How many sites? Methods to assist design decisions when collecting multivariate data in ecology

Ben Maslen1 | Gordana Popovic1,2 | Michelle Lim1 | Ezequiel Marzinelli3,4,5 | David Warton1,2

Abstract
1. Sample size estimation through power analysis is a fundamental tool in planning an ecological study, yet there are currently no well-established procedures for when multivariate abundances are to be collected. A power analysis procedure would need to address three challenges: designing a parsimonious simulation model that captures key community data properties; measuring effect size in a realistic yet interpretable fashion; and ensuring computational feasibility when simulation is used both for power estimation and significance testing.

2. Here, we propose a power analysis procedure that addresses these three challenges by: using for simulation a Gaussian copula model with factor analytical structure, fitted to pilot data; assuming a common effect size across all taxa, but applied in different directions according to expert opinion (to “increaser”, “decreaser” or “no effect” taxa); using a critical value approach to estimate power, which reduces computation time by a factor of 500 (if we would otherwise use 999 resamples to estimate each $p$-value) with minor loss of accuracy.

3. The procedure is demonstrated on pilot data from fish assemblages in a restoration study, where it was found that the planned study design would only be capable of detecting relatively large effects (change in abundance by a factor of 1.7 or more).

4. The methods outlined in this paper are available in accompanying R software (the ecopower package), which allows researchers with pilot data to answer a wide range of design questions to assist them in planning their studies.

KEYWORDS
copula, multivariate, power analysis, restoration, sample size, simulation, software, statistics, study design
1 | INTRODUCTION

In planning any study, it is important to consider how large an effect ("effect size") can be detected by the intended study design (Fritz & MacKinnon, 2007; Kelley & Preacher, 2012; Rosenthal et al., 1994). Such information is often obtained via a power analysis of preliminary data, which can help design a study that has a good chance of detecting effects of ecological interest (Cohen, 1992, 2013). While texts for ecologists frequently discuss the importance of power analysis, and related techniques, in sample size determination (Gerrodette, 1987; Green & MacLeod, 2016; Johnson et al., 2015), there is little guidance for the ecologist on how to use power analysis to inform the design of studies collecting multivariate data, that is, multiple response variables measured per sample. This is particularly the case when we look beyond classical statistics (Green, 1989) at the situation where abundance or presence-absence data is simultaneously collected for many different taxa; hereafter multivariate abundances.

There are three key challenges to be addressed in developing a power analysis procedure that could be used to inform the design of studies collecting multivariate abundance data. Challenge 1 is to design a simulation model to randomly generate realistic multivariate abundance data, reflecting key properties of the data to be collected. It should be possible to use pilot data, when available, to tune the settings of the simulation model, to generate data that "looks like" the pilot data. Some simulation approaches are available in the literature (e.g., Xu et al., 2010), however with limited ability to tune using pilot data. Other methods have also been developed using a Bray-Curtis distance based approach (Irvine et al., 2011) or by considering the abundances as continuous random variables and calculating Euclidean distances (Angeler et al., 2009; Collins et al., 2000). These approaches however ignore important mean–variance relationships and therefore subsequent analyses can be misleading (Warton et al., 2012). Multivariate modelling approaches have been developed recently that could be used to address this, including Gaussian copulas (Anderson et al., 2019; Popovic et al., 2018) and hierarchical models (Ovaskainen et al., 2017; Warton, Foster, et al., 2015). Challenge 2 is measuring effect size. When many taxa are to be sampled, many parameters need to be specified a priori that will capture the size and nature of the effect the study has been designed to detect. Decisions about these parameters are to be made based on relatively little information, and need to be captured by a simple, interpretable effect size measure in order for results to be useful for study design (Kelley & Preacher, 2012). Challenge 3 is making the power analysis procedure computationally efficient. Hypothesis testing procedures for multivariate abundance data typically use resampling (Anderson, 2001; Wang et al., 2012), conventionally recalculating a test statistic at least 1000 times across resampled datasets. A power analysis would require this to be done for each of say 1000 simulated datasets, such that estimating power for a typical multivariate abundance dataset might take hours or days.

This paper proposes a power analysis procedure for multivariate abundance data that addresses each of the above three challenges. Challenge 1 is addressed using a Gaussian copula-factor analysis model to simulate data in a parsimonious fashion, an approach for which methods to tune the model using training data were only recently proposed (Popovic et al., 2018). Challenge 2 is addressed using a simple parameterisation for effect size that requires a priori information only concerning taxa that are likely to be affected, and the direction of the effect. Challenge 3 is addressed using a novel implementation of the critical value approach to hypothesis testing, which can reduce computation time for power simulation of resampling-based hypothesis tests from days to minutes. The procedure will be illustrated on a marine habitat restoration project which involves regular monitoring of ecological communities, where we are interested in the number of samples required for a future monitoring period in order to likely detect community differences in abundance across treatments, as well as the size of effects that are able to be detected under different sampling designs.

2 | MATERIALS AND METHODS

2.1 | Operation Crayweed Restoration Project

Researchers within the Operation Crayweed Restoration Project in Sydney are restoring the locally extinct macro-algae Phyllospora comosa ("crayweed": see Coleman et al., 2008; Campbell et al., 2014; Vergès et al., 2020) and are interested in the effect of this restoration on associated ecological communities (Marzinelli et al., 2014, 2016; Wood et al., 2019). Pilot data have already been collected, where the abundance of fish species in nine open ocean sites have been recorded. We are interested in observing if there is a change in mean fish abundance between control sites (those in Sydney without crayweed) and restored sites (similar sites where crayweed has recently been transplanted, as in Figure 1). There are plans to collect

**FIGURE 1** Images of characterised environments from control and restored sites. Control sites (a) are characterised as urchin barren habitats in Sydney shallow rocky reefs where crayweed has been lost, whilst restored sites (b) are locations which are similar to control sites where crayweed has recently been transplanted.
more data in the future, however there is an upper bound of approximately 24 possible spatially independent restored or control coastal bays/sites within the Sydney region and surroundings.

We are interested in answering the following experimental design questions from the pilot data, which will act as the motivating examples for this paper:

1. How many sites are required to likely detect 20% differences between treatments?
2. Under the maximum independent sampling design of 24 sites, what are the size of effects that are likely to be detected?

Precisely which species change in mean abundance across treatments, and in what direction, will be discussed in detail below when addressing Challenge 2.

2.2 Power analysis

Consider a situation where the objective of the study is to test a specific null hypothesis \( H_0 \) by looking for evidence of some alternative \( H_1 \). The effectiveness of any given study design and testing procedure can be evaluated using power, that is, the probability of rejecting the null hypothesis \( H_0 \) given that a particular alternative hypothesis \( H_1 \) is true. Power generally increases as sample size and effect size increases, and variability decreases (Cohen, 2013). Thus, given an understanding of the variability to be expected in the data, power can be used to estimate how large a sample size is needed to detect an effect of a given size, or what sizes of effects can be detected in study of a given sample size (Cohen, 1992, 2013).

To undertake a power analysis, it is necessary to work through the following steps:

1. Specify a model for data, which captures key properties of the data that are expected to be collected. For multivariate abundances, as in Operation Crayweed, this is a non-trivial task (Challenge 1).
2. Decide on a measure of effect size that is ecologically meaningful. When there are many taxa, as for fish assemblages sampled in Operation Crayweed, many different effect size parameters need to be considered (Challenge 2).
3. Decide on the testing procedure. Hypothesis tests of multivariate abundance data typically use resampling to ensure valid inference (Anderson, 2001; Wang et al., 2012); this paper will make use of a generalised estimating equations approach (Wang et al., 2012, using the mvabund package) that is increasingly common in ecology. For the crayweed fish abundance data, we assumed abundances have a negative binomial distribution, with diagnostic plots suggesting this adequately captured the mean–variance trend in the data (Figures S6 and S7 in Supporting Information).
4. Estimate power. In some simpler settings, this can be done analytically (Cohen, 2013), however for multivariate abundances, this needs to be done by simulation, generating data under the assumed model, then applying the testing procedure for each simulated dataset and recording the proportion of times the null hypothesis was rejected. When using a resampling-based testing procedure, this involves two levels of simulation and will be very computationally intensive (Challenge 3).

Solutions to the three challenges identified above are proposed in the following sections.

2.3 Challenge 1—Data generating model

A data generating model is needed that can capture key properties of multivariate abundance data that will be collected. Multivariate abundances are discrete, with many zeros, and highly dimensional, with a large number of responses relative to the sample size. Parametric statistical models that can be fitted to multivariate abundances are a relatively recent development (Popovic et al., 2018; Warton, Blanchet, et al., 2015). This is an important advance because it allows the simulation model to be tuned to pilot data in order to generate data statistically similar to what will actually be observed when the study is undertaken. In particular, power can be strongly affected by mean abundance, variability and correlation across taxa (Warton, 2011), all of which can vary considerably from one study to another. For a power analysis to be informative, these data properties need to be tuned to the study in question.

This paper adopts a Gaussian copula approach. To date, copulas have rarely been used in ecology (Anderson et al., 2019; Popovic et al., 2018, 2019). A specific advantage of a copula approach is that it specifies a marginal (unconditional) model, making parameters more interpretable. For example, if we set an effect size parameter such that there is a two-fold change between treatment and control groups, we can be sure that data will in fact be generated with a two-fold change in mean abundance. In hierarchical models this does not always happen, because they operate as conditional models, which can induce some surprising behaviour when interpreted marginally (Breslow & Lin, 1995; Gurka et al., 2011; Lin & Breslow, 1996). A nice feature of the copula model is that it assumes the same marginal model as the testing procedure to be used here (Wang et al., 2012), and so is a suitable simulation model when using a Generalised Estimating Equation (GEE) procedure. Copulas are also used in Anderson et al. (2019), although to compare different test procedures, rather than for sample size determination.

Let \( y_{ij} \) be the abundance of the \( j \)th observation of the \( i \)th taxon in our pilot data, where we have taken \( N_{\text{pilot}} \) observations \((i = 1, \ldots , N_{\text{pilot}})\) on a total of \( D \) taxa \((j = 1, \ldots , D)\). The abundances \((y_{1j}, \ldots , y_{Dj})\) are typically correlated across taxa because they interact, directly and indirectly, and because some taxa may respond to unmeasured predictors (Popovic et al., 2019). They are also often discrete, with abundance being measured as a count (or presence/absence), and we will assume that the marginal distribution of \( y_{ij} \) is \( F_{y_i}(y) \), for some cumulative distribution function \( F \), which is usually related to the exponential family (binomial, Poisson, negative
A discrete Gaussian copula (Popovic et al., 2018, 2019) can model correlated discrete data \( y_j \) using latent Gaussian variables \( z_j \):

\[
Y_j = F_{\eta_j}^{-1}(\Phi(z_j)),
\]

where \( \Phi \) is the distribution function of the standard normal distribution. We further assume that the \( z_j \) are jointly normal with zero mean and covariance structure \( \Sigma \) (which has all diagonal elements equal to one),

\[
(z_1, \ldots, z_D) \sim \mathcal{N}_D(0, \Sigma).
\]

This model can be estimated from the pilot data using maximum likelihood to obtain \( \hat{\Psi}_j \) and \( \hat{\Sigma} \). To simulate new multivariate abundances, we simply generate new latent variables \( z_j \) from \( \mathcal{N}_D(0, \hat{\Sigma}) \) and then transform these to abundances using \( y_j = F_{\eta_j}^{-1}(\Phi(z_j)) \), where \( \Theta_j = (\hat{\Psi}_j, \hat{\beta}_j) \) has been parameterised via an effect of interest \( \beta_j^* \).

One difficulty with this approach however is estimating a covariance structure across responses. With \( D \) taxa, there are \( D(D - 1)/2 \) pairwise covariances to estimate, and these will typically be estimated from a small amount of pilot data. Thus, a parsimonious method of modelling covariance is needed if the simulation model is to be trained using pilot data. This issue is addressed here by assuming the \( D \) variables are driven by a shared response to a few \( (Q \ll D) \) unobserved latent variables through the use of factor analysis. These latent variables can be interpreted as unobserved environmental covariates (Warton, Blanchet, et al., 2015). This approach requires \( D(Q + 1) - Q(Q - 1)/2 \) elements to be estimated to formulate a covariance matrix \( \Sigma_{FA} \) which can be much smaller than \( D(D - 1)/2 \). For our fish abundance data set with \( D = 34 \) species, if we take 1, 2 or 3 factors we have 68, 101 or 133 covariance parameters to estimate, which is much less than the 561 parameters that would otherwise have been needed.

The simulation process has been implemented within the ecopower package, using an internal function called extend. This function takes a cord object (obtained by fitting a Gaussian copula to a manyglm object using the cord function from the ecocopula package; Popovic et al., 2022) and simulates \( N \) multivariate abundances using the above procedure. The function then refits the simulated responses to a manyglm object with a data frame that is ‘extended’ in a manner that preserves the original or pre-specified design (Box 2).

### 2.4 Challenge 2—Specifying an interpretable measure of effect size

With a large number of taxa, there are a large number of ways that these taxa can respond to a treatment. To conduct a power analysis, an effect needs to be specified in a way that is interpretable, such that the researcher can understand how large the effect is in the context of their study, thus helping to understand whether or not it is necessary to increase the sample size. This interpretable measure of effect size must also be specified with relatively little a priori information, as this is generally lacking in an ecological setting (at least in comparison to the possible complex relations that the taxa can respond to a treatment).

A simple approach proposed here is to decide on:

1. Which species/response variables are expected to be (i) positively related to a given treatment (e.g. species that increase in abundance; “increasers”), (ii) those expected to be negatively related (species that decrease in abundance; “decreasers”), or (iii) those not related to the treatment at all (“no-effect”).
2. The size of effect (ϕ) that is negatively or positively related to the mean abundance υ of taxa, on the proportional scale. That is, ϕ = 2 implies that the mean abundance doubles for increasers, and halves for decreasers.

This is a relatively simplistic scenario, however it enables the effect size to be captured in a single coefficient ϕ, and for expert opinion to inform the way in which different taxa are likely affected. Being on the proportional scale it also allows regression coefficients for simulated models to be easily specified, for example with log link: \[ \log \text{mean} = \log \text{mean}_{\text{base}} = \beta_1 \text{ for species that increase in abundance and } \log \text{mean}/\text{mean}_{\text{base}} = -\beta_2 \text{ for species that decrease in abundance in the treatment group, relative to the control.} \]

By using existing data of fish surveys along the NSW coastline, as well as results from Curley et al. (2002), assumptions were made regarding which fish species would be expected to change in sites with crayweed (restored), compared to those without crayweed (control). This allowed the desired parameterisation of three types of fish species, whose mean abundances increase, decrease or do not change in restored sites relative to control sites (Figure 2), with the magnitude of these effects being specified with ϕ.

By using pilot data to estimate mean abundance of each species, then taking for example ϕ = 1.2 to specify effects across different treatments, the mean abundances of data to be simulated from marginal distributions F_i can be specified as in Figure 2, ϕ − 1 can also be interpreted as the % change in mean abundance across treatments, so to answer our first experimental design question, we simply specify ϕ = 1.2.

This approach has been implemented within the ecopower package, with the effect_alt function. Users input a manyglm object, the name of the predictor of interest (term), an effect size of interest (effect_size) and a list of taxa that are "increasers" or "decreasers". The function then returns a parameterised coefficient matrix that can be used in ensuing power simulations, where taxa not specified as "increasers" or "decreasers" are assumed to not be affected. There are also options to specify more complicated effect sizes for effects that change over multiple levels of a categorical predictor. To produce the coefficient matrix in Figure 2 we would use the code in Box 1.

2.5 Challenge 3—Managing computation time

Power at significance level α can be estimated from a set of \( n_{\text{sim}} \) simulated datasets as follows:

\[
\text{Power} = \frac{1}{n_{\text{sim}}} \sum_{i=1}^{n_{\text{sim}}} \{I(P_i \leq \alpha)\},
\]

where \( I \) is the indicator function and \( P_i \) is the p-value computed for the \( i \)-th dataset. That is, we estimate the power at significance level α as the sample proportion of \( P \)-values \( P_i \) less than or equal to \( \alpha \) from \( n_{\text{sim}} \) simulated datasets. To get a reliable Monte Carlo estimate of power we need to set \( n_{\text{sim}} \) to a large number (\( n_{\text{sim}} = 1000 \) is often used). In a multivariate abundance setting, the significance of multivariate Wald or score GEE test statistics is calculated via resampling techniques (Wang et al., 2012), where we estimate the p-value from \( n_{\text{res}} \) samples of the observed data (where we usually use \( n_{\text{res}} = 999 \) following Manly, 2018, for tests at \( \alpha = 0.05 \)). However, using a resampling-based hypothesis test like this in a power simulation could be computationally intensive. For each of \( n_{\text{res}} \) simulated data sets, we would have to resample \( n_{\text{crit}} \) times to estimate each simulated \( P \)-value \( P_i \). For example, for simulated data sets of size \( N = 100 \) and \( D = 34 \) taxa, with \( n_{\text{sim}} = 1000 \) and \( n_{\text{res}} = 999 \), this process takes over 3.5h using parallel computing on a computer with 20 logical processes.

A key innovation proposed in this paper is to use a critical value approach to estimate the power of resampling-based hypothesis tests, globally testing the significance of \( n_{\text{sim}} \) test statistics simulated under our parameterised alternative hypothesis \( T_0 \in H_2 \), using a single critical value \( \hat{c}_a \). The critical value can be estimated as the upper \( (1 - \alpha) \)-quantile of simulated test statistics under the null hypothesis \( T_0 \in H_2 \) satisfying

\[
\frac{1}{n_{\text{crit}}} \sum_{j=0}^{n_{\text{crit}}} \{I(P_j > \hat{c}_a)\} = \alpha,
\]

where we include the test statistic computed for the first simulated dataset, \( T_{01} \), in this calculation (denoted as \( T_{01} \)). Power can then be estimated as

\[
\text{Power}_{\text{crit}} = \frac{1}{n_{\text{sim}}} \sum_{i=1}^{n_{\text{sim}}} \{I(T_i > \hat{c}_a)\}.
\]

That is, we are estimating power as the proportion of times our test statistics simulated under the alternative hypothesis \( T_0 \in H_2 \) exceed
our single critical value $\hat{\xi}_c$. Power is now being estimated using test statistics $T_1$ rather than $P$-values calculated via resampling, leading to a huge computational saving—we only need $n_{\text{crit}} + n_{\text{sim}}$ simulated or resampled datasets, rather than $(n_{\text{crit}} + 1) \times n_{\text{sim}}$. When $n_{\text{sim}} = 1000$ and $n_{\text{crit}} = 999$, this leads to a reduction by a factor of 500, reducing computation time for the Crayweed restoration dataset from over 3.5 h to just 2 min.

Using our proposed approach ($\text{Power}_{\text{crit}}$) involves approximations. First, we are assuming that $\xi_c$ is constant across all simulated datasets, when the critical value should actually vary across simulated datasets, because the resampling distribution of a statistic varies as the dataset being resampled varies. This variation gets smaller as sample size $N$ increases, as the resampling distribution converges to the true sampling distribution. Second, Monte Carlo error is also introduced and in two places—first in estimating the critical value $\hat{\xi}_c^s$, then in estimating $\text{Power}_{\text{crit}}$. It is important to estimate the critical value $\hat{\xi}_c^s$ well, which can be achieved by using a large value of $n_{\text{crit}}$. The number of resamples required to accurately estimate $\hat{\xi}_c^s$ increases as the significance level decreases (as is the case for $P$-values). For hypothesis testing using resampling, Manly (2018) recommended that a minimum of $n_{\text{crit}} = 999$ be used when testing at the $\alpha = 0.05$ significance level, and $n_{\text{crit}} = 4999$ be used for testing at $\alpha = 0.01$, based on analysis of binomial variation. However, in our setting, there is additional variation beyond the usual binomial error, due to our estimation of a common critical value as $\hat{\xi}_c^s$, so we would suggest erring on the conservative side, for example, multiplying the values of $n_{\text{crit}}$ recommended by Manly (2018) by about five. Hence in the following, we use $n_{\text{sim}} = 1000$ and $n_{\text{crit}} = 4999$, and use simulations to investigate how well this approximation works in our context.

Our critical value approach to estimating power is implemented in the ecopower package, through the powersim function. It takes a cord object, a coefficient matrix (coeffs) that can be specified using the effect_alt function, a total sample size $N$, the type I error rate (alpha - defaults to 0.05), the name of the predictor of interest (term), number of simulations (nsim) and returns a power estimate. Users can also separately choose the number of simulated null hypothesis test statistics using the ncrit command (defaults to nsim). Computation time is reduced by running simulations in parallel over a series of clusters (ncore), which defaults to one less then the number of cores available on your machine. To estimate power, for a sample size of $N = 100$ and $\alpha = 0.05$, using the pre-specified effect of $\rho = 1.2$ from coeff.alt, we would use the code in Box 2.

### BOX 2  ecopower code to obtain a power estimate over npow = 1000 simulations for a sample size of N = 100 using the powersim function by first fitting a copula model from our manyglm object using the cord function and then using the effect size (coeff. alt) generated in Box 1

```r
> fit.cord <- cord(fit)
> powersim(fit.cord, N=100, coeff.alt, term="Site.Type", nsim=1000, ncrit=4999)
Time elapsed: 0 hr 3 min 6 sec
Power: 0.626
```

### 3 | RESULTS

#### 3.1 | Validating method

The effectiveness of estimating $\text{Power}_{\text{crit}}$ as $\text{Power}_{\text{crit}}$ can be assessed by comparing results from power simulations. We used each of these two methods to answer our first experimental design question: to find the number of samples required to likely detect an effect size of 20% change in mean abundance between control and restored sites.
(where some species have been specified to increase, decrease or not change across the treatments, Figure 2). We repeated Power_{crit} estimation five times to get a sense for how variable estimates of power were when using our proposed critical value method.

As seen in Figure 3, the use of a critical test statistic approach provides a good approximation of power, even for small study designs. Both approaches recommended at least \( N \approx 150 \) sites to likely detect 20% changes in mean abundances, using a conventional power target of 80%. The critical value approach also reduced computation time from over 9 days to 50 min. It tended to result in slightly more variable power estimates, which could be overcome by increasing the number of resamples (\( n_{crit} \)) or simulations (\( n_{sim} \)) at the cost of increased computation time. The critical value approach would still however be appreciably faster (unless \( n_{crit} \) and \( n_{sim} \) were increased to 2,500,000). As is usual when dealing with sample proportions, there is more variation towards the middle of these estimates (when power is around 0.5) than in the extremes.

### 3.2 Example

While Figure 3 suggests \( N \approx 150 \) sites in total, we note this is not feasible under a balanced experimental design, because only 24 control or restored sites are available within the region where this restoration project was undertaken. Hence, it is unlikely that we will detect 20% changes in the mean abundances of fish species between the treatment levels. This is most likely due to the variability of the observed fish abundances creating a lot of noise in the data and the large (66%) proportion of zero counts observed in the data.

The second experimental design question we are interested in is: what effect sizes are likely to be observed at 12 sites per treatment? In order to answer this question, we can simply plot power curves over a range of effect size specifications. For a broader perspective, we plotted power curves as a function of sample size \( N \) (Figure 4), for effect size specifications \( \rho_{10\%} = 1.1, \rho_{20\%} = 1.2, \rho_{30\%} = 1.3, \rho_{40\%} = 1.4, \rho_{50\%} = 1.5, \rho_{60\%} = 1.6, \rho_{70\%} = 1.7, \rho_{80\%} = 1.8, \rho_{90\%} = 1.9, \rho_{100\%} = 2 \). Computing each power curve took approximately 40 min of computation time on a computer with 20 logical processors. Figure 3 took less time, because the sample sizes under consideration were smaller. As expected, we observed that increasing the effect size specifications also increased the power of each experimental design (Figure 4), since larger effect sizes are easier to detect. Importantly, we observed that under the maximum balanced sample size of \( N = 24 \) sites, the smallest effect size to be observed under a conventional power target of 80% is \( \rho_{90\%} = 1.7 \), or equivalently 70% changes in the mean abundances of fish species. The pilot survey included \( N_{pilot} = 9 \) sites, which would not be able to reliably detect any of the simulated effect sizes.

### 4 Discussion

Here, we have described a sample size estimation procedure for multivariate abundance data. Techniques and software for this are much needed in ecology; however, the problem is rather difficult technically, given the three aforementioned challenges. The main innovations of our procedure are to: specify a joint data generating model through discrete margin copulas that can be tuned to the properties of pilot data; implement a simple and interpretable approach to specifying an effect size for complex multivariate effects; reduce computation time by a factor of 500, which was achieved using a critical value approach to power estimation. The procedure has been coded in R with general purpose functions in the ecopower package, which can be downloaded from CRAN (Maslen & Lim, 2023). This procedure can be applied to a vast range of ecological studies with multivariate abundance data to answer experimental design questions and give sample size recommendations, an important yet technically difficult task previously unaddressed in the literature.

![Figure 4](image-url)  
**Figure 4**: Power curves for the Operation Crayweed Restoration Project under a range of effect size specifications: \( \rho_{10\%} = 1.1, \rho_{20\%} = 1.2, \rho_{30\%} = 1.3, \rho_{40\%} = 1.4, \rho_{50\%} = 1.5, \rho_{60\%} = 1.6, \rho_{70\%} = 1.7, \rho_{80\%} = 1.8, \rho_{90\%} = 1.9, \rho_{100\%} = 2 \). ‘\( N \)’ is the total sample size.
The ecopower software could be used for power analysis of many designs and scenarios beyond what has been presented here. While we analysed counts, any distribution compatible with ecoCopula:cord could be used, which includes most distributions related to the exponential family, including distributions suitable for counts, presence/absence, biomass and semi-quantitative measures of abundance. Power can also be estimated for more complex designs, provided that they can be specified as fixed effects models that are linear in the mean parameters (that is, the model for the mean can be specified using the usual formula interface for generalised linear models). For instance, the methods proposed here (and the software tools in the ecopower package) can estimate power for effects on both categorical and quantitative predictors, interactions, and can handle designs with multiple covariates. It can also investigate unbalanced designs. Ecopower can also be used to explore different types of scenarios. To estimate the power to detect complete removal of a species during restoration, for example, we could set the effect size coefficient to a large negative number. If desired, more complicated effect size structures could be specified rather than using the increaser/decreaser approach proposed in this paper. Note however that all effect size parameters must be set a priori, and the more intricate the scenario, the more a priori decisions would be needed.

Power analyses are highly sensitive to a priori decisions about the size and nature of effect size, and the amount of variability assumed to be in the data (Cohen, 1992, 2013). As such, we encourage researchers to explore sensitivity to these factors using multiple power simulation runs. For multivariate abundance data we can expect power to be especially sensitive to decisions made for species with high abundance, because these tend to be more influential on power (for example, see Warton et al., 2012, Figure 4c). When using our approach to specifying effect size, we expect power to be more sensitive to choice of increaser species, because such species will then become more abundant in simulated datasets. For example, in our power simulation example, the most abundant species was classified as an increaser, and if it were changed to a “no-effect” species then power would in some cases decrease by 20% (Figure 5).

When modelling count data, we can also expect power to be sensitive to assumptions about overdispersion, because this controls how variable the simulated data are (especially for abundant species). In our example, halving the amount of overdispersion in our simulations increased power by as much as 25% (Figure 5). Hence, while we concluded from our example that 12 sites per treatment could only really detect an effect size of \( \rho = 1.7 \), meaning a 70% change in mean abundance for increasers and decreasers, perhaps this design could detect slightly smaller effect sizes had we overestimated overdispersion from our pilot data, or if some abundant species labelled as “no-effect” or decreaser species were actually to increase in numbers following restoration.

It would be interesting to apply this procedure to other study designs and consider whether the patterns seen here apply elsewhere too. Specifically, this study’s maximum balanced sampling design \( (N = 24) \) was underpowered for the size of effects that we would like to detect. Is this more generally true—do ecological monitoring studies tend not to sample enough sites to detect effects of practical interest?

**AUTHOR CONTRIBUTIONS**

Ben Maslen designed the power analysis procedure and built the R package ecopower with Michelle Lim. David Warton and Gordana Popovic provided expertise and direction in copula modelling, multivariate analysis and power simulation. Ezequiel Marzinelli provided the fish abundance data, the motivating example and used his domain expertise to help define the effect size of interest. Ben Maslen wrote the first draft of the manuscript, and all authors contributed substantially to revisions.

**ACKNOWLEDGEMENTS**

David Warton and Gordana Popovic are supported by the Australian Research Council Discovery Grant DP210101923. Ezequiel
Marzinelli is supported by the Australian Research Council Linkage Grant LP160100836, the NSW DPI Recreational Fishing Trust, and Environmental Trust. Ezequiel Marzinelli is also supported by Australian Research Council Discovery Grants DP180104041 and DP190102030, respectively. We also thank Alexandra Campbell for field sampling and Adriana Verges for providing domain expertise to help define an effect size of interest. Finally, we thank the reviewers for improving the manuscript through their suggestions and insight.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

PEER REVIEW

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1111-2041-210X.14094.

DATA AVAILABILITY STATEMENT

The data and code used to produce the results presented within this manuscript can be found through the R package ecopower (Maslen & Lim, 2023). The package can be installed from CRAN (install.packages("ecopower"), and the development version can be installed through github (remotes::install_github("BenMaslen/ecopower", dependencies = T, build_vignettes = T)).

ORCID

Ben Maslen https://orcid.org/0000-0003-3475-6247
Gordana Popovic https://orcid.org/0000-0002-1376-1058
Michelle Lim https://orcid.org/0000-0002-4891-7761
Ezequiel Marzinelli https://orcid.org/0000-0002-3762-3389
David Walton https://orcid.org/0000-0001-9441-6645

REFERENCES

Anderson, M. J. (2001). A new method for non-parametric multivariate analysis of variance. Austral Ecology, 26(1), 32–46.
Anderson, M. J., de Valpine, P., Punnett, A., & Miller, A. E. (2019). A pathway for multivariate analysis of ecological communities using copulas. Ecology and Evolution, 9(6), 3276–3294.
Angeler, D. G., Viedma, O., & Moreno, J. M. (2009). Statistical performance and information content of time lag analysis and redundancy analysis in time series modeling. Ecology, 90(11), 3245–3257.
Breslow, N. E., & Lin, X. (1995). Bias correction in generalised linear mixed models with a single component of dispersion. Biometrika, 82(1), 81–91.
Campbell, A. H., Marzinelli, E. M., Vergés, A., Coleman, M. A., & Steinberg, P. D. (2014). Towards restoration of missing underwater forests. PLoS ONE, 9(1), e84106.
Cohen, J. (1992). Statistical power analysis. Current Directions in Psychological Science, 1(3), 98–101.
Cohen, J. (2013). Statistical power analysis for the behavioral sciences. Routledge.
Coleman, M. A., Kelaher, B. P., Steinberg, P. D., & Millar, A. J. (2008). Absence of a large brown macroalga on urbanized rocky reefs around Sydney, Australia, and evidence for historical decline. Journal of Phycology, 44(4), 897–901.
Collins, S. L., Micheli, F., & Hartt, L. (2000). A method to determine rates and patterns of variability in ecological communities. Oikos, 91(2), 285–293.
Curley, B. G., Kingsford, M. J., & Gillanders, B. M. (2002). Spatial and habitat-related patterns of temperate reef fish assemblages: Implications for the design of marine protected areas. Marine and Freshwater Research, 53(8), 1197–1210.
Fritz, M. S., & MacKinnon, D. P. (2007). Required sample size to detect the mediated effect. Psychological Science, 18(3), 233–239.
Gerrodette, T. (1987). A power analysis for detecting trends. Ecology, 68(5), 1364–1372.
Green, P., & MacLeod, C. J. (2016). Simr: An r package for power analysis of generalized linear mixed models by simulation. Methods in Ecology and Evolution, 7(4), 493–498.
Green, R. H. (1989). Power analysis and practical strategies for environmental monitoring. Environmental Research, 50(1), 195–205.
Gurka, M. J., Edwards, L. J., & Muller, K. E. (2011). Avoiding bias in mixed model inference for fixed effects. Statistics in Medicine, 30(22), 2696–2707.
Irvin, K. M., Dinger, E. C., & Sarr, D. (2011). A power analysis for multivariate tests of temporal trend in species composition. Ecology, 92(10), 1879–1886.
Johnson, P. C., Barry, S. J., Ferguson, H. M., & Müller, P. (2015). Power analysis for generalized linear mixed models in ecology and evolution. Methods in Ecology and Evolution, 6(2), 133–142.
Kelley, K., & Preacher, K. J. (2012). On effect size. Psychological Methods, 17(2), 137–152.
Lin, X., & Breslow, N. E. (1996). Bias correction in generalized linear mixed models with multiple components of dispersion. Journal of the American Statistical Association, 91(435), 1007–1016.
Manly, B. F. (2018). Randomization, bootstrap and Monte Carlo methods in biology: Texts in statistical science. Chapman and Hall/CRC.
Marzinelli, E., Campbell, A., Vergés, A., Coleman, M., Kelaher, B. P., & Steinberg, P. (2014). Restoring seaweeds: Does the declining fucoid phyllospora comosa support different biodiversity than other habitats? Journal of Applied Phycology, 26(2), 1089–1096.
Marzinelli, E. M., Leong, M. R., Campbell, A. H., Steinberg, P. D., & Vergés, A. (2016). Does restoration of a habitat-forming seaweed restore associated faunal diversity? Restoration Ecology, 24(1), 81–90.
Maslen, B., & Lim, M. (2023). ecopower. https://doi.org/10.5281/zenodo.7679303
Ovaskainen, O., Tikhonov, G., Norberg, A., Guillaume Blanchet, F., Duan, L., Dunson, D., Roslin, T., & Abrego, N. (2017). How to make more out of community data? A conceptual framework and its implementation as models and software. Ecology Letters, 20(5), 561–576.
Popovic, G. C., Hui, F. K., & Warton, D. I. (2018). A general algorithm for covariance modeling of discrete data. Journal of Multivariate Analysis, 165, 86–100.
Popovic, G. C., Hui, F. K. C., & Warton, D. I. (2022). Fast model-based ordination with copulas. Methods in Ecology and Evolution, 13(1), 194–202.
Popovic, G. C., Warton, D. I., Thomson, F. J., Hui, F. K., & Moles, A. T. (2019). Untangling direct species associations from indirect mediator species effects with graphical models. Methods in Ecology and Evolution, 10(9), 1571–1583.
Rosenthal, R., Cooper, H., & Hedges, L. (1994). Parametric measures of effect size. The Handbook of Research Synthesis, 621, 231–244.
Vergés, A., Campbell, A. H., Wood, G., Kajlich, L., Eger, A. M., Cruz, D., Langley, M., Bolton, D., Coleman, M. A., Turpin, J., Crawford, M., Combes, N., Camilleri, A., Steinberg, P. D., & Marzinelli, E. M. (2020). Operation crayweed: Ecological and sociocultural aspects of restoring Sydney’s underwater forests. Ecological Management & Restoration, 21(2), 74–85.
Wang, Y., Naumann, U., Wright, S. T., & Warton, D. I. (2012). Mvabund–an r package for model-based analysis of multivariate abundance data. Methods in Ecology and Evolution, 3(3), 471–474.
Warton, D. I. (2011). Regularized sandwich estimators for analysis of high-dimensional data using generalized estimating equations. Biometrics, 67(1), 116–123.
Warton, D. I., Blanchet, F. G., O’Hara, R. B., Ovaskainen, O., Taskinen, S., Walker, S. C., & Hui, F. K. (2015). So many variables: Joint modelling in community ecology. Trends in Ecology & Evolution, 30(12), 766–779.
Warton, D. I., Foster, S. D., De’ath, G., Stoklosa, J., & Dunstan, P. K. (2015). Model-based thinking for community ecology. Plant Ecology, 216(5), 669–682.
Warton, D. I., Wright, S. T., & Wang, Y. (2012). Distance-based multivariate analyses confound location and dispersion effects. Methods in Ecology and Evolution, 3(1), 89–101.
Wood, G., Marzinelli, E., Coleman, M., Campbell, A. H., Santini, N., Kajlich, L., Verdura, J., Wodak, J., Steinberg, P., & Vergés, A. (2019). Restoring subtidal marine macrophytes in the anthropocene: Trajectories and future-proofing. Marine and Freshwater Research, 70(7), 936–951.
Xu, J., Nelson, B. L., & Hong, J. (2010). Industrial strength compass: A comprehensive algorithm and software for optimization via simulation. ACM Transactions on Modeling and Computer Simulation (TOMACS), 20(1), 3.

**SUPPORTING INFORMATION**

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**Figure S6.** Mean–variance plot of fish abundances for D = 34 species. The red line depicts the mean–variance assumption under a Poisson model (mean = variance). The variance appears larger than the mean for species with larger abundances, indicating over-dispersion relative to the Poisson distribution, with the negative binomial distribution being preferred for this dataset.

**Figure S7.** Diagnostic plot of a negative binomial model fit to the fish abundance data from the Crayweed Restoration Project using manyglm. The large cluster of data on the left hand side of this graph refers to observations predicted to have zero abundance in a treatment. The lack of fan shape in the residual versus fitted plot implies the negative binomial distribution has adequately accounted for the mean–variance relationship in the data shown in Figure S6.

**How to cite this article:** Maslen, B., Popovic, G., Lim, M., Marzinelli, E., & Warton, D. (2023). How many sites? Methods to assist design decisions when collecting multivariate data in ecology. Methods in Ecology and Evolution, 14, 1564–1573. [https://doi.org/10.1111/2041-210X.14094](https://doi.org/10.1111/2041-210X.14094)