J., SAVELIEV, A.A. & SMITH, G.M. 2009. Mixed effects models and extensions in R. Springer

Table 11. Statistical analyses for investigating relationships and trends.

Analytical objective	What type of response (indicator) data?	How many predictor variables?	What type of predictor variable?	Do response data meet simple regression assumptions?	Recommended Tests or Models
Identify whether a relationship exists between the indicator and another variable.	Continuous	1	Continuous /Discrete	Yes	Pearson's correlation (where relationship is linear)
	Discrete Continuous	1		No	Spearman's Rank correlation (& Spearman's rho) or Man-Kendall test (& Kendal's Tau) (where relationship is monotonic)
Identify whether predictor variable/s have a causative relationship with the response variable.	Continuous	1	Continuous / Discrete	Yes	Simple linear regression
		>1		Yes	Multiple regression
Identify whether predictor variable/s have a causative relationship with a response variable which is not normally distributed.	Continuous Discrete Bernoulli (i.e. 0,1)	≥1	Continuous / Discrete	No	Generalized Linear Model (GLM) with different link functions for various distributions (e.g. Poisson, Negative Binomial, Gamma, Gaussian, Binomial)
Display patterns where the relationships between predictor and response variables are non-linear.	Continuous Discrete Bernoulli (i.e. 0,1)	≥1	Continuous / Discrete	No	Generalized Additive Model (GAM)
Identify whether predictor variable/s have a causative relationship where random effects are present, or where dependency is present in the response variable.	Continuous Discrete Bernoulli (i.e. 0,1)	≥1	Continuous / Discrete	No	Generalized Linear Mixed Model (GLMM)
Identify whether predictor variable/s have a causative relationship with the response variable where the same subjects (e.g. individuals within fixed plots) have been measured repeatedly.	Continuous Discrete Bernoulli (i.e. 0,1)	≥1	Continuous / Discrete	No	Repeated Measures Generalized Linear Model (GLM)

10.3.3 Identifying differences between groups

Analyses which identify differences between groups are most likely to be used for investigative monitoring to identify the effect of management measures or a manipulation, by comparison of control and impact sites before and after the event. These types of analyses are also appropriate for sentinel monitoring where different strata have been sampled (i.e. substrate types or pressure categories), or when the number of sampling events is not sufficient to conduct time-series trend analysis. Commonly used analyses are summarised below and in Table 13.

Differences between groups are determined using linear models (as described in Section 10.3.2), where one or more predictors are categorical (factors), and continuous and discrete predictors may be added to improve the fit of the model.

Where parametric assumptions are met, the **independent T-test** can be used to identify whether a statistically significant difference exists between two groups of a single factor, for example:

- Control vs Impact (factor = treatment)
- Before vs After (factor = time)
- sand vs mud (factor = sediment type)
- high vs low pressure (factor = pressure category)
- area 1 vs area 2 (factor = area)

One-way analysis of variance (ANOVA) is used to identify whether statistically significant differences exist in a single response variable between two or more groups of a single factor.

Examples of the use of one-way ANOVA are:

- sand vs mud vs mixed sediment (factor = sediment type)
- high vs low vs moderate pressure (factor = pressure category)
- area 1 vs area 2 vs area 3 (factor = area)

Two-way ANOVA adds an extra factor to the standard ANOVA model, enabling measurement of interaction effects in BACI designs; in effect determining whether the ‘impact’ group changes differently to the ‘control’ group between the ‘before’ and ‘after’ events. The basic two-way ANOVA can be adapted to accommodate the addition of more control sites and sampling events by specifying fixed and random effects in the model. Schwartz (2015) suggests four BACI models, including BACIPS and Beyond BACI designs, which are presented in Table 12.

Where data do not meet parametric assumptions, rank-sum tests such as the **Mann-Whitney U-test** (two groups in a single factor) or the **Kruskal-Wallis test** (more than two groups in a single factor) can be used. These tests do not require the data to be fitted to a distribution, but they also do not allow for addition of an additional factor, and therefore interaction terms cannot be modelled (with the exception of the **Friedman test**, which can be used for blocking factors). Where the data correspond to a distribution such as Poisson, Negative Binomial or Binomial, **Generalized Linear Models (GLM)** can be used, and extra random or fixed factors can be added as predictors using a **Generalized Linear Mixed Model (GLMM).**

Where the same individuals have been measured repeatedly over time (e.g. coral density within a fixed plot), a repeated measures analysis should be used (see Table 13).

Table 12. Four two-way ANOVA models for analysis of BACI data with fixed and random effects (adapted from Schwartz 2015).

BACI Design	Number of sampling sites				Analysis	Model (R = random effect)
	Before	After	Control	Impact		
Simple BACI	1	1	1	1	Two-way ANOVA (fully randomised)	Impact Time Impact*Time
BACI with multiple control sites	1	1	>1	1	Two-way mixed effects ANOVA	Impact Site (**R**) Time Impact*Time Site*Time (**R**)
BACIPS	>1	>1	1	1	Two-way mixed effects ANOVA	Impact Time Impact*Time SampleTime (**R**)
Beyond BACI	>1	>1	>1	1	Two-way mixed effects ANOVA	Impact Time Impact*Time SampleTime (**R**)

R = random effect, Impact = control vs impact effect, Time = before vs after effect, Site = control sites, SampleTime = sampling event

Recommended reading:

FARAWAY, J.J. 2009. Linear models with R. Chapman & Hall/CRC

RUTHERFORD, A. 2011. ANOVA and ANCOVA: A GLM approach (2nd Ed). Wiley & Sons

SCHWARTZ, C.J. 2015. Analysis of BACI experiments. In Course Notes for Beginning and Intermediate Statistics. Available at http://www.stat.sfu.ca/~cschwarz/CourseNotes. Retrieved 2015-11-23

Table 13. Statistical analyses for investigating differences between groups.

Analytical objective	What type of response (indicator) data?	How many factors?	Number of groups in factor/s	Do the response data meet parametric assumptions?	Recommended Tests / Models	Recommended Tests / Models for repeated measures design*
Identify differences between groups or treatments.	Continuous Discrete Ordinal	1	2	No	Mann-Whitney U test	Wilcoxon Matched Pairs test
		1	>2	No	Kruskal-Wallis test	Friedman test (can also be used for 2 factors)
		≥1	≥2	No	Generalized Linear Model / Generalized Linear Mixed Model (fixed and random effects)	Repeated Measures GLM / GLMM
	Continuous	1	2	Yes	Independent t-test	Paired T-test
		1	>2	Yes	One-way ANOVA	One-way Repeated Measures ANOVA
		2	>2	Yes	Two-way ANOVA	Factorial Repeated-Measures ANOVA

* Repeated measures designs involve taking successive measures at the exact same sampling locations or of the same individuals.

10.4 Summary of key points and recommendations

Section 10: Conducting statistical analyses

Key Points:

- It is essential that data are thoroughly explored prior to analysis. This stage improves understanding of the data and allows identification of potential issues which could influence the outcome of the analysis.
- Statistical analyses have been broadly classed into three groups;
    - Identifying patterns in multivariate community data,
    - Identifying relationships and trends,
    - Identifying differences between groups.

Recommendations:

- The protocol presented by Zuur *et al* (2010) provides a framework by which to conduct data exploration (Table 9).
- Statistical analyses from all three groups can be used for each monitoring type. Some techniques will be more appropriate for specific monitoring types, whilst some are more interchangeable.
- Multivariate community analyses are particularly appropriate for sentinel monitoring, where analysis may be more descriptive and exploratory than hypothesis-driven, especially for the first sampling event in a time series. Table 10 summarises the attributes of multivariate community analyses.
- Analyses which identify relationships and trends are most relevant to operational monitoring, and sentinel monitoring where a substantial time-series exists. Relationship and trend analyses may also be used for single-event sentinel and investigative monitoring datasets to explore relationships between indicators and environmental variables. Table 11 summarises the characteristics of commonly used tests.
- Analyses which identify differences between groups are most likely to be used for investigative monitoring, to identify whether management measures have been effective. This group of analyses is also appropriate for sentinel monitoring where different strata have been sampled or when the number of sampling events is not sufficient to conduct trend analysis. Common tests for differences between groups are summarised in Table 13.

11 Acknowledgements

We would like to thank those from the following organisations who contributed to this document by providing comments and technical advice.

JNCC: Kerstin Kröger, Paul Whomersley, Hayley Hinchen, Emma Verling, Simone Pfeifer, Joe Turner, Amy Ridgeway, Becky Hitchin, Henk van Rein, Yessica Griffiths, Laura Robson, Gareth Johnson, Mike Nelson, Neil Golding, Hannah Carr, David Vaughan, Becky Phillips, Joey O'Connor, Helen Lillis, Siobhan Vye & Jennifer Lawson.

Natural England: Mike Young, Ben Green, Emma Foster, Ross Bullimore, Dylan Todd & Maija Marsh.

Natural Resources Wales (Cyfoeth Naturiol Cymru): Mike Camplin, Paul Brazier, Kirsten Ramsey & Natasha Lough

Scottish Natural Heritage (Dualchas Nàdair na h-Alba): Ben James, Lisa Kamphausen & Flora Kent.

Department of Agriculture, Environment and Rural Affairs (Northern Ireland): Hugh Edwards.

Cefas: Jon Barry.

12 References

ADDISON, P. 2011. A global review of long-term Marine Protected Area monitoring programmes: The application of a good framework to marine biological monitoring. JNCC Report No. 455. Available from: http://jncc.defra.gov.uk/pdf/jncc455_Vol1_Vol2%20combined_web.pdf [Accessed 03 June 2017].

ALBERT, C.H., YOCCOZ, N.G., EDWARDS, T.C., GRAHAM, C.H., ZIMMERMANN, N.E. & THUILLER, W. 2010. Sampling in ecology and evolution – bridging the gap between theory and practice. Ecography, **33**, 1028-1037.

ALEXANDER, D., COLCOMBE, A., CHAMBERS, C. & HERBERT, R.J.H. 2014. Conceptual ecological modelling of shallow sublittoral coarse sediment habitats to inform indicator selection. JNCC Report No. 520. Available from: http://jncc.defra.gov.uk/pdf/Report%20520_web.pdf [Accessed 03 June 2017].

ALEXANDER, D., COATES, D.A., TILLIN, H. & TYLER-WALTERS, H. 2015. Conceptual ecological modelling of sublittoral rock habitats to inform indicator selection. JNCC Report No. 560. Available from: http://jncc.defra.gov.uk/pdf/Report_560_web.pdf [Accessed 03 June 2017].

ANDERSON, M.J. 2001. A new method for non-parametric multivariate analysis of variance. Austral Ecology, **24**, 32-46.

ANDERSON, T.W. 2003. An introduction to multivariate statistical analysis, 3rd Edition. Wiley.

ANDERSON, D.R., BURNHAM, K.P. & THOMPSON, W.L. Null hypothesis testing: problems, prevalence and an alternative. Journal of Wildlife Management **64** (4) 912-923

ANDREW, N.L. & MAPSTONE, B.D. 1987. Sampling and the description of spatial pattern in marine ecology. Oceanography & Marine Biology Annual Review, **25**, 39-90

BAINBRIDGE, T.R. 1985. The committee on standards: precision and bias. ASTM Standardization News, **13**, 44 -46.

BARRIO FROJÁN, C.R.S. & MASON, C. 2010. A quantitative comparison between the sampling efficacy of the mini Hamon grab and the Costerus twin grab. Marine Aggregate Levy Sustainability Fund, MEPF Ref. No. MEPF REC 08/PO4, 9 pp.

BARRY, J. & MAXWELL, D. 2017. emon: tools for environmental and ecological survey design and analysis. https://cran.r-project.org/.

BARRY, J. & NICHOLSON, M.D. 1993. Measuring the probabilities of patch detection for four spatial sampling schemes. Journal of Applied Statistics, **20**, 353-362.

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

BLOMQVIST, S. 1991. Quantitative sampling of soft-bottom sediments: problems and solutions. Marine Ecology Progress Series, **72**, 295-304.

BOURGERON, P.S., FORTIN, M-J. & HUMPHRIES, H.C. 2001. Elements of spatial data analysis in ecological assessments. In: JENSEN, M.E. & BOURGERON, P. S. (eds.) A guidebook for integrated ecological assessments XIII, p. 536.

BOYD, S.E., BARRY, J. & NICHOLSON, M. 2006. A comparative study of a 0.1m^2 and 0.25 m^2 Hamon grab for offshore marine gravels. Journal of the Marine Biological Association of the United Kingdom, **86**, 1315-1328.

BISHOP, J.D.D. & HARTLEY, J.P. 1986. A comparison of the fauna retained on 0.5 mm and 1.0 mm meshes from benthic samples taken in the Beatrice Oilfield, Moray Firth, Scotland. Proceedings of the Royal Society of Edinburgh. Section B. Biological Sciences, **91**, 247-262.

BREEN, P., VANSTAEN, K. & CLARK, R.W.E. 2014. Mapping inshore fishing activity using aerial, land, and vessel-based sighting information. ICES Journal of Marine Science. First published online July 16, 2014 doi:10.1093/icesjms/fsu115doi: 10.1093/icesjms/fsu115.

BROWN, A. 2000. Habitat monitoring for conservation management and reporting. 3: Technical guide. Countryside Council for Wales, Bangor.

BUHL-MORTENSEN, L. 1996. Type-II statistical errors in environmental science and the precautionary principle. Marine Pollution Bulletin, **32**, 528-531.

BYRNES, M.E. 2000. Sampling and surveying radiological environments. CRC Press LLC, Florida.

CABRAL, H.N. & MURTA, A.G. 2004. Effect of sampling design on abundance estimates of benthic invertebrates in environmental monitoring studies. Marine Ecology Progress Series, **276**, 19-24.

CHAO, A. 2005. Species richness estimation. p. 7909-7916 In: BALAKRISHNAN, N., READ, C.B. & VIDAKOVIC, B. (eds.) Encyclopedia of Statistical Sciences. Wiley, New York.

CHATFIELD, C. & COLLINS, A.J. 1980. Introduction to multivariate analysis. Chapman & Hall, London.

CHURCH, N.J., CARTER, A.J., TOBIN, D., EDWARDS, D., EASSOM, A., CAMERON, A., JOHNSON, G.E., ROBSON, L.M. & WEBB, K.E. 2016. JNCC Pressure Mapping Methodology. Physical Damage (Reversible Change) - Penetration and/or disturbance of the substrate below the surface of the seabed, including abrasion. *JNCC Report No. 515*. JNCC, Peterborough.

CLARKE, K.R. 1993. Non-parametric multivariate analyses of changes in community structure. Australian Journal of Ecology, **18**, 117-143.

CLARKE, K.R. & WARWICK, R.M. 2001. Change in marine communities: an approach to statistical analysis and interpretation. 2nd Edition. PRIMER-E, Plymouth.

CLARKE, K.R. & GORLEY, R.N. 2006. PRIMER v6: User manual/tutorial. PRIMER-E, Plymouth.

COATES, D.A., ALEXANDER, D., STAFFORD, R. & HERBERT, R.J.H. 2015. Conceptual ecological modelling of shallow sublittoral mud habitats to inform indicator selection. Marine Ecological Surveys Ltd - A report for the Joint Nature Conservation Committee, JNCC Report No. 557. Available from: http://jncc.defra.gov.uk/pdf/Report%20557_web.pdf [Accessed 03 June 2017].

COCHRAN, G.C. & COX, G.M. 1957. Experimental designs, 2nd Edition. Wiley, New York.

COCHRAN, W.G. 1977. Sampling techniques, 3rd Edition. Wiley, New York.

COGGAN, R., MITCHELL, A., WHITE, J. & GOLDING, N. 2007. Recommended operating guidelines (ROG) for underwater video and photographic imaging techniques. MESH Project Available from: http://www.emodnet-seabedhabitats.eu/pdf/GMHM3_Video_ROG.pdf [Accessed 03 June 2017].

COLLIE, J.S., HALL, S.J., KAISER, M.J. & POINTER, I.R. 2000. A quantitative analysis of fishing impacts on shelf-sea benthos. Journal of Animal Ecology, **69**, 785-798.

CRAWLEY, M.J. 2013. The R Book. 2nd Edition, John Wiley & Sons Ltd, Chichester, UK

CROWDER, M.J. & HAND, D.J. 1990. Analysis of repeated measures. Chapman and Hall, London.

CURTIS, M. & COGGAN, R. 2007. Recommended operating guidelines (ROG) for MESH trawls and dredges. MESH Project. Available from: http://www.emodnet-seabedhabitats.eu/PDF/GMHM3_Trawls_and_Dredges_ROG.pdf [Accessed 03 June 2017].

DAVIES, J., BAXTER, J., BRADLEY, M., CONNOR, D., KHAN, J., MURRAY, E., SANDERSON, W., TURNBULL, C. & VINCENT, M. 2001. Marine monitoring handbook. Section 2: Establishing monitoring programmes for marine features. Available from http://jncc.defra.gov.uk/PDF/MMH-Section%202.pdf [Accessed 03 June 2017].

DAVIES, G.M. & GRAY, A. 2015. Don't let spurious accusations of pseudoreplication limit our ability to learn from natural experiments (and other messy kinds of ecological monitoring). Ecology & Evolution, **5**, 5295–5304.

DERNIE, K.M., KAISER, M.J. & WARWICK, R.M. 2003. Recovery of soft sediment communities and habitats following physical disturbance. Journal of Experimental Marine Biology and Ecology, **285-286**, 415–434.

DIAMOND, J.M. 1975. Assembly of species communities. In: Cody, M. L., Diamond, J. M. Ecology and evolution of communities. pp. 342-344, Belknap Press.

DI STEFANO, J. 2001. Power analysis and sustainable forest management. Forest Ecology and Management, **154**, 141-153.

DI STEFANO, J. 2003. How much power is enough? Against the development of an arbitrary convention for statistical power calculations. Functional Ecology, **17**, 707-709.

DOBSON, A.J. & BARNETT, A. 2008. An introduction to Generalized Linear Models. Chapman & Hall/CRC

DODGE, Y. 2003. The Oxford Dictionary of Statistical Terms. The International Statistical Institute. Oxford University Press.

DUDA, R.O., HART, P.E. & STORK, D.G. 2000. Pattern Classification, 2nd Edition. Wiley Interscience.

DYTHAM, C. 2011. Choosing and using statistics: a biologists guide. 3rd Edition. Wiley-Blackwell.

ELEFTHERIOU, A. (ed.) 2013. Methods for the study of marine benthos. John Wiley & Sons, Chichester

ELEFTHERIOU, A. & MOORE, D.C. 2013. Macrofauna Techniques. In: Eleftheriou, A. (ed.) Methods for the study of marine benthos. John Wiley & Sons, Chichester.

FAIRWEATHER, P.G. 1991. Statistical power and design requirements for environmental monitoring. Australian Journal of Marine and Freshwater Research, **42**, 555-567.

FARAWAY, J.J. 2009. Linear models with R. Chapman & Hall/CRC.

FARAWAY, J.J. 2006. Extending the linear model with R. Chapman & Hall/CRC.

FISH, J.D. & FISH, S. 1996. A student's guide to the seashore. 2nd Edition. Cambridge: Cambridge University Press.

FISHER, R.A. 1925. Statistical methods for research workers. Oliver and Boyd, Edinburgh.

FORTIN, M-J., DALE, M.R.T. & VER HOEF, J. 2002. Spatial analysis in ecology. In: El-Shaarawi, A.H., & Piegorsch, W.W. Encyclopedia of Environmetrics (ISBN 0471 899976), John Wiley & Sons Ltd, Chichester.

FOWLER, J., COHEN. L. & JARVIS, P. 1998. Practical statistics for field biology. 2nd Edition. Wiley & Sons.

GELMAN, A. 2008. Objections to Bayesian statistics. Bayesian Analysis, **3** (3), 445-449.

GEORGE, C.L. & WARWICK, R.M. 1985. Annual macrofauna production in a hard-bottom reef community. Journal of the Marine Biological Association of the United Kingdom, **65**, 713-735.

GILI, J.M. & HUGHES, R.G. 1995. The ecology of marine benthic hydroids. Oceanography & Marine Biology: an Annual Review, **33**, 351-426.

GILLISON, A.N. 1984. Gradient orientated sampling for resource surveys – the gradsect method. In: MYERS, K.R., MARGULES, C.R. and MUSTO, I. (eds.) Survey Methods for Nature Conservation pp. 349–74. Proc. Workshop held at Adelaide Univ. 31 Aug. to 31 Sept. 1983. (CSIRO (Aust.) Division of Water and Land Resources, Canberra).

GRAY, J.S. 1990. Statistics and the precautionary principle. Marine Pollution Bulletin, **21**, 174-176.

GRAY, J.S. 1996. Environmental science and a precautionary approach revisited. Marine Pollution Bulletin, **32**, 532-534.

GREEN, R.H. 1979. Sampling Design and Statistical Methods for Environmental Biologists. John Wiley & Sons, New York.

GREEN, R.H. 1989. Power analysis and practical strategies for environmental monitoring. Environmental Research, **50**,195-205.

GROSS, J.E. 2003. Developing conceptual models for monitoring programmes. NPS Inventory and Monitoring Programme, USA. Available from: http://science.nature.nps.gov/im/monitor/docs/Conceptual_modelling.pdf [Accessed 03 June 2017].

GUERRA, M.T. & FREITAS, R. 2012. Recommended Operational Guidelines (ROG) for grab sampling and sorting, and treatment of samples. MeshAtlantic v1.4. Available from: http://www.emodnet-seabedhabitats.eu/pdf/MeshA_ROG_grab_sampling_v1.4.pdf [Accessed 03 June 2017].

HALPERN, B.S. 2003. The impact of marine reserves: do reserves work and does reserve size matter? Ecological Applications, **13**, 117-137.

HARTNOLL, R.G. 1975. The annual cycle of *Alcyonium digitatum*. Estuarine and Coastal Marine Science, **3**, 71-78.

HASTIE, T. & TIBSHIRANI, R. 1986. Generalized additive models. Statistical Science, **1** (3) 297-318.

HAYWARD, P.J. & RYLAND, J.S. 1998. Cheilostomatous Bryozoa. Part 1. Aeteoidea - Cribrilinoidea. Synopses of the British Fauna (New Series), BARNES, R.S.K. & CROTHERS, J.H. (eds) The Linnean Society of London. Shrewsbury: Field Studies Council. Synopses of the British Fauna, No. 10. 2nd Edition.

HEFFNER, R.A., BUTLER IV, M.J. & REILLY, C.K. 1996. Pseudoreplication revisited. Ecology, **77** (8), 2558-2562.

HIDDINK, J.G., JENNINGS, S., KAISER, M.J., QUEIRÓS, A.M., DUPLISEA, D.E. & PIET, G.J. 2006.Cumulative impacts of seabed trawl disturbance on benthic biomass, production, and species richness in different habitats. Canadian Journal of Fisheries and Aquatic Sciences, **63** (4), 721-736.

HILL, J. & WILKINSON, C. 2004. Methods for ecological monitoring of coral reefs. Australian Institute of Marine Science, Townsville, 117 pp.

HITCHIN, R., TURNER, J. & VERLING, E. 2015. Epibiota remote monitoring from digital imagery: operational guidelines. NE Atlantic Marine Biological Analytical Quality Control Scheme (NMBAQC). Available from: http://www.nmbaqcs.org/media/1591/epibiota_operational_guidelines_final.pdf [Accessed 03 June 2017].

HOLT, T.J., REES, E.I., HAWKINS, S.J. & SEED, R. 1998. Biogenic Reefs (Volume IX). An overview of dynamic and sensitivity characteristics for conservation management of marine SACs. Scottish Association for Marine Science (UK Marine SACs Project).

HOLTROP, G. & BREWER, M. 2013. Provision of statistical advice to the Marine Protected Sites monitoring project. JNCC Report No. 493.

HOPKINS, A. 2007. Recommended operating guidelines (ROG) for swath bathymetry. MESH Project. Available from: http://www.emodnet-seabedhabitats.eu/PDF/GMHM3_Swath_Bathymetry_ROG.pdf [Accessed 03 June 2017].

HOWSON, C.M. & DAVISON, A. 2000. Trials of monitoring techniques using divers and ROV in Loch Maddy cSAC, North Uist. Scottish Natural Heritage, Edinburgh.

HURLBERT, S.H. 1984. Pseudoreplication and the design of ecological field experiments. Ecological Monographs, **54** (2), 187-211.

JENKINS, C., NELSON, M., WHOMERSLEY, P., JOHNSON, G., CAMERON, A., BARRY, J., EGGLETON, J., CHURCH, N. & WEBB, K. 2015. Developing ecologically significant sampling units for fishing pressure using Vessel Monitoring System data for the purpose of planning broadscale benthic monitoring surveys. *JNCC/Cefas Partnership Report Series No.1*. Available from: http://jncc.defra.gov.uk/pdf/JNCC-CefasPartnershipReportSeries_No%201_FINALweb.pdf [Accessed 03 June 2017].

JENSEN, A.C., COLLINS, K.J., LOCKWOOD, A.P.M., MALLINSON, J.J. & TURNPENNY, A.W.H. 1994. Colonisation and fishery potential of coal waste artificial reef in the United Kingdom. Bulletin of Marine Science, **55**, 1263-1276.

JNCC. 2004a. Common standards monitoring guidance for marine features. Version August 2004, ISSN 1743-8160.

JNCC. 2004b. Common standards monitoring guidance for inshore sublittoral sediment habitats. Available from: http://jncc.defra.gov.uk/PDF/CSM_marine_sublittoral_sediment.pdf [Accessed 03 June 2017].

JNCC. 2004c. Common standards monitoring guidance for littoral rock and inshore sublittoral rock habitats. Available from: http://jncc.defra.gov.uk/PDF/CSM_marine_rock.pdf [Accessed 03 June 2017].

KAISER, M.J., CLARKE, K.R., HINZ, H., AUSTEN, M.C.V., SOMERFIELD, P.J. & KARAKASSIS, I. 2006. Global analysis of response and recovery of benthic biota to fishing. Marine Ecology Progress Series, **311**, 1–14.

KEOUGH, M.J. & MAPSTONE, B.D. 1995. Protocols for Designing Marine Ecological Monitoring Programs Associated with BEK Mills. National Pulp Mills Research Program, Technical Report No. 11. CSIRO, Canberra.

KINGSFORD, M. & BATTERSHILL, C. 1998. Studying temperate marine environments: a handbook for ecologists. Canterbury University Press, Christchurch, 335 pp.

KRÖGER, K. & JOHNSTON, C. 2016. The UK marine biodiversity monitoring strategy v4.1. Available from: http://jncc.defra.gov.uk/pdf/Marine_Monitoring_Strategy_ver.4.1.pdf [Accessed 03 June 2017].

LEGENDRE, P. & FORTIN, M-J. 1989. Spatial pattern and ecological analysis. Vegetatio, **80**, 107-138

LEWIS III, F.G. & STONER, A.W. 1981. An examination of methods for sampling macrobenthos in seagrass meadows. Bulletin of Marine Science, **31**,116-124

LIMPENNY, D.S., FOSTER-SMITH, R.L., EDWARDS, T.M., HENDRICK, V.J., DIESING, M., EGGLETON, J.D., MEADOWS, W.J., CRUTCHFIELD, Z., PFEIFER, S. & REACH, I.S. 2010. Best methods for identifying and evaluating *Sabellaria spinulosa* and cobble reef.

Aggregate Levy Sustainability Fund Project MAL0008. JNCC, Peterborough, 134 pp., ISBN - 978 0 907545 33 0.

LINCOLN, L.J., BOXHALL, G.A. & COX, P.F. 1982. A dictionary of ecology, evolution and systematic. Cambridge University Press.

LINDENMAYER, D.B. & LIKENS, G.E. 2010. The science and application of ecological monitoring. Biological Conservation, **143**, 1317-1328.

LONG, D. 2005. Recommended operating guidelines (ROG) for side scan sonar. MESH Project. Available from: http://www.emodnet-seabedhabitats.eu/PDF/GMHM3_Sidescan_Sonar_ROG.pdf [Accessed 03 June 2017].

MACARTHUR, R.H. & WILSON, E.O. 1967. The theory of island biogeography. Princeton, N.J. Princeton University Press.

MADDOX, D., POIANI, K. & UNNASCH, R. 1999. Evaluating management success: using ecological models to ask the right management questions. In SEXTON, W.T., MALK, A.J., SZARO, R.C. & JOHNSON, N.C. (eds.) Ecological Stewardship. Elsevier Science, Oxford, UK, 563-584

MAGGS, C.A. 1983. A seasonal study of seaweed communities on subtidal maerl (unattached coralline algae). Progress in Underwater Science, **9**, 27-40.

MANLEY, P., ZIELINSKI, W.J., STUART, C.M., KEANE, J.J., LIND, A.J., BROWN, C., PLYMALE, B.L. & NAPPER, C.O. 2000. Monitoring ecosystems in the Sierra Nevada: the conceptual model foundation. Environmental Monitoring and Assessment, **6**, 139-152.

MANLY, B.F.J. & NAVARRO, J.A. 2015. Introduction to ecological sampling. CRC Press, Florida.

MAPSTONE, B.D. 1995. Scalable decision rules for environmental-impact studies: effect size, type-I, and type-II errors. Ecological Applications, **5**, 401-410.

MILLAR, R.B. & ANDERSON, M.J. 2004. Remedies for pseudoreplication. Fisheries Research, **70**, 397-407.

MUNKITTRICK, K.R., ARENS, C.J., LOWELL, R.B. & KAMINSKI, G.P. 2009. A review of potential methods of determining critical effect size for designing environmental monitoring programs. Environmental Toxicology and Chemistry, **28** (7), 1361-1371.

NEHLS, G. & THIEL, M. 1993. Large-scale distribution patterns of the mussel *Mytilus edulis* in the Wadden Sea of Schleswig-Holstein: Do storms structure the ecosystem? Netherlands Journal of Sea Research, **31**, 181-187.

NUZZO, R. 2014. Statistical errors: P values, the 'gold standard' of statistical validity, are not as reliable as many scientists assume. Nature, **506**, 150-153.

OLEA, R.A. 1984. Sampling design optimization for spatial functions. Mathematical Geology, **16**, 369-392.

OSENBERG, C.W., SCHMITT R.J., HOLBROOK S.J., ABUSABA, K.E. & FLEGAL, A.R. 1994. Detection of environmental impacts: natural variability, effect size, and power analysis. Ecological Applications, **4**, 16-30.

OSENBERG, C.W., BOLKER, B.M., WHITE, J.S., ST MARY, C.M. & SHIMA, J.S. 2006. Statistical issues and study design in ecological restorations: lessons learned from marine reserves. In: FALK, D.A., PALMER, M.A. & ZEDLER, J.B. (eds.) Foundations of restoration ecology. Society for Ecological Restoration International, 280-302 pp.

OSENBERG, C.W., SHIMA, J.S., MILLER, S.L. & STIER, A.C. 2011. Assessing effects of marine protected areas: confounding in space and possible solutions, 144-167. In CLAUDET, J., ed. Marine Protected Areas: a multidisciplinary approach. Cambridge University Press.

OSPAR. 2004. OSPAR guidelines for monitoring the environmental impact of offshore oil and gas activities. Reference number: 2001-11.

OSPAR. 2012. MSFD Advice Manual and Background Document on Biodiversity: Approaches to determining good environmental status, setting of environmental targets and selecting indicators for Marine Strategy Framework Directive descriptors 1, 2, 4 and 6. Version 3.2 (5 March 2012). Prepared by the OSPAR Intersessional Correspondence Group on the Coordination of Biodiversity Assessment and Monitoring (ICG COBAM) under the responsibility of the OSPAR Biodiversity Committee (BDC), OSPAR Commission, London.

PARRY, M., TIERNEY, M., WOOD, L., STANWELL-SMITH, D., NORTHEN, K., ABDULLA, A., CORRIGAN, C., GASSNER, P. & FLETCHER, L. 2012. Review of international Marine Protected Area seabed monitoring and assessment of 'good practice' to inform application within UK waters. *JNCC Report No. 460*. JNCC, Peterborough. Available from: http://jncc.defra.gov.uk/pdf/jncc460_Web.pdf [Accessed 03 June 2017].

PETERMAN, R.M. 1990. Statistical power analysis can improve fisheries research and management. Canadian Journal of Fisheries and Aquatic Science, **47**, 2–15.

PETERMAN, R.M. & M'GONIGLE, A.B. 1992. Statistical power analysis and the precautionary principle. Marine Pollution Bulletin, **24**, 231-234.

PHILLIPS, G.R., ANWAR, A., BROOKS, L., MARTINA, L.J., MILES, A.C. & PRIOR, A. 2014. Infaunal quality index: WFD classification system for marine benthic invertebrates. Environment Agency Report SC080016.

POPPER, K. 1935. Logik der Forschung. Springer, Vienna, Austria.

PORTMANN, J.E. 2000. Review of current UK marine observation in relation to present and future needs. Interagency Committee for Marine Science and Technology, publ. No. 7.

QUINN, G.P. & KEOUGH, M.J. 2002. Experimental design and data analysis for biologists. Cambridge, U.K.

REISH, D.J. 1959. A discussion of the importance of screen size in washing quantitative marine bottom samples. Ecology, **40**, 307-309.

REISS, H. & KRÖNCKE, I. 2005. Seasonal variability of infaunal community structures in three areas of the North Sea under different environmental conditions. Estuarine, Coastal and Shelf Science, **65** (1-2), 253-274.

RUTHERFORD, A. 2011. ANOVA and ANCOVA: A GLM approach, 2nd Edition. Wiley & Sons.

RUXTON, G.D. & NEUHÄUSER, M. 2010. When should we use one-tailed hypothesis testing? Methods in Ecology and Evolution, **1** (2), 114-117.

RYLAND, J.S. 1976. Physiology and ecology of marine bryozoans. Advances in Marine Biology, **14**, 285-443.

SCHWARTZ, C.J. 2015. Analysis of BACI experiments. In Course Notes for beginning and intermediate statistics. Available from: http://www.stat.sfu.ca/~cschwarz/CourseNotes [Accessed 03 June 2017].

SHORT, F.T., IBERLINGS, B.W. & DEN HARTOG, C. 1988. Comparison of a current eelgrass disease to the wasting disease in the 1930s. Aquatic Botany, **30**, 295-307.

SIMONSOHN, U., NELSON, L.D. & SIMMONS, J.P. 2014. Journal of Experimental Psychology: General, **143** (2), 534-547.

SOKAL, R.R. & ROHLF, F.J. 1981. Biometry: The principles and practice of statistics in biological research, 2nd Edition. W.H. Freeman, San Francisco.

STEELE, B.M. 2001. Sampling design and statistical inference for ecological assessment. In: JENSEN, M.E. & BOURGERON, P.S. (eds.) A guidebook for integrated ecological assessments XIII, p. 536.

STEWART-OATEN, A., MURDOCH, W.W. & PARKER, K.R. 1986. Environmental impact assessment: “pseudoreplication” in time? Ecology, **67** (4), 929-940.

SUTHERLAND, W.J. 2006. Ecological census techniques: a handbook. Cambridge University Press.

TABACHNICK, B.G. & FIDELL, L.S. 2013. Using multivariate statistics. Pearson Education Limited

TAYLOR, B.L. & GERRODETTE, T. 1993. The uses of statistical power in conservation biology – the Vaquita and Northern Spotted Owl. Conservation Biology, **7**, 489–500.

THUILLER, W., BROTONS, L., ARAÚJO, M.B., LAVOREL, S. 2004. Effects of restricting environmental range of data to project current and future species distributions. Ecography, **27**, 165-172.

UNDERWOOD, A.J. 1990. Experiments in ecology and management: their logics, functions and interpretations. Australian Journal of Ecology, **15**, 365-389.

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

UNDERWOOD, A.J. 1997a. Environmental decision-making and the precautionary principle: what does this mean in environmental sampling practice? Landscape and Urban Planning, **37**, 137-146.

UNDERWOOD, A.J. 1997b. Experiments in ecology: Their logical design and interpretation using analysis of variance, Cambridge, U.K.

UNDERWOOD, A.J. & CHAPMAN, M.G. 2013. Design and analysis in benthic surveys in environmental sampling. In ELEFTHERIOU, A. (ed.) Methods for the study of marine benthos. John Wiley & Sons, Chichester.

VAN DENDEREN, P.J., BOLAM, S.G., HIDDINK, J.G., JENNINGS, S., KENNY, A., RIJNSDORP, A.D. & VAN KOOTEN, T. 2015. Similar effects of bottom trawling and natural disturbance on composition and function of benthic communities across habitats. Marine Ecology Progress Series, **541** (31), 31-43.

VAN DER MEER, J. 1997. Sampling design of monitoring programmes for marine benthos: A comparison between the use of fixed versus randomly selected stations. Journal of Sea Research, **37** (1–2), 167–179.

WALTHER, B.A. & MOORE, J.L. 2005. The concepts of bias, precision and accuracy, and their use in testing the performance of species richness estimators, with a literature review of estimator performance. Ecography, **28**, 815 - 829.

WARWICK, R.M. 1986. A new method for detecting pollution effects on marine macrobenthic communities. Marine Biology, **92**, 557-62.

WASSERSTEIN, R.L. & LAZAR, N.A. 2016. The ASA's Statement on p-Values: Context, Process, and Purpose. The American Statistician, **70** (2), 129-133

WESSELS, K.J., VAN JAARSVELD, A.S., GRIMBEEK, J.D. & VAN DER LINDE, M.J. 1998. An evaluation of the gradsect biological survey method. Biodiversity & Conservation, **7** (8), 1093-1121.

WOOD, S. 2006. Generalized Additive Models: An introduction with R. Chapman & Hall/CRC Texts in Statistical Science.

WILDING, T.A., NICKELL, T.D., HUGHES, D.J., NARAYANASWAMY, B.E., BURROWS, M.T. & HAUSRATH, J. 2015. Statistical advice to the Marine Habitats Monitoring project under Framework Agreement C10-206-0387. JNCC Report No. 545. Available from: http://jncc.defra.gov.uk/pdf/545_web.pdf [Accessed 03 June 2017].

WILSON, D.P. 1970. The larvae of *Sabellaria spinulosa* and their settlement behaviour. Journal of the Marine Biological Association of the UK, **50**, 33-52.

WILSON, D.P. 1971. Sabellaria colonies at Duckpool, North Cornwall, 1961-1970, Journal of the Marine Biological Association of the UK, **51**, 509-580.

ZUUR, A., IENO, E.N. & SMITH, G. 2007. Analysing ecological data. Springer-Verlag, New York.

ZUUR, A.F., IENO, E.N., WALKER, N.J., SAVELIEV, A.A. & SMITH, G.M. 2009. Mixed effects models and extensions in R. Springer

ZUUR, A.F., IENO, E.N. & ELPHICK, C.S. 2010. A protocol for data exploration to avoid common statistical problems. Methods in Marine Ecology and Evolution, **1** (1), 3-14.

Annex I: Sources of existing UK data

Data Source	Description	Link
UK Government open data portal	Online repository for data published by government departments and agencies, public bodies and local authorities. Includes a wide range of marine datasets from organisations such as Environment Agency, JNCC, Cefas, Environment Agency, Natural England, UK Hydrographic Office, British Geological Survey.	https://data.gov.uk/
Marine Recorder & Snapshot	Database application used by JNCC, SNCBs and other organisations to store marine benthic sample data, such as species, physical attributes and biotopes. Marine Recorder is fully compatible with the National Biodiversity Network (NBN) data model. Data extracted from a Marine Recorder database into a queryable format is known as a Marine Recorder Snapshot.	http://jncc.defra.gov.uk/page-1599
Statutory Nature Conservation Body (SNCB) Interactive Mapping Portals	Spatial datasets can be downloaded from various online portals maintained by UK SNCBs:	
	United Kingdom: JNCC Interactive Map of Marine Protected Areas	http://jncc.defra.gov.uk/page-5201
	Great Britain: MAGIC mapper	http://magic.defra.gov.uk/
	Scotland: National Marine Plan Interactive	http://www.gov.scot/Topics/marine/seamanagement/nmpihome
	Wales: Wales Marine Planning Portal Lle Geo-Portal	http://lle.gov.wales/apps/marineportal/ http://lle.gov.wales/home?lang=en
National Biodiversity Network (NBN) Gateway	Database and interactive mapping tool collating and making accessible information stored by the Biological Records Centre (BRC). The Gateway provides access to >100 million terrestrial, freshwater and marine species records from >100 data providers.	https://data.nbn.org.uk/
UK Directory of the Marine-observing Systems (UKDMOS)	Searchable meta-database holding information on marine monitoring programmes across a range of more than 45 organisations and is maintained and updated by MEDIN, with data being stored by MEDIN Data Archive Centres (DACs).	http://www.ukdmos.org/v_ukdmos_edios_v2/search.asp
European Marine Observation and Data Network (EMODnet) portal	The portal provides substrate maps, habitat maps, bathymetry data and a range of other biological, geological and physical parameters for western European waters.	http://emodnet.eu
British Geological Survey (BGS) Offshore GeoIndex	Application providing access to geological data, including sediment sample data.	http://www.bgs.ac.uk/GeoIndex/home.html
The National Network of Regional Coastal Monitoring Programmes	The Network consists of six regional monitoring programmes which collect and distribute data to underpin evidence-based decisions regarding flood and coastal erosion risk management. Provides open access to aerial photography,	http://www.channelcoast.org/

Data Source	Description	Link
	swath bathymetry and other coastal data such as Digital Terrain Models and sediment distribution maps	
Crown Estate Marine Data Exchange	The exchange provides access to survey data and reports collated during the planning, building and operating of offshore renewable energy projects.	http://www.marinedataexchange.co.uk
UK Oil & Gas Data	Information on oil and gas exploration and development licenses	https://www.ukoilandgasdata.com/dp/controller
UK Benthos	The UK Benthos database and desktop application are available on the Oil & Gas UK (formerly UKOOA) website, and holds biological and physico-chemical data from >600 baseline and monitoring surveys within the UK Exclusive Economic Zone (EEZ).	http://www.oilandgasuk.co.uk/knowledgecentre/uk_benthos_database.cfm
Strategic Environmental Assessment (SEA) data portal	Data portal administered by BGS, which provides access to a wide range of environmental data acquired and collated to inform SEAs.	http://www.bgs.ac.uk/data/sea/home.html
Marine Aggregate Regional Environmental Assessment (REA) document repository	Data portal providing the results of REAs conducted to describe the baseline environmental characteristics in aggregate licensed areas, and evaluate the potential cumulative and in-combination effects of existing and planned dredging operations.	http://www.marine-aggregate-rea.info/

Annex II: Abbreviations and Glossary

Abbreviations

ABC	Abundance-Biomass Comparison
ANOSIM	Analysis of Similarity
ANOVA	Analysis of Variance
BA	Before-After design
BACI	Before-After-Control-Impact design
BACIPS	Before-After-Control-Impact Paired Series design
BGS	British Geological Survey
BRC	Biological Records Centre
BTA	Biological Traits Analysis
CEM	Conceptual Ecological Model
CI	Control-Impact design
DAC	Data Archive Centres
EEZ	Exclusive Economic Zone
ES	Effect size
EMODnet	European Marine Observation and Data Network
GAM	Generalized Additive Model
GES	Good Environmental Status
GIS	Geographical Information System
GLM	Generalized Linear Model
GLMM	Generalized Linear Mixed Model
HBDSEG	Healthy and Biologically Diverse Seas Evidence Group
ICES	International Council for the Exploration of the Sea
ISO	International Organisation for Standardization
JNCC	Joint Nature Conservation Committee
LIDAR	Light Detection and Ranging
MANOVA	Multivariate Analysis of Variance
MCZ	Marine Conservation Zone
MDAC	Methane-Derived Authigenic Carbonate
MDS	Multidimensional Scaling
MEDIN	Marine Environmental Data and Information Network
MESH	Mapping European Seabed Habitats
MPA	Marine Protected Area
MSFD	Marine Strategy Framework Directive
NCMPA	Nature Conservation Marine Protected Area
NGO	Non-Governmental Organisation
NMBAQC	National Marine Biological Analytical Quality Control Scheme
OSPAR	Oslo-Paris Convention
PCA	Principal Components Analysis
PERMANOVA	Permutation-based ANOVA
PRIMER	Plymouth Routines in Multivariate Ecological Research
PSD	Particle Size Distribution
R&D	Research and Development
ROG	Recommended Operational Guidelines
SAC	Special Area of Conservation
SCI	Site of Community Importance
SEA	Strategic Environmental Assessment
SIC	Site Information Centre
SIMPER	Similarity Percentages analysis
SIMPROF	Similarity Profile analysis
SNCB	Statutory Nature Conservation Body

SPUE	Sightings-per-unit-effort
UKMBMP	UK Marine Biodiversity Monitoring Programme
UKDMOS	UK Directory of the Marine-Observing Systems
UKMMAS	UK Marine Monitoring and Assessment Strategy
VMS	Vessel Monitoring System
WFD	Water Framework Directive

Glossary

Accuracy	The closeness of a measurement or estimate to the true value of the population, as related to the bias and precision of the measurement.
Bias	The difference between a measured (sample) population mean and an accepted true population value.
Dependence	A condition in which two random variables, or sampling units, are not independent of each other (i.e. the occurrence of one affects the other).
Effect size (ES)	The magnitude of an effect on a response variable.
Experimental unit	One member of a set of objects (e.g. discrete sampling areas) that are initially equivalent, with each object then subjected to experimental treatments.
Factor	An explanatory (predictor) variable which has two or more levels (or categories).
Independence	A condition in which two random variables, or sampling units, are independent of each other (i.e. the occurrence of one does not affect the other).
Indicator	'…any measurable feature or condition of the marine environment that is relevant to the stability and integrity of habitats and communities, the sustainability of ecosystem goods and services (e.g. primary productivity, maintenance of food chains, nutrient cycling, biodiversity), the quality and safety of seafood, and the status of amenities of socio-economic importance.' (OSPAR, 2012).
Inference	The process of deducing properties of an underlying population by analysis of sample data.
Inshore	The area of sea and seabed between the mean high water spring tide, and 12 nautical miles from the mean high water spring tide.
Interaction	The presence of a significant interaction indicates that the effect of one predictor variable on the response variable is different at different levels of the other predictor variable.
Investigative monitoring	Monitoring to investigate the cause of change. This type of monitoring is conducted to determine management needs and effectiveness, and includes manipulative experiments.
Judgement sampling	A type of non-random sampling that is designed based on the opinion of an expert.
Monitoring	An activity by which evidence necessary to meet the aims of the monitoring programme is collected (UKMBMP definition, Kröger & Johnston 2016).

Noise	Unexplained variation or randomness in a dataset.
Observation	The value of a variable taken from a specific sampling unit.
Operational monitoring	Monitoring to measure state and relate observed change to possible causes, through investigation of pressure-state relationships.
Offshore	The area of sea and seabed between 12 nautical miles from the mean high water spring tide and the limit of the Exclusive Economic Zone.
Population	A collection of elements, objects or organisms of interest, to which the findings of a study are extrapolated.
Power (1-β)	The probability that a test will reject the null hypothesis when it is false. Power is inversely related to β, or the probability of making a Type II error.
Power analysis	A technique used to calculate the sample size needed to detect a given effect size (ES), where the degree of variance is predicted, and levels of power (1-β) and significance (α) are specified.
Precision	The degree of concordance among a number of measurements or estimates for the same population, precision is reflected by the variability of an estimate.
Predictor variable	An independent variable that represents causes of variation in the response variable.
Pressure	An adverse environmental effect caused by human activities.
Pressure unit	A standardised experimental unit within which the intensity of a specific pressure is known.
Response variable	A variable of interest in monitoring, i.e. an indicator metric, which may be affected by predictor variables. Also referred to as the dependent variable.
Sample (N)	A part of a population, or subset from a set of sampling units, about which generalised conclusions can be drawn about the population by inference.
Sampling unit	A sampling unit is one of the units into which an aggregate (i.e. a population) is divided for the purpose of sampling, each unit being regarded as individual and indivisible when the selection is made.
Sentinel monitoring	Monitoring to measure the rate and direction of long-term change.
Serial correlation	The relationship between a given variable and itself over various time intervals. Also referred to as temporal autocorrelation.
Significance level (α)	The probability of rejecting the null hypothesis when it is true, or of committing a Type I error.
Simple random sampling	A method of sampling design where sampling points are randomly distributed within a survey area.
Spatial autocorrelation	The positive or negative correlation of a variable with itself through space.
Strata	The divisions into which a population can be separated to increase precision in a sampling design, e.g. different habitat types.
Stratified random	A method of sampling design where sampling points are randomly

sampling	distributed within strata representing different environmental conditions.
Systematic sampling	A method of sampling design where sampling points are distributed using a fixed periodic interval.
Treatment	A combination of different factor levels in an experiment.
Type I error	The incorrect rejection of the null hypothesis when it is true (a false positive).
Type II error	The incorrect acceptance of the null hypothesis when it is false (a false negative).
Variance (σ)	The distribution of data around their mean value.