Fewer sites but better data? Optimising the representativeness and statistical power of a national monitoring network

Matthew T. O'Hare [a], Iain D. M. Gunn [a], Nathan Critchlow-Watton [b], Robin Guthrie [b], Catriona Taylor [b], Daniel S. Chapman [a,c*]

[a] Centre for Ecology & Hydrology, Penicuik, Midlothian, EH26 0QB, UK

[b] Scottish Environment Protection Agency, Stirling, FK9 4TZ, UK

[c] Biological and Environmental Sciences, University of Stirling, Stirling, FK9 4LA, UK

* Corresponding author: daniel.chapman@stir.ac.uk

Abstract

Indicators of large-scale ecological change are typically derived from long-term monitoring networks. As such, it is important to assess how well monitoring networks provide evidence for ecological trends in the regions they are monitoring. In part, this depends on the network's representativeness of the full range of environmental conditions occurring in the monitored region. In addition, the statistical power to detect trends and ecological changes using the network depends on its structure, size and the intensity and accuracy of monitoring. This paper addresses the optimisation of representativeness and statistical power when re-designing existing large-scale ecological monitoring networks, for example due to financial constraints on monitoring programmes. It uses a real world example of a well-established river monitoring network of 254 sites distributed across Scotland. We first present a novel approach for assessing a monitoring network's representativeness of national habitat and pressure gradients using the multivariate two-sample Cramér's *T* statistic. This compares multivariate gradient distributions among sites inside and outside of the network. Using this test, the existing network was found to over-represent larger and more heavily polluted sites, reflecting earlier research priorities when it was originally designed. Network re-design was addressed through stepwise selection of individual sites to remove from or add to the network to maximise multivariate representativeness. This showed that combinations of selective site retention and addition can be used to modify existing monitoring networks, changing the number of sites and improving representativeness. We then investigated the effect of network re-design on the statistical power to detect long-term trends across the whole network. The power analysis was based on linear mixed effects models for long-term trends in three ecological indicators (ecological quality ratios for diatoms, invertebrates and macrophytes) over a ten-year period. This revealed a clear loss of power in smaller networks with less accurate sampling, but sampling schedule had a smaller effect on power. Interestingly, more representative networks had slightly lower trend detection power than the current unrepresentative network, though they should give a less biased estimate of national trends. Our analyses of representativeness and statistical power provide a general framework for designing and adapting large-scale ecological monitoring networks. Wider use of such methods would improve the quality of indicators derived from them and improve the evidence base for detecting and managing ecological change.

Keywords: Environmental change; Ecological monitoring; Monitoring network; Spatial prioritisation; Power analysis; Water Framework Directive.

1. Introduction

Long-term monitoring allows the assessment of the state of the environment, detection of ecological change and evaluation of the effects of stressors or management interventions on ecological systems (Lindenmayer and Likens, 2010; Lovett et al., 2007). Indeed indicators derived from monitoring data often provide the evidence that informs management. As such, achieving adequate data quality whilst controlling costs and resource requirements are core challenges for the design of monitoring networks. These are generic issues applicable to aquatic and terrestrial systems, individual and multi-species monitoring programmes and they have attracted significant research interest (Carvalho et al., 2016; Munkittrick et al., 2009; Rhodes and Jonzén, 2011; Stegman et al., 2017; Wikle and Royle, 1999). However, most previous studies have considered initial network design rather than strategies for revising or modifying existing long-term monitoring networks (Levine et al., 2014). This omission is important since long-term monitoring networks will be periodically reviewed and may be revised both for scientific and budgetary reasons.

An important consideration for the design and modification of monitoring networks is representativeness, i.e. the network's proportionate coverage of the full range of environmental conditions occurring in the monitored region (Urquhart and Kincaid, 1999). Monitoring networks should be representative because indicators from unrepresentative networks may provide a biased representation of patterns across the monitored region. Stratified random sampling of sites is generally advocated as an approach to produce monitoring networks an unbiased representative sample of the range of sites in the area of interest (Vos et al., 2000). However, for various reasons, this is not always done. Monitoring networks often grow and evolve over time and at each step the priorities for representativeness may change, so it is not uncommon to end up with networks that are not fully representative. For example river monitoring networks often have an original sampling design focused on comparable sites upstream and downstream of point sources of pollution, such as sewage treatment works (SEPA, 2007). These may be useful for determining the effects of pollution, but do not represent high elevation rivers that are not generally impacted by pollution and so the network is less useful for estimating national-scale trends resulting from, for example, climate change. For both network design and re-design there is a need for statistical tools and algorithms to prioritise sites for inclusion or removal from monitoring networks in order to improve representativeness.

The design and revision of monitoring programs should also take account of their statistical power to detect trends and ecological changes. For this, generic power analysis tools (Cohen, 2013; Johnson et al., 2015; Thomas, 1997) can be applied to statistical models fitted to data from monitoring networks (Irvine et al., 2012; Peterman, 1990). In general, statistical power will depend on the size of the monitoring network, its sampling intensity and the accuracy of data collection (Levine et al., 2014; Osenberg et al., 1994). It is often advocated that pilot datasets are used to investigate power to detect a specific level of change in advance of establishing a monitoring programme (Osenberg et al., 1994; Peterman, 1990; Toft and Shea, 1983). In practice this is rarely implemented, especially for long-term monitoring of systems that change slowly over time. Nevertheless, when long-term monitoring programs are periodically revised, retrospective power analysis on existing monitoring data (Thomas, 1997) is a pragmatic approach to evaluate the effect of proposed network redesign or revision to sampling strategies.

This paper addresses the optimisation of representativeness and statistical power of a large-scale long-term ecological monitoring network – Scotland’s national river surveillance network of 254 monitoring sites (SEPA, 2007). The network is a European Union (EU) Water Framework Directive (WFD) surveillance network (European Commission, 2000). Similar networks exist in all EU member states and their purpose is to allow the ecological status of rivers within Europe to be compared between nation states on a similar basis. Substantial effort was expended by regulators and academics in developing the national networks. For example, national sampling methodologies were assessed and intercalibrated to provide harmonised information on ecological condition across Europe (Birk et al., 2013, 2012; Friberg et al., 2006; Furse et al., 2006). The system was developed to the point where multimetric indices, created by combining data on a number of biological groups, could be used to indicate ecological status (Hering et al., 2006; Johnson et al., 2006; Kennard et al., 2006). Sources of uncertainty were well quantified, such as inter-sampler error (Clarke et al., 2006) but it was not possible to integrate statistical power analysis into the design and no methods were in common use that could optimise representativeness across multiple environmental gradients.

As WFD surveillance networks have been operational across Europe for approximately ten years, it is timely to review the performance of current networks (Levine et al., 2014). The major habitat gradients controlling ecological communities in rivers are well known, e.g. River InVertebrate Prediction And Classification System (RIVPACS) predictors (Wright et al., 2000), as are the major pressure gradients that determine the anthropogenic impact on freshwater systems. Data on these gradients are often available across an entire country, potentially allowing an up to date assessment of network representativeness and the identification of sites to remove from or add to the network in order to create a truly representative monitoring network. This type of analysis is especially important in countries where the landscape is heterogeneous, and the habitats and anthropogenic influences on them vary spatially, such as Scotland (Carey et al., 1995; O'Hare et al., 2012).

Here we use this well described monitoring network to address two questions:

1. How representative of natural environmental gradients and pressure gradients is the existing river monitoring network and how can its representativeness be improved?

2. How powerful is the current network at detecting trends and how would this be affected by modifications to network structure designed to improve representativeness?

2. Materials and methods

2.1 Analytical overview

We performed a series of analyses that together form an approach to assess and improve the representativeness and statistical power of ecological monitoring networks (Figure 1). We first assessed the representativeness of an existing monitoring network by comparing its coverage of important habitat and pressure gradients to the national distributions of those gradients. We then created an algorithm that modified the network size to maximise its representativeness across all gradients. Using this algorithm we revised the current network to a range of different sizes, to investigate potential options for network re-design.

Following this, we conducted a power analysis on models for trends in the monitoring data over a recent ten-year period. These estimated the strength of the recent trends and characterised the noise obscuring the trend arising through seasonality, variation among water bodies and types of water body, variation among years and other unexplained (residual) sample-level variance. Based on these models, power analysis simulation techniques (Johnson et al., 2015) were used to estimate the minimum detectable trends in the recent monitoring data from the river network, and also to estimate the effect of revised networks and sampling regimes on power to detect trends.

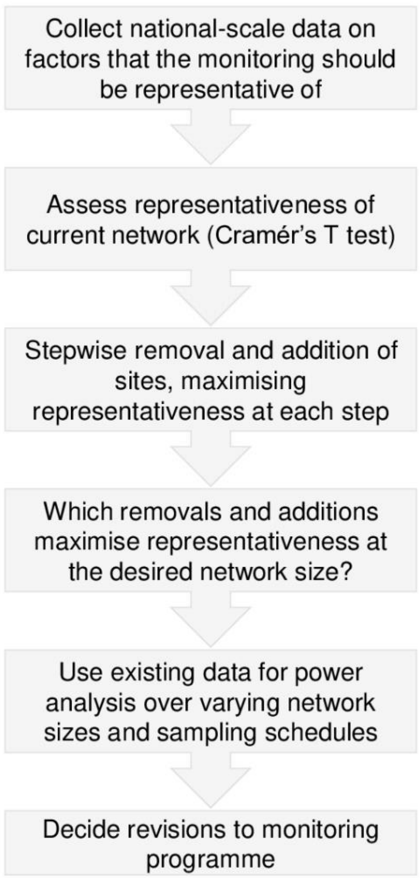

Figure 1. Overview of the proposed scheme to assess and update the size and sampling schedule of an ecological monitoring network.

2.2 Data

The study focused on Scotland, UK, where suitable data were readily available (SEPA, 2007). The representativeness analysis used 'river water bodies' as the unit of analyses, defined by the Scottish Environment Protection Agency (SEPA) as sub-catchment polygons containing connected sections of the river network and excluding lakes (Figure 2). SEPA has defined 2273 river water bodies nationally, of which 254 form the current surveillance monitoring network in which SEPA regularly monitor diatoms, benthic invertebrates and macrophytes (Figure 2).

Data representing major anthropogenic pressures and habitat factors influencing ecosystem sensitivity was available for nearly all WBs (Table 1). Other factors such as climate, which can be important for determining water chemistry (Le et al., 2019), likely co-varied with these gradients, e.g. land use, elevation, easting and northing strongly correlate to climate in Scotland.

Figure 2. Map of river water body (WB) polygons in Scotland, capturing the unique (inter) catchment of a section of main stem rivers. Shading highlights WBs within the river surveillance network.

Pressure gradients	**Habitat gradients**
Phosphate concentration from diffuse sources (mg l^{-1})	Sub-catchment mean elevation (m)
Phosphate concentration from point sources (mg l^{-1})	Sub-catchment area (km^2)
Nitrate concentration from diffuse sources (mg l^{-1})	Sub-catchment peat coverage (%)
Nitrate concentration from point sources (mg l^{-1})	Sub-catchment siliceous bedrock coverage (%)
Phosphate load from diffuse sources (kg day^{-1})	Sub-catchment calcareous bedrock coverage (%)
Phosphate load from point sources (kg day^{-1})	Mean channel slope (%)
Nitrate load from diffuse sources (kg day^{-1})	Natural Q_{mean} flow rate (Ml day^{-1})
Nitrate load from point sources (kg day^{-1})	River sinuosity index
Morphology pressure to channel (%)	Easting (m)
Morphology pressure to bank and riparian zone (%)	Northing (m)
Low and medium flow modification pressure	
High flow modification pressure	

Table 1. Pressure and habitat gradients for which the representativeness of the river surveillance network was assessed. Nutrient concentrations and loads were estimated by SEPA using the Source Apportionment-GIS (SAGIS) modelling framework (Comber et al., 2013). Morphology pressures were assessed on the ground by SEPA as the percentage of the bank or channel under pressure. Flow pressures were scored from one to five based on estimated reductions in natural flow previously estimated by SEPA hydrologists using Low Flow Enterprise modelling (LFE). Habitat gradients were derived from SEPA GIS databases.

Ecological monitoring data from the above surveillance network was obtained for the ten-year period, 2007-2016. The data comprised Ecological Quality Ratios (EQRs), calculated by

SEPA for individual samples. EQRs are a prescribed methodology under the EU Water Framework Directive (WFD) (European Commission, 2000) and are indicators of the degree to which an observed assemblage represents the assemblage that would be expected in unstressed conditions, given the particular type of water body present (Van de Bund and Solimini, 2007; Wright et al., 2000). The SEPA surveillance network monitors EQRs for WFD compliance, and as such they are the appropriate indicator to analyse in this study. However, in other monitoring networks the same approaches could be applied to other ecological metrics (e.g. diversity indices).

We analysed EQRs for communities of diatoms (River Trophic Diatom Index, TDI4) (Kelly and Whitton, 1995), benthic invertebrates (Average Score Per Taxon, ASPT abundance) (Walley and Hawkes, 1997), and macrophytes (River Macrophyte Nutrient Index, RMNI) (Willby et al., 2009). WBs in the surveillance network were monitored for these three communities, though slightly different numbers of WBs were monitored for each community. Diatoms were typically sampled every two or three years, with two samples collected per sampling year. For benthic invertebrates, the typical sampling schedule was to sample every other year with two samples collected per sampling year. Macrophyte sampling generally occured once every six years with only one survey per sampling year. Note that these schedules were asynchronous between sites, i.e. some sites were sampled in every year. The total numbers of samples available for analysis were 3662 for diatoms, 3202 for benthic invertebrates, and 488 for macrophytes.

2.3 Representativeness of the existing network

The representativeness of the river surveillance network for each gradient in Table 1 was assessed by comparing the gradient distributions among water bodies within the network with gradient distributions among water bodies outside the network. For individual gradients, this was tested using two-sample Kolmogorov-Smirnov (KS) tests. The KS test is a non-parametric test for the equality of distributions among two samples, based on the maximum absolute difference between the empirical cumulative density functions of both samples. *P* values were estimated by a permutation test with 10^6 random permutations (Good, 2013) that accounts for ties in the data and the discrete nature of two of the gradients (both flow pressure scores).

In addition, representativeness across all gradients was assessed in a similar way using the non-parametric multivariate two-sample Cramér test (Baringhaus and Franz, 2004). The test statistic T is based on the sum of all Euclidean distances between all data points in the two samples, minus half of the corresponding sums of distances within each sample. As such, T is sensitive to differences in the locations, variances and covariances of two multivariate datasets, and in the context of our analysis larger values of T indicate a less representative network. To standardise the influence of each variable on T, we applied a rank-transformation on each gradient so that they conformed to Gaussian distributions with means of zero and standard deviations of one. As above, we assessed the statistical significance of T using 10^6 permutations.

2.4 Improving network representativeness

An algorithm for prioritising the removal or addition of water bodies to maximise network representativeness was developed in R (R Core Team, 2019). Network representativeness was assessed with the Cramér's T statistic, comparing water bodies inside and outside of the network. Specifically, in a removal step, all possible removals of single water bodies were tried and the one resulting in the lowest value of T was chosen. Likewise, in an addition step, all possible single water body additions were tried and the one causing the lowest T value was selected. The orders of water body removal and addition provide prioritisation rankings for restructuring the monitoring network.

SEPA are planning to reduce the size of the surveillance network due to budget constraints. Therefore, using the stepwise algorithm the existing network was first iteratively reduced in size from its current 254 water bodies to 10 WBs. Then a stepwise water body addition was simulated starting from the existing network and from networks of sites reduced in size to 50, 100, 150 and 200 water bodies. This resulted in a range of networks of up to 300 water bodies in size. The representativeness of each was compared based on the resulting values of Cramér's *T*.

2.5 Power analysis for long-term ecological trends

Power analysis simulation methods (Johnson et al., 2015) were used to test the effect of network structure, measurement errors and network sampling strategy on the ability of the surveillance programme to detect long-term ecological trends. As the basis for power analysis, linear mixed effects (LME) models for long-term trends across the whole network were fitted to ecological indicators (EQRs) monitored from 2007-2016. LMEs provide a suitable analytical framework because of their ability to accommodate multiple levels of variation as random effects as well as trends of interest as fixed effects (Bolker et al., 2009). Separate LME models were fitted to the monitored EQRs for diatoms, benthic invertebrates and macrophytes using the lme4 R package (Bates et al., 2015). Model fitting used restricted maximum likelihood (REML) and fixed effect statistical significance was estimated using Satterthwaite's approximation of the numbers of degrees of freedom, as implemented in the lmerTest R package (Kuznetsova et al., 2017). Prior to model fitting, the invertebrate EQR was $\log_{10}$ transformed, as it has a lower bound >0. Diatom and macrophyte EQRs also had a lower bound >0 but were only available to us as 'capped' values with an upper bound of 1 imposed, so an empirical logit transformation was applied (Warton and Hui, 2011).

In the LMEs, a fixed effect of year (values centred on their midpoint) was included to model the long-term trend in the EQRs. To improve interpretability, the fitted trend coefficients were converted into the proportion change over a 10-year period. To model seasonality, linear fixed effects were included for the first two harmonics of the Fourier series for day of year (centred on zero and scaled to the same variance as the year variable). Seasonal terms were not included in models for macrophytes since these were sampled only once per year and sampling dates were not available. As random effects, we included random intercepts for year, to model annual divergence from the trend, and for WFD river typology and water body nested within typology, to model spatial variability. The LMEs for diatoms and benthic invertebrates also included random trends at typology and water body level, to model spatial variability in the trend. It was not possible to include random trends for macrophytes, as there was insufficient data. In lme4 format the full model formula was: EQR ~ year + h_1 + h_2 + h_3 + h_4 + (year_f | typology / water body) + (1 | year_f), where h_i is the *i*th harmonic of the day of year and year_f is year treated as a discrete factor.

To perform power analysis, equivalently-specified LMEs were fitted to simulated data generated from the original LME (Johnson et al., 2015). Data simulation involved randomly generating new EQR values specifying the network structure (water body identities and typologies), sampling rate (which years and days are samples taken), overall trend, seasonality, random effect variances and residual errors. For a simple assessment of power, LMEs were fitted to 1000 simulated response variables and the power calculated as the proportion giving a statistically significant trend ($P < 0.05$).

First, we evaluated the effect of trend size on power for the current network. Data were simulated from the LMEs with a range of trend values and for the water bodies in the current network, the exact dates they had been sampled, and the estimated random effect and residual variance. By varying the trend values, we established a ‘power curve’ showing how power varies as a function of trend size (Johnson et al., 2015; Thomas, 1997).

Second, a power experiment was used to investigate the effect of improved sampling accuracy on the power curve of the current network by repeating the above analysis with LME residual errors reduced to 75% of their current magnitude.

Third, a power experiment was used to investigate the effect of altered network size, representativeness and sampling rate on detection of trends of current magnitude. Power simulations were performed for simulated monitoring programmes across all combinations of: (1) network size of 50, 100, 150, 200 or the current number of water bodies monitored for each EQR (~254); (2) the network is a random sample of sites in the current network, or is a more representative network produced by our stepwise algorithm described above (the networks were produced by applying stepwise site removal and then stepwise site addition, with the number of removal steps selected as the fewest leading to a representative network with $P > 0.05$); (3) water body sampling rate is once per year every year, twice per year every two years or three times per year every third year. For power analysis of each simulated monitoring program, EQRs were simulated using their current trend coefficient, the estimated random effect variances and residual errors and with sampling seasonality following the observed distribution of days of year.

3. Results

3.1 Representativeness of the existing network

The existing river surveillance network does not provide a representative sample of the pressure and habitat gradients found across Scotland according to the two-sample KS tests on individual gradients ($P < 0.04$ for all gradients) and the multivariate two-sample Cramér test on all gradients ($T = 219.8$, $P < 0.001$). Among the pressures, the network was least representative of nutrient loads, with a major bias towards high loads (Figure 3). The network was also very strongly biased towards water bodies with large catchments and high natural flow rates. There were less strong, but still clear, biases towards higher nutrient concentrations from point sources, higher morphological pressures, shallower slopes, higher sinuosity, more peat, less siliceous bedrock and more calcareous bedrock. Lower biases for higher nutrient concentrations from diffuse sources and higher flow modification pressures were evident, while catchment elevation was relatively well represented by the river surveillance network.

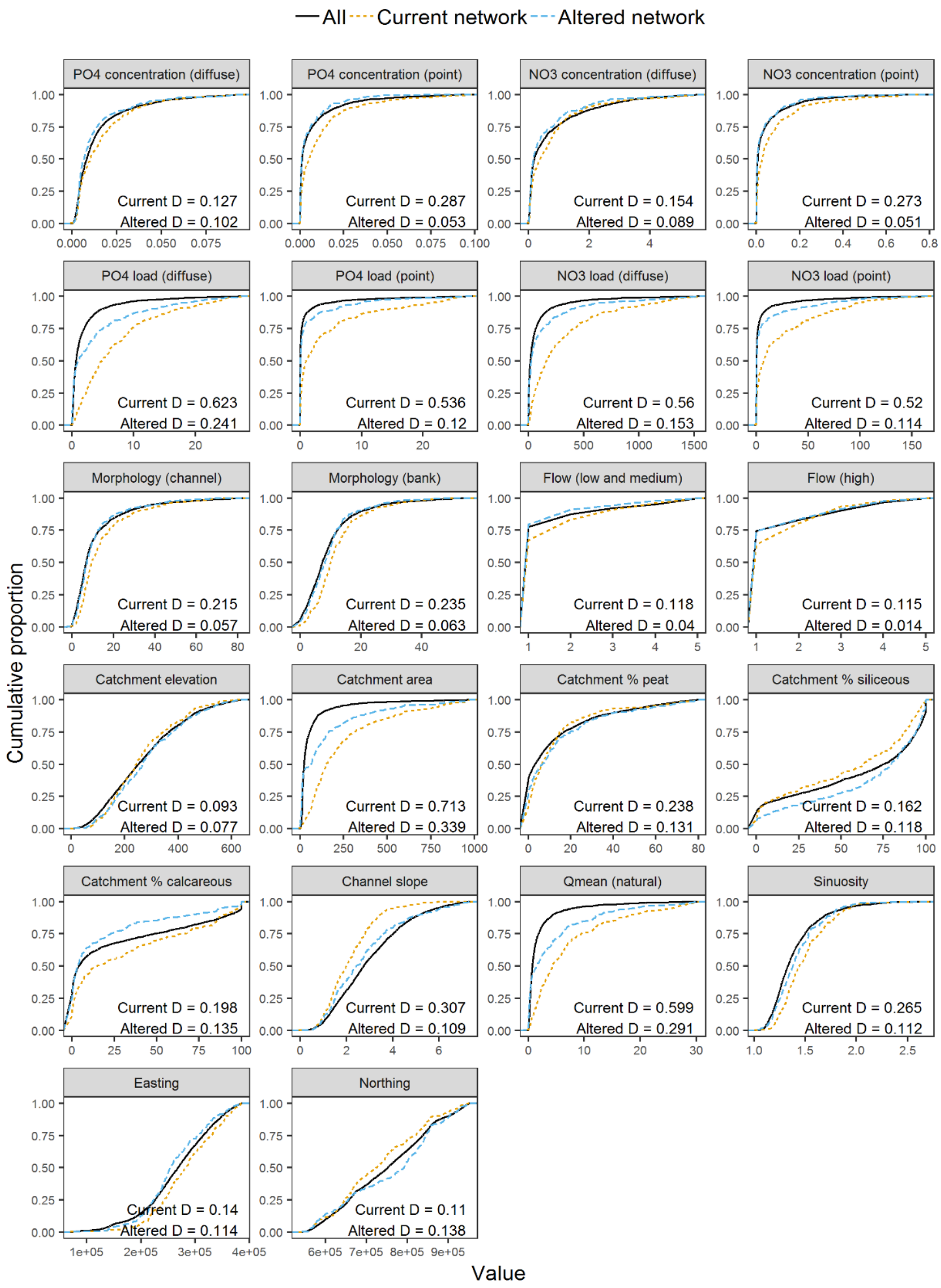

Figure 3. Cumulative gradient distributions (Table 1) in all Scottish river water bodies (WBs), the current surveillance network and a more representative network. The latter was generated by reducing the current network to 100 WBs and then adding 154 WBs, so that it was the same final size as the current network. For most gradients, the altered network was more representative of the overall Scottish distribution than was the current network, indicated by lower D values (Kolmogorov Smirnov test statistics). To aid visualisation, upper extreme values beyond the 97.5th percentile are omitted.

3.2 Improving network representativeness

Selective water body removal progressively improved representativeness but did not result in a statistically representative network (Cramér's T with $P > 0.05$) until the network was reduced to 58 or fewer water bodies (Figure 4). As such, to achieve a large and statistically representative network, it was necessary to combine water body removal with stepwise water body addition. For example, stepwise reduction in the size of the current network to 200 water bodies gave a highly unrepresentative network, while producing a 200 water body network, by first reducing to 100 water bodies and then selectively adding 100 new water bodies, resulted in a statistically representative network (Figure 4).

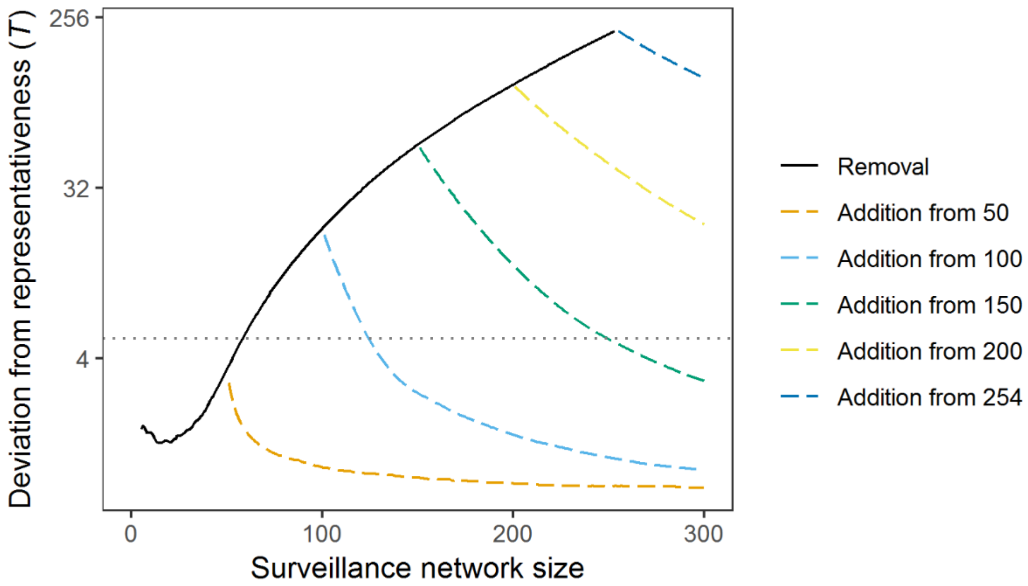

Figure 4. Effect of modifications to the river surveillance network to increase its representativeness by minimising the Cramér's T statistic. Lines show the results of stepwise water body removal and stepwise addition from different starting points. The horizontal dotted line is at the critical value of T, below which the network cannot be distinguished statistically from a random sample of Scotland's water bodies.

3.3 Power analysis for long-term ecological trends

Linear mixed effects (LME) models fitted to ecological indicators from 2007 to 2016 detected significant increasing trends in the EQRs for diatoms (7.0% increase, $P = 0.022$) and benthic invertebrates (2.2% increase, $P = 0.010$), while there was a marginally non-significant increasing trend in the macrophyte EQR (5.0% increase, $P = 0.074$).

The power analysis for trend detection, based on data simulated from the LMEs with different trend values showed there was a greater power to detect stronger trends, as was expected (Figure 5). It also demonstrated that there was relatively low power to detect trends of the observed magnitude for diatoms and macrophytes (Figure 5). For both of these groups the observed power was below 80%, often considered a reasonable target for effect detection (Di Stefano, 2003). The only group for which the network apparently provided adequate power to detect the observed level of change was benthic invertebrates, for which we estimated an 85% power to detect the current trend (Figure 5).

To simulate improved accuracy and consistency of sampling, the power analysis described above was repeated with residual errors reduced to 75% of their current level. This increased power for any trend magnitude (Figure 5). The network now gave more than adequate power for diatoms as well as benthic invertebrates, while macrophytes fell just short of the 80% power target.

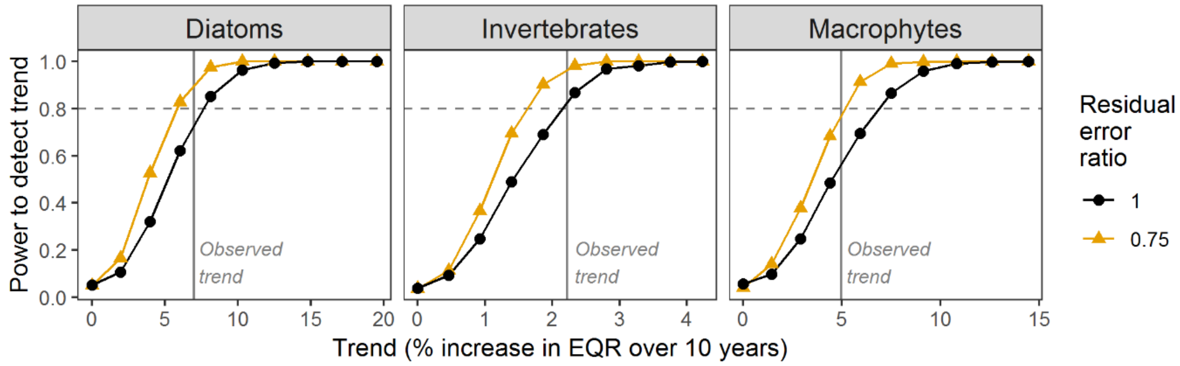

Figure 5. Power to detect trends in three ecological indicators over 10 years, estimated from data simulated with varying trend sizes and the current monitoring network and sampling regime. Simulations either used the observed residual error standard deviation (residual error ratio = 1) or reduced this to 75% of its observed value (residual error ratio = 0.75), simulating an increase in sampling accuracy. Vertical solid lines show the observed trends and the dashed horizontal line is at 80% power, often considered a reasonable target (Di Stefano, 2003). Power to detect the observed trends in diatoms and macrophytes was <80%, suggesting the network is under-powered.

Power analyses using modified surveillance networks and the current ecological trends monitored over ten years revealed a clear loss of power in smaller networks (Figure 6). Interestingly, representative networks appeared to be slightly less powerful than the current network, especially at small network sizes. Sampling strategy had a smaller influence on power, although annual sampling every year usually gave marginally higher power than the other strategies (Figure 6).

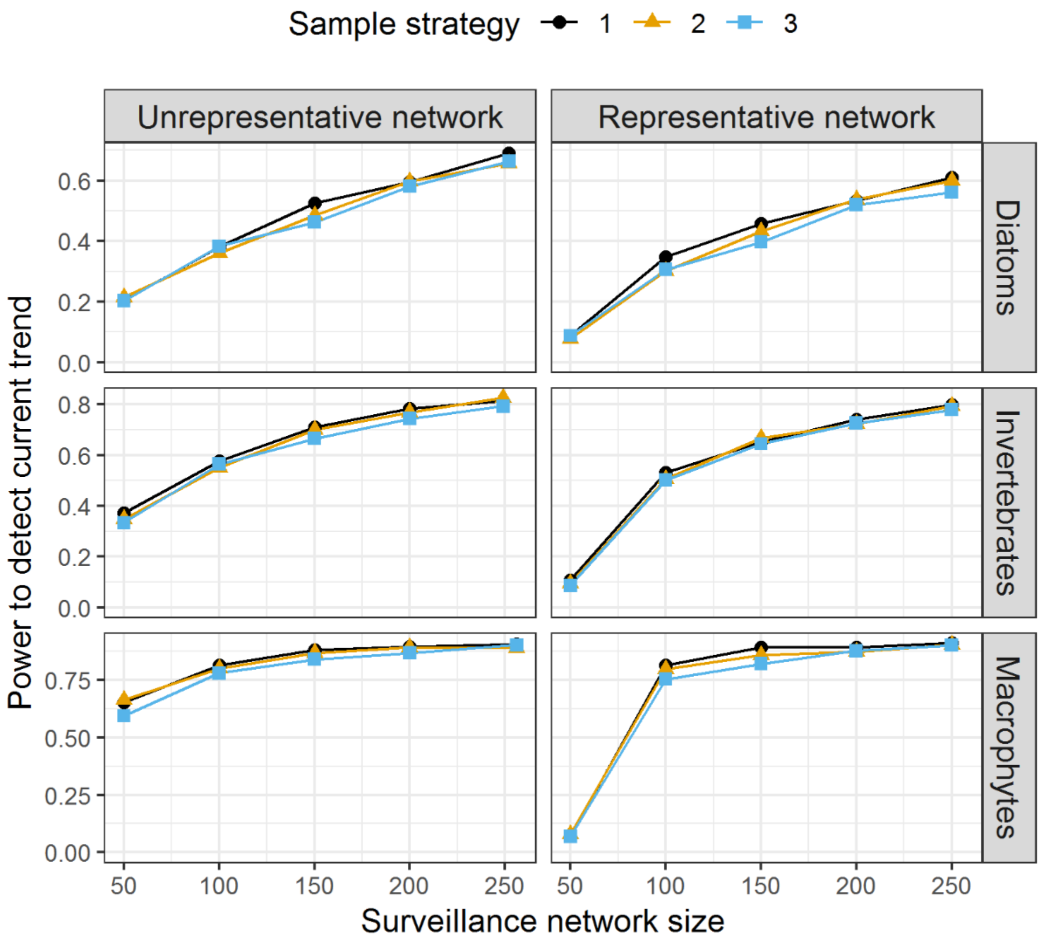

Figure 6. Effect of modified surveillance network structure and sampling strategy on power to detect current ecological trends in diatom, invertebrate and macrophyte EQRs. Power was evaluated for different network sizes generated either as a random sample of the current unrepresentative network or to improve representativeness and assuming ten years of monitoring following three equal effort sampling strategies (1 = sample once every year, 2 = sample two times every second year, 3 = sample three times every third year).

4. Discussion

This study provides a framework for improving the quality of indicators derived from ecological monitoring networks. The framework involves a novel application of Cramér's test for the equality of two multivariate distributions (Baringhaus and Franz, 2004) to assess network representativeness and a novel stepwise algorithm for prioritising site removal or addition to improve representativeness. In addition, power analysis simulation methods were implemented to investigate the consequences of network redesign for the performance of the monitoring network. Together, our approach can be used to find ways to restructure monitoring networks to improve representativeness and optimise sampling strategies for detecting ecological trends. As such, it makes a novel contribution to the literature on design and performance of ecological monitoring networks (Carvalho et al., 2016; Eyre et al., 2011; Levine et al., 2014; Weatherhead et al., 2017; Wikle and Royle, 1999).

4.1 Network representativeness

Using univariate and multivariate tests, we showed that the river surveillance network exhibits statistically significant deviations from representativeness of national gradients in a large number of pressure and habitat gradients. Most significantly, the current network over-represented sites with greater discharge and those more heavily polluted by nutrients. By contrast, smaller channels, higher up the river networks were under-represented in the network. These under-represented sites are generally subject to different combinations of pressures than the larger, more lowland rivers. For example, they often sit in commercial forestry, peatlands, semi-natural grasslands or unimproved grazing land, have much lower nutrient fluxes and their channel morphology is only occasionally engineered (Maitland et al., 1994). However, these rivers may be impacted by other stressors such as impoundments, intensive grazing, riparian vegetation management and upland drainage, which have altered many of these systems from their natural state. Since the network does not provide a representative sample of national river types or pressure patterns, it is likely to provide a biased evidence base for the overall status and trends in Scotland's rivers.

These findings reflect SEPA's original design of the surveillance network in 2007 to over-represent anthropogenically-impacted lowland rivers, including those that were monitored historically prior to 2007. As many other countries in Europe also wanted to maintain existing long-term monitoring sites, they also built their surveillance networks around pre-existing networks. Therefore, it is possible that they too may have similar sampling biases. However, we suggest this should be done with careful consideration of how well they represent the specific habitat and pressure gradients found in that country. These considerations would help to tailor monitoring networks to the specific conditions found in each individual country and help to avoid similar sampling biases. Although this may lead to differences in network structure between countries, these differences will be quantifiable in a transparent, measurable fashion. The Cramér's test and our algorithms for changing network structure provides the kind of general framework for harmonised application in different countries.

We have also developed a novel algorithm for improving network representativeness by selectively removing or adding new monitoring locations to the network, in a stepwise fashion to minimise the Cramér's T statistic. This algorithm provides a generic tool for re-designing existing monitoring networks that could allow harmonised application for monitoring networks in different countries or ecosystem types. For the river surveillance network studied here, we found that because we started from a highly unrepresentative network it was necessary to combine water body addition with removal in order to make substantive improvements to representativeness. For example, stepwise removal of approximately 40% of water bodies, followed by stepwise addition of the same number results in a new network whose profiles of environmental and pressure gradients are statistically indistinguishable from those across the whole of Scotland (see Figures 3 and 4). Importantly, the new representative network retains 60% of the currently monitored sites. As such, the stepwise algorithm developed here provides a solution for improving monitoring network design while also preserving a large proportion of the legacy of long-term monitoring. This is beneficial both for analysis of trends across the whole network and for analysis of site-specific trends.

This highlights a more general point that re-design of monitoring networks likely requires balancing the trade-off between improving representativeness by replacing unrepresentative sites and the loss of historical long-term monitoring data at those sites. Network managers must decide on how much weight is given to both of those criteria in order to determine the best option for updating the network. Indeed, it may be possible to extend the current stepwise algorithm to factor in multiple criteria with user-defined weightings, in order to automate the process.

4.2 Power analysis

When considering changes to existing monitoring networks or sampling regimes, power analysis informed by a base of existing monitoring data can be used to evaluate how these changes may influence the ability of the monitoring programme to detect change (Levine et al., 2014; Stegman et al., 2017; Toft and Shea, 1983). Here, retrospective power analysis (Thomas, 1997) indicated that the existing network was under-powered for detecting trends of the observed magnitude in diatoms and macrophytes over a ten-year period. However, the network was adequately powered for detecting trends in benthic invertebrates. The power analysis also suggested that improved sampling methodologies that yield more consistent and less noisy data would lead to major improvements in the quality of ecological monitoring. Indeed, the adequate power for benthic invertebrates may reflect substantial past efforts into minimising sampling noise by testing different field protocols and auditing standards (Clarke, 2013; Clarke et al., 2006, 2002; Clarke and Hering, 2006; Wright et al., 2000).

For freshwater macrophytes, low power to detect trends may reflect a combination of low sampling intensity, high sampling variance and the effects of unrecorded human impacts on macrophyte assemblages, such as those from routine maintenance of channels. There have been some attempts to standardise and test macrophyte sampling the but the effort has not been sustained (Staniszewski et al., 2006). In countries such as Denmark, macrophytes are recorded in a more standardised way, routine maintenance is known and macrophyte data has proven a reliable and diagnostic measure of river quality (Baattrup-Pedersen et al., 2016; Baattrup-Pedersen et al., 2015). For diatoms, inadequate power may have arisen because their assemblages are strongly influenced by short-term events, such as minor floods, that will have contributed to large variability in trends. A practical solution, implemented by SEPA, is to screen data and remove measurements that are likely to have been unduly influenced by short-term events. Additionally, new automated or rapid diatom monitoring methods are in development that would provide high temporal resolution data that could produce more statistical power (Kelly et al., 2016).

The power analysis also demonstrated that reductions in network size result in substantial losses of trend detection power for all three ecological indicators and that it was marginally preferable to sample once per year every year rather than sample multiple times per year but in fewer years. This likely reflects the lack of independence of samples taken within years, even after accounting for seasonality (Rhodes and Jonzén, 2011) and provides useful guidance for deciding how to sample the monitoring network. The more surprising result from the power analysis was that the more representative networks had slightly lower power than the current unrepresentative network. The likely explanation is that the representative networks contained a greater range of water body types in closer proportion to their national frequency, but with less replication of the rarer types. As a result, between-type variability may have obscured the overall trend in the ecological indicator. Nevertheless, moving towards more representative monitoring networks is still desirable as reducing bias is at least as important as signal detection power for the quality of evidence from monitoring.

Although useful, power analysis is always approximate and subject to a number of caveats (Hoenig and Heisey, 2001; Johnson et al., 2015). For example, one caveat comes from the assumption that the trends and structure of noise in future monitoring data will follow patterns from the last ten years. This may not be true because emerging technologies for monitoring may improve accuracy (i.e. reduce sample-level residual variation), there may be better standardisation of sampling and laboratory methods, or factors such as climate change may alter patterns of variability among seasons, years, sites or site types. Nevertheless, the power analysis approaches developed and applied here should be considered an important element in the design of environmental monitoring programmes.

4.3 Conclusions

This study provides a framework for informing the re-design of monitoring networks and revision of sampling strategies, combining assessment and improvement of network representativeness and power analysis to evaluate trend detection power of alternative networks and sampling strategies. We suggest that this approach will be useful for the periodic appraisal and updating of multi-site ecological monitoring networks, helping to ensure they remain fit for purpose and cost effective over the long term. Indeed, the relevant monitoring authority, SEPA, intends to review their river surveillance network following this study. In addition, the stepwise algorithm to add sites in a representative way could be applied to design new monitoring networks, including in developing countries with fewer historical monitoring networks and stronger budget constraints. A key advantage of our framework is that it adapts rather than replaces existing networks, maximising retention of historical monitoring data while improving network structure. It can also inform decisions over the size of the network, intensity of sampling, balance between monitoring of different indicators, and where to make investments to improve data quality. Overall therefore, moving towards more representative networks that are optimised for representativeness and statistical power will allow monitoring agencies to better understand the challenges facing the environment, and ensure that they can more effectively provide evidence that drives improvements.

Acknowledgements

Support for this study was provided by the Scottish Environment Protection Agency (grant CR/2016/07). We thank colleagues at SEPA and CEH who assisted data processing and commented constructively on the study. We also thank three anonymous reviewers for feedback on previous drafts. R scripts for implementing the Cramér's test of representativeness and stepwise site removal or addition can be found in the Supplementary Material.

References

Baattrup-Pedersen, A., Göthe, E., Riis, T., O'Hare, M.T., 2016. Functional trait composition of aquatic plants can serve to disentangle multiple interacting stressors in lowland streams. Sci. Total Environ. 543, 230–238.

Baattrup- Pedersen, A., Göthe, E., Larsen, S.E., O'Hare, M., Birk, S., Riis, T., Friberg, N., 2015. Plant trait characteristics vary with size and eutrophication in European lowland streams. J. Appl. Ecol. 52, 1617–1628.

Baringhaus, L., Franz, C., 2004. On a new multivariate two-sample test. J. Multivar. Anal. 88, 190–206.

Bates, D., Mächler, M., Bolker, B., Walker, S., 2015. Fitting linear mixed-effects models using lme4. J. Stat. Softw. 67, 1–48.

Birk, S., Bonne, W., Borja, A., Brucet, S., Courrat, A., Poikane, S., Solimini, A., Van De Bund, W., Zampoukas, N., Hering, D., 2012. Three hundred ways to assess Europe's surface waters: an almost complete overview of biological methods to implement the Water Framework Directive. Ecol. Indic. 18, 31–41.

Birk, S., Willby, N.J., Kelly, M.G., Bonne, W., Borja, A., Poikane, S., Van de Bund, W., 2013. Intercalibrating classifications of ecological status: Europe's quest for common management objectives for aquatic ecosystems. Sci. Total Environ. 454, 490–499.

Bolker, B.M., Brooks, M.E., Clark, C.J., Geange, S.W., Poulsen, J.R., Stevens, M.H.H., White, J.-S.S., 2009. Generalized linear mixed models: a practical guide for ecology and evolution. Trends Ecol. Evol. 24, 127–135.

Carey, P.D., Preston, C.D., Hill, M.O., Usher, M.B., Wright, S.M., 1995. An environmentally defined biogeographical zonation of Scotland designed to reflect species distributions. J. Ecol. 83, 833–845. https://doi.org/10.2307/2261420

Carvalho, S.B., Gonçalves, J., Guisan, A., Honrado, J.P., 2016. Systematic site selection for multispecies monitoring networks. J. Appl. Ecol. 53, 1305–1316.

Clarke, R.T., 2013. Estimating confidence of European WFD ecological status class and WISER Bioassessment Uncertainty Guidance Software (WISERBUGS). Hydrobiologia 704, 39–56.

Clarke, R.T., Furse, M.T., Gunn, R.J.M., Winder, J.M., Wright, J.F., 2002. Sampling variation in macroinvertebrate data and implications for river quality indices. Freshw. Biol. 47, 1735–1751.

Clarke, R.T., Hering, D., 2006. Errors and uncertainty in bioassessment methods—major results and conclusions from the STAR project and their application using STARBUGS, in: Furse, M.T., Hering, D., Brabec, K., Buffagni, A., Sandin, L., Verdonschot, P.F.M. (Eds.), The Ecological Status of European Rivers: Evaluation and Intercalibration of Assessment Methods. Springer, pp. 433–439.

Clarke, R.T., Lorenz, A., Sandin, L., Schmidt-Kloiber, A., Strackbein, J., Kneebone, N.T., Haase, P., 2006. Effects of sampling and sub-sampling variation using the STAR-AQEM sampling protocol on the precision of macroinvertebrate metrics, in: The Ecological Status of European Rivers: Evaluation and Intercalibration of Assessment Methods. Springer, pp. 441–459.

Cohen, J., 2013. Statistical power analysis for the behavioral sciences. Routledge.

Comber, S.D.W., Smith, R., Daldorph, P., Gardner, M.J., Constantino, C., Ellor, B., 2013. Development of a chemical source apportionment decision support framework for catchment management. Environ. Sci. Technol. 47, 9824–9832.

Di Stefano, J., 2003. How much power is enough? Against the development of an arbitrary convention for statistical power calculations. Funct. Ecol. 17, 707–709.

European Commission, 2000. Directive 2000/60/EC of the European Parliament and of the Council of 23 October 2000 establishing a framework for Community action in the field of water policy. Off. J. Eur. Union L 327 43, 1–72.

Eyre, T.J., Fisher, A., Hunt, L.P., Kutt, A.S., 2011. Measure it to better manage it: a biodiversity monitoring framework for the Australian rangelands. Rangel. J. 33, 239–253.

Friberg, N., Sandin, L., Furse, M.T., Larsen, S.E., Clarke, R.T., Haase, P., 2006. Comparison of macroinvertebrate sampling methods in Europe, in: The Ecological Status of European Rivers: Evaluation and Intercalibration of Assessment Methods. Springer, pp. 365–378.

Furse, M.T., Hering, D., Brabec, K., Buffagni, A., Sandin, L., Verdonschot, P.F.M., 2006. The ecological status of European rivers: evaluation and intercalibration of assessment methods BT - The Ecological Status of European Rivers: Evaluation and Intercalibration of Assessment Methods, in: Furse, M.T., Hering, D., Brabec, K., Buffagni, A., Sandin, L., Verdonschot, P.F.M. (Eds.), . Springer Netherlands, Dordrecht, pp. 1–2. https://doi.org/10.1007/978-1-4020-5493-8_1

Good, P., 2013. Permutation tests: a practical guide to resampling methods for testing hypotheses. Springer Science & Business Media.

Hering, D., Feld, C.K., Moog, O., Ofenböck, T., 2006. Cook book for the development of a Multimetric Index for biological condition of aquatic ecosystems: experiences from the European AQEM and STAR projects and related initiatives, in: The Ecological Status of European Rivers: Evaluation and Intercalibration of Assessment Methods. Springer, pp. 311–324.

Hoenig, J.M., Heisey, D.M., 2001. The abuse of power: the pervasive fallacy of power calculations for data analysis. Am. Stat. 55, 19–24.

Irvine, K.M., Manlove, K.R., Hollimon, C., 2012. Power analysis and trend detection for water quality monitoring data: An application for the Greater Yellowstone Inventory and

Monitoring Network.

Johnson, P.C.D., Barry, S.J.E., Ferguson, H.M., Müller, P., 2015. Power analysis for generalized linear mixed models in ecology and evolution. Methods Ecol. Evol. 6, 133–142.

Johnson, R.K., Hering, D., Furse, M.T., Clarke, R.T., 2006. Detection of ecological change using multiple organism groups: metrics and uncertainty. Hydrobiologia 566, 115–137.

Kelly, M.G., Krokowski, J., Harding, J.P.C., 2016. RAPPER: A new method for rapid assessment of macroalgae as a complement to diatom-based assessments of ecological status. Sci. Total Environ. 568, 536–545.

Kelly, M.G., Whitton, B.A., 1995. The Trophic Diatom Index: a new index for monitoring eutrophication in rivers. J. Appl. Phycol. 7, 433–444. https://doi.org/10.1007/BF00003802

Kennard, M.J., Harch, B.D., Pusey, B.J., Arthington, A.H., 2006. Accurately defining the reference condition for summary biotic metrics: a comparison of four approaches. Hydrobiologia 572, 151–170.

Kuznetsova, A., Brockhoff, P.B., Christensen, R.H.B., 2017. lmerTest package: tests in linear mixed effects models. J. Stat. Softw. 82.

Le, T.D.H., Kattwinkel, M., Schützenmeister, K., Olson, J.R., Hawkins, C.P., Schäfer, R.B., 2019. Predicting current and future background ion concentrations in German surface water under climate change. Philos. Trans. R. Soc. B 374, 20180004.

Levine, C.R., Yanai, R.D., Lampman, G.G., Burns, D.A., Driscoll, C.T., Lawrence, G.B., Lynch, J.A., Schoch, N., 2014. Evaluating the efficiency of environmental monitoring programs. Ecol. Indic. 39, 94–101.

Lindenmayer, D.B., Likens, G.E., 2010. The science and application of ecological monitoring. Biol. Conserv. 143, 1317–1328.

Lovett, G.M., Burns, D.A., Driscoll, C.T., Jenkins, J.C., Mitchell, M.J., Rustad, L., Shanley, J.B., Likens, G.E., Haeuber, R., 2007. Who needs environmental monitoring? Front. Ecol. Environ. 5, 253–260. https://doi.org/10.1890/1540-9295(2007)5[253:WNEM]2.0.CO;2

Maitland, P.S., Boon, P.J., McLusky, D.S., 1994. The fresh waters of scotland: A national resource of international significance, Aquatic Conservation: Marine and Freshwater Ecosystems. John Wiley & Sons, Ltd, Chichester.

Munkittrick, K.R., Arens, C.J., Lowell, R.B., Kaminski, G.P., 2009. A review of potential methods of determining critical effect size for designing environmental monitoring programs. Environ. Toxicol. Chem. 28, 1361–1371. https://doi.org/10.1897/08-376.1

O'Hare, M.T., Gunn, I.D.M., Chapman, D.S., Dudley, B.J., Purse, B. V., 2012. Impacts of space, local environment and habitat connectivity on macrophyte communities in conservation lakes. Divers. Distrib. 18, 603–614. https://doi.org/10.1111/j.1472-4642.2011.00860.x

Osenberg, C.W., Schmitt, R.J., Holbrook, S.J., Abu-Saba, K.E., Flegal, A.R., 1994. Detection of Environmental Impacts: Natural Variability, Effect Size, and Power Analysis. Ecol. Appl. 4, 16–30. https://doi.org/10.2307/1942111

Peterman, R.M., 1990. Statistical Power Analysis can Improve Fisheries Research and Management. Can. J. Fish. Aquat. Sci. 47, 2–15. https://doi.org/10.1139/f90-001

R Core Team, 2019. R: A language and environment for statistical computing.

Rhodes, J.R., Jonzén, N., 2011. Monitoring temporal trends in spatially structured populations: how should sampling effort be allocated between space and time? Ecography (Cop.). 34, 1040–1048. https://doi.org/10.1111/j.1600-0587.2011.06370.x

SEPA, 2007. Scotland's WFD aquatic monitoring strategy Scottish Environment Protection Agency. Stirling, UK.

Staniszewski, R., Szoszkiewicz, K., Zbierska, J., Lesny, J., Jusik, S., Clarke, R.T., 2006. Assessment of sources of uncertainty in macrophyte surveys and the consequences for river classification. Hydrobiologia 566, 235–246. https://doi.org/10.1007/s10750-006-0093-4

Stegman, L.S., Primack, R.B., Gallinat, A.S., Lloyd-Evans, T.L., Ellwood, E.R., 2017. Reduced sampling frequency can still detect changes in abundance and phenology of migratory landbirds. Biol. Conserv. 210, 107–115. https://doi.org/https://doi.org/10.1016/j.biocon.2017.04.004

Thomas, L., 1997. Retrospective Power Analysis. Conserv. Biol. 11, 276–280.

Toft, C.A., Shea, P.J., 1983. Detecting Community-Wide Patterns: Estimating Power Strengthens Statistical Inference. Am. Nat. 122, 618–625.

Urquhart, N.S., Kincaid, T.M., 1999. Designs for Detecting Trend from Repeated Surveys of Ecological Resources. J. Agric. Biol. Environ. Stat. 4, 404–414. https://doi.org/10.2307/1400498

Van de Bund, W., Solimini, A., 2007. Ecological Quality Ratios for ecological quality assessment in inland and marine waters. Inst. Environ. Sustain. (ed), Italy.

Vos, P., Meelis, E., Ter Keurs, W.J., 2000. A Framework for the Design of Ecological Monitoring Programs as a Tool for Environmental and Nature Management. Environ. Monit. Assess. 61, 317–344. https://doi.org/10.1023/A:1006139412372

Walley, W.J., Hawkes, H.A., 1997. A computer-based development of the Biological Monitoring Working Party score system incorporating abundance rating, site type and indicator value. Water Res. 31, 201–210. https://doi.org/https://doi.org/10.1016/S0043-1354(96)00249-7

Warton, D.I., Hui, F.K.C., 2011. The arcsine is asinine: the analysis of proportions in ecology. Ecology 92, 3–10. https://doi.org/10.1890/10-0340.1

Weatherhead, E.C., Bodeker, G.E., Fassò, A., Chang, K.-L., Lazo, J.K., Clack, C.T.M., Hurst, D.F., Hassler, B., English, J.M., Yorgun, S., 2017. Spatial Coverage of Monitoring Networks: A Climate Observing System Simulation Experiment. J. Appl. Meteorol. Climatol. 56, 3211–3228. https://doi.org/10.1175/JAMC-D-17-0040.1

Wikle, C.K., Royle, J.A., 1999. Space-Time Dynamic Design of Environmental Monitoring Networks. J. Agric. Biol. Environ. Stat. 4, 489–507. https://doi.org/10.2307/1400504

Willby, N., Pitt, J.A., Phillips, G., 2009. The ecological classification of UK rivers using aquatic macrophytes. UK Environ. Agency Sci. Reports. Proj. SC010080/SR1. Environ. Agency, Bristol.

Wright, J.F., Sutcliffe, D.W., Furse, M.T., 2000. Assessing the biological quality of fresh waters: RIVPACS and other techniques. Freshwater Biological Association, Ambleside, UK.