JNCC Report

No: 545

Statistical advice to the Marine Habitats Monitoring project under Framework Agreement C10-206-0387

Wilding, T.A., Nickell, T.D., Hughes, D.J., Narayanaswamy, B. E., Burrows, M.T., & Hausrath, J.

March 2015

ISSN 0963 8901

For further information please contact:

Joint Nature Conservation Committee
Monkstone House
City Road
Peterborough PE1 1JY
http://jncc.defra.gov.uk

This report should be cited as:

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 Number 545, JNCC, Peterborough.

Acknowledgements and Data Intellectual Property Rights (IPR)

The surveys carried out in 1996 and 1998 were undertaken on behalf of the Atlantic Frontier Environmental Network (AFEN). AFEN comprises: Agip (UK) Ltd, Amerada Hess Ltd, Amoco (UK) Exploration Company Ltd, ARCO British Ltd, BG E&P Ltd, BP Exploration Operating Company Ltd, Chevron UK Ltd, Conoco (UK) Ltd, Deminex UK Oil and Gas Ltd, Elf Exploration UK plc, Enterprise Oil plc, Esso Exploration and Production UK Ltd, Fina Exploration Ltd, Marathon Oil UK Ltd, Mobil North Sea Ltd, Phillips Petroleum Company UK Ltd, Saga Petroleum Ltd, Shell UK Exploration and Production, Statoil Ltd, Texaco Britain Ltd, Total Oil Marine plc, Joint Nature Conservation Committee, Fisheries Research Services (now Marine Scotland), and the Department of Trade and Industry (now the Department for Business, Innovation and skills). The survey project was scoped and agreed between the Department of Trade and Industry, SOAEFD, the Southampton Oceanography Centre and the AFEN consortium, and was commissioned and funded by the AFEN consortium.

All remaining data were acquired during surveys undertaken as part of the UK Department of Energy and Climate Change's (DECC) Energy Strategic Environmental Assessment programme. All intellectual property rights (including, without limitation, copyrights, database rights and all other rights which subsist or may at any time in the future subsist in the Dataset(s)) in the Dataset(s) ('Intellectual Property Rights') are owned by DECC (formerly the Department of Trade and Industry, and the Department for Business, Enterprise and Regulatory Reform).

The infaunal taxa abundance data were collated and prepared by SAMS as part of JNCC contract C13-0223-0670 'Definition of deep-sea infaunal assemblages for inclusion in a deep-sea section of the Marine Habitat Classification for Britain and Ireland' (Hughes *et al.*, 2014). The processed epifaunal data have been provided to JNCC by Plymouth University as part of JNCC contract C13-0223-0670 'Definition of epifaunal assemblages for inclusion in a deep-sea section of the Marine Habitat Classification for Britain and Ireland' (Piechaud and Howell, 2013).

Summary

The Joint Nature Conservation Committee (JNCC) is leading a partnership of the Statutory Nature Conservation Bodies working on developing options for an integrated approach to monitoring marine biodiversity across UK waters as part of a strategy to fulfil national and international monitoring and assessment obligations. There is increasing concern with regard to deep-sea biodiversity change, particularly as a consequence of activities such as commercial fishing and via anthropogenic climate change. There is a requirement to assess the direction and the magnitude of this biodiversity change.

JNCC has access to a number of related benthic data sets from the Scottish Continental Slope and Faroe-Shetland Channel. They consist of counts of epifauna identified from video, collected through transects over the seabed in 2006, and of infaunal macrobenthic counts, identified to family level, derived from remotely collected sediment samples taken biennially over the period 1996-2002.

This JNCC report is based on the findings of a research contract to:

1. Identify the most appropriate response metric (as derived from the multivariate data) and method of spatially distinguishing (stratifying) differing areas within the sampled areas and to use parameter estimates based on these to determine the likely statistical power of a range of sampling scenarios

2. Examine the relationship between inter-sample distance and similarity in terms of measured metrics.

3. Comment on any data limitations

4. Make recommendations regarding future monitoring programmes.

The authors conclude that the epifaunal data analysed were insufficient to support power analysis in relation to change over time. A further analysis of these data is recommended. In terms of the macrobenthic data, it was found that the Number of Families and Pielou's Evenness were the optimal univariate measures for describing the assemblage structure and identified two optimal stratification methods from which the degree of change, over time, between strata, and the unexplained variability in the response metrics, were determined (these were the necessary parameters for power analysis). However, the data analysed cannot be used to inform monitoring strategies capable of quantifying long-term trends, over the entirety of the study area. In order to assess such changes historical data, typically of 50-200 years, would be required.

The statistical power of 'before-after-control-impact' (BACI) designs could be evaluated using the current data. The data were characterised by high variability (which varied between response metric and stratification system) over space and time. Power curves were produced for a diverse range of parameters and indicated relatively low power given the parameters estimated from these data. It is recommended that macrobenthic samples be taken with a separation distance of at least 20km when characterising the area, and that BACI designs should only be employed when the appropriate spatio-temporal variability can be shown to be sufficiently low. In order to better inform BACI designs, and estimate their power, it is recommended that sampling be conducted to quantify small-scale temporal and spatial variability in the response metrics. More generally, assessments should be undertaken to determine the influence of sampling-methodology on derived diversity/evenness metrics. Further consideration should be given to the efficacy of

diversity/evenness metrics derived from family-level identification in assessing change in the deep-sea.

Contents

1 General introduction

1.1 Background

The Joint Nature Conservation Committee (JNCC) is leading a partnership of the Statutory Nature Conservation Bodies (SNCBs), working on developing options for an integrated approach to monitoring marine biodiversity across UK waters. This is part of the UK Marine Monitoring and Assessment Strategy (UKMMAS) and aims to encompass existing policy and statutory obligations for monitoring and assessment, such as those required under the European Habitats and Birds Directives and the EU Marine Strategy Framework Directive (MSFD), in the most effective and cost-efficient manner.

The driving objective behind the integrated UK biodiversity monitoring programme is to provide the necessary evidence to support timely and scientifically robust advice for the management of human activities, as well as to fulfil the national and international obligations for monitoring and assessment. To achieve this need, the monitoring scheme requires the selection of metrics that are capable of separating impacts from human activities from naturally occurring cycles and trends. At the same time, a sampling design is needed that provides representative estimates of the selected metrics.

As part of the work to develop monitoring options for UK biodiversity, JNCC is currently looking to develop sampling design options to measure the status, and the rate and direction of long-term change in the condition of bathyal habitats along the Scottish Continental Slope and within the Faroe-Shetland Channel (FSC).

To achieve this objective, JNCC requires a sampling design that provides robust information to distinguish directional trends (natural and human-induced) from short-scale variability in space and time, and hence allows background variation in habitat condition to be measured so that any change detected can be put within the context of the natural variation of the system.

The aim of the research described in this report was to assess whether and how existing data collected in the UK deep-sea environment can be used to inform decisions on sampling designs to monitor the status and trends in the condition of benthic habitats along the Scottish Continental slope and within the Faroe-Shetland Channel.

The JNCC set four objectives for this project and these are summarised here, as amended following discussions with JNCC.

1. Identify what can be extracted from existing data to help identify the optimal stratification methodology and identify the optimal diversity/evenness index.
2. Identify the optimal sampling design, within the proposed strata for the metrics of interest.
3. Suggest sampling designs based on the optimal stratification system and the metrics of interest identified above and compare their statistical power.
4. Provide recommendations with regards to the optimal sampling design given the issues identified during the process.

1.2 Data

JNCC has access to large infaunal and epifaunal datasets derived from data collected during a number of surveys. These datasets, used in the present analysis, were collected to the West of Shetland and North of Shetland along the continental slope and the FSC. Macrobenthic infaunal samples were collected by the National Oceanography Centre, Southampton, in 1996, 1998, 2000 and 2002. The macrofauna were collected using either a Day Grab, Box Corer or Megacorer, depending on sediment type. The sediment samples were washed over a 0.5mm mesh sieve and the retained fauna were identified, where possible (Bett, 2001; Narayanaswamy *et al.*, 2005; Narayanaswamy *et al.*, 2010). The epibenthic megafauna were sampled in 2006 by the University of Plymouth along a series of transects using a Seatronics drop frame system with a Kongsberg 5 megapixel digital stills camera (Howell *et al.*, 2010). The sample locations and years collected are shown in Figure 1.

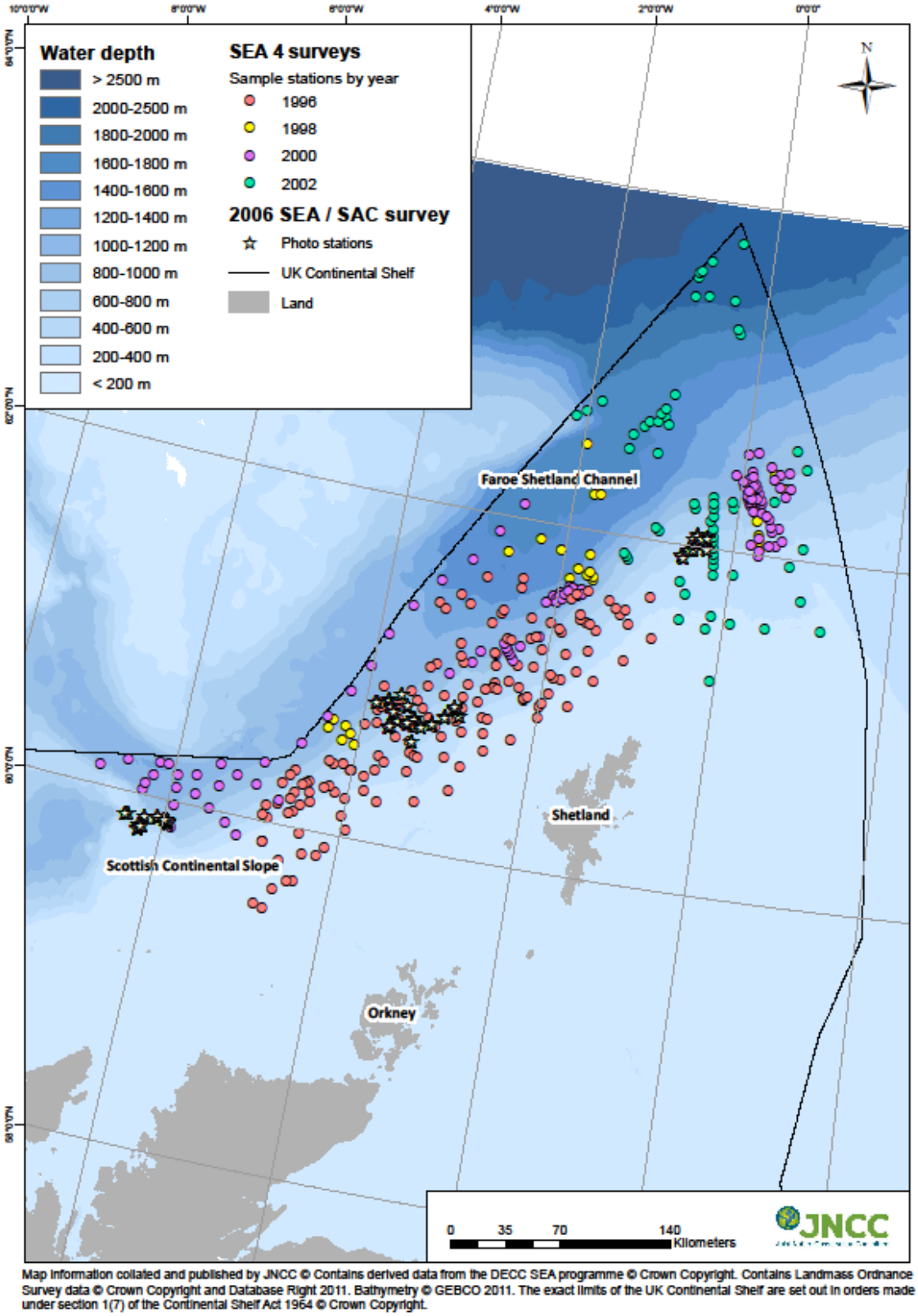

Figure 1. Locations of benthic infaunal and epifaunal samples collected along the Scottish Continental Slope and within the Faroe-Shetland Channel included in this report. Infaunal data were collected during surveys between 1996 and 2002; epifaunal data were obtained during the 2006 SEA/SAC survey; Number of locations (N) = 366.

1.3 Benthic monitoring and diversity indices

The investigation and monitoring of sedimentary habitats involves passing samples of sediment, obtained using a remotely deployed grab or coring device from the sample location, through a sieve (usually 0.5 or 1.0mm mesh size to investigate macrofaunal assemblages) and the subsequent identification and enumeration of the retained organisms (Gage and Bett, 2005). Covariates (e.g. environmental data) are often simultaneously recorded, the motivation being to link any observed patterns in response variable (e.g. macrofaunal abundance, biomass or diversity) with environmental drivers. In the present context these can include water depth, water temperature and latitude. Multivariate data, as here, consisted of counts of numerous species at each sampling location. Such data can be visualised and interpreted using a range of multivariate techniques (Clarke, 1993; Shaw, 2003) which aim to assess biotic similarity between samples and potentially relate these to environmental drivers (Shaw, 2003). The results from multivariate analyses are interpreted as 'associations', because such observational studies are seldom 'inferential' in the true statistical sense (manipulative studies are required to prove causation - and these are difficult to undertake in the deep-sea). Multivariate techniques are relatively sensitive (exhibit high-power) and are able to identify differences in assemblage composition between sample locations and/or times better compared with their univariate counterparts (see below).

Multivariate data can be converted to a 'single-number summary' which can be followed, over time, to assess change. In terms of single-number summaries of assemblages there are numerous options available but, in general, they consist of richness and diversity or evenness indices (Magurran, 1988). These indices measure different aspects of the original data; at the most basic of definitions, the diversity of a sample is simply the number of species present (richness). Measures of diversity that incorporate more information are commonly used to compare samples. Some of these indices attempt to measure how individuals are distributed among species, and not merely the richness, although heterogeneity indices contain both metrics. Other indices attempt to measure the degree of phylogenetic separation present in the taxa comprising a sample. The conversion of multivariate data to univariate metrics is a simplification which, inevitably, loses information and analytical sensitivity. However, univariate measures (e.g. diversity per sample) have the advantage that they can be used in a linear modelling framework in order to quantify the relative importance of any modelled environmental drivers (via parameter estimates). The conversion to a univariate measure also enables power analysis and can therefore more easily inform sampling design. Multivariate power analysis is available for data that can be analysed using MANOVA; that is, where the underlying distribution of the response variables is multivariate normal (Gaussian). However, assemblage data (i.e. the data we have here) cannot be analysed using MANOVA (because of the dominance of zeros counts and the subsequent non-Gaussian distribution) and, currently, no techniques are available for conducting multivariate power analysis on assemblage data.

1.4 Change in the deep-sea

The deep-sea will, inevitably, change over time as part of natural and anthropogenic trends occurring at a range of temporal and spatial scales. For the purposes of this research, these natural changes are those that would have occurred in the absence of man, particularly in relation to relatively recent industrial activities which are changing global-scale systems. The concept of no-change (with either 'natural' and/or 'anthropogenic' drivers) is not plausible in terms of impact monitoring, as change is a natural part of global systems (Schmitt and Osenberg, 1996; Johnson, 1999; Anderson *et al.*, 2000; Gigerenzer, 2004; Nakagawa and Cuthill, 2007). Accepting that changes are inevitable makes the standard

hypothesis test, which has underpinned much of modern biological and ecological literature, redundant (Gigerenzer, 2004). The standard null hypothesis (H_0) in many monitoring programmes is generally in terms of 'H_0 – no change (from whatever cause) has occurred' and the 'objective' of many monitoring programmes is to gather data to determine the extent of evidence against this null hypothesis (which is inevitably false) (Schmitt and Osenberg, 1996). However, 'no evidence of change is *not* evidence of no change' and change will always be detected (in a monitoring programme) if enough sampling effort is applied (Schmitt and Osenberg, 1996). In monitoring, the failure to reject the null hypothesis (of no change) is a 'Type II error' and merely means the sampling protocol adopted had insufficient 'power' (Schmitt and Osenberg, 1996). Power analysis is an assessment of how likely a Type II error is likely to occur, that is, how likely differences that are actually occurring will be detected.

2 Macrobenthic infaunal analysis

The macrobenthic infaunal analysis formed the bulk of the presented analysis. The data were collected from the sample locations shown in Figure 1, which defines the sampling domain and, consequently, the spatial extent to which any inferences from this work could apply. All the macrofaunal sampling cruises were led by the National Oceanography Centre (NOC), Southampton. The first two research cruises in 1996 and 1998 were undertaken on behalf of the Atlantic Frontier Environmental Network (AFEN), a consortium of oil companies, UK government environmental advisers (Fisheries Research Services, JNCC) and the UK's Department of Energy and Climate Change (DECC) (formerly known as the Department for Trade and Industry). A further two research cruises, in 2000 and 2002, were undertaken as part of DECC's Strategic Environmental Assessment (SEA) process.

The contract required the identification of the optimal response metric and stratification system with consideration of spatial and power analysis. The process of identifying the optimal response metric and sample stratification system is described in the next section together with the method by which the spatial and power analysis were conducted.

2.1 Methods

2.1.1 Response metrics

The first objective was to select diversity metrics that were stable, informative and easily interpretable. JNCC suggested a range of response metrics, which can be divided into two categories: 1) diversity measures and 2) evenness measures. Of these candidate metrics, some were omitted from analyses for two primary reasons (i) correlation with other metrics, and (ii) they were inappropriate for the taxonomic resolution of the data. The Berger-Parker diversity index was excluded because, being based on abundance, it was likely to be correlated with other abundance-based metrics. Hill's Index (as described in Heip *et al.*, 1998) gives either species richness or Shannon Weiner H' depending on the order of the index, and thereby the information provided is captured by the other indices calculated. Chao's index was not selected because its reliance on rare species made it inappropriate to apply it to the available family level data. Rarefied species counts were not undertaken as this process requires sampling from the same location (to account for the relationship between effort and the metric being derived). In the present case there was no replication of samples from the same location.

The following indices were calculated for the benthic data sets supplied:

- Margalef's *D*, a measure of species richness, $\frac{(S-1)}{\ln N}$, where *N* is the total number of individuals and *S* is the number of species;

- Shannon's *H'* (a measure of diversity, $-\sum\left(P_i \times ln\, P_i\right)$, where P_i is the fraction of the i_{th} species of the total population);

- Pielou's *J'*, a measure of evenness, **Error! Bookmark not defined.** $\frac{H'}{H'_{\max}}$, where *H'* is the Shannon index, $H'_{max} = ln\ S$ and *S* is the total number of species); Brillouin's H_B (an abundance measure, $[\ln(N!) - S \ln \frac{(n_i!)}{N}]$, where *N* is the total number of individuals, and n_i is the number of individuals of the i_{th} species; and

- Simpson's 1-λ' (a diversity measure of both species and abundance, $1 - \frac{\sum_{i=1}^{S} n_i(n_i-1)}{N(N-1)}$, where *S* is the number of species, *N* is the total number of individuals, and n_i is the number of individuals of the i_{th} species).

- Taxonomic distinctness Δ^+ (Clarke and Warwick, 1998) is a diversity measure of the length of Linnean phylogenetic separation between taxa in a sample $\Delta^+ = \left[\Sigma\Sigma_{i<j}\omega_{ij}\right]/[m(m-1)/2]$, where *m* is the number of species and ω_{ij} the path length between species *i* and *j*.

The list of metrics selected for analysis, and their abbreviations are shown in Table 1.

Table 1. Selected metrics for analysis and abbreviations.

Name	Abbreviation
Number of families	NosFam
Margalef's d	Marg
Pielou's J	Piel
Brillouin's H_B	Brill
Fisher's α	Fish
Shannon's H' (log_e)	Shan
Simpson's 1-λ'	Simp
Taxonomic distinctness Δ^+	Tdi.

Diversity indices were calculated for each sample using PRIMER v 6.1.6.[1] Abundance data from separate cruises were provided by JNCC in Excel format. Older versions of Excel are limited to 65,536 rows by 256 columns. Due to the large numbers of columns (stations) encountered when working with combined data sets, Excel 2010 (Windows 7) or Excel 2011 (Mac) were used to compile the data (1,048,576 rows by 16,384 columns). As the data were provided with family names as column headings, and sample location names as row

[1] www.primer-e-.com

headings, the 'copy and paste special' function was used and the data transposed into a new sheet.

PRIMER allows the importation of Excel data, or creation of new data files by copying and pasting from the Excel spreadsheet. The copy/paste method was used on the transposed data, after counting the number of columns and rows the abundance data occupied in Excel, and this value entered in the relevant areas in the PRIMER dialog box. The abundance data button was selected.

Labels for sample locations and variables (families) were entered at this stage, by copying and pasting from the Excel abundance spreadsheet (N.B. as sample location labels were at the top of columns in the transposed Excel data, they were transposed into a new excel sheet before copying and pasting into PRIMER).

The PRIMER Diverse routine was selected from the Analyse menu to obtain number of taxa, individual abundance, Margalef's d, Pielou's J, Brillouin's H, Fisher's α, Shannon's H' ($\log_e$), Simpson's 1-λ', and taxonomic distinctness (Δ*). The taxonomy master data aggregation was created from the WoRMS database, which provided taxonomic classification from Family to Kingdom of the abundance data.[2] This aggregation file was created by selecting File/New in PRIMER and selecting Aggregation data, and entering the number of columns and rows on the next dialog page. These numbers corresponded to the taxonomic classification from Kingdom to Family of the abundance data, which were then copy/pasted from Excel into the new PRIMER aggregation file. Labels were also copied (e.g. Kingdom, Phylum, etc.). The diversity indices thus calculated were set to output to Excel worksheet format.

The diversity indices (untransformed) were correlated and scatter plots for each combination were produced (using the pairs function in R). The assessment of correlations was done visually to assess information-redundancy (one of either of the highly correlated variables contains all the information when used as a response variable, and so there is no reason to include both). The objective was to eliminate one of each pair of metrics, choosing the metric that was either most intuitive (to simplify interpretation) or most commonly used (which would help to explain the relevance of the metric in any future literature reviews, e.g. into what constitutes a meaningful change). Potential response metrics were then plotted against levels of different strata (see below) to assess the presence of outliers. In order to select a response variable, it must not be characterised by outliers, as outliers skew parameter estimates that are required for the BACI power analysis.

2.1.2 Development and testing of optimum stratification system

The objective of this analysis was to identify the optimal stratification system, that is, the system that best partitions (accounts for) the variance between the predictors in the model. In the current case, the predictors are stratum, time and the stratum time interaction.

In order to monitor change over time it is necessary to understand how the metric under study varies over time and how this variation differs over space. This temporal-spatial variation is an 'interaction' term and, in order to assess statistical power, must be estimated from the data. The assessment of change over the entire spatial domain (using these objectives) requires sampling over the entire spatial domain (see recommendations in Section 5). Sampling over the entire domain can be purely random (i.e. taking no account of depth or location) or within pre-defined groupings (stratified). The advantage of sampling

[2] http://www.marinespecies.org/aphia.php?p=match

within a stratified system is that the variance, per grouping, can be reduced (depending on the utility of the stratification system employed). The disadvantage of sampling within strata is that, necessarily, each stratum-level must be sampled in order to have covered the entire spatial domain and, for a given maximum number of possible samples (e.g. within a budget-limited scenario), the effort has to be split between all the strata-levels.

In the present context, power analysis requires estimates of how the metric being studies interest varies over time and how this varies in space (the interaction term). A stratification system which maximises the partitioning in space and time (i.e. accounts for as much of the variance as possible) is necessarily superior. An assessment of the ability of each stratification system supplied was undertaken by running a 2-factor analysis of variance (ANOVA) with NosFam and Piel as the response variables and stratum, time and the stratum x time interaction terms as the predictors. A major caveat in this necessary approach is that, for some stratification systems, there are nil, or very few data for some stratification level-year combinations. This does not preclude the ANOVA but it does reduce the confidence in the parameter estimates. The reliability of the parameter estimates increases for stratification systems with fewer levels because the number of replicates, per level, is greater.

The extent of spatio-temporal 'overlap' in the data depends on the resolution of the spatial stratification used. There are numerous ways of classifying the macrobenthic infaunal sampling locations and the JNCC supplied eight different stratification systems for testing (Table 2). Different stratification systems could be created by combining existing stratification systems. Discussions within the SAMS team suggested that a stratification system should, at a minimum, include sample location and depth information (from an ecological perspective) but that the optimal stratification approach should maximise the number of replicates per level (from a statistical perspective). The requirement to maximise the replicates is to increase the robustness (reliability) of any parameter estimates subsequently derived using that stratification system. Further detail is provided in Table 2.

Table 2. Brief description of stratification systems. NL - number of levels[3]

Name	Brief description/ source	NL
S1	Simple depth-based classification. Based on MSFD predominant habitat definitions	3
S2	Depth: substratum type classification. Based on depth zones of S1 and sediment distribution as per JNCC working definition of rock/sediment habitats.	5
S3	Depth: substratum type classification. Based on depth zones of S1 and sediment distribution as per UK SeaMap substrate layer.	11
S4	Depth-based. From Bett (2012)	5
S5	Simple geographic. From Bett (2012)	2
S6	Depth/location based. From Bett (2012)	8
S7	Location/depth/substratum type based. From Narayanaswamy *et al.* (2014)	10
S8	Depth-based, from Piechaud and Howell (2013) and Hughes (2014)	4
S9 = S1:S5	Combination of S1 and S5	6

2.1.3 Parameter estimation

Each stratification system was examined to determine the extent to which the diversity metrics differed between the different strata with the objective of identifying the stratification system where apparent differences were maximised. For each stratification system a 2-factor analysis of variance (ANOVA) was conducted, including all terms (i.e. stratum, year and stratum:year interaction). This effectively partitions the variance in the metric (NosFam and Piel) between these three sources. The objective was to identify the stratification system which best accounted for the variance (minimised the residual term in the ANOVA). We also wished to consider how best to minimise the number of resultant categories (stratum levels), thus maximising the number of samples in each level, which, in our opinion, were likely to be the main drivers of assemblage. A brief description of each stratification system, and its source, is given in Table 2.

2.1.4 Spatial analysis

Analysis of the spatial scale of variation in this deep-sea data gives an idea of the best spatial configuration of sampling within a monitoring programme (i.e. one that provides the most information about the study area). Here we use a statistical approach that shows how the changing variation between sample pairs separated by a specified distance can be used to guide the choice of a minimum separation between sampling locations.

In order to design a rigorous monitoring programme there has to be an understanding of the optimal spatial distribution of sample locations. Traditionally, independence between sample locations would have been assured by randomisation, but this may result, particularly where

[3] Information on the strata boundaries of each stratification system is included in the data inventory prepared for this contract, which is available from JNCC on request.

sampling density is low, in samples being taken from a limited spatial area (e.g. where the randomisation process just happens to place the sample locations in close proximity). This problem could occur at any level, for example, in any one of a chosen stratum level. In such circumstances, a pseudo-random sampling approach should be adopted following an assessment of the similarity between a given sample-metric as a function of the distance between the sampling locations. The relatedness of samples, as a function of distance between them, gives an indication of their spatial autocorrelation (degree of independence). The degree of spatial autocorrelation indicates, for the metric considered, the minimum distance between sample locations that is required in order to ensure independence of samples and maximise the information that can be gained from the monitoring programme.

One method of assessing spatial autocorrelation is to produce semi-variograms. In producing a semi-variogram, the variance in the metric under study, between all possible points (sample locations in the current context) is determined (so with four points you would have six possible pairs, in our actual data set we have number of records = 336 which equates to 56,280 possible pair combinations) and these are then plotted against the distance between the pairs (x axis) (a detailed explanation of semivariograms is given in Rossi *et al.* (1992)). Semi-variograms have three components which can be usefully interpreted: (1) the nugget, which is the semi-variance at a distance of zero and indicates the inherent variability between replicates samples taken from the same location, (2) the sill, which represents the total amount of variability present in the data and is represented in semi-variograms by a levelling off of any observed trend beyond a certain distance and (3) the range, which is the distance beyond which the variance ceases to increase (i.e. the minimum distance between sample locations where they can be considered independent) (Crawley, 2007). In terms of sampling-design the semi-variogram gives an indication of the required distance between sample locations to ensure independence (and, thereby, maximise the information from the sampling programme) and the degree to which replicates samples are likely to differ.

Other methods of assessing spatial autocorrelation include hierarchical ANOVA which independently quantifies variance on each spatial scale (Burrows *et al.*, 2009). The hierarchical spatial ANOVA approach gives the spectrum of spatial variation. It shows the relative importance of large-scale variation, such as that due to the effects of gradients in temperature or productivity on abundance or distributions, versus small-scale variation, associated with habitat-scale effects such as topography. The results of such analysis can guide monitoring programmes: if large-scale variation dominates then fewer samples may be needed to characterise a given area. Conversely, more small-scale variation ('patchiness') relative to large-scale variation suggests more samples may be needed to characterise a region. However, hierarchical spatial ANOVA, has a limited capacity to detect the 'sill' as described above.

The reliability of the semi-variance estimate increases with sample size for any given pair-wise distance. This depends on the overall sampling effort (number of samples) and the spatial distribution of the samples (Crawley, 2007). In the current case, as samples were collected over time, as well as over space and, given the inherent variability over time and space, at least in terms of numbers of families, it is sensible to split the variograms into year-classes and, where appropriate, into different strata/year combinations. We wanted to investigate the likely spatial autocorrelation within levels of a stratum because sampling might, reasonably, be stratified by stratum levels. Crawley (2007) recommends 30 as the minimum number of samples pairs necessary to determine semi-variance and this, therefore, excludes the 1998 data from the analysis as there were only 27 samples (in total) and some levels within a given stratification system (which ones depends on the stratification system used).

2.1.5 Power analysis

Power, in the present sense, is the ability of a monitoring programme and data analysis to reject the null hypothesis when it should be rejected (Sokal and Rohlf, 1995). A number of factors influence the power of a monitoring programme including the number of samples, the inherent variability in the system (high variance leads to low power) and the magnitude of the change occurring ('effect size'). A large effect size, in a system exhibiting low inherent variability, will require only limited sampling effort to detect the effect; the converse applies - a small effect, in a system showing high variability, will be difficult to detect (Di Stefano, 2003). Inadequately or poorly considered questions and unknown statistical power continue to blight ecological and environmental research programmes (Peterman, 1990; Mapstone, 1995; Schmitt and Osenberg, 1996; Johnson, 1999; Anderson *et al.*, 2000; Fidler *et al.*, 2004).

When looking at change in space, over time, the ideal scenario is to have a long historical record of the response metric and one which extends back sufficiently far to enable an assessment of natural variability in the absence of substantial human interference. The historical record (time-series data) indicates the natural variability in the metric under consideration. If this variability is known, then trends occurring during any part of that historical record, or identified following future monitoring programmes, can be put into perspective. The power of such an approach depends, in part, on the length of the historical record - the power is the ability to be able to state just how unusual an observation, or trend, is (Schmitt and Osenberg, 1996). In the absence of historical data it is impossible to detect change that is occurring over the entire study region (i.e. the sampling domain shown in Figure 1) (e.g. because of climate change, see Section 4.1) because the entire spatial area could be changing. Assessing spatial-temporal trends (as opposed to just temporal trends, as above) can only detect change that is occurring within specific areas within the general area of interest (for example, at the level of different strata within a stratification system). In the absence of time-series data one can only assess change by comparing locations, over time. This comparison can occur at various scales (between strata and within strata, see below), depending on the survey objectives.

In the present case, the status of a particular stratum level (for example, Upper Bathyal, North of Shetland) could be monitored by comparing how metrics derived from samples taken within that stratum level change over time compared with metrics derived from samples taken from other strata (over time). In terms of assessing change, by comparing metrics derived for different strata, it is critical to understand how the response metric changes over time and how this change (over time) differs between different strata in the absence of the impact. For example, in the absence of impact, if the metric being studied increases over time within some strata and decreases, over the same time-period, in other strata then the chances of a monitoring programme detecting differences following an impact will be low because any effect (within a single strata) would be hidden by the inherent location-time variability that characterises the system. The data provided do allow an assessment of the location:time interaction at the scale of between strata, and power analyses, based on assessing the location:time interaction, are conducted (see below).

Monitoring change on a finer-scale (e.g. within strata) would require the comparison of the impacted site (e.g. on the scale of an oil-well), over time, with other areas *within the same stratum* over time. The impact would be assessed by analysing how the metric being studied changed over time at the impacted site compared with all the other sites within the same stratum. If the 'impacted' site changed, over time, differently compared to the other sites, this would be indicative of an impact (Schmitt and Osenberg, 1996). To conduct analysis on this scale requires an understanding of how the metric under consideration (e.g. NosFam) naturally changes, over time, *within strata* (not between strata). The data analysed

under the current contract do not allow an assessment of the within-stratum variability over time, because there are no subdivisions of strata with replication, over time. However, the power analysis described is as applicable to point impacts (provided the parameter estimates are available) as it is to between strata (as described below).

The dataset provided does allow (albeit with caveats, see Section 4) the estimation of parameters to allow power analysis to be conducted (see below). The power analysis is most applicable to testing the power of monitoring programmes where samples are taken in different strata, over time. This would be applicable, for example, to assess fishing impacts where the impact is occurring within a single stratum and where the reference locations are located in different strata. This type of monitoring falls into the broad category of methods called 'before-after-control-impact' (BACI) designs. Power analysis can be conducted for BACI designs (see below).

BACI designs have been used for several decades to detect changes occurring around point-sources (such as a single oil-well or sub-sea sewage outfall), changes occurring at larger scales (e.g. between strata in the current context) and at multiple impacted sites (Underwood, 1991; 1992; 1994). There are several BACI designs (Schmitt and Osenberg, 1996) but two (the BACI-paired series approach and the beyond-BACI approach) are particularly relevant and potentially appropriate in the current case. In BACI designs there should be one or more impacted 'locations' and more than one 'reference location' and, within both of these randomly designated sampling stations. The BACI-paired series approach is explained in Stewart-Oaten and Bence (2001) and makes the assumption that temporal changes in the response variable do not occur. In the present case, the assumption of no spatial variability in the nature of change, over time, is untenable and this means that the 'beyond-BACI' approach (Underwood, 1991; 1992; 1994) should be adopted. This approach requires several reference locations to be monitored, in addition to the impacted location(s), both before, and after, the possible impact has occurred. This beyond-BACI approach has the advantage that there is no requirement for samples to be taken at the same time, unlike the paired-series approach (Underwood, 1994) (although sampling should occur at the same time of year to reduce variability if only sampling annually). The basis of detecting change, and attributing that to the impact, is a change in the pattern, over time, between the reference locations and the 'impact' location(s). This is the 'interaction' term and is core to BACI analysis (Underwood, 1991; 1992; 1994).

2.2 Results and discussion

The data range over the period 1996–2002, which is insufficient to allow a characterisation of the long-term variability in the NosFam or any other metrics within the study region (see Section 4.1 for a fuller discussion of this issue). Without an understanding of the long-term, historical, variability one cannot assess future changes in terms of how usual they are (i.e. whether they are likely to constitute a deterioration in quality). In order to undertake long-term monitoring, and to identify long-term trends, there needs to be a historical data set to put the monitoring results into a relevant temporal context. It is possible to initiate a time-series study now but this would not be able to determine if ecologically significant changes were occurring in the near future (because the time-series needs to be extensive, for example > 50 years as is used in Tett *et al.* (2013)) and any assessment of future change would be confounded by the current unknown baseline condition. The unknown baseline condition refers to the fact that the current status of the environment, in relation to man-induced change, is not understood.

2.2.1 Response metric

A total of eight diversity/evenness metrics were determined (see Table 1 for list and abbreviations) and the correlations between all were evaluated (Figure 2). Those metrics showing a high correlation were identified (visually) and, from any pair of highly correlated metrics (see also Table 1 for a summary of reasons for selection of metrics), one was selected on the basis of its interpretability (Zuur *et al.*, 2010). From the correlation plot Marg, NosFam and Fish were highly correlated (r > 0.97) and NosFam was retained because of its ease of interpretation. Shan and Brill were highly correlated (r > 1.00) and both showed a high correlation with NosFam (r > 0.7) and were, therefore, excluded. Taxonomic distinctness did not show any correlation with the other metrics, but showed a counter-intuitive, and opposite, trend in relation to depth, compared with NosFam. The behaviour of Tdi, when based on family-level identification, in relation to other metrics, or as an indicator of any aspect of environmental status, has not been investigated by the scientific community and, consequently, it was excluded from the analyses.

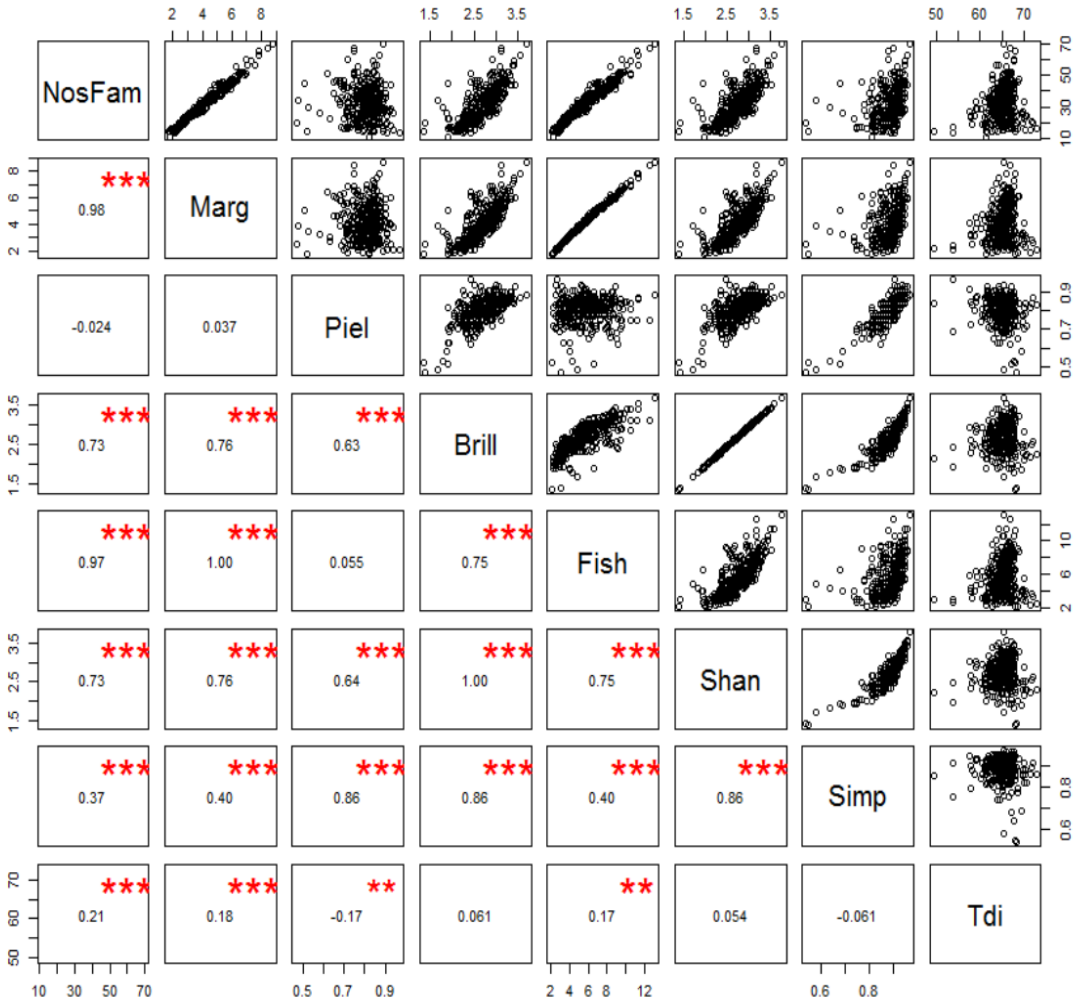

Figure 2. Correlation between diversity and evenness metrics derived from macrobenthos multivariate data (based on family-level identification only). One of each pair of highly correlated metrics was eliminated from further consideration. Key to abbreviations: see Table 1. The correlation coefficients are given in the lower diagonal cells. The number of stars associated with each coefficient indicates the significance of the correlation: *< 0.05, **0.01-0.05, ***< 0.01.

After this elimination process, the following were considered potentially useful metrics: NosFam, Piel and Simp. These response variables were standardised (mean subtraction and divided by their standard deviations) in R, and then plotted against different levels of each of the stratification systems assessed during the study. The standardisation process expresses all metrics in units of their own standard deviations and allows direct comparison between them. This process identified those metrics that were characterised by large numbers of outliers. The results from this process, plotted against stratification system 7 and 9 are shown in Figure 4 and Figure 5 respectively (see Sections 2.1.2 and 2.2.2 for a quantitative justification of this stratification system based on the ability to explain the variation in the data).

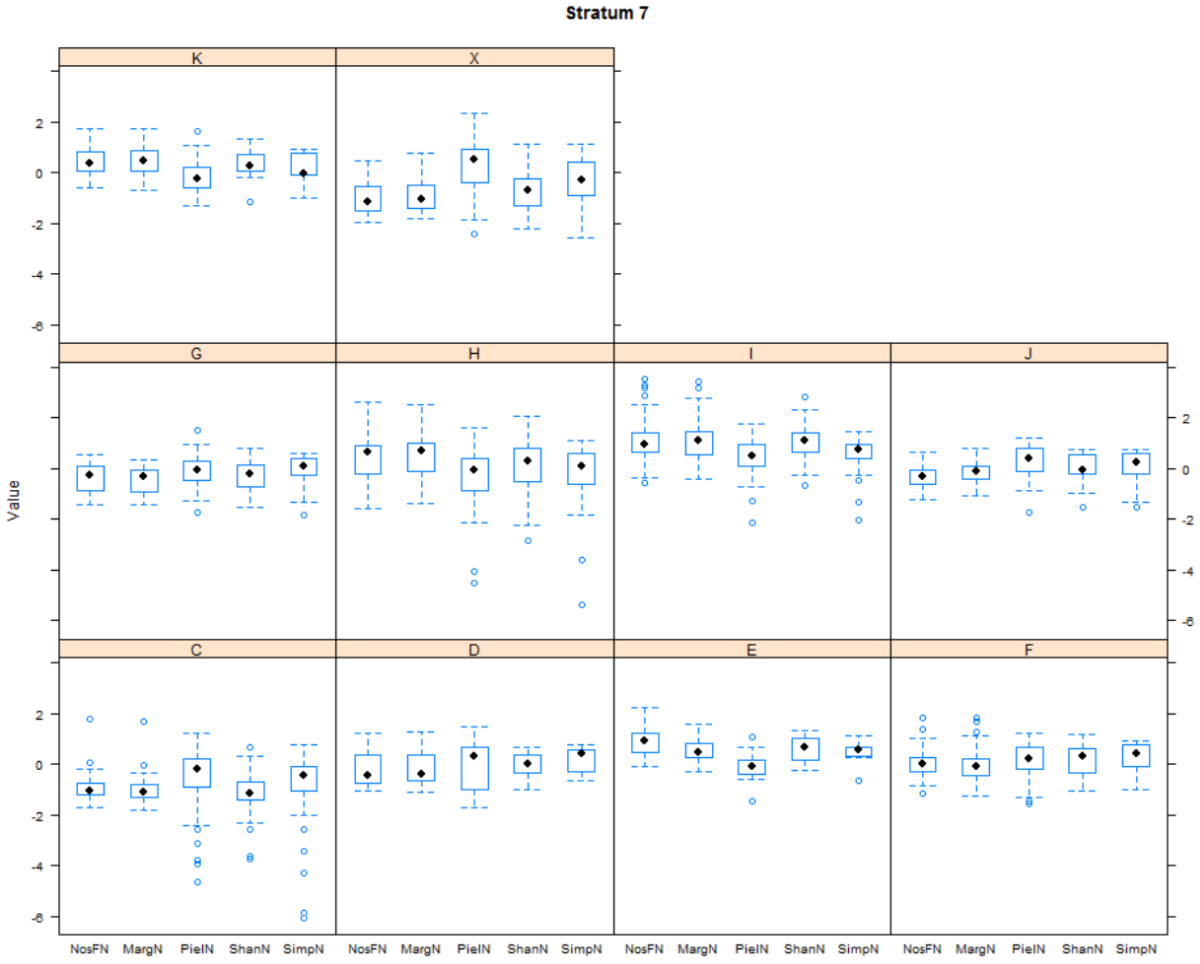

Figure 3. Standardised metrics against differing levels of stratification system 7. The strip-key indicates the level of stratification system.

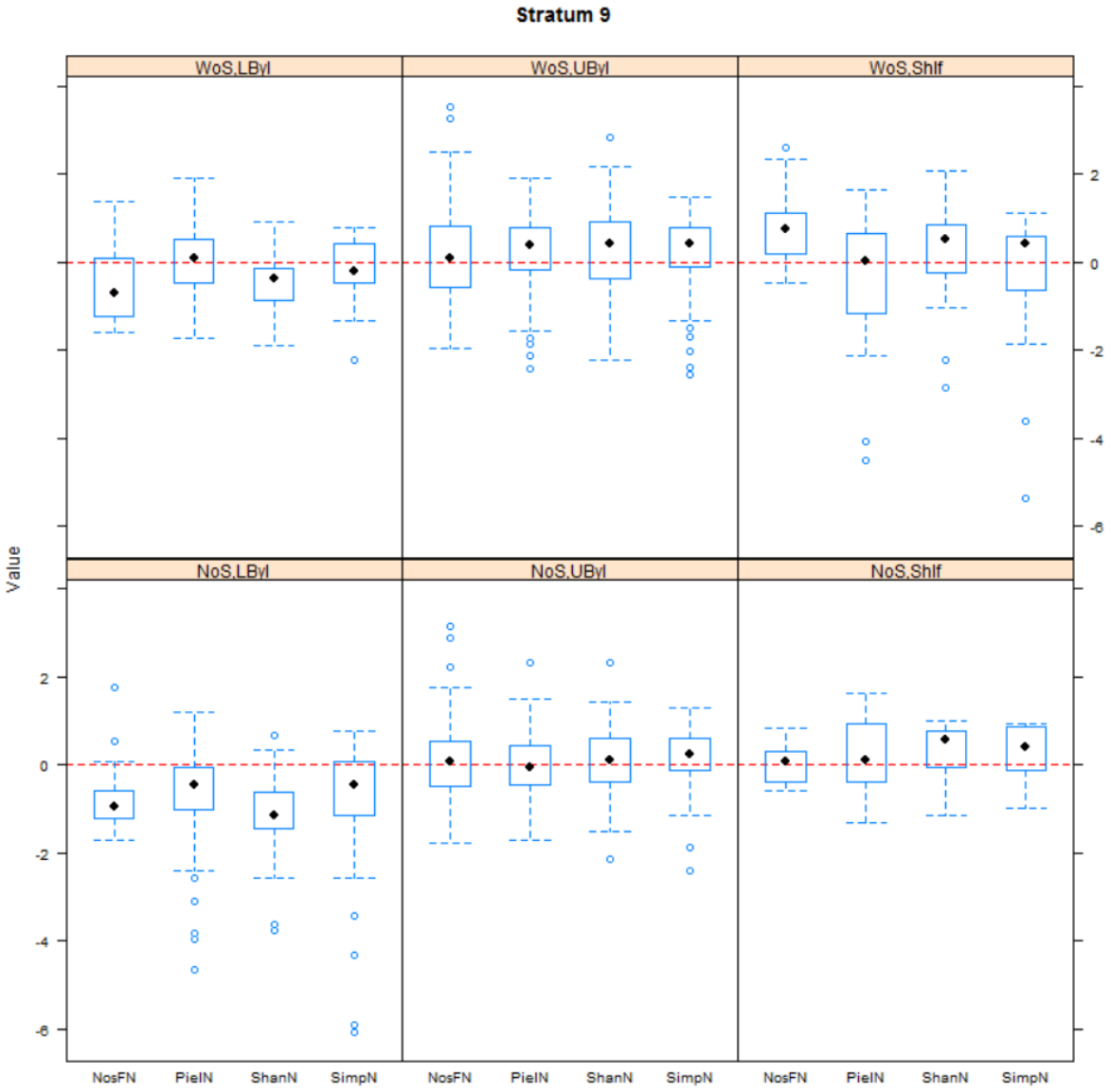

Figure 4. Standardised metrics against differing levels of stratification system 9. The units (Y axis) are standard deviations. The red line, at zero, indicates the average of the given metric. Key: NoS - North of Shetland, WoS – West of Shetland, LByl - lower bathyal (200-1100 m water depth), UByl - upper bathyal (> 1100 m), Shlf - shelf (≤ 200 m). For key to X axis abbreviations see Table 1 (the suffix 'N' indicates these data were standardised).

The sample sizes for the different levels of strata in S7 and S9 are shown in the following tables (Table 3 and Table 4).

Of the remaining three metrics (NosFam, Piel and Simp), Simp was characterised by a greater number and magnitude of outliers compared with Piel, with which it was highly correlated ($r > 0.8$) (Figure 5), and Simp was therefore excluded. The two metrics retained cover two important aspects of the assemblage, diversity (NosFam) and evenness (Piel).

Table 3. Numbers of samples for each level of S7 (these are given single-letter codes). X is the number of unclassified samples.

C	D	E	F	G	H	I	J	K	X
43	18	12	53	40	37	64	12	14	43

Table 4. Numbers of samples for each level of S9. For key to stratum levels see Figure 5.

NoS,LByl	NoS,UByl	NoS,Shlf	WoS,LByl	WoS,UByl	WoS,Shlf
37	68	7	34	160	30

Table 5. Summary of univariate metrics considered and rationale for their rejection or inclusion in the subsequent analyses. For metric abbreviations see Table 1.

Metric	**Included (Y or N)**	**Reasoning**
NosFam	Y	A simple, easily understood metric with obvious relevance
Rarefied number of taxa	N	Rarefaction is a method of establishing the number of taxa as a function of sampling effort. No replicate samples exist in the dataset, precluding use of this measure.
Marg	N	Very highly correlated with NosFam.
Brill	N	Very highly correlation with Shan.
Fish	N	High correlation with Shan.
Shan	N	High correlation with number of families and Piel.
Piel	Y	Little correlation with number of families so provides additional insights and few outliers.
Simp	N	Highly correlated with Piel. Plots indicated numerous outliers.
Tdi	N	A different approach to determining a single metric defining station similarity but insufficient background in its suitability when determined for family-level identification.

2.2.2 Development and testing of optimum stratification options

The requirement for a given stratification system to contain spatial and depth-based information and to be represented in as many years as possible (to allow an assessment of how the chosen metrics vary over location and time) resulted in the generation of a ninth stratification system (S9). The replicates per strata year combination are shown in Table 6.

Table 6. Number of samples per combination of S9 and Year. Key: NoS and WoS - North and West of Shetland respectively, LByl and UByl - lower and upper bathyal respectively, Shlf - shelf. Strata for which temporal variability was assessed are highlighted in grey.

Year	NoS, LByl	NoS, UByl	NoS, Shlf	WoS, LByl	WoS, UByl	WoS, Shlf
1996	0	1	0	12	117	30
1998	11	10	0	0	6	0
2000	3	25	0	22	37	0
2002	23	32	7	0	0	0

Each stratification system (S1-S9) was then tested using a 2-way ANOVA. The objective was to determine which stratification system accounted for the variance in the system the best. ANOVA models assess the degree to which explanatory factors partition the variance. In the current context the explanatory factors are stratum (i.e. the degree to which the response variable varies between different levels of the stratum, irrespective of time), time (i.e. the degree to which changes occur over time, irrespective of space) and the stratum-time interaction which is an assessment of how the response varies over time at different levels of the stratum. The residual is a measure of the variance that is not accounted for by the factors in the model. In most statistical modelling scenarios the objective (as here) is to minimise the sum-of-squares attributable to the residual term and maximise those attributable to the main effects. Further guidance to the interpretation of two-factor ANOVA is given in Sokal and Rohlf (1995) and Quinn and Keough (2002).

The results of the ANOVA analyses (Appendix 7.1**.**) showed that stratification system S7 was the best (the mean square attributable to the Stratum main effect is much larger for S7 compared with any of the other stratification systems, meaning that this stratification system accounts for the variance in the data better than the others (see Table 9). However, in S7 several stratum level-year combinations were absent (because of the spatio-temporal distribution of sampling effort) and this raised doubts about the robustness of parameter estimates based on this analysis. For this reason, we also determined parameter estimates based on the results from S9 (justification for the creation of S9 stratification system is given in Section 2.1.2). The results from these two analyses are shown in Table 7 and Table 8 and parameter estimates from both are used in the subsequent BACI analysis (see Section 2.2.6).

Table 7. Results from 2-way ANOVA (Stratification system S7). These results were used to inform power analysis.

```
Response: NosFam
           Df  Sum Sq Mean Sq F value  Pr(>F)
S7          9 20942.6 2326.96 45.4157 < 2e-16 ***
Year        3   284.0   94.66  1.8474 0.13853
S7:Year    15  1307.3   87.16  1.7010 0.04961 *
Residuals 308 15781.0   51.24

Response: Piel
           Df  Sum Sq   Mean Sq F value    Pr(>F)
S7          9 0.18896 0.0209956  4.5557 1.189e-05 ***
Year        3 0.06763 0.0225431  4.8915  0.002455 **
S7:Year    15 0.04553 0.0030354  0.6586  0.824230
Residuals 308 1.41946 0.0046086
---
```

Table 8. Results from 2-way ANOVA (Stratification system S9). These results were used in the parameterisation of the BACI power models.

```
Response: NosFam
           Df  Sum Sq Mean Sq F value    Pr(>F)
S9          3  4233.7 1411.23 14.8685 5.042e-09 ***
Year        3   970.1  323.37  3.4069  0.018059 *
S9:Year     5  1775.3  355.07  3.7409  0.002689 **
Residuals 287 27240.4   94.91

Response: Piel
           Df  Sum Sq  Mean Sq F value    Pr(>F)
S9          3 0.15458 0.051526 13.1291 4.664e-08 ***
Year        3 0.06132 0.020441  5.2084  0.001622 **
S9:Year     5 0.00791 0.001582  0.4031  0.846515
Residuals 287 1.12636 0.003925
```

Table 9. Summary results from ANOVA analysis for stratification system S7. The mean square for each term is taken from Appendix 7.1**.** and the proportion of the variance for each term is calculated.

Source	Mean square	% total
Stratum levels	2326	90
Year	95	3.7
Stratum: Year	87	3.4
Residual	51	2.0

2.2.3 Parameter estimation

Benthic communities will change over time and across space. Temporal change may occur at a number of scales, from seasonal to multi-decadal whilst spatial variability will occur at scales ranging from within sample to changes between samples taken at the same location, to latitudinal scales relevant to changes in overlying water column. In terms of assessing long-term change (decadal scale) at large spatial scales (e.g. the whole region to the north and west of Shetland) understanding and quantifying these trends is critical. The current data do not support such an analysis because they are of insufficient duration (four sampling periods) and sampling has not occurred in the same places. This aspect of data limitation is further discussed in Section 4.1.

To assess temporal variability, repeated measures of the response variable, over time, at the same location are required. The requirement for the same sampling location is to distinguish (separate) temporal and spatial variability. In the present case, such data were not available and, in order to estimate temporal stability, samples from the same stratum were considered to be from the same location. This necessary approach has the disadvantage that it effectively combines temporal and spatial variation. In the present case the parameter required for power analysis is the location:time interaction term. This is an assessment of how the response metric varies over time and how this differs between locations. Ideally, from a monitoring perspective, this interaction term would be zero, that is the same temporal patterns would occur irrespective of location. The ANOVA interaction term gives an estimate of this interaction (which gives us our parameter estimates for the BACI analysis) but, by way of example, it is also illustrated in Table 10. In Table 10 a considerable degree of location:time interaction can be seen, for example within the NoS, Ubyl stratum the mean NosFam increases from 27 and 38 over the period 2000 to 2002 whilst, over the same time period, it decreases from 28 to 22 within the NoS, Lbyl stratum. The trend, over time, between the two strata, is going in opposite directions and is of considerable magnitude. This means that detecting change using the metric NosFam in this environment will be challenging.

The interaction, and residual terms (necessary for model parameterisation) were based on 2-way ANOVA models with NosFam and Piel as the response variables and Stratum (S7 or S9), Year and their interaction as the predictors (see Table 7 and Table 8). For S9 the stratum level ‘shelf’ was excluded because so few measurements were based on the shelf. The necessary standard deviations were derived by taking the square-root of the Mean square values from the model. Only the standard deviations from the stratum: year interaction and Residual terms affect the power in BACI designs.

Table 10. Mean, standard deviation (S) and sample size (N) of NosFam and Piel as a function of Stratum and Year for S9.

Metric	Stratum	Year	Mean	S	N
NosFam	NoS, Ubyl	1996	30	NA	1
NosFam	NoS, Ubyl	1998	34.2	4.7	10
NosFam	NoS, Ubyl	2000	27.4	5.8	25
NosFam	NoS, Ubyl	2002	37.8	12.7	32
NosFam	WoS, Ubyl	1996	34.4	11.5	117
NosFam	WoS, Ubyl	1998	32.8	4.8	6
NosFam	WoS, Ubyl	2000	30.4	8.1	37
NosFam	NoS, Lbyl	1998	23.7	6.3	11
NosFam	NoS, Lbyl	2000	27.7	20.2	3
NosFam	NoS, Lbyl	2002	22.3	4.7	23
Piel	NoS, Ubyl	1996	0.87	NA	1
Piel	NoS, Ubyl	1998	0.827	0.03	10
Piel	NoS, Ubyl	2000	0.787	0.054	25
Piel	NoS, Ubyl	2002	0.803	0.053	32
Piel	WoS, Ubyl	1996	0.829	0.055	117
Piel	WoS, Ubyl	1998	0.832	0.037	6
Piel	WoS, Ubyl	2000	0.793	0.064	37
Piel	NoS, Lbyl	1998	0.772	0.035	11
Piel	NoS, Lbyl	2000	0.700	0.075	3
Piel	NoS, Lbyl	2002	0.745	0.117	23

The critical parameters requiring quantification when designing ‘Before-after-control impact’ studies are the inherent variability in the metric under study (within relevant time and space limitations) and the extent to which time-related changes differ between locations that could be used as reference locations (location:time interaction).

Small-scale temporal patterns

The data did not support an assessment of the small-scale (same location) sampling variability because there was no sampling overlap in space and time (i.e. there was no repeated sampling from the same location). Exactly what constitutes the same location will vary in practice, depending on the water depth, sampling gear, weather conditions etc. Note that sampling within the same strata (as opposed to at the same sample location), over time, is available from these data (the degree to which this overlap occurs is dependent on the stratification system used) and that the assessment of the location:time interaction (as described in Section 2.2.3), on this spatial scale, has been achieved.

2.2.4 Spatial patterns

There was a clear negative association between number of families (per sample) and sample depth, as shown by the south-east to north-west trend from shallow to deep areas in Figure 5 and extending across both the north and west of Shetland regions. The depth gradient extends from shallow in the SE (where large blue dots indicate greater than

average NosFam) to deep in the NW (mostly large green dots indicative of less than average NosFam in Figure 5).

There was no apparent relationship between Pielou's evenness and location (Appendices, Figure A1.) or depth.

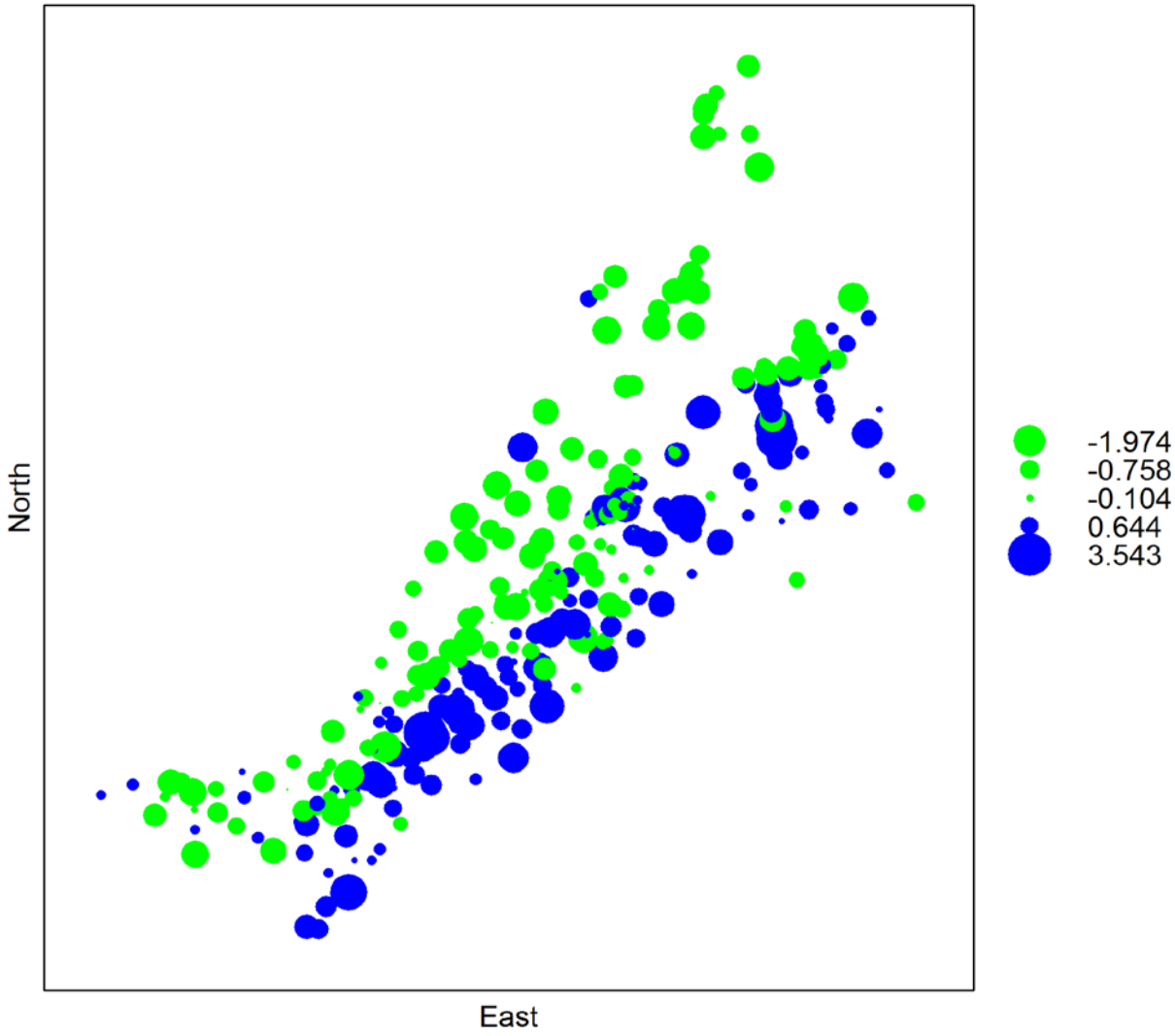

Figure 5. Relative numbers of families across locations. These are normalised data and the units are standard deviations. Blue and green represent greater-than and less-than mean respectively, the size of the circle is indicative of magnitude. It cannot be determined to what extent this apparent spatial trend is temporally driven.

The trends in numbers of families, as a function of depth (Figure 5) is also seen by splitting across different levels of S9 (see Table 2) and is shown in Figure 6.

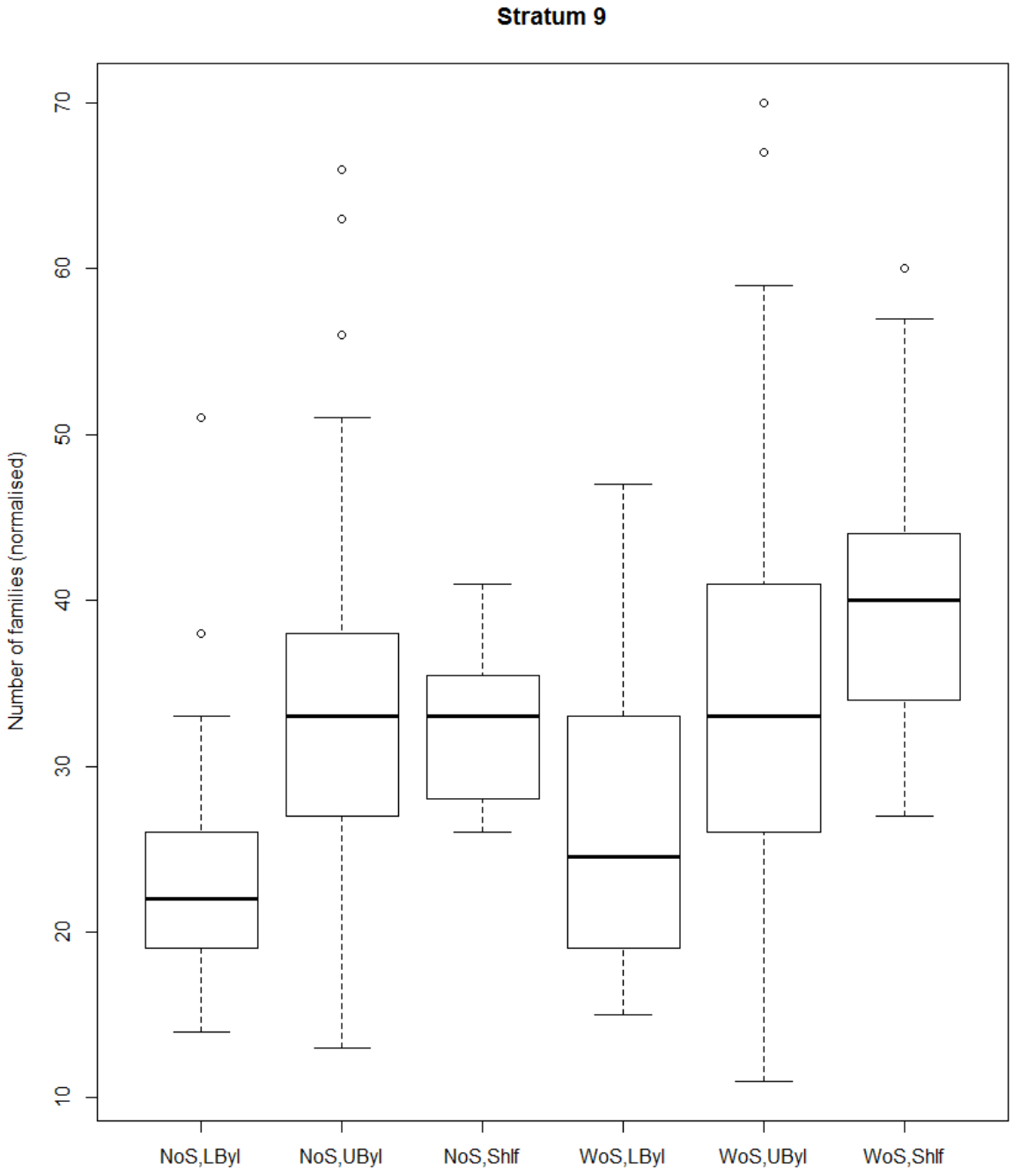

Figure 6. Numbers of families as a function of levels of S9. The modified boxplot shows minimum (lower limit), quartiles (25th and 75th percentiles, bounding the box), median, and maximum. Where high and low values exceed 1.5 times the interquartile range, these data are shown as circles, denoting possible outliers.

The semi-variograms showed variable and inconsistent patterns over time in both number of families and Pielou's evenness. These data are spatially confounded with time (different areas were sampled at different times) and, as a consequence, semi-variograms produced for different years also represent different areas. The sample size for the 1996 survey is much larger than in both 2000 and 2002 and, consequently, is likely to be more reliable, and is interpreted here. In terms of NosFam, and when based on the entire spatial range of the 1996 data, the nugget semi-variance is approximately 70 (Figure 7). This means that, when

considering all samples simultaneously (i.e. not separated into strata, see below), samples taken in very close proximity (~4km) may be expected to differ by approximately 12 families (12 being the approximate square-root of 2 x the semi-variance (=140) at distance 4km). Within the context of the spatial area covered by the 1996 samples the semi-variance increases almost linearly. This indicates that samples were taken from an environment where there is an ongoing trend, over distance, in number of families. In the present case, this may be attributable to trends across depth and/or latitude (note that the 1996 samples extended over both depth and latitude).

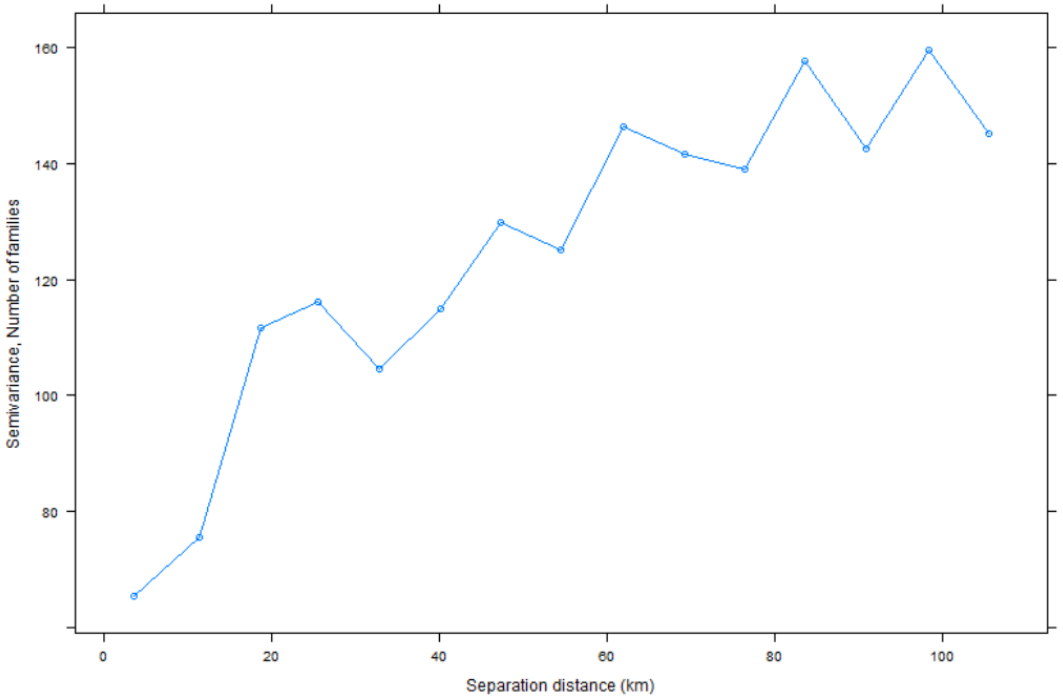

Figure 7. Semi-variogram for the number of families (NosFam) showing the relationship between semi-variance (y axis), representing half the average squared difference between pairs of samples, and the distance (x axis) between the members of each pair (N = 160) in the unstratified dataset as a whole. Semi-variance was calculated using all possible pairs of data and averaged over 8km distance classes.

In terms of Piel there was a steep rise in the semi-variance over the distance 4-16km indicating spatial autocorrelation, that is, data that are taken from locations within 16km are much more similar to each other than those separated by greater distances. However, after ~16km there was a confused pattern which is difficult to interpret. These data indicate that sample locations, in relation to Piel, should be at least 16km apart in order to maximise the information from samples taken.

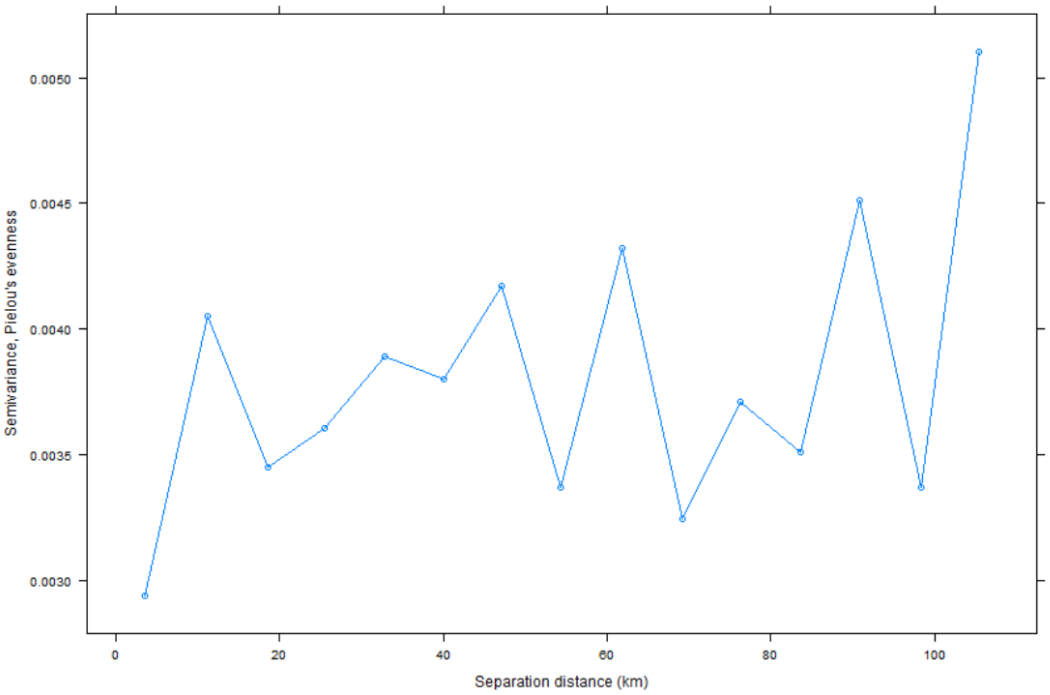

Figure 8. Semi-variogram for Pielou's index (Piel) showing the relationship between semi-variance (y axis) and distance (x axis) (N = 160) in the dataset as a whole.

The sill was not observed in the semi-variograms produced (over the interpretable horizontal range, that is up to half the maximum distance between samples (Crawley, 2007). Consequently, this analysis does not indicate a spatial scale at which the variability in number of families ceases to increase (Figure 7). In terms of sampling design, this analysis indicates a high degree of variability at small scales (< 4km). In terms of monitoring design this indicates that replication should occur at small spatial scales (< 4km) in order to adequately quantify this variability. In our opinion, the absence of a sill and the high degree of small-scale variability is probably due to the strong effect of depth on NosFam, imposing a large difference between samples up and down the slope, but not necessarily among samples at the same depth. In order to investigate this possibility, semi-variograms were constructed based on single strata, the logic being that single-strata (e.g. from S7) will already take into account the relationship between depth and/or sediment effects and Piel/ NosFam. The stratification system S7, level 'I', 1996 was chosen, as this had the largest sample size (N = 64) and S7 had been identified as minimising the variance within stratum levels (see Section 2.2.2). The resultant semi-variance plot is shown in Figure 9. This indicates that, at least for the specified stratum S7 level and year, that the semi-variance increases up until a distance of ~20km. These data indicate that to properly characterise this stratum level and maximise the information gained, sampling locations should be located with a minimum separation distance of ~20km.

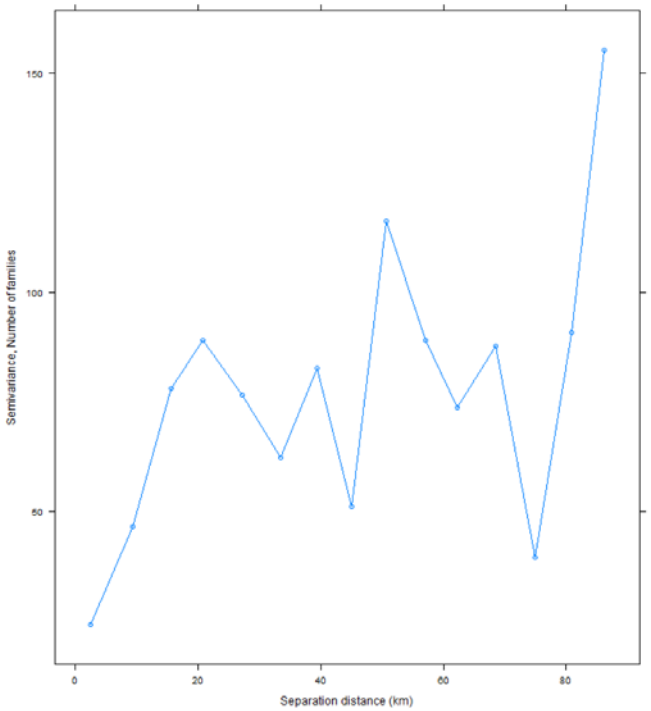

Figure 9. Semi-variance plot for Numbers of families, for level 'I' in Stratification system S7 for 1996 only (N = 64).

2.2.5 BACI power analysis

BACI power analysis requires the input of a number of parameters that are either estimated or inferred from existing data (Table 11 and Table 12**.**). A number of values for each parameter can be trialled but clearly as the parameter estimates become less reliable so does the power analysis (Carey and Keough, 2002a). BACI power analysis also requires 'user-input' on constant values (or a range of constant values), such as the number of years' sampling pre-and post-impact, the number of reference locations, the number of replicates per reference/impact location and the effect size. There are an infinite number of permutations possible for power analysis; those selected were agreed with the JNCC. The BACI power analysis used here was developed by Tony Booth (Department of Ichthyology and Fisheries Science, Rhodes University, Grahamstown, South Africa) with the permutations based on Stroup (2002).

The following power analyses could be applied to two situations:

1. An impact occurring at the level of an individual stratum (parameter estimates derived from S7 and S9 are used) or

2. A point-source impact occurring within an individual stratum. The parameter estimates are based on between-strata levels (for example, the change in NosFam, over time, in differing strata).

In this circumstance scenario 1 is more directly relevant unless point impacts were going to be assessed by comparing temporal trends across strata. The statistical power of a monitoring programme is tested where there are varying degrees of sampling effort (independent replicates per stratum) and where the metric being studied is compared to other samples taken from other strata where the 'impact' has not been applied (see below). The monitoring scenario being assessed, in terms of statistical power, makes the

assumption that sampling occurred biennially, over eight years, before and after the impact occurred. This choice was made because it underpins the sampling design that was used to derive the parameter estimates used in the power analysis model. The parameters used in the model, and the rationale for their use, are given in Table 11 and Table 12.

Table 11. BACI power model parameterisation for NosFam. In terms of BACI analysis, the overall (within strata) estimated standard deviation (S) and the variability of this between locations determines power. Only a single measure of Residual was trialled because of the time taken to run the models.

	NosFam		Piel	
Variable	**SD estimate**	**Range trialled**	**SD estimate**	**Range trialled**
Site*year	From S7 = 9. From S9 = 18	0, 5, 10, 15.	S7 = 0.0551 S9 = 0.0398	0.000, 0.020, 0.040, 0.06
Residual	For S7 = 7.1 For S9 = 9.7	10	S7 = 0.068 S9 = 0.063	0.07

Table 12. BACI power model parameterisation: rationale behind choice of constants used in model. These values were used in both the Number of families (NosFam)- and Pielou's-evenness (Piel) power analyses. These analyses, for NosFam, were based on stratification systems 7 and 9.

Variable	**Values adopted**	**Rationale**
Number of replicates per stratum level.	2 – 24	Covers a wide range, from the minimum to a maximum that is probably more than practically achievable.
Number of reference locations	2, 4, 8, 16, 32	Two is the minimum whilst 32 reference locations, within deep-sea research, is very large.
Number of years pre- and post-'impact'	4 years each	Four years data (over 6 years) was considered realistic in the current context. This conceived monitoring programme was based on the data set provided.
Effect size (% reduction)	50, 40, 30, 20 (NosFam); 30, 20, 10 and 5 (Piel)	This range included extreme impacts (50%) to moderate impacts (20%). The magnitude tested for NosFam and Piels differed because of the lower location:time interaction term in Piel compared with NosFam.
Mean abundance before impact.	The mean value over the entire data set	Logical decision but mean is irrelevant to power analysis.
Alpha (type I error rate)	0.05	'Industry' standard that is considered a reasonable compromise between Type I and Type II errors.

The results from the power analysis for NosFam and Piel are shown in Figure 10 and Figure 11 respectively and illustrate several main features in relation to the probability of detecting the effect in both NosFam and Piel:

1. As the level of stratum-by-year interaction increases the power decreases markedly,

2. Where the stratum:year interaction is low increasing the number of replicates, even at low numbers of reference locations, results in reasonable power and,

3. As the stratum:year interaction term increases there is more experimental power by sampling a larger number of reference locations rather than by increasing replicates per sampling location.

In terms of NosFam, the power analysis also shows that where the stratum:year interaction is >10 (which is indicated by the current data) then the probability of determining even a large (e.g. 50%) reduction in family numbers, even where numerous (>20) replicates are taken, is low ($P < 0.85$). In terms of Pielou's evenness the location:year interaction term is relatively small (compared to numbers of families). This means that the power of tests based on Pielou's evenness is higher. Figure 11 shows that at a location:time variability level that was actually determined (sd = 0.04) there is a high chance ($P > 0.9$) of detecting a 20% decrease in Pielou's evenness even with two reference locations and 5-10 replicates per reference location (other parameters being based on those outlined in Table 12).

The BACI power analysis illustrated some very important points. The location:time interaction, which is a measure of how the response metric varies over location and time (i.e. the degree to which a trend occurring over time in one place occurs in other places) is critical to the power to detect changes, occurring over time, in the region of interest. The analysis presented here was based on two estimates of the location:time interaction (from S7 and S9). These estimates were quite different because of the way in which the differing stratification methods partitioned the data (see Section 2.2.2). In consideration of the NosFam BACI power analysis even using a smaller location:time interaction term (~10 families, from S7) the analysis indicates that with 32 reference locations, a 50% reduction in the number of families, using up to 20 replicates per reference location, only gives a power of ~0.80 (this is seen in Figure 10 - labelled 32:10). If the variability, in numbers of families, over time, within the same stratum is as high as this then detecting change, even within levels of Stratum 7, will be very difficult and expensive (note that the parameter estimate that we actually can derive from the data is an assessment of how the metrics vary over time *between* strata).

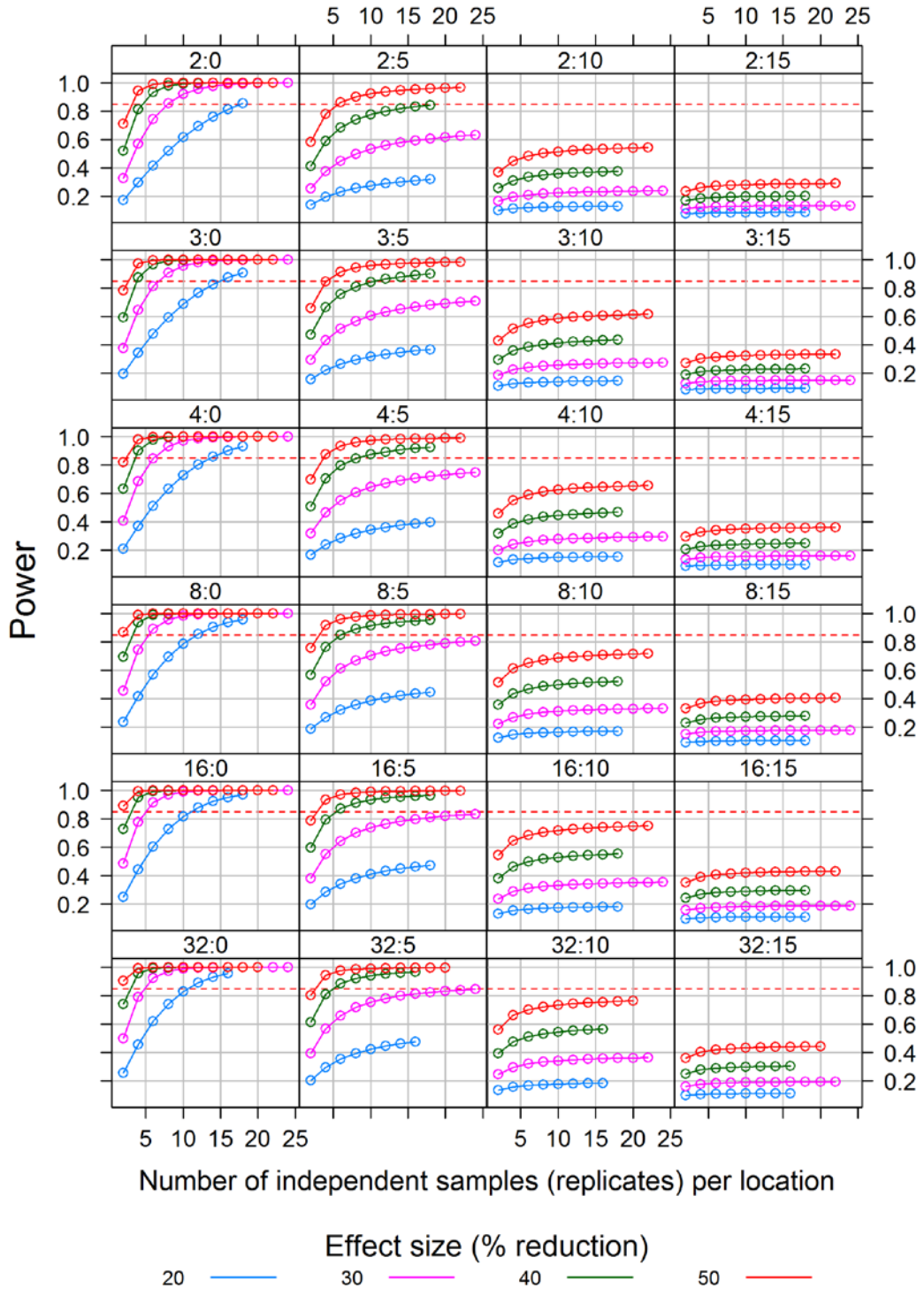

Figure 10. Results from power analysis based on NosFam. Power (alpha=0.05) represents the probability of rejecting a hypothesis of no-change when change is occurring. Key to strip: first number is the number of reference locations (2 to 32), second number is the standard deviation associated with the location:time interaction (0 to 15). The x axis represents the number of replicates per reference location ('site').

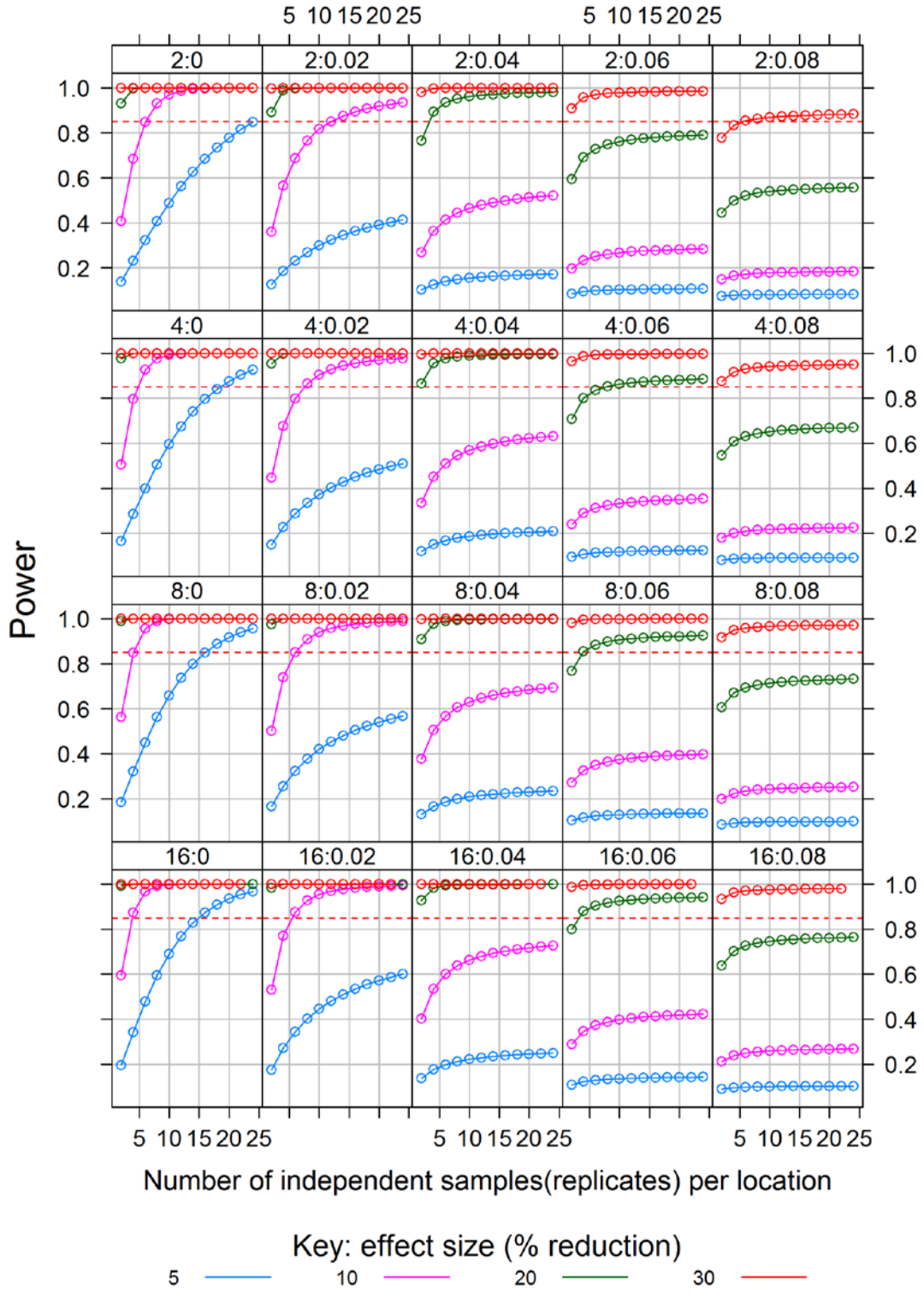

Figure 11. Results from power analysis based on Piel. Power (alpha = 0.05) represents the probability of rejecting a hypothesis of no-change when change is occurring. Key to strip: first number is the number of reference locations (2 to 32), second number is the standard deviation associated with the location:time interaction (0 to 0.08).

This power analysis has highlighted that a major data gap is the lack of understanding of how NosFam varies, over time, at the same location. This is a critical data requirement to

predict power. The BACI power analysis for Piel indicates similar patterns (reduced power as the interaction term increases). The power to detect a 30% reduction in Piel is quite high in all scenarios, including when using the location:time interaction term as derived from the data. Assuming the location:time interaction term is 0.04 (the approximate value as derived from the S9 stratification system) then detecting a 20% reduction in Piel is also likely even where there are only two reference locations. However, the power analysis indicates that detecting a 10% reduction is unlikely under any of the scenarios tested here, for example, based on 16 reference locations, with 25 replicates per reference location, would result in a power of 0.75 (looking at cell with strip value given as 16:0.04 in Figure 11). Whether NosFam or Piel is the better metric requires an understanding of how they respond to change (see Recommendations in Section 5).

3 Epifaunal data analysis

The epifaunal data available for this contract were collected in 2006 on behalf of DECC by researchers from NOC in three separate areas within the region of interest (Figure 12). Details of the epifaunal collection and analysis methods are given in Howell *et al.* (2010) and Jacobs and Howell (2007). Two broad regions were studied within the context of this research - Wyville Thomson Ridge (WTR) and the West-Shetland Channel (WSC). The WSC surveys were split into two study areas; West of Shetland (WoS) and North of Shetland (NoS) - Figure 12).

At each survey area a number of transects were covered. The data consisted of records of the presence or absence of fauna as observed in a number of still photographs within each of these transects. The number of still photographs analysed, per transect, varied from 3 to 26 (median 10). The total number of transects, over the three areas covered, was 40.

Within the timeframe of the project, only one area could be investigated in detail. Following discussion with JNCC, the West-Shetland Channel (WSC), west of Shetland (WoS) data were further analysed as these data were more numerous (19 transects) and covered a greater depth range compared with the other two areas (Figure 13). Associated with each photograph were a range of sample location/transect descriptors including continuous (e.g. depth, oxygen, substrate hardness) and categorical variables (location, water mass and seven different stratification systems, suggested by the JNCC, among others).

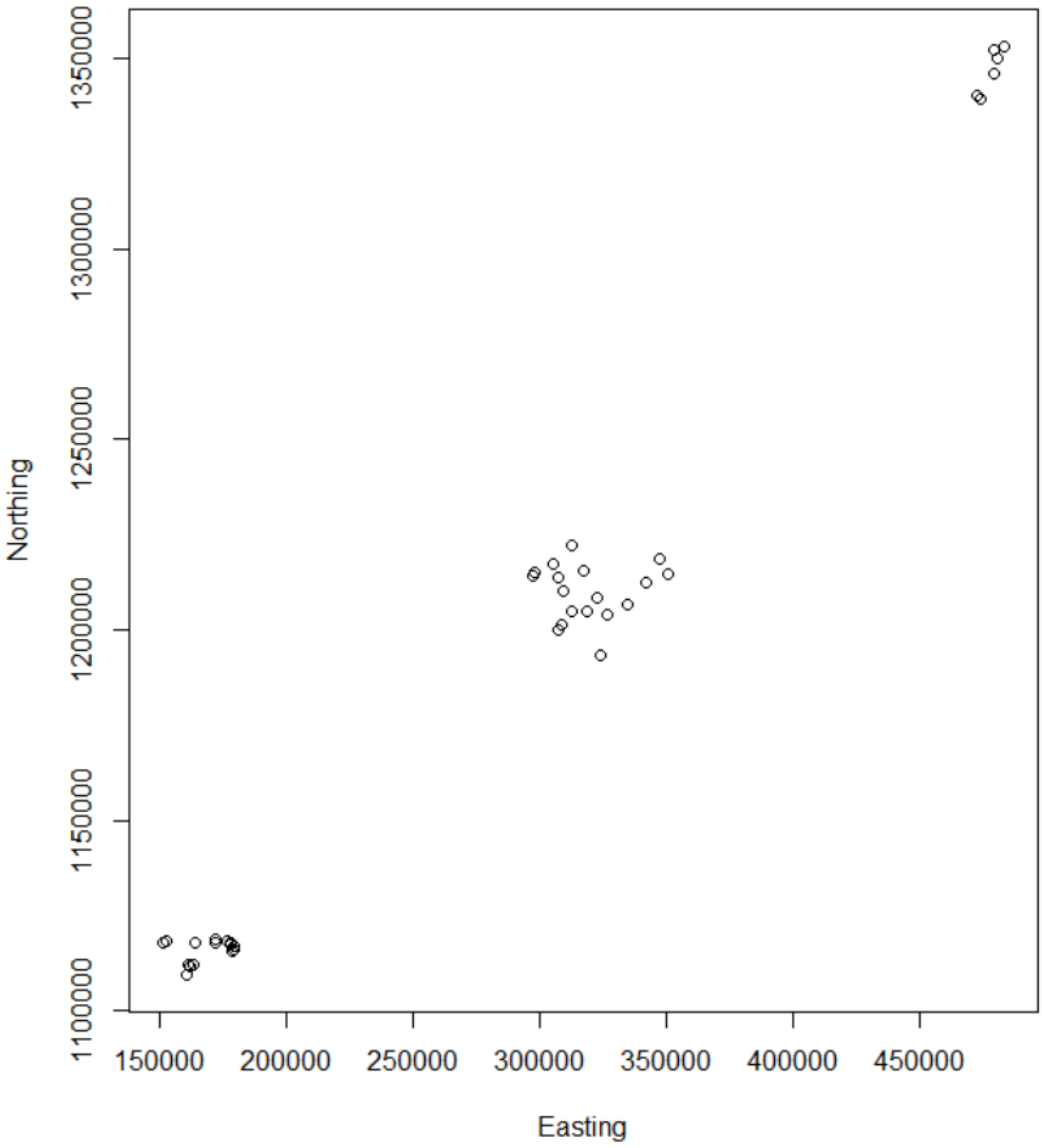

Figure 12. Location of the epifaunal transects, split between three areas (upper right - WSC NoS, middle - WSC WoS, lower left - WTR). Eastings and Northings are British National Grid (BNG). The central group of sample locations (WSC WoS, N = 19 transects) were selected for further analysis.

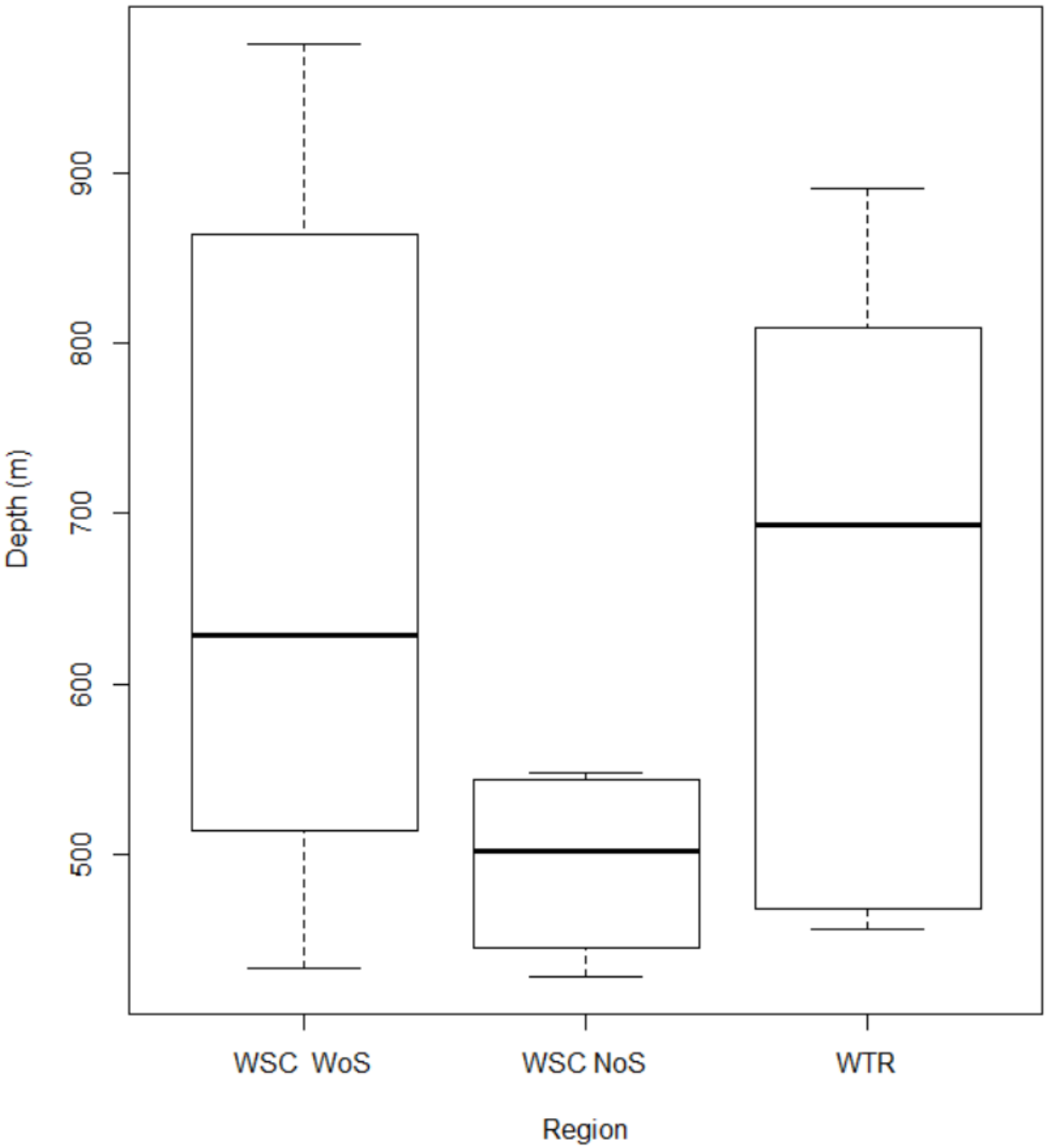

Figure 13. Depth range of the three separate epifaunal surveys. Key: WSC West Shetland Channel (split into WoS, west of Shetland and NoS, North of Shetland) and WTR - Wyville Thomson Ridge.

Each photograph was associated with a record of the presence/absence of a range of biota (118 different types of organisms). The biota were recorded at several levels, from species to general descriptions, for example, 'Porifera, orange, encrusting'. In addition, 10 unknown taxa were recorded. The presence of unidentified species does not have any bearing on the resultant diversity indices provided each is unambiguously and consistently defined and recorded. The conversion of a mixed identification-level (i.e. including species level and general-descriptive level identification) also has no bearing on the generation of the diversity indices but these, obviously, will not relate to species diversity/evenness.

The data were dominated by zero scores (i.e. most taxa, in most locations, were absent) and, consequently, the results from individual photographs were collated across transect and

presence/absence in the transect re-determined (note that a semi-quantitative indication of abundance could have been produced during this process, i.e. a count of how many photographs contained one or more of the species concerned, however, this is outside the scope of this project, see comment at the end of this section). The dominance of zeros within assemblage data causes problems in correlation analysis and, because it represents virtually no information, does not enhance the analysis. The means of all continuous variables, and modal value of any categorical variables, per transect, were collated and this formed the 'raw' data. For each transect a range of diversity and evenness metrics were determined (Figure 14). Figure 14 shows a high correlation between all the metrics because these metrics were based on presence/absence data. All further analyses were based on the number of taxa (this is a mix of species-level and general descriptive-level identification, as above as this is the easiest metric to interpret of those determined).

The NosTaxa were plotted against all the stratification systems supplied by JNCC in order to assess patterns of variability across different factors. These preliminary analyses indicate that the number of taxa is more variable on sediment compared with rock (Figure 15).

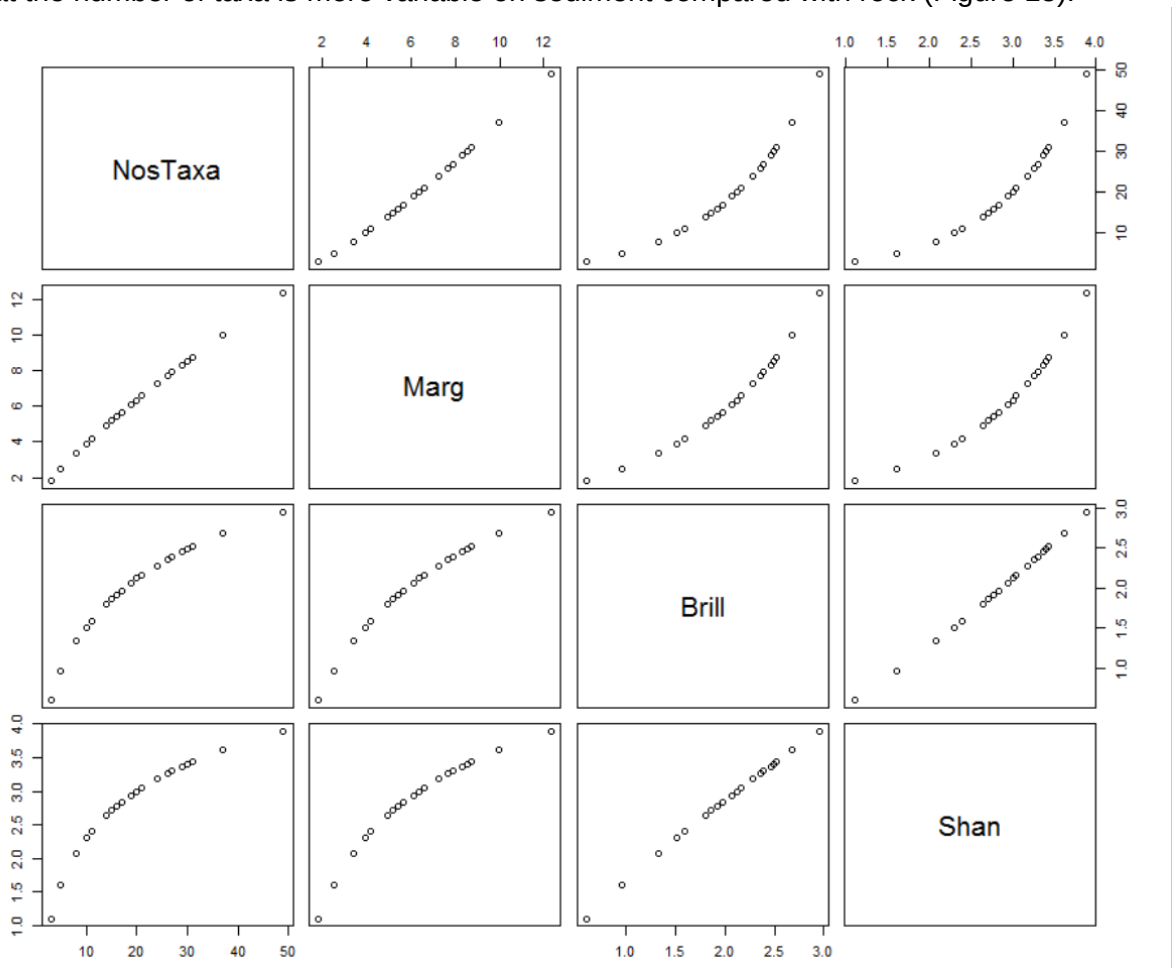

Figure 14. Association between number of taxa (NosTaxa), Marg, Brill and Shannon from the epifaunal surveys.

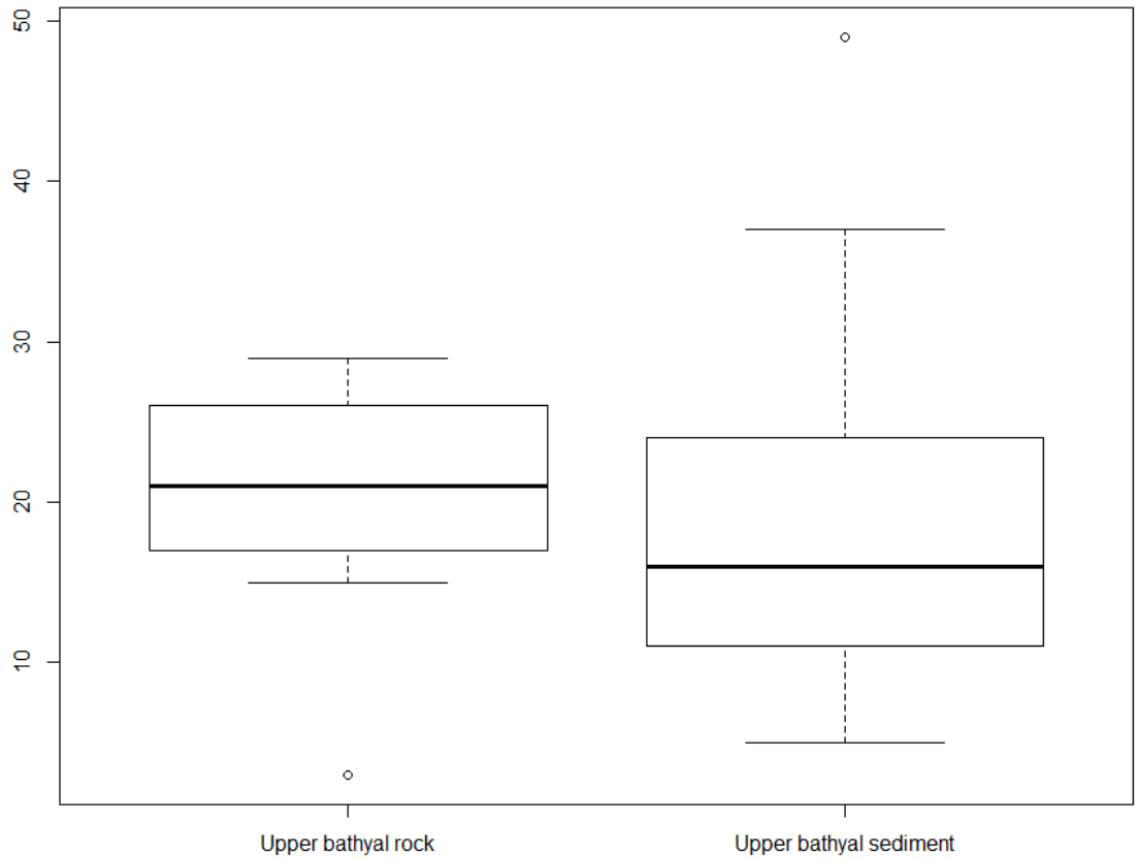

Figure 15. Comparison of Numbers of Taxa (Y axis) versus the two levels of JNCC stratification system S1. These data suggest that the diversity of epifauna is more variable on upper bathyal sediment (N = 25) compared with upper bathyal rock (N = 15).

JNCC stratification systems (S4 and S5) were based on depth. There was no clear relationship between depth and number of taxa observed in these data. Other JNCC stratification systems were based on location and, given that these analyses were based on WoS only, were not relevant.

The semi-variogram plot for NosTaxa (Figure 16) indicated that, in terms of NosTaxa, there was no spatial pattern in terms of similarities over distance. Thus NosTaxa does not appear to show any spatial dependence within the area investigated, so adjacent transects are no more similar (in terms of NosTaxa) than ones further away. However, the semi-variogram must be interpreted with caution given the small sample size. Also note that interpretation of variograms at distances over half the total range (in this case the total range is ~50km, as indicated on the x axis) is not recommended (Crawley, 2007).

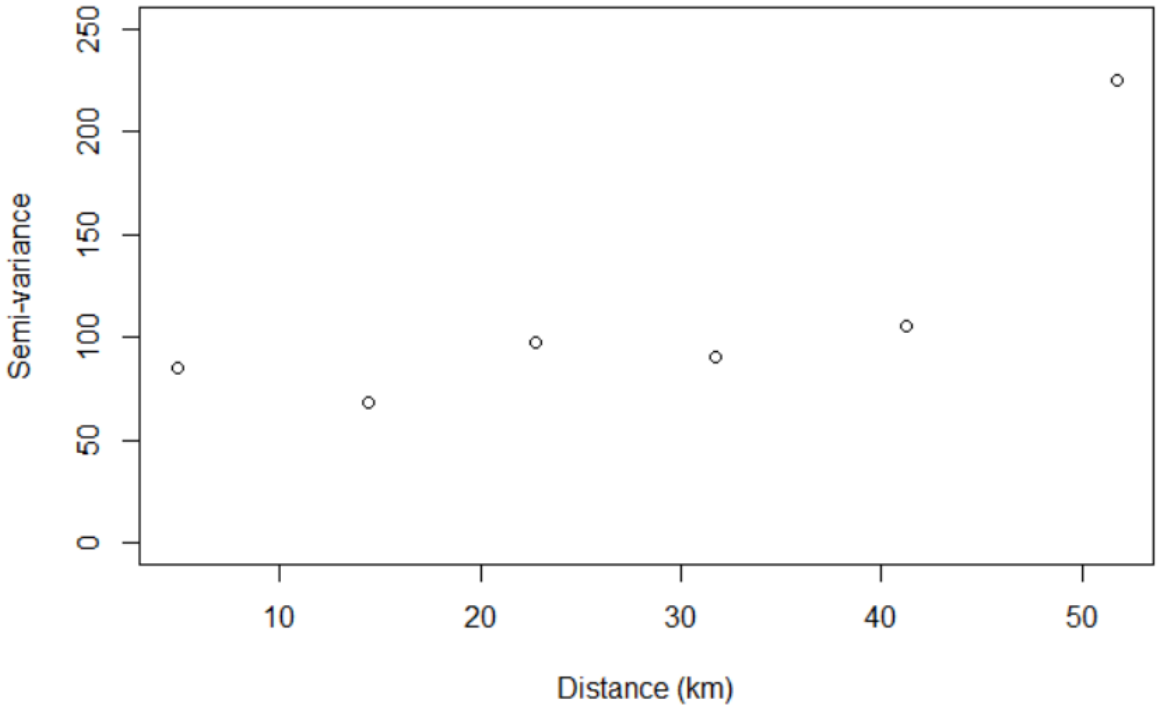

Figure 16. Semi-variogram for numbers of taxa (NosTaxa), as a function of distance. Semi-variance (y axis), representing half the average squared difference between pairs of samples, is plotted against and the distance (x axis) between the members of each pair (N = 160) in the unstratified dataset as a whole. Semi-variance was calculated using all possible pairs of data and averaged over 8km distance classes. The plot indicates that the NosTaxa observed in transects that are in close proximity are as different as those that are further apart.

Underwater video and stills imaging provides an invaluable insight into the habitat types and dominant species present in the deep-sea environment and aids an intuitive understanding of the dominant conditions and potential forcing factors. The data used here make a valuable contribution to a description on deep-sea epifaunal assemblages. However, the data are not of sufficient breadth, over time or in space, to sustain a comprehensive statistical analysis. The collation of photographs, across different transects, resulted in a semi-quantitative assessment of abundance (an example could be 5 out of 20 photographs contained a certain species) and this metric could be used in more sophisticated analyses based on counts (though the widely varying numbers of photographs, per transect, would make this technically demanding). Such analyses would be superior because they would include more data (a count rather than a presence/absence). Further details of the limitations in these data, and the sampling/monitoring programme that would be necessary for a sufficiently thorough analysis are given in Section 4.

4 Data limitations and knowledge gaps

The FSC and WSC are among the most studied areas of the deep-sea and the data supplied by JNCC constitute one of the most comprehensive data sets concerning the deep-sea that are available. These data were gathered in order to assess the faunal distribution as a function of the convergence of two water masses (cold Arctic and warmer NE Atlantic water), where temperatures vary by up to 8°C over relatively small temporal (few months) and spatial scales (~100m depth). These data, much of which is published (see references by Narayanaswamy and/or Bett), have enabled a characterisation of this part of the deep-sea fauna. However, these data were not gathered with a view to initiating a long-term monitoring programme or, indeed, for assessing the 'status' of the deep-sea from any statutory perspective. The data analysed here do enable a critical first evaluation of the issues, challenges and possible solutions to deep-sea monitoring and constitute essential pilot-studies in this region in terms of developing monitoring protocols. There are three main interrelated deficiencies in terms of the current data set in terms of fulfilling the objectives as set out in the tender document (see Section 1.5). These are: 1) the lack of time-series data, 2) the coarse taxonomic resolution (identification to family level only), and 3) spatial/temporal confounding. Other data deficiencies are also discussed in the following sections.

4.1 Temporal and spatial variability

One of the principal objectives of this work is to evaluate the utility of the present macrobenthic dataset as the basis for developing a statistically-robust long-term monitoring programme for the UK deep-sea. In this respect, a significant limitation is the brief temporal extent of the dataset, which as a result provides very limited information on patterns of interannual (and intra-annual) change. Discussions with JNCC indicated an interest in detecting directional long-term trends and separating anthropogenic and natural causes of this change, through the development of a robust monitoring programme in the future. This would require formal time-series analysis and the adoption of sophisticated time-series models (e.g. auto-regressive indexed moving average, ARIMA, models). To employ such models requires a considerable volume of historic data (the actual amount depending on the model employed and the number of parameters that are being estimated). Accurate assessments of macrobenthic diversity stretching back many years are, obviously, not available.

The scarcity of long-term datasets is a recognised problem in deep-sea ecology (Glover *et al.*, 2010) and one which seriously limits our ability to distinguish natural change from anthropogenic impacts, and to predict the likely impacts of a changing climate (Smith *et al.*, 2008; Smith *et al.*, 2009). The few ongoing studies of long-term dynamics of deep-sea benthic communities have focused largely on epibenthic megafauna (Glover *et al.*, 2010) and even less information is available for the macrobenthic infauna. The best available data on long-term change in deep-sea macrobenthic communities come from two abyssal sites in the north-east Atlantic (1991-1999) and north-east Pacific (1991-2005) (Laguionie-Marchais *et al.*, 2013). Polychaete communities at both sampling locations showed interannual variation in density, family evenness and rank abundance distributions. In both time-series the greatest changes occurred in 1998, when polychaete densities peaked, accompanied by changes in the rank abundance of the major families and functional groups. No meaningful associations were found between polychaete density and particulate organic matter flux or climate indices, and the authors were also unable to identify ecological factors driving the family-level changes. The two studies reported by Laguionie-Marchais *et al.* (2013) refer to abyssal plain environments of uniform seabed topography and substratum type and largely stable benthic hydrography. If the ecological drivers of interannual change are hard to determine in a setting such as this, the challenge is likely to be far greater along the

continental margin of the FSC, an environment characterised by steep depth gradients, wide variations in substrata, highly complex hydrography, and which spans the biogeographic boundary between Atlantic and Arctic faunas. Detecting any type of change, including anthropogenic change, in this environment will be extremely difficult.

4.2 Taxonomic resolution

Species identification is problematic in deep-sea samples and, consequently, the analyses presented here were based on identification to family level only. This means that diversity indices are based on families, not species. The family-level identification approach has been used in multivariate analyses of intertidal fauna where there was no meaningful loss of analytical sensitivity (Warwick, 1988) and it is worthwhile considering a similar approach in terms of deep-sea data.

The polychaete community analyses of Laguionie-Marchais *et al.* (2013) were performed at the family level. Family-level analysis may be necessary in deep-sea studies where accurate species-level identifications may not be available, or are not standardised. The problem of non-standardisation is likely to occur where datasets have been produced by different research groups. Family-level analysis entails a loss of information in comparison with using species-level data, and this may reduce the degree of confidence that can be assigned to any observed patterns. Narayanaswamy *et al.* (2014) attempted to define macrofaunal assemblages by multivariate analysis of family-level FSC and Rockall Trough datasets. Eleven macrofaunal assemblages were defined by cluster analysis, but it was necessary to set a relatively low threshold level of similarity (~50%), and a large number of sample locations were not included within the defined categories. Given the problems of species-level standardisation and the relatively coarse resolution of family-level data, analysis to the genus level need to be considered. Bett and Narayanaswamy (2014) compared genus- and species-level studies of the diversity and ecology of deep-sea macrobenthos on the West Shetland Slope. They concluded that genus-level α- and β-diversity measures are highly correlated and are good predictors of their species-level equivalents and that community ecology is very well-described by genus-level data. Given the complexity of the West Shetland Slope environment, it may be reasonable to expect these conclusions to hold in other deep-sea environments.

4.3 Other data limitations

There are a number of confounding issues in relation to the supplied data. The samples were collected at different months in different years (Table 13) meaning that any apparent differences between years might be attributable to differences between seasons. The intra-annual variability, in terms of diversity and evenness in this part of the deep-sea, is not known.

Table 13. Number of samples taken during different years highlighting seasonal differences. Differences between different years might be accounted for by differences in season (this applies particularly to the 1998 data).

Year	May	June	July	August	September
1996	0	0	84	76	0
1998	21	6	0	0	0
2000	0	0	12	71	4
2002	0	0	13	49	0

The samples were also collected using different techniques, normally as a function of the water depth and/or substratum type (Table 14), for example, 93% of samples taken from deeper than 600 m were collected using a mega-corer. There are mechanisms for correcting for gear-type, in terms of benthic biomass (Narayanaswamy and Bett, 2011), and noted issues in terms of comparing meiobenthic assemblages based on samples collected using different gears (Bett *et al.*, 1994). Consideration should be given to methods which standardise the species/family counts across differing sampling methods and basing any analysis on the corrected data.

Table 14. Numbers of samples taken by different gear types at different depths. Differences between depths could be a function of gear type (rather than depth).

Depth (m)	Box corer (BC)	Day grab	Megacorer (MgC)	BC+MgC
100-300	0	60	0	0
300-600	66	8	3	0
600-1200	8	0	132	4
1200-1500	4	0	21	0
> 1500	0	0	30	0

In addition, sampling during different years tended to occur in different areas and, to a certain extent, different depths; consequently, some areas/locations have been sampled only during a single year. This means any apparent difference between years might be due to differences in space or *vice versa* and means that, in effect, we have neither spatial nor temporal data available (just data from different sample locations collected during different years.

4.4 Conclusions in relation to an optimal monitoring programme

An assessment of change, of whatever type, necessitates a comparison with either another time or space. In the absence of historical data nothing can be said about the present state in relation to the historical 'norm'. An alternative method for detecting change is to compare the system under investigation with other 'similar' areas that are not subject to the source of impact that is under assessment. This method necessarily excludes impacts that have a global scale (e.g. climate change) because all areas will/ may be subject to this stressor. However, this approach would allow the assessment of the entire FSC in, for example, an assessment of fishing-damage (provided the other areas were not subject to this fishing pressure). In the absence of historical data, and data from other areas of the deep-sea, the

only remaining course of action is to assess changes within the FSC. This approach allows the assessment of small-scale changes, for example, point-source impacts or spatially discrete impacts, for example, from bottom-contact gear fishing occurring within a particular stratum.

In summary, the data supplied under this contract do not permit the design of a monitoring programme, with a quantified statistical power, to detect long-term change in deep-sea benthic status within the study area and/or to distinguish the cause (anthropogenic or natural) to any change identified. There are two major gaps in our understanding which would need to be filled before such a design could be made:

1. There is no record of natural variability, over time, at the same location (sampling location), in the benthic assemblages within the surveyed area; and

2. There is an insufficient understanding of how the metrics derived from the data supplied (number of families and evenness) respond in relation to impacts (of whatever cause) in the deep-sea.

The data do allow an assessment (with limitations, see below) of the power of monitoring programmes designed to assess impacts occurring at a single point (e.g. oil well) or occurring at the level of single strata (as defined using stratification system 7). The power of beyond-BACI monitoring was determined from estimates of how the metrics under consideration vary, over time, between different strata. In an actual monitoring situation, the reference stations would be repeatedly sampled, over time, to compare with the impacted station(s). In all likelihood, the variability, at the same location over time, would be less than that determined from the data: in the current case an estimate of how the number of families varied, over time, *between different strata* was used in the model. This may represent an overestimate because there is likely to be greater temporal variability in assemblage structure over larger spatial scales. In order to better estimate sampling effort for point-source impacts, or impacts occurring at the scale of a single stratum, there needs to be greater understanding of small-scale (e.g. within a few km and at the same depth) temporal variability in the metric being studied. Currently, these data are not available.

5 Recommendations

Carrying out observational activities in the deep-sea is challenging and expensive so there is a requirement to maximise the value of existing knowledge in order to inform subsequent monitoring programmes. The objective in the present study was to, where possible, use the data available to inform the design of future monitoring programmes and to identify the nature of limitations to the present data that need to be addressed before a more-refined monitoring programme could be designed.

From these results, it is thought that there is currently no way to understand how system-wide changes (e.g. climate change) are affecting the FSC and, consequently, no way to distinguish anthropogenic and natural trends. In order to make such an assessment, historical data is required and this would, ideally, need to extend to a period preceding the industrial revolution. This problem is widely acknowledged in deep-sea research.

Making recommendations is complicated for a number of reasons. These reasons include that we currently do not understand how the diversity metrics (including NosFam and Piel) respond to change, or what constitutes a meaningful departure from baseline conditions (threshold values). Our recommendations for BACI designs would differ depending on the metric (NosFam or Piel) chosen because different stratification systems resulted in different

parameter estimates (see below) and these may change if the data were re-analysed based on species-level identification (which is one of our recommendations). In designing a monitoring programme several factors need to be considered including the expected sampling effort (numbers of replicates), the Type I/II error rate and, in terms of monitoring over time, the size of the location:time interaction terms and all of these need to be specified in order to estimate power.

Allowing for the caveats above, there are three broad categories of inter-related recommendations. These are:

1. Sampling recommendations in relation to the data analysed;
2. Recommendations in relation to filling the data gaps identified; and
3. Broader recommendations that should be considered in any future deep-sea monitoring programme.

Caveats apply to our recommendations and these are detailed in the main text and are not repeated here.

1. Sampling recommendations in relation to the data analysed
    a. There were clear differences in NosFam between strata when S7 was used. S7 should be used as a basis for separating areas in assessing impacts that are occurring on the per-stratum scale when assessing change using NosFam. If Piel is chosen as the response metric (see below) then stratification system S9 is a better way of stratifying the sampling approach.
    b. It was found that spatial autocorrelation was negligible at distances > 20 km. It is, therefore, recommended that minimum separation distance of 20 km between sampling locations.
    c. NosFam should be used as the metric to determine small-scale impacts (e.g. on the scale of fishing or oil-well) where it can be shown that relevant location:time interaction term is less than 10.
    d. Any monitoring programme should be conducted annually (once the intra-annual variability is understood) in order to assess change in assemblage composition. This is because many members of the macrobenthic and epibenthic assemblages show annual recruitment. This programme should start immediately whilst acknowledging that that the FSC may already be impacted.

2. Recommendations in relation to filling data gaps.
    a. Within-year sampling should be conducted to assess the extent of seasonable variability in the metrics be considered. Random samples (in time) should be collected within that period to assess temporal stability. The random samples should be collected within all levels of any adopted stratification system (e.g. across a range of habitat types such as those defined under S7).
    b. In relation to 1 c. above, it is recommended that variability over small scales (same sampling location) be assessed over time. This should be done for three years in the first instance and an assessment of the temporal variability then undertaken.
    c. An assessment of the relationship between diversity /evenness metrics and sampling technique should be conducted in order to show whether direct comparisons, in diversity /evenness, from samples taken using different gear-types, is reasonable.
    d. A review of available data from the shelf areas should be conducted as these areas are highly under-represented in the dataset investigated in this contract

and no information about how benthic assemblages change over time could be extracted.

e. In terms of the epifaunal data, the potential of generating semi-quantitative assessments from collations across the available photographs should be assessed.

3. Broader recommendations
    a. Consider the spatial domain for monitoring e.g. in monitoring the deep-sea is it optimal (within, for example, a limited budget) to sample all strata. If not, then attention should be focussed on those strata where the temporal variability in the metric being studied is lowest. In the current context this would be the deeper parts (e.g. lower bathyal strata of S9).
    b. Critically evaluate the credibility of basing any monitoring and assessment programme on diversity indices (of any type) and determine the sensitivity of such assessments based on family-level identification.
    c. Evaluate existing data (AFEN data), which consists of genus-level identification, in terms of sensitivity (e.g. whether strata differences show a greater degree of separation).
    d. Evaluate the potential of combining samples, at various spatial scales of sampling, to reduce variability (this requires an understanding of fine-scale variability and is called compositing; Carey and Keough (2002b)).
    e. Conduct a thorough review and assessment of analogous shallow-water long-term monitoring programmes to evaluate sampling options. This would include an assessment of sampling design in terms of sampling locations (e.g. fully random v. initially random then repeated visits, or systematic) (van der Meer, 1997).
    f. Evaluate technological advances which may reduce sampling processing costs (e.g. molecular methods (Pawlowski *et al.*, accepted)).

6 References

ANDERSON, D. R., BURNHAM, K. P. & THOMPSON, W. L. 2000. Null hypothesis testing: Problems, prevalence, and an alternative. *Journal of Wildlife Management,* **64**: 912-923.

BETT, B. J. 2001. UK Atlantic Margin Environmental Survey: Introduction and overview of bathyal benthic ecology. *Continental Shelf Research,* **21**: 917-956.

BETT, B. J. 2012. Seafloor biotope analysis of the deep waters of the SEA4 region of Scotland's seas. Report No. 472. JNCC.

BETT, B. J. & NARAYANASWAMY, B. E. 2014. Genera as proxies for species α- and β-diversity: tested across a deep-water Atlantic–Arctic boundary. *Marine Ecology*: n/a-n/a.

BETT, B. J., VANREUSEL, A., VINCX, M., SOLTWEDEL, T., PFANNKUCHE, O., LAMBSHEAD, P. J. D., GOODAY, A. J., FERRERO, T. & DINET, A. 1994. Sampler bias in the quantitative study of deep-sea meiobenthos. *Marine Ecology Progress Series,* **104**: 197-203.

BURROWS, M. T., HARVEY, R., ROBB, L., POLOCZANSKA, E. S., MIESZKOWSKA, N., MOORE, P., LEAPER, R., HAWKINS, S. J. & BENEDETTI-CECCHI, L. 2009. Spatial scales of variance in abundance of intertidal species: effects of region, dispersal mode, and trophic level. *Ecology,* **90**: 1242-1254.

CAREY, J. M. & KEOUGH, M. J. 2002a. The variability of estimates of variance, and its effect on power analysis in monitoring design. *Environmental Monitoring and Assessment,* **74**: 225-241.

CAREY, J. M. & KEOUGH, M. J. 2002b. Compositing and subsampling to reduce costs and improve power in benthic infaunal monitoring programs. *Estuaries,* **25**: 1053-1061.
CLARKE, K. R. 1993. Nonparametric multivariate analyses of changes in community structure. *Australian Journal of Ecology,* **18**: 117-143.

CLARKE, K. R. & WARWICK, R. M. 1998. A taxonomic distinctness index and its statistical properties. *Journal of Applied Ecology,* **35**: 523-531.

CRAWLEY, M. J. 2007. *The R Book.,* John Wiley and Sons Ltd., Chichester, UK.

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.

FIDLER, F., THOMASON, N., CUMMING, G., FINCH, S. & LEEMAN, J. 2004. Editors Can Lead Researchers to Confidence Intervals, but Can't Make Them Think: Statistical Reform Lessons From Medicine. *Psychological Science,* **15**: 119-126.

GAGE, J. D. & BETT, B. 2005. Chapter 7 Deep-sea benthic sampling. *In: Methods for the study of marine benthos* (eds A. Eleftheriou & A. D. McIntyre). Blackwell Science Ltd, Oxford, London, Edinburgh.

GIGERENZER, G. 2004. Mindless statistics. *Journal of Socio-Economics,* **33**: 587-606.

GLOVER, A. G., GOODAY, A. J., BAILEY, D. M., BILLETT, D. S. M., CHEVALDONNE, P., COLACO, A., COPLEY, J., CUVELIER, D., DESBRUYERES, D., KALOGEROPOULOU, V., KLAGES, M., LAMPADARIOU, N., LEJEUSNE, C., MESTRE, N. C., PATERSON, G. L. J., PEREZ, T., RUHL, H., SARRAZIN, J., SOLTWEDEL, T., SOTO, E. H., THATJE, S., TSELEPIDES, A., VAN GAEVER, S. & VANREUSEL, A. 2010. Temporal change in deep-sea benthic ecosystems: a review of the evidence from recent time-series studies. *In: Advances in Marine Biology, Vol 58* (ed. M. Lesser), pp. 1-95.

HEIP, C. H. R., HERMAN, P. M. J. & SOETAERT, K. E. R. 1998. Indices of diversity and evenness. *Oceanis,* **24**.

HOWELL, K. L., DAVIES, J. S. & NARAYANASWAMY, B. E. 2010. Identifying deep-sea megafaunal epibenthic assemblages for use in habitat mapping and marine protected area network design. *Journal of the Marine Biological Association of the United Kingdom,* **90**: 33-68.

HUGHES, D. J. 2014. Benthic habitat and megafaunal zonation across the Hebridean Slope, western Scotland, analysed from archived seabed photographs. *Journal of the Marine Biological Association of the United Kingdom,* **FirstView**: 1-16.

JACOBS, C. L. & HOWELL, K. L. 2007. MV Franklin Cruise 0206, 03-23 Aug 2006, habitat investigations within the SEA4 and SEA7 areas of the UK continental shelf, Research & Consultancy Report No. 24. National Oceanography Centre, Southampton.
JOHNSON, D. H. 1999. The Insignificance of Statistical Significance Testing. *The Journal of Wildlife Management,* **63**: 763-772.

LAGUIONIE-MARCHAIS, C., BILLETT, D. S. M., PATERSON, G. L. D., RUHL, H. A., SOTO, E. H., SMITH, K. L., JR. & THATJE, S. 2013. Inter-annual dynamics of abyssal polychaete communities in the North East Pacific and North East Atlantic-A family-level study. *Deep-Sea Research Part I-Oceanographic Research Papers,* **75**: 175-186.
MAGURRAN, A. E. 1988. *Ecological diversity and its measurement,* Chapman and Hall, London.

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

NAKAGAWA, S. & CUTHILL, I. C. 2007. Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biological Reviews,* **82**: 591-605.

NARAYANASWAMY, B. E. & BETT, B. J. 2011. Macrobenthic Biomass Relations in the Faroe-Shetland Channel: An Arctic-Atlantic Boundary Environment. *Plos One,* **6**: e18602.
NARAYANASWAMY, B. E., BETT, B. J. & GAGE, J. D. 2005. Ecology of bathyal polychaete fauna at an Arctic-Atlantic boundary (Faroe-Shetland Channel, North-east Atlantic). *Marine Biology Research,* **1**: 20-32.

NARAYANASWAMY, B. E., BETT, B. J. & HUGHES, D. J. 2010. Deep-water macrofaunal diversity in the Faroe-Shetland region (NE Atlantic): a margin subject to an unusual thermal regime. *Marine Ecology,* **31**: 237-246.

NARAYANASWAMY, B. E., NICKELL, T. D. & HUGHES, D. J. 2014. Definition of infaunal assemblages for inclusion in a deep-sea section of the Marine Habitat Classification of Britain & Ireland. Report to the JNCC. Scottish Association for Marine Science, Oban.

PAWLOWSKI, J., ESLING, P., LEJZEROWICZ, F., CEDHAGEN, T. & WILDING, T. accepted. Environmental monitoring through protist NGS metabarcoding: assessing the impact of fish farming on benthic foraminiferal communities. *Molecular Ecology Resources*.

PETERMAN, R. M. 1990. Statistical power analysis can improve fisheries research and management. *Canadian Journal of Fisheries and Aquatic Sciences,* **47**: 2-15.
PIECHAUD, N. & HOWELL, K. L. 2013. Definition of epifaunal assemblages for inclusion in a deep-sea section of the Marine Habitat Classification of Britain and Ireland: Methods report. JNCC.

QUINN, G. P. & KEOUGH, M. J. 2002. *Experimental Design and data analysis for biologists,* Cambridge University Press.

ROSSI, R. E., MULLA, D. J., JOURNEL, A. G. & ELDON, H. F. 1992. Geostatistical Tools for Modeling and Interpreting Ecological Spatial Dependence. *Ecological Monographs,* **62**: 277-314.

SCHMITT, R. J. & OSENBERG, C. W. 1996. Detecting ecological impacts caused by human activities. *In: Detecting Ecological Impacts* (eds R. J. Schmitt & C. W. Osenberg), pp. 3 - 15. Academic Press, San Diego.

SHAW, P. J. A. 2003. *Multivariate statistics for the environmental sciences,* Hodder Arnold, London. .

SMITH, C. R., DE LEO, F. C., BERNARDINO, A. F., SWEETMAN, A. K. & ARBIZU, P. M. 2008. Abyssal food limitation, ecosystem structure and climate change. *Trends in Ecology & Evolution,* **23**: 518-528.

SMITH, K. L., JR., RUHL, H. A., BETT, B. J., BILLETT, D. S. M., LAMPITT, R. S. & KAUFMANN, R. S. 2009. Climate, carbon cycling, and deep-ocean ecosystems. *Proceedings of the National Academy of Sciences of the United States of America,* **106**: 19211-19218.

SOKAL, R. R. & ROHLF, F. J. 1995. *Biometry: the principles and practice of statistics in biological research,* W. H. Freeman and Company, New York.

STEWART-OATEN, A. & BENCE, J. R. 2001. Temporal and spatial variation in environmental impact assessment. *Ecological Monographs,* **71**: 305-339.

STROUP, W. W. 2002. Power analysis based on spatial effects mixed models: A tool for comparing design and analysis strategies in the presence of spatial variability. *Journal of Agricultural, Biological, and Environmental Statistics,* **7**: 491-511.

TETT, P., GOWEN, R., PAINTING, S., ELLIOTT, M., FORSTER, R., MILLS, D., BRESNAN, E., CAPUZZO, E., FERNANDES, T., FODEN, J., GEIDER, R., GILPIN, L., HUXHAM, M., MCQUATTERS-GOLLOP, A., MALCOLM, S., SAUX-PICART, S., PLATT, T., RACAULT, M., SATHYENDRANATH, S., VAN DER MOLEN, J. & WILKINSON, M. 2013. Framework for understanding marine ecosystem health. *Marine Ecology Progress Series,* **494**: 1-27.

UNDERWOOD, A. J. 1991. Beyond BACI - experimental designs for detecting human environmental impacts on temporal variations in natural populations. *Australian Journal of Marine and Freshwater Research,* **42**: 569-587.

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

UNDERWOOD, A. J. 1994. On Beyond BACI: Sampling Designs that Might Reliably Detect Environmental Disturbances. *Ecological Applications,* **4**: 3-15.

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**: 167-179.

WARWICK, R. M. 1988. The level of taxonomic discrimination required to detect pollution effects on marine benthic communities. *Marine Pollution Bulletin,* **19**: 259-268.

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

7 Appendices

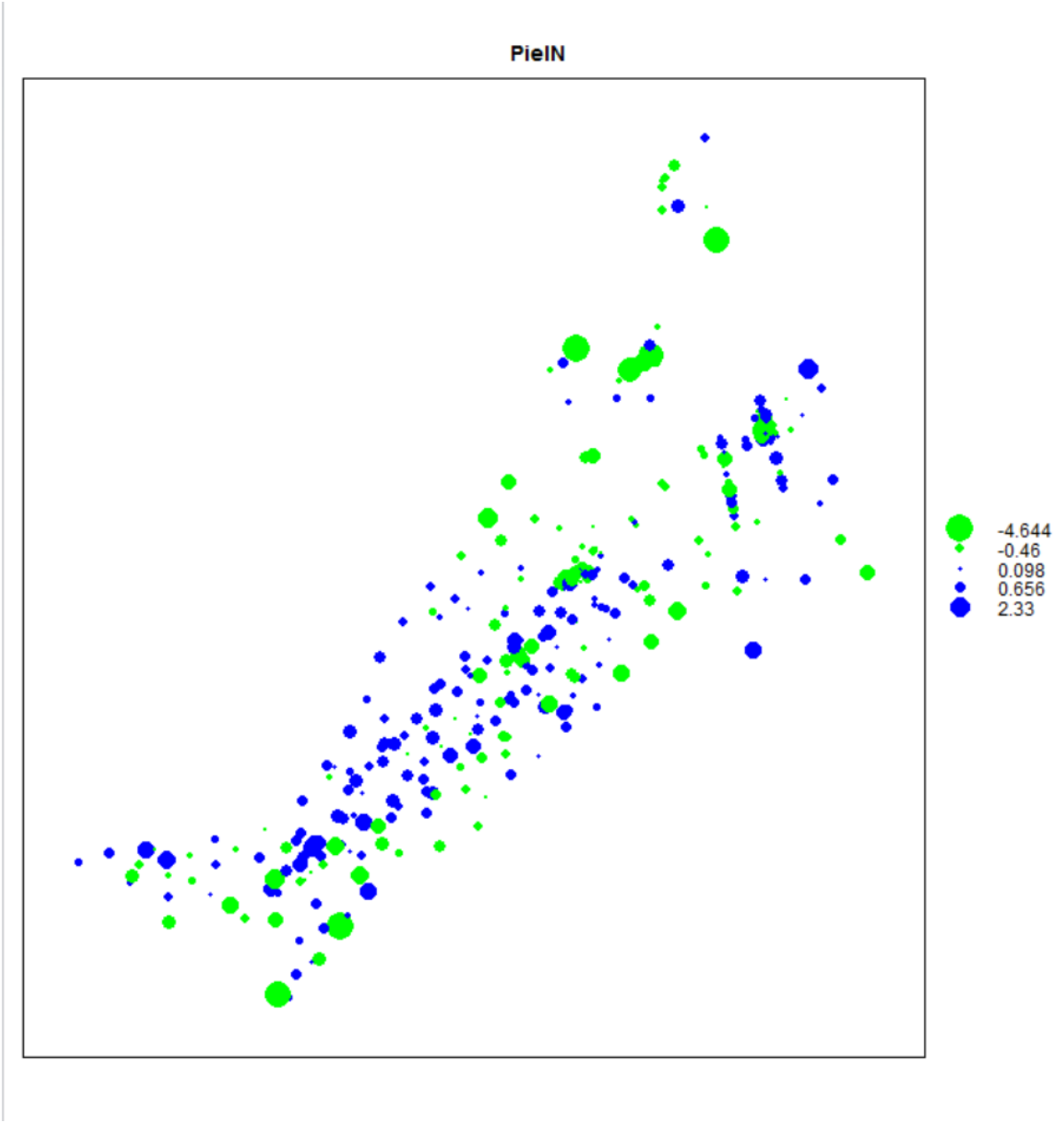

Figure A1. Piel as a function of location. Note the mix of green and blue dots which indicates that Piel is relatively consistent across the spatial domain.

Table A.1. Results from the ANOVA. For each ANOVA model the total variance is the same but some partition this better between levels of strata, time and their interaction.

```
Response: NosFam
           Df  Sum Sq Mean Sq F value    Pr(>F)
S1          2  5627.8 2813.90 30.9760 4.827e-13 ***
Year        3   370.1  123.38  1.3582    0.2555
S1:Year     4  2702.8  675.69  7.4381 9.608e-06 ***
Residuals 326 29614.2   90.84
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> S2=lm(NosFam~S2*Year,data=DS);anova(S2);
Analysis of Variance Table

Response: NosFam
           Df  Sum Sq Mean Sq F value    Pr(>F)
S2          4  6782.2 1695.56 18.9984 4.955e-14 ***
Year        3   208.8   69.59  0.7798 0.5059489
S2:Year     8  2764.7  345.59  3.8723 0.0002253 ***
Residuals 320 28559.2   89.25
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> S3=lm(NosFam~S3*Year,data=DS);anova(S3);
Analysis of Variance Table

Response: NosFam
           Df  Sum Sq Mean Sq F value    Pr(>F)
S3         10  8099.3  809.93  9.6000 6.196e-14 ***
Year        3   259.9   86.62  1.0267     0.381
S3:Year    14  3970.7  283.62  3.3617 4.671e-05 ***
Residuals 308 25985.1   84.37
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> S4=lm(NosFam~S4*Year,data=DS);anova(S4);
Analysis of Variance Table

Response: NosFam
           Df  Sum Sq Mean Sq F value  Pr(>F)
S4          4  9814.0 2453.49 29.6900 < 2e-16 ***
Year        3   484.4  161.48  1.9541 0.12081
S4:Year    11  1820.6  165.51  2.0028 0.02767 *
Residuals 317 26195.9   82.64
```

```
           Df Sum Sq Mean Sq F value   Pr(>F)
S5          1    800  799.54  7.2289 0.007539 **
Year        3   1118  372.59  3.3687 0.018831 *
S5:Year     2      9    4.74  0.0429 0.958043
Residuals 329  36388  110.60
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> S6=lm(NosFam~S6*Year,data=DS);anova(S6);
Analysis of Variance Table

Response: NosFam
           Df  Sum Sq Mean Sq F value  Pr(>F)
S6          7 10121.9 1445.98 17.6808 < 2e-16 ***
Year        3   769.8  256.59  3.1375 0.02567 *
S6:Year     8  1498.2  187.27  2.2899 0.02140 *
Residuals 317 25925.1   81.78
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> S7=lm(NosFam~S7*Year,data=DS);anova(S7);
Analysis of Variance Table

Response: NosFam
           Df  Sum Sq Mean Sq F value  Pr(>F)
S7          9 20942.6 2326.96 45.4157 < 2e-16 ***
Year        3   284.0   94.66  1.8474 0.13853
S7:Year    15  1307.3   87.16  1.7010 0.04961 *
Residuals 308 15781.0   51.24
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> S8=lm(NosFam~S8*Year,data=DS);anova(S8);
Analysis of Variance Table

Response: NosFam
           Df  Sum Sq Mean Sq F value    Pr(>F)
S8          3  9681.4  3227.1 39.0510 < 2.2e-16 ***
Year        3   308.3   102.8  1.2437  0.293891
S8:Year     9  1880.8   209.0  2.5289  0.008191 **
Residuals 320 26444.4    82.6
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> S9=lm(NosFam~S9*Year,data=DS);anova(S9);
Analysis of Variance Table

Response: NosFam
           Df  Sum Sq Mean Sq F value    Pr(>F)
S9          5  6122.7 1224.54 13.3904 7.181e-12 ***
Year        3   970.1  323.37  3.5360  0.015082 *
S9:Year     5  1775.3  355.07  3.8827  0.001974 **
Residuals 322 29446.8   91.45
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

ANOVA results for Piel

```
Response: Piel
           Df  Sum Sq  Mean Sq F value    Pr(>F)
S1          2 0.10561 0.052804 11.5184 1.467e-05 ***
Year        3 0.05750 0.019167  4.1811  0.006344 **
S1:Year     4 0.06399 0.015997  3.4896  0.008283 **
Residuals 326 1.49448 0.004584
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> S2=lm(Piel~S2*Year,data=DS);anova(S2);
Analysis of Variance Table

Response: Piel
           Df  Sum Sq   Mean Sq F value    Pr(>F)
S2          4 0.10937 0.0273435  5.9928 0.0001161 ***
Year        3 0.06083 0.0202773  4.4442 0.0044615 **
S2:Year     8 0.09132 0.0114144  2.5017 0.0119898 *
Residuals 320 1.46006 0.0045627
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> S3=lm(Piel~S3*Year,data=DS);anova(S3);
Analysis of Variance Table

Response: Piel
           Df  Sum Sq   Mean Sq F value    Pr(>F)
S3         10 0.20402 0.0204019  4.3919 8.896e-06 ***
Year        3 0.02580 0.0086012  1.8516    0.1378
S3:Year    14 0.06100 0.0043572  0.9380    0.5180
Residuals 308 1.43076 0.0046453
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> S4=lm(Piel~S4*Year,data=DS);anova(S4);
Analysis of Variance Table

Response: Piel
           Df  Sum Sq  Mean Sq F value    Pr(>F)
S4          4 0.20095 0.050236 11.1758 1.709e-08 ***
Year        3 0.05926 0.019752  4.3942  0.004777 **
S4:Year    11 0.03643 0.003312  0.7368  0.702885
Residuals 317 1.42495 0.004495
---
```

```
Response: Piel
           Df  Sum Sq  Mean Sq F value   Pr(>F)
S5          1 0.05093 0.050926 10.2556 0.001496 **
Year        3 0.03082 0.010274  2.0691 0.104142
S5:Year     2 0.00612 0.003060  0.6162 0.540611
Residuals 329 1.63371 0.004966
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> S6=lm(Piel~S6*Year,data=DS);anova(S6);
Analysis of Variance Table

Response: Piel
           Df  Sum Sq  Mean Sq F value    Pr(>F)
S6          7 0.22968 0.032812  7.3094 3.893e-08 ***
Year        3 0.06072 0.020239  4.5087  0.004095 **
S6:Year     8 0.00817 0.001022  0.2276  0.985711
Residuals 317 1.42301 0.004489
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> S7=lm(Piel~S7*Year,data=DS);anova(S7);
Analysis of Variance Table

Response: Piel
           Df  Sum Sq   Mean Sq F value    Pr(>F)
S7          9 0.18896 0.0209956  4.5557 1.189e-05 ***
Year        3 0.06763 0.0225431  4.8915  0.002455 **
S7:Year    15 0.04553 0.0030354  0.6586  0.824230
Residuals 308 1.41946 0.0046086
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> S8=lm(Piel~S8*Year,data=DS);anova(S8);
Analysis of Variance Table

Response: Piel
           Df  Sum Sq  Mean Sq F value   Pr(>F)
S8          3 0.19195 0.063984 14.3341 8.72e-09 ***
Year        3 0.05942 0.019806  4.4370 0.004505 **
S8:Year     9 0.04181 0.004645  1.0407 0.407184
Residuals 320 1.42840 0.004464
---
```

```
Response: Piel
           Df  Sum Sq  Mean Sq F value    Pr(>F)
S9          5 0.18527 0.037055  8.1329 3.034e-07 ***
Year        3 0.06132 0.020441  4.4865  0.004211 **
S9:Year     5 0.00791 0.001582  0.3472  0.883881
Residuals 322 1.46708 0.004556
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```