ECOGRAPHY

Research

SSP: an R package to estimate sampling effort in studies of ecological communities

Edlin J. Guerra-Castro, Juan Carlos Cajas, Nuno Simões, Juan J. Cruz-Motta and Maite Mascaró

E. J. Guerra-Castro and J. C. Cajas, Escuela Nacional de Estudios Superiores, Unidad Mérida, Univ. Nacional Autónoma de México, Mérida, Yucatán, México. – N. Simões and M. Mascaró (https://orcid.org/0000-0003-3614-4383) ✉ (mmm@ciencias.unam.mx), Unidad Multidisciplinaria de Docencia e Investigación, Facultad de Ciencias, Univ. Nacional Autónoma de México, Sisal, Yucatán, México. EJG-C, NS and MM also at: Laboratorio de Resilencia Costera (LANRESC, CONACYT), Sisal, Yucatán, México. NS, International Chair for Coastal and Marine Studies in Mexico, Harte Research Inst. for Gulf of Mexico Studies, Texas A&M Univ.-Corpus Christi, USA. – J. J Cruz-Motta, Dept of Marine Sciences, Univ. of Puerto Rico, Mayagüez, Puerto Rico.

Ecography
44: 561–573, 2021
doi: 10.1111/ecog.05284

Subject Editor: Brody Sandel
Editor-in-Chief: Miguel Araújo
Accepted 16 December 2020

SSP (simulation-based sampling protocol) is an R package that uses simulations of ecological data and dissimilarity-based multivariate standard error (*MultSE*) as an estimator of precision to evaluate the adequacy of different sampling efforts for studies that will test hypothesis using permutational multivariate analysis of variance. The procedure consists in simulating several extensive data matrixes that mimic some of the relevant ecological features of the community of interest using a pilot data set. For each simulated data, several sampling efforts are repeatedly executed and *MultSE* calculated. The mean value, 0.025 and 0.975 quantiles of *MultSE* for each sampling effort across all simulated data are then estimated and standardized regarding the lowest sampling effort. The optimal sampling effort is identified as that in which the increase in sampling effort does not improve the highest *MultSE* beyond a threshold value (e.g. 2.5%). The performance of **SSP** was validated using real data. In all three cases, the simulated data mimicked the real data and allowed to evaluate the relationship *MultSE* – *n* beyond the sampling size of the pilot studies. **SSP** can be used to estimate sample size in a wide variety of situations, ranging from simple (e.g. single site) to more complex (e.g. several sites for different habitats) experimental designs. The latter constitutes an important advantage in the context of multi-scale studies in ecology. An online version of **SSP** is available for users without an R background.

Keywords: community ecology, dissimilarities, multivariate analyses, PERMANOVA, resampling, sampling design, simulation, standard error

Background

Defining sample size is a key decision in the planning of ecological research. In the context of hypothesis testing, a decision to take too few samples could produce misleading information about the statistical population, imprecise statistics or a high probability of retaining a false null hypothesis. Instead, increasing sampling size improves the precision of estimations and the power of statistical tests, but will also increase its costs (Mapstone 1995, Underwood 1997, Underwood and Chapman 2003).

There is a wide variety of methodologies aimed at optimizing the use of resources to obtain the best possible sampling design at the lowest cost (a.k.a. cost–benefit optimization). However, most methods used for this purpose are based on statistical theories that consider only one response variable; that is when the variable of interest can be represented as a single descriptor. In early ecological studies at community level, the sample size was usually determined based on single merging variables such as species richness or some diversity index (Green 1979, Clarke and Green 1988). However, the study of communities has evolved from the use of univariate descriptors to the use of multivariate/dissimilarity-based statistics (Clarke 1993, Anderson et al. 2006, Legendre and De Cáceres 2013). Therefore, the analytical methods for estimation of sample size should consider the highly-dimensional structure of ecological data (Anderson and Santana-Garcon 2015, Blanchet et al. 2016).

Most of the conventional approaches in experimental and sampling design consist in evaluating the precision of the arithmetic mean of a response variable obtained in a previous pilot study, which allows one to estimate the number of replicates that are needed to improve that precision. These strategies consider the expected random error, a previous estimate of the natural variability of the variable, the effort with which such information was obtained, and some theoretical distribution as a reference (e.g. standardized normal distribution) (Quinn and Keough 2002). Statistical precision refers to the level of concordance between multiple estimators of the same parameter under the same sampling procedure (Underwood 1997). To estimate the precision of a mean obtained from a random sample, the standard error of the mean must be calculated ($s_{\bar{x}} = s / \sqrt{n}$), where s represents the standard deviation of the sample and n the number of sampling units. Data from a previous pilot study providing such information is often required. Because the denominator of the standard error is the sample size, the standard error will always decrease (and precision will improve) as the sampling effort increases.

Anderson and Santana-Garcon (2015) developed a computationally intensive statistical approach that allows the estimation of a multivariate *pseudo* standard error (*MultSE*) as a proxy of precision to identify an optimal sample size (Eq. 3 Anderson and Santana-Garcon 2015). Their procedure consists of a double resampling (with and without replacement) of a data matrix obtained during a pilot study. For each resampling, the dissimilarities between each pair of samples of a randomly chosen subset of the main matrix are estimated, and the *MultSE* is calculated. The procedure is repeated several times with $n_i = 2, 3, 4$ to n, where n refers to the original sample size in the pilot study. The behavior of *MultSE* is projected on a plot of means and error bars, with the abscissa showing the sampling effort and the ordinate showing *MultSE* values (Fig. 1 in Anderson and Santana-Garcon 2015). The plot shows the way precision relates to sample size, thereby helping to identify the sampling effort for an acceptable measure of standard error. The R scripts for this approach are available as supplementary material in Anderson and Santana-Garcon (2015).

The method proposed by Anderson and Santana-Garcon (2015) is pioneer in assessing sample size in community studies where the variability in the composition of species or structure of the assemblage is the focus of the research. This method has recently been proposed as a framework to define sampling effort in coastal dunes and coral reefs monitoring programs (Maccherini et al. 2020, Montilla et al. 2020). However, we identified two limitations in the original approach: 1) this method cannot extrapolate the *MultSE* – *n* relationship beyond the sampling effort used during the pilot study (i.e. by definition, the chosen *n* may only be less than or equal to that used in the pilot study), and 2) any random deviation of estimates obtained with the original sampling will be reiteratively reflected on the *MultSE* because permutations are restricted to the same sampling space delimited by the pilot survey. **SSP** addressed these limitations using simulations. Furthermore, to define the magnitude of the *MultSE* at which variation in species composition is ecologically significant is challenging because of the complexity involved in interpreting change in a multivariate context. Whilst such difficulty is both intrinsic and sensitive to each case study, deciding sample size on the basis of precision (as measured by *MultSE*) provides a standardized, hence repeatable procedure that leads to cost–benefit optimization.

We present **SSP ver. 1**, an R package (<www.r-project.org>) designed to estimate sample effort in studies of ecological communities using intensive sampling over several sets of simulated data. **SSP** can also be accessed from a web application, designed for users without R skills. Our procedure is based on the previous definition of *MultSE* but eludes the double resampling over a unique pilot data set (Anderson and Santana-Garcon 2015). In general, the protocol consists of 1) simulating several data matrixes that retain observed properties of the community of interest, 2) obtaining independent estimates of *MultSE* from those simulated data matrixes, for different sample sizes and number of sites and 3) a quantitative identification of the optimal sampling effort as well as the graphic representation of the *MultSE* – *n* relationship and the optimal effort. Data collected in a standard pilot survey are used to simulate the data matrixes but are not included in the resampling procedure. With **SSP**, users will be able to objectively identify the number of samples and sites necessary to characterize the community of interest with sufficient precision at a reasonable cost. The use of several simulated larger data matrixes as a central element of the procedure will assure a better appreciation of the *MultSE* – *n* relationship over a wider range of sample sizes. **SSP** was evaluated using three sets of real data: 1) micromollusk of marine shallow sandy bottoms, 2) coral reef sponges and 3) epibenthic assemblages on Caribbean mangrove roots.

Methods and features

Workflow of SSP

SSP was divided into seven stages: 1) extrapolation of assemblage parameters using pilot data, 2) simulation of several

data sets based on extrapolated parameters, 3) evaluation of plausibility of simulated data, 4) repeated estimations of *MultSE* for different sampling designs in simulated data sets, 5) summary of the behavior of *MultSE* for each sampling design across all simulated data sets, 6) identification of the optimal sampling effort and 7) graphical representation of the *MultSE* and the sampling effort (Fig. 1). For each of these steps, we provide the following seven functions:

i. **assempar**: The following ecological properties of the assemblage are estimated: potential number of species, probability of occurrence of each species within and among sites, the pattern of abundance of each species and the pattern of spatial aggregation of species. The potential number of species in the assemblage (S_{est}) is estimated with any of the incidence-based nonparametric methods available in the `specpool` function of the `vegan` package (Oksanen et al. 2015). The probability of occurrence of each species is calculated between and within sites. The former is computed as the frequency of occurrence of each species against the number of sites sampled (f_b); the latter is computed as the weighted average frequencies in sites where the species were present (f_w, or just f if the pilot data is restricted to one site) (Gaston 1994, Magurran and Henderson 2011). The mean and variance ($\bar{x}$ and s^2, respectively) of the abundance of each species are also estimated. The degree of spatial aggregation of species (only for real counts of individuals) is identified with the index of dispersion D (Clarke et al. 2006). The corresponding properties of unseen species are approximated using the information on observed species: the probability of occurrence is assumed to be equal to the rarest species of pilot data. The mean (and variance) of species abundance are defined using random Poisson values with lambda as the overall mean. **assempar** returns an object of class list, to be used by **simdata**.

ii. **simdata**: The simulation starts by setting the dimensions of the data matrix $\hat{Y}$ (i.e. number of columns and rows). The number of columns was programmed to be equal to the potential number of species, while the number of rows (N_r) is defined arbitrarily as the potential number of sampling units per site (N) multiplied by the potential number of sites. The presence/absence of each species at each site is simulated with Bernoulli trials where the probability of success equals to the empirical frequency of occurrence of each species among sites in the pilot data (f_b). For sites with the simulated presence of the species, Bernoulli trials are used again with the probability of success equal to the empirical frequency estimated within the sites in pilot data (f_w). If required, the presence/absence matrixes are converted to matrixes of abundance replacing species presence with random values from an adequate statistical distribution and parameters equals to those estimated in the pilot data (McArdle and Anderson 2004). Counts of individuals are generated using Poisson or negative binomial distributions, depending on the degree of species aggregation in the pilot data (McArdle and Anderson 2004, Anderson and Walsh 2013). When abundances were measured as a continuous variable (i.e. coverage, biomass), data are generated using the lognormal distribution. The simulation procedure is repeated to create as many simulated data matrixes as needed. It is important to highlight that the procedure assumes 'similar environmental conditions across samples; this is, that the multivariate structure of the assemblage is produced by intrinsic properties of species (e.g. patterns of gregariousness/dispersion) and is not influenced by environmental constraints. Whilst this assumption is common to other statistical methods examining community structure, it is rarely considered and its implications are often overlooked. The assumption of similar environmental conditions across samples, however, requires that simulations do not combine data from different habitats (e.g. mixing quadrats from high and low tides in a rocky shore, forest plots from different altitudes in a vegetation study). For these cases, simulations should be performed independently for each habitat or strata in the environmental gradient. **simdata** returns an object of class list that will be later used by **sampsd** and **datquality**.

iii. **datquality**: The quality of the simulated data matrixes is assessed by their resemblance to the pilot data considering the following estimations: 1) the average number of species per sampling unit, 2) the average species diversity (Simpson diversity index) per sampling unit and 3) the multivariate dispersion (MVD), measured as the average dissimilarity from all sampling units to the main centroid in the space corresponding to the dissimilarity measure of interest (Anderson 2006). In general, 1), 2) and 3) should be similar in simulated and pilot data. **datquality** returns a table with these estimates.

iv. **sampsd**: If several virtual sites have been simulated, subsets of sites of size 2 to m are sampled, followed by the selection of sampling units (from 2 to n) using inclusion probabilities and self-weighted two-stage sampling (Tillé 2011). Each combination of sampling effort (number of sample units and sites) is repeated several times (e.g. 100) for all simulated matrixes. If simulated data correspond to a single site, sampling without replacement is performed for each sample size (from 2 to n) within each simulated matrix. This approach is computationally intensive, especially when k is high (e.g. 100), and should be considered when time availability and computational resources are scarce. For each sample, suitable pre-treatments are applied and distance/similarity matrixes estimated using the appropriate coefficient. When simulations are done for a single site, the *MultSE* is calculated as $\sqrt{V / n}$, being V the *pseudo* variance measured at each sample of size n. When several sites are generated, *MultSE* are calculated using estimates of the *pseudo* component of variation of residuals ($\sqrt{CV_{\text{residual}} / n}$) and sites ($\sqrt{CV_{\text{site}} / m}$) from a distance-based multivariate analysis of variance (Anderson 2017).

v. **summary_ssp**: This function is required to estimate an average of all *MultSE* obtained with the k repetitions for each sampling effort within each simulated data. It also estimates an overall mean, together with the lower and

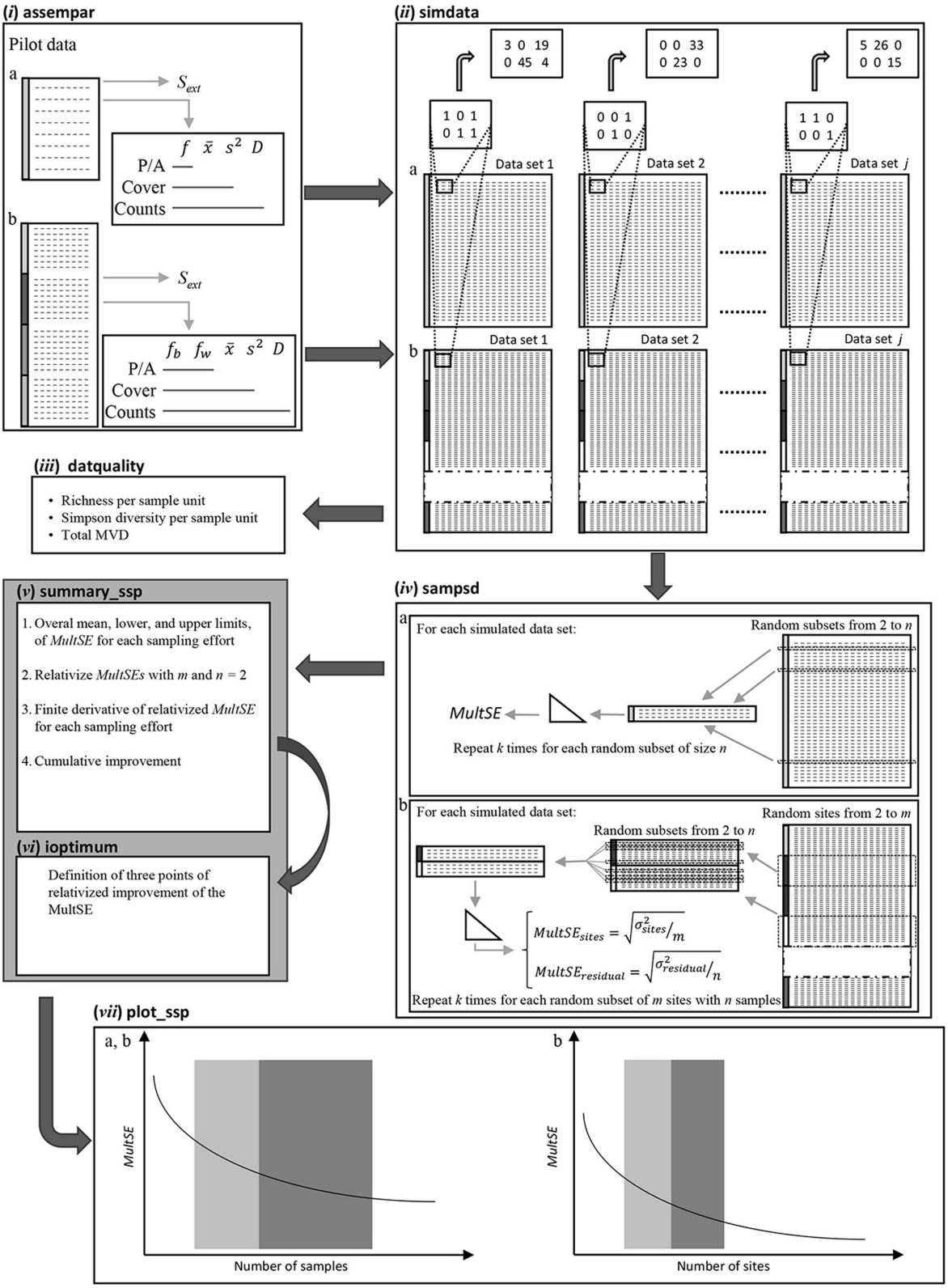

Figure 1. Flowchart of **SSP v1**. The procedure begins with the estimation of the parameters using a pilot data and **assempar**. The pilot data may correspond to a single (a) or many sites (b). Simulations are done with **simdata**. The output is evaluated with **datquality**. The resampling and *MultSE* estimations are generated with **sampsd**. Summary of results are obtained with **summary_ssp**, and the cut-off points are identified with **ioptimum**. The plot showing the behavior of the *MultSE* is generated with **plot_ssp**.

upper intervals of means (0.025 and 0.975 quantiles) for each sampling effort among all simulated data. To evaluate the rate of change of the averaged *MultSE* according to the sampling effort, a relative measure of the maximum *MultSE* value (obtained with the lowest sampling effort: 2) is calculated; then, a standard forward finite derivation is computed. All results are summarised in a table (object of class data frame) used later to plot *MultSE* and the sampling effort.

vi. **ioptimum**: This function identifies three cut-off points based on the finite derivatives between the standardized *MultSE* and the sampling effort (as the percentage of improvement in precision per sample unit, by default 10%, 5% and 2.5%), thus allowing to identify: 1) the minimum improvement required, 2) sub-optimal improvement and 3) optimal improvement. It is possible that the cut-off points defined by the default settings are not achieved (e.g. if the arguments *n* or *m* of **sampsd** were set low). In such cases, a warning message will specifically indicate which cut-off point was not achieved and the current maximum effort will be returned. Functions **sampsd** and **summary_ssp** must then be run with higher values of *n* or *m*. Alternatively, the cut-off points can be made flexible by setting it to for example 15, 10, 5%, respectively, or higher.

vii. **ssp_plot**: This function allows the user to visualize the behavior of the *MultSE* as sampling effort increases. When the simulation involves two sampling scales, a plot for samples and a plot for sites are generated. Above the *MultSE* ~ sampling effort projection, two shaded areas are shown. These areas reflect the sampling effort that improves the precision to acceptable (light gray) or desirable levels (dark grey), but gains beyond the latter could be considered unnecessary. In addition, the relative improvement (considering the *MultSE* estimated with the lower sampling effort) is presented for each sampling effort as a cumulative percentage. This feature is especially useful because it indicates quantitative measure of how much the precision is improved per increase in sampling unit. The plot is generated with `ggplot2`, and the resulting object can be further modified using the functions of that package.

Examples

1. Micromollusks of marine shallow sandy bottoms: The presence/absence of 68 species were recorded using six cores of 10 cm diameter and 10 cm deep taken in sandy bottoms at Cayo Nuevo, Gulf of Mexico, Mexico (a small reef cay located 240 km off the North-Western coast of Yucatan). Data correspond to a study on the biodiversity of marine benthic reef habitats off the Yucatan shelf (Ortigosa et al. 2018). The main objective was to estimate an adequate sampling effort for further quantitative studies to characterize temporal changes in species composition. To speed up the process, only 20 data sets were simulated. Each data matrix consisted of $N = 100$ potential sampling replicates in one site, and subsets ranging in size from 2 to 50 were repeated 10 times. The Jaccard index was used as the similarity measure between sample units. **SSP** indicated that the lowest value of precision would improve from 37% (suboptimal) to 55% (optimal) with a sampling effort between 5 and 10 samples, a remarkable precision gained with each additional sample (Fig. 2). After 11 samples, the improvement in precision obtained with increased sampling effort is small enough to consider the extra effort unnecessary. The R script for this example is in Box 1.

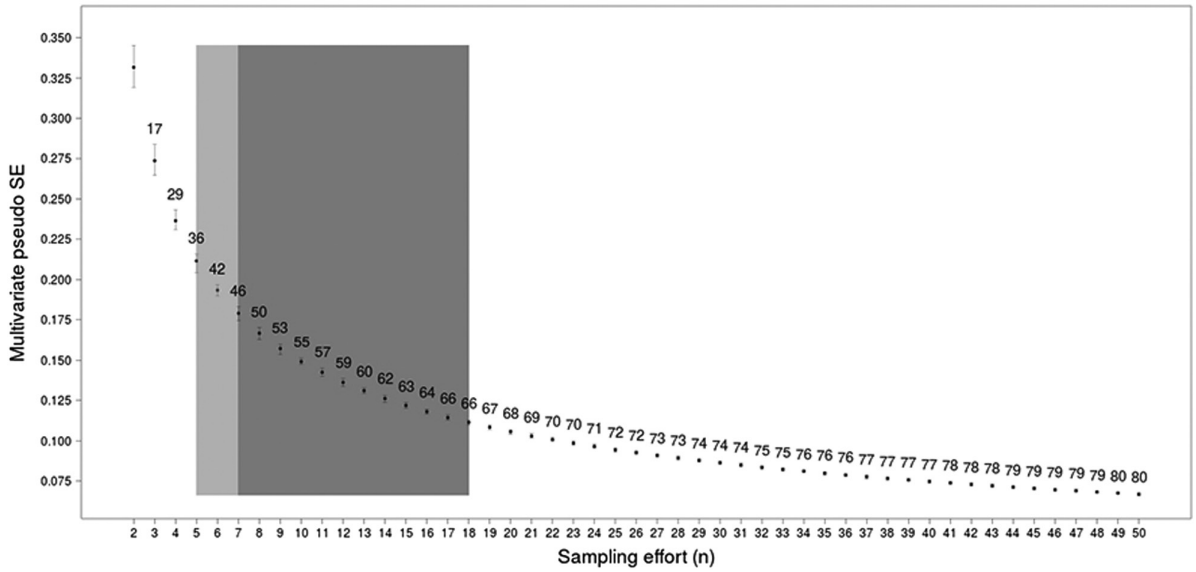

Figure 2. Relation between *MultSE* (in Jaccard dissimilarity) and sampling effort using micromollusk simulated data from Example 1. Shaded areas indicate the range of samples in which an increase in sampling effort provided a suboptimal (light grey) and optimal improvement in precision (dark grey). The cumulative relative improvement is projected over each error bar.

Box 1. SSP applied to data on micromollusks assemblages from sandy bottoms at Cayo Nuevo, Gulf of Mexico, Mexico (Ortigosa et al. 2018). The execution of these codes took 15 s in RStudio Cloud (<https://rstudio.cloud/>) with the default resource settings (1 GB RAM, 1 CPU).

```
library(SSP)
data(micromollusk)

#Estimation of parameters
par.mic <- assempar(data=micromollusk, type="P/A")

#Simulation of data
sim.mic <- simdata(Par=par.mic, cases=20, N=100, site=1)

# Quality of simulated data
qua.mic <- datquality(data=micromollusk, dat.sim=sim.mic, Par=par.mic, transforma-
tion="none", method="jaccard")

#Sampling and estimation of MultSE
samp.mic <- sampsd(sim.mic, par.mic, transformation="P/A", method="jaccard",     n=50,
m=1, k=10)

#Summarizing results
sum.mic <- summary_ssp(results=samp.mic, multi.site=FALSE)

#Identification of optimal effort
opt.mic <- ioptimum(xx=sum.mic, multi.site=FALSE, c1=10, c2 =5, c3= 1)

#plot
fig.2 <- plot_ssp(xx=sum.mic, opt=opt.mic, multi.site=FALSE)
```

2. Coral reef sponges: The structure and composition of sponge assemblages associated to Alacranes Reef National Park (ARNP), Gulf of Mexico, Mexico, was estimated in 36 transects of 20 × 1 m across six sites (≈ 4–8 transect per site). In each transect, the colonies of 41 species of sponges were counted. This data corresponds to a pilot study on sponge biodiversity in reef habitats of the Yucatán shelf (Ugalde et al. 2015). The main objective was to estimate an adequate sampling effort at two spatial scales (i.e. transect and sites) for further quantitative studies. The studied area represented the leeward area of the reef with similar geomorphology; hence, differences in sponge diversity due to environmental heterogeneity at this spatial scale could not be argued a priori. Therefore, we considered valid to simulate data for the entire leeward area using the information of the six sites. Here again, to speed up the process, only 10 data sets were simulated, each consisting of 20 virtual sites and 20 virtual transects per site. Combinations of n (from 2 to 20) and sites (from 2 to 20) were repeatedly sampled 10 times each. The Bray–Curtis index was used as the similarity measure between sample units after a square root transformation of simulated abundances. The R script for this second example is in Box 2.

Results showed a noticeable decrease of the *MultSE* between 5 and 11 sites (Fig. 3). A suboptimal improvement of 44% in sampling effort was attained with seven sites, whereas an optimal improvement of 55% was achieved with 11 sites. Sampling nine sites would improve precision in the lowest value of approximately 51%. A suboptimal improvement in sampling effort among transects is accomplished with 5 or 6 replicates, whereas an optimal improvement is attained with more than 7 but less than 11 replicates. Each additional sample increased the highest *MultSE* value by 2–3%, achieving a 55% improvement in sampling effort with 10 transects.

A noticeable feature of this simulation is the marked differences in *MultSE* obtained for the two sources of variation. The magnitude of the difference in variation corresponds to that obtained with the *pseudo*-components of variation estimated for sites and residuals in a distance-based multivariate analysis of variance of the pilot data $cv_{sites} = 26.4$, $cv_{transects} = 34.7$ (R script in the Supporting information). These results, together with considerations of sampling costs and the relative contribution of each spatial scale to total variation, suggest the convenience of keeping the number of sites within the range of suboptimal improvement (i.e. 5–7) and setting the number of transects to 8 (Fig. 3).

3. Epibenthic assemblages on Caribbean mangrove roots: Data consists of the coverage (by point-intercept) of 116 taxa identified in 180 mangrove roots sampled under a hierarchically nested spatial design (Guerra-Castro et al. 2011). The design included six random sites within each of three sectors of the lagoon system corresponding to a strong environmental gradient: external (E), intermediate (M) and internal (I). The

Box 2. SSP applied to data on sponge assemblages associated to Alacranes Reef National Park (ARNP), Gulf of Mexico, Mexico (data from Ugalde et al. 2015). The execution of these codes took 13.5 min in RStudio Cloud (<https://rstudio.cloud/>) with the default resource settings (1 GB RAM, 1 CPU).

```
library(SSP)
data(sponges)

#Estimation of parameters
par.spo <- assempar(data=sponges, type="counts")

#Simulation of data
sim.spo <- simdata(Par=par.spo, cases=10, N=20, sites=20)

# Quality of simulated data
qua.spo  <-  datquality(data=sponges,  dat.sim=sim.spo,  Par=par.spo,  transforma-
tion="square root", method="bray")

#Sampling and estimation of MultSE
samp.spo <- sampsd(sim.spo, par.spo, transformation="square root",
          method="bray", n=20, m=20, k=10)

#Summarizing results
sum.spo <- summary_ssp(results=samp.spo, multi.site=TRUE)

#Identification of optimal effort

opt.spo <- ioptimum(xx=sum.spo, multi.site=TRUE)

#plot
fig.3 <- plot_ssp(xx=sum.spo, opt=opt.spo, multi.site=TRUE)
```

abundance of epibenthic organisms of 10 roots were described within each site, producing a total of 60 roots in each sector. One of the main objectives of this pilot study was to define the sampling effort needed to evaluate spatiotemporal patterns of variation in species composition among sectors considering the environmental gradient. To achieve this, it was necessary to identify a sampling effort (number of sites and roots) that guaranteed the highest precision at the lowest cost. For each sector, 20 data sets were simulated, each with 30 virtual sites and 30 virtual roots. A two-stage random sampling was then simulated using sites from 2 to 20 and roots from 2 to 20 with each combination repeated 10 times. The Bray–Curtis index was used as the similarity measure between sample units once a fourth root transformation of abundance had been applied. The R script for all operations is in Box 3.

Results of **SSP** indicated that a similar number of sites are required to obtain a consistent level of precision among sectors (*MultSE* value below 0.1). Considering the relatively low *MultSE* value among sites, only a small improvement in precision would be required at this spatial scale (suboptimal region). In contrast, the *MultSE* among roots was considerably high among all sectors (between 0.245 and 0.287). Consequently, at least eight roots per site would be required to reduce the *MultSE* by half and attain an optimal improvement in sampling effort (Fig. 4).

Discussion

The R package described in the present study expands the method developed by Anderson and Santana-Garcon (2015). It maintains its purpose as a quantitative tool to define sample size for studies of communities with data to be analyzed using distance-based methods. Results herein show that **SSP** improves the usefulness of the original procedure regarding the following aspects: 1) *MultSE* – *n* relationship could be evaluated beyond the original sampling size of the pilot study; 2) *MultSE* estimates are obtained by sampling over several simulated data sets, ensuring statistical independence of the estimations; 3) sampling effort can be defined in terms of suboptimal and optimal improvement regarding the highest *MultSE*, 4) this protocol can be used to estimate sample size in a wide variety of situations from simple (i.e. sampling a single or few sites) to more complex experimental designs (i.e. sampling several sites for different habitats). The latter constitutes an important advantage, since it offers new possibilities for planning complex sampling designs, as it has been advised for multi-scale studies in ecology (Underwood and Chapman 1998, Leibold et al. 2004, Chase et al. 2018).

The data sets used in this study comprised a variety of sampling designs, in all of which the **SSP** package provided

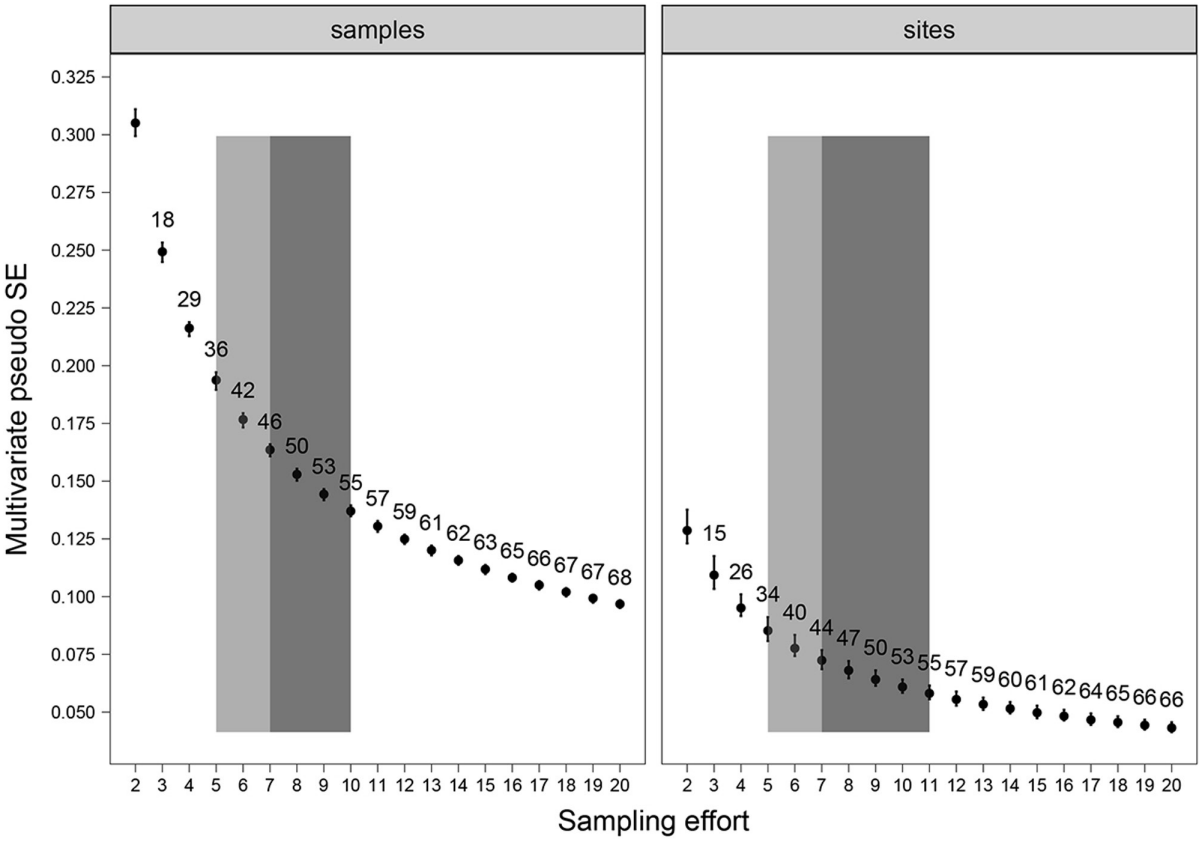

Figure 3. Relation between *MultSE* (in Bray–Curtis dissimilarity and square root transformation of species abundance) and sampling effort using sponge simulated data. Shaded areas indicate the range of samples in which an increase in sampling effort provided a suboptimal (light grey) and optimal improvement in precision (dark grey).

an appropriate and useful visualization tool to identify the optimal sampling effort. The **SSP** protocol applied to data from the simplest case in micromollusks of Cayo Nuevo showed that the sampling effort should be increased compared to the pilot study (Fig. 3). The sponge case brings the opportunity to evaluate the precision required considering a more complex situation, since the **SSP** protocol can estimate *MultSE* in two different spatial scales: transects and sites. The possibility of targeting sampling effort in this way constitutes a key advantage in experimental design, since it allows a revaluation of the number of sampling sites, hence the costs of the corresponding fieldwork. Similarly, the results of the simulation on the data from epibenthic fauna in mangrove roots demonstrated the potential of **SSP** to define sampling effort in studies of communities in heterogeneous environments. In the original research, using a very primitive version of **SSP**, the sampling effort was defined with four sites per sector and eight roots per site. This sampling effort was used by Guerra-Castro et al. (2016) and had the statistical power to detect differences along the gradient despite reduced sampling effort. Decisions such as changing the number or distribution of sampling effort are central since they directly impact the costs of a sampling project. An adequate combination of effort at different spatial scales will help optimize the allocation of resources that are often limited. Optimization of sampling designs using cost–benefit procedures described by previous authors (Underwood 1997) could easily be combined with the protocol presented here, further improving cost–effective allocations of sampling effort.

One aspect that requires theoretical development is the one referring to the behavior of the *MultSE* and its scale dependency. Even though the *MultSE* decreases as $\sim 1/\sqrt{n}$, its magnitude and interpretation are strongly associated with the dissimilarity index and transformation of species abundance used in each case. By standardizing to the maximum, it is possible to improve the level of precision to a desired level, but this does not solve the differences in scales. Therefore, it is not appropriate to make comparisons of precision levels between studies that use different dissimilarity coefficients. This implies that users must carefully consider the level of optimization required for any particular study case, while taking into account the dissimilarity coefficient of interest, the magnitude of multivariate dispersion, its meaning and the cost associated with each sampling unit.

It is important to mention that **SSP** is not free from assumptions. First, simulations in the present study do not consider environmental constraints, neither the co-occurrence of species nor their joint distribution functions; the procedure essentially assumes that any combination of species is possible. Therefore, we recommend avoiding

Box 3. SSP applied to data on epibenthic assemblages on Caribbean mangrove roots (data from Guerra-Castro et al. 2011). Considering the environmental differences between each sector, SSP was applied independently to each sector. Results were merged and analyzed simultaneously. The execution of these codes took 1.86 h in RStudio Cloud (<https://rstudio.cloud/>) with the default resource settings (1 GB RAM, 1 CPU).

```
library(SSP)
library(tidyr)
library(ggplot2)
library(dplyr)

data(pilot)
sectors <- levels(pilot$Sector)

#Defining arguments for simulation and sampling
N=30
sites=30
cases=20
n=20
m=20
k=10

#Lists to store results
sum.l <- opt.l <- qua.l <- vector(mode="list", length=3)

#Loop SSP at each sector
for (i in 1:length(sectors)){
  dat <- pilot[pilot$Sector==sectors[i],2:length(pilot)]

  #parameters for simulation
  par <- assempar(data=dat, type="cover", Sest.method="chao")

  # Simulation of data
  sim <- simdata(Par=par, cases=cases, N=N, sites=sites)

  # Quality of simulated data
  qua <- datquality(data=dat, dat.sim=sim, Par=par, transformation="fourth root",
method="bray")
  qua$sector <- rep(sectors[i], nrow(qua))
  qua.l[[i]] <- qua

  # Sampling and estimation of multse for each data set
  samp <- sampsd(dat.sim=sim, Par=par, transformation="fourth root",
          method="bray", n=n, m=m, k=k)

  # average of multse for each potential sampling design
  sum <- summary_ssp(samp, multi.site=TRUE)

  #Optimal sample sizes
  opt <- ioptimum(sum)
  opt <- as.data.frame(opt)
  opt$sv <- c("sites", "samples")
  opt <- pivot_longer(opt, cols=c("c1", "c2", "c3"), names_to="cut", values_to="effort")
  opt$sector <- rep(sectors[i], nrow(opt))
  opt.l[[i]] <- opt

  #arrangement to plot
  sum$sector <- rep(sectors[i], nrow(sum))
  sum.l[[i]] <- sum
```

```
  }

#combine summary into a data frame
sum.df <- do.call(rbind.data.frame, sum.l)
sum.df$sector <- factor(sum.df$sector, levels=c("E", "M", "I"))

#combine optimal sample sizes into a data frame
opt.df <- do.call(rbind.data.frame, opt.l)
opt.df$sector <- factor(opt.df$sector, levels=c("E", "M", "I"))

#Combine quality features into a data frame
qua.df <- do.call(rbind.data.frame, qua.l)

# Generation of plot
my_breaks <- function(x) {
  y <- seq(min(x), max(x), 1)
  y <- round(y,0)
}

#Definition of values for shade areas
shade.opt <- opt.df %>%
        group_by(sector, sv) %>%
        filter(cut != "c1") %>%
        summarise(xmin=min(effort), xmax=max(effort))

shade.sub <- opt.df %>%
        group_by(sector, sv) %>%
        filter(cut != "c3") %>%
        summarise(xmin=min(effort), xmax=max(effort))

#plot
fig.4 <- ggplot(sum.df, aes(x=samples, y=mean))+
     geom_point()+
     geom_errorbar(aes(ymin=lower, ymax=upper), width=.1)+
     facet_grid(sector~sv, scales="free_x")+
     theme_bw(base_size=16) +
     ylab ("Multivariate pseudo SE")+
     xlab("Sampling effort")+
     scale_y_continuous(breaks=seq(0.0, max(sum.df$upper), 0.025))+
     scale_x_continuous(breaks=my_breaks)+
     theme(axis.text.x=element_text(colour="black", size=rel(0.7)),
     axis.text.y=element_text(colour="black", size=rel(0.7)),
     axis.title.x=element_text(colour="black", size=rel(0.9)),
     axis.title.y=element_text(colour="black", size=rel(0.9)),
     panel.grid.major=element_blank(),
     panel.grid.minor=element_blank(),
     panel.border=element_rect(size=0.4),
     axis.ticks= element_line(size=0.2))+
     geom_rect(data=shade.opt, aes_(x=NULL,y=NULL,
               xmin=~xmin,    xmax=~xmax,    ymin=min(sum.df$lower),    ymax=max(sum.
df$upper)), alpha=0.5, fill="grey10")+
     geom_rect(data=shade.sub, aes_(x=NULL,y=NULL,
               xmin=~xmin,    xmax=~xmax,    ymin=min(sum.df$lower),    ymax=max(sum.
df$upper)), alpha=0.5, fill="grey50")+
     geom_text(aes_(x=~samples, y=~upper+0.01, label=~cum), na.rm=TRUE)
```

the combination of data from environmentally different habitats when using simulations with **SSP**. Instead, associations among species (i.e. co-occurrence or repulsion) considering differences between assemblages can be modelled using copulas (Anderson et al. 2019, Tang et al. 2019). The advantage of copulas is that it allows simulating species arrangements like those observed in the pilot data. The **SSP** protocol assumes that potential differences in species

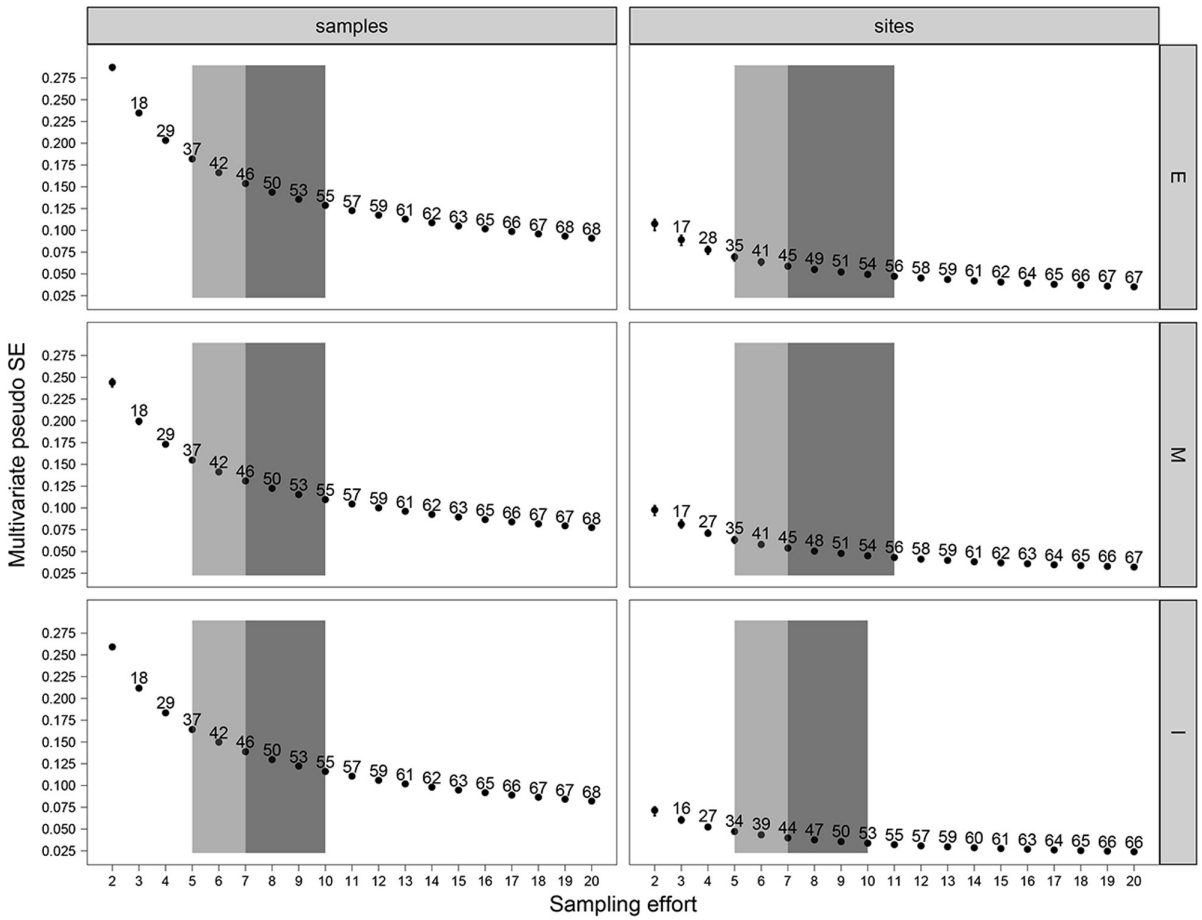

Figure 4. Relation between *MultSE* (in Bray–Curtis dissimilarity and fourth root transformation of abundances) and sampling effort using epibenthic assemblages on Caribbean mangrove roots. Shaded areas indicate the range of samples in which an increase in sampling effort provided a suboptimal (light grey) and optimal improvement in precision (dark grey).

composition among sites are due to spatial aggregation of species measured in the pilot data. Thus, any spatial structure of species that was not captured by the pilot data will not be reflected by **SSP**. The simulation of associations among species and the use of copulas are aspects to be considered in a future version of **SSP**.

Whilst the protocol performs well with small pilot data, the pilot sampling should not be restricted to a small number of samples or sites but should capture the greatest possible variability in the system under study. After evaluating the quality of the simulated data in the present study, it became clear that simulated data did not always have properties

Table 1. Output of **datquality** for the three examples: relevant features of original and simulated data. Features include the number of sample units (*n*), the number of sites (*m*), average number species/sample ($\bar{S}$), average Simpson diversity index (aSDI) and range of multivariate dispersion (MVD).

		n	*m*	$\bar{S}$	aSDI	MVD
Micromolusk from Cayo Nuevo, Gulf of Mexico						
	Pilot	6	1	25.3 (±7.2)	–	0.25
	Simulated	1000	1	30.5 (±3.9)	–	0.22–0.24
Sponges from Alacranes Reef, Gulf of Mexico						
	Pilot	4–8	6	12.4 (±4.3)	0.84 (±0.08)	0.16
	Simulated	20	20	9.1 (±2.4)	0.72 (±0.13)	0.22–0.23
Epibionts on Caribbean mangrove roots						
External sector	Pilot	60	6	18.3 (±5.7)	0.82 (±0.07)	0.19
Intermediate sector	Pilot	60	6	16.1 (±5.25)	0.80 (±0.10)	0.18
	Simulated	30	30	16.2 (±2.9)	0.93 (±0.01)	0.14–0.15
Internal sector	Pilot	60	6	14.3 (±2.6)	0.83 (±0.07)	0.14
	Simulated	30	30	14.8 (±2.7)	0.86 (±0.05)	0.14–0.15

identical to those in the pilot data (Table 1). In Example 1, the mean number of species per sample was 20% higher than the pilot data. This could be a consequence of the limited number of samples used to estimate the probability of occurrence of each species ($n = 6$, only six possible values of f, the lowest equal to 1/6). In Example 2 (with n between 4 and 8 transects at each of six sites), the mean number of species per sample was 26% lower than the pilot data. Unlike the first example, the MVD of the sponge simulations exceeded 38% the original dispersion. By contrast, the simulated data of Example 3 closely resembled the corresponding pilot data, probably because the pilot data was extensive.

Despite these limitations, if the properties of the simulated data resemble the community of interest and show ecological plausibility, the extrapolations derived from the procedure presented here will hold valid to define the sampling size of any study based on dissimilarity-based multivariate analysis. Overall, this procedure simulates data that satisfies the key features on which inferences are to be made, thereby allowing for independent and multiple estimations of multivariate standard errors to be drawn from simulated data matrixes and the unbiased construction of a *MultSE* – *n* relationship. Finally, the versatility of **SSP** can be used to the advantage of researchers without R background as a powerful decision-making tool to define adequate sampling effort by using the online app.

Software availability

SSP ver. 1 is free and open source, distributed under GNU Public License ver. 2 (GPL-2). This package is available on the comprehensive R archive network (CRAN) <https://cran.r-project.org/web/packages/SSP/index.html> and is also hosted in GitHub <https://github.com/edlinguerra/SSP>. The online version can be accessed at <https://edlin.shinyapps.io/ssp_web/>. Data from all three examples are available in the package. To cite **SSP** or acknowledge its use, cite this software note including the appropriate version number.

Data availability statement

Data available from the Dryad Digital Repository: <http://dx.doi.org/10.5061/dryad.3bk3j9kj5> (Guerra-Castro et al. 2020).

Acknowledgements – Special thanks to M. J. Anderson and several anonymous reviewers for providing constructive critical comments that substantially improved an early version of this manuscript. We extend our gratitude to Diana Ugalde, Lilian Hernández, Deneb Ortigosa, Cesar Herrera, Luis Montilla, Edgar Torres and Jorge Montero for their valuable comments on the preliminary versions of **SSP**.

Funding – EGC was supported by a DGAPA Post-doctoral Fellowship at the Univ. Nacional Autónoma de México (UNAM). This research was financed by the project PE207416 (PAPIME, UNAM) under the supervision of MM. Funding was also provided by project #325 (CÁTEDRAS, CONACYT), the HARTE Research Inst. for Gulf of Mexico Studies and the Harte Charitable Foundation. NS holds the Furgason Fellowship International Chair for Coastal and Marine Studies in Mexico. The publication fee was covered by Escuela Nacional de Estudios Superiores, Unidad Mérida, UNAM.

Author contributions

Edlin Guerra-Castro: Conceptualization (equal); Data curation (equal); Formal analysis (equal); Funding acquisition (equal); Investigation (equal); Methodology (equal); Software (equal); Visualization (equal); Writing – original draft (equal). **Juan Carlos Cajas**: Formal analysis (equal); Funding acquisition (equal); Methodology (equal); Software (equal); Writing – review and editing (equal). **Nuno Simoes**: Conceptualization (equal); Data curation (equal); Funding acquisition (equal); Validation (equal); Writing – review and editing (equal). **Juan Jose Cruz-Motta**: Conceptualization (equal); Data curation (equal); Methodology (equal); Validation (equal); Writing – review and editing (equal). **Maite Mascaro**: Conceptualization (equal); Funding acquisition (equal); Investigation (equal); Methodology (equal); Project administration (equal); Supervision (equal); Writing – original draft (equal).

References

Anderson, M. J. 2006. Distance-based tests for homogeneity of multivariate dispersions. – Biometrics 62: 245–253.

Anderson, M. J. 2017. Permutational multivariate analysis of variance (PERMANOVA). – Wiley StatsRef: Statistics Reference Online. Wiley.

Anderson, M. J. and Santana-Garcon, J. 2015. Measures of precision for dissimilarity-based multivariate analysis of ecological communities. – Ecol. Lett. 18: 66–73.

Anderson, M. J. and Walsh, D. C. I. 2013. PERMANOVA, ANOSIM and the Mantel test in the face of heterogeneous dispersions: what null hypothesis are you testing? – Ecol. Monogr. 83: 557–574.

Anderson, M. J. et al. 2006. Multivariate dispersion as a measure of beta diversity. – Ecol. Lett. 9: 683–693.

Anderson, M. J. et al. 2019. A pathway for multivariate analysis of ecological communities using copulas. – Ecol. Evol. 9: 3276–3294.

Blanchet, F. G. et al. 2016. A new cost–effective approach to survey ecological communities. – Oikos 125: 975–987.

Clarke, K. R. 1993. Non-parametric multivariate analyses of changes in community structure. – Aust. J. Ecol. 18: 117–143.

Clarke, K. R. and Green, R. H. 1988. Statistical design and analysis for a 'biological effects' study. – Mar. Ecol. Prog. Ser. 46: 213–226.

Clarke, K. R. et al. 2006. Dispersion-based weighting of species counts in assemblage analyses. – Mar. Ecol. Prog. Ser. 320: 11–27.

Chase, J. et al. 2018. Embracing scale-dependence to achieve a deeper understanding of biodiversity and its change across communities. – Ecol. Lett. 21: 1737–1751.

Gaston, K. 1994. Rarity. – Chapman and Hall.

Green, R. H. 1979. Sampling design and statistical methods for environmental biologists. – Wiley.
Guerra-Castro, E. et al. 2011. Cuantificación de la diversidad de especies incrustantes asociadas a las raíces de Rhizophora mangle L. en el Parque Nacional Laguna de La Restinga. – Interciencia 36: 923–930.
Guerra-Castro, E. J. et al. 2016. Scales of spatial variation in tropical benthic assemblages and their ecological relevance: epibionts on Caribbean mangrove roots as a model system. – Mar. Ecol. Prog. Ser. 548: 97–110.
Guerra-Castro, E. J. et al. 2020. Data from: SSP: an R package to estimate sampling effort in studies of ecological communities. – Dryad Digital Repository, <http://dx.doi.org/10.5061/dryad.3bk3j9kj5>.
Legendre, P. and De Cáceres, M. 2013. Beta diversity as the variance of community data: dissimilarity coefficients and partitioning. – Ecol. Lett. 16: 951–963.
Leibold, M. A. et al. 2004. The metacommunity concept: a framework for multi-scale community ecology. – Ecol. Lett. 7: 601–613.
Maccherini, S. et al. 2020. Enough is enough? Searching for the optimal sample size to monitor European habitats: a case study from coastal sand dunes. – Diversity 12: 138.
Magurran, A. E. and Henderson, P. A. 2011. Commonness and rarity. – In: Magurran, A. E. and McGill, B. J. (eds), Biological diversity: frontiers in measurement and assessment. Oxford Univ. Press, pp. 97–104.
Mapstone, B. D. 1995. Scalable decision rules for environmental impact studies: effect size, type I and type II errors. – Ecol. Appl. 5: 401–410.
McArdle, B. H. and Anderson, M. J. 2004. Variance heterogeneity, transformations and models of species abundance: a cautionary tale. – Can. J. Fish. Aquat. Sci. 61: 1294–1302.
Montilla, L. M. et al. 2020. The use of pseudo-multivariate standard error to improve the sampling design of coral monitoring programs. – PeerJ 8: e8429.
Oksanen, J. et al. 2015. Vegan: community ecology package. – R package ver. 2.3-0, <https://cran.r-project.org/web/packages/vegan/index.html>.
Ortigosa, D. et al. 2018. First survey of interstitial molluscs from Cayo Nuevo, Campeche Bank, Gulf of Mexico. – ZooKeys 779: 1–17.
Quinn, G. P. and Keough, M. J. 2002. Experimental design and data analysis for biologists. – Cambridge Univ. Press.
Tang, Y. et al. 2019. Copula-based semiparametric models for spatio-temporal data. – Biometrics 75: 1156–1167.
Tillé, Y. 2011. Sampling algorithms. – In: Lovric, M. (ed.), International encyclopedia of statistical science. Springer, pp. 1273–1274.
Ugalde, D. et al. 2015. Marine sponges (Porifera: Demospongiae) from the Gulf of México, new records and redescription of *Erylus trisphaerus* (de Laubenfels, 1953). – Zootaxa 3911: 151–183.
Underwood, A. J. 1997. Experiments in ecology: their logical design and interpretation using analysis of variance. – Cambridge Univ. Press.
Underwood, A. J. and Chapman, M. G. 1998. A method for analysing spatial scales of variation in composition of assemblages. – Oecologia 117: 570–578.
Underwood, A. J. and Chapman, M. G. 2003. Power, precaution, Type II error and sampling design in assessment of environmental impacts. – J. Exp. Mar. Biol. Ecol. 296: 49–70.