The following supplements accompany the article

Species surrogacy in environmental impact assessment and monitoring: extending the BestAgg approach to asymmetrical designs

Stanislao Bevilacqua*, Antonio Terlizzi

*Corresponding author: stanislao.bevilacqua@unisalento.it

Marine Ecology Progress Series 547: 19–32 (2016)

Supplementary Information

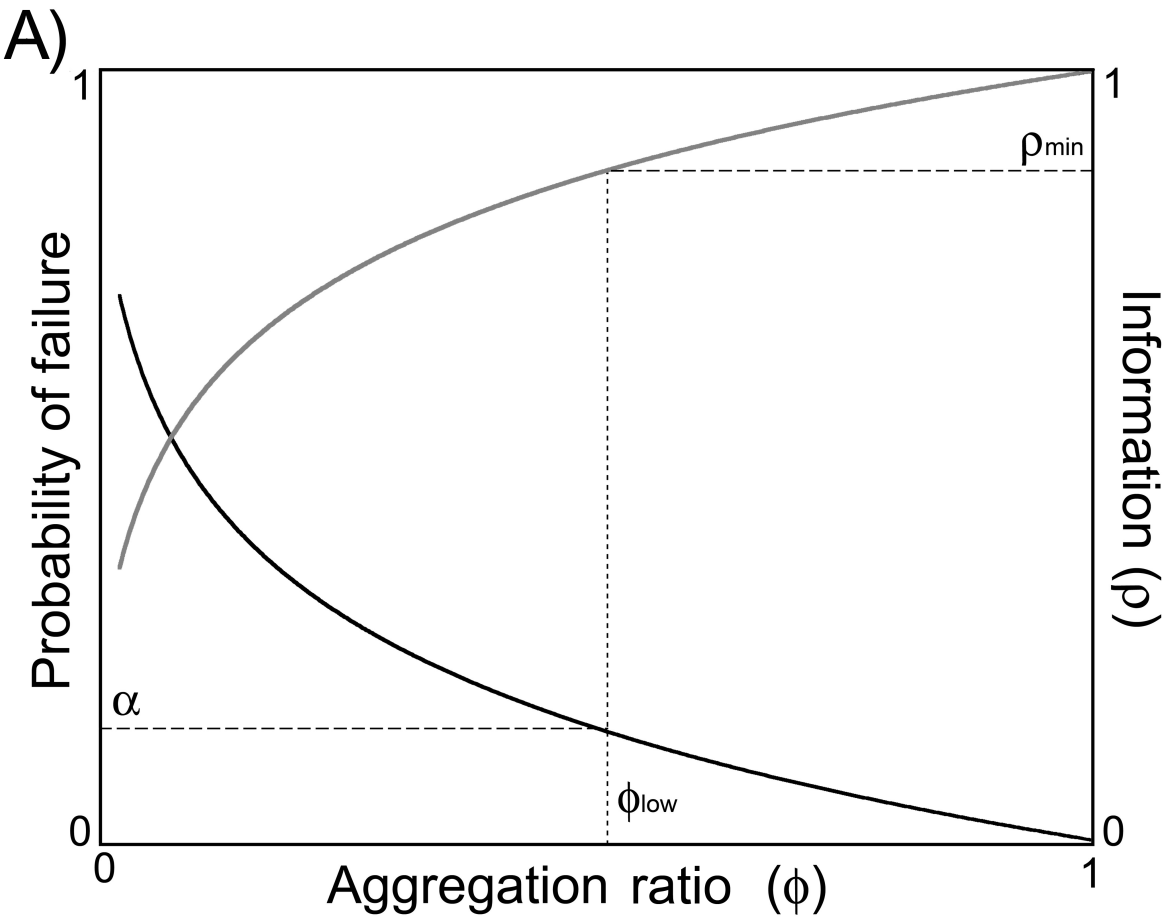
Additional material complementing the article "*Species surrogacy in environmental impact assessment and monitoring: extending the BestAgg approach to asymmetrical designs*" by S. Bevilacqua and A. Terlizzi is provided in this section. Supplementary Information consists of:

Supplement 1
Figure S1. Summary of the theoretical framework underlying the BestAgg approach.
Table S1. Taxonomic list of taxa recorded;
Table S2. Selected surrogates for BestAgg;
Table S3. Results of SIMPER analysis;
Table S4. Summary of PERMANOVA tests on simulated aggregations.

Supplement 2
R code for BestAgg analyses with asymmetrical designs (2 nested factors) and user guide
Sample File 1. Example data from Time 1 (T1) (as separate .csv file);
Sample File 2. Factors for example data (as separate .csv file);

Supplement 1.

Figure S1. (A) Theoretical model underlying the BestAgg approach. As the original species (or taxa, etc.) in a multivariate data matrix are gradually aggregated (i.e. grouped and summed) into a decreasing number of new variables (surrogates), ϕ (i.e. the ratio of the number of groups to the number of original variables) decreases, and the information (here expressed as the Spearman's correlation ρ between the original and the aggregated matrix) on species-level patterns is progressively lost (dark grey curve). As a consequence of this loss of information, the ability of surrogates to reflect multivariate responses as at species level will decrease. In other words, at decreasing ϕ (which means an increasing packaging of the original variables) the probability of surrogates to fail in detecting multivariate patterns as at species level will progressively increase (black curve). There will be, therefore, a given value of ϕ, namely ϕ_{low}, at which the information will achieve the minimum values (ρ_{min}) below which the probability of surrogates to fail will be higher than a fixed level of significance (α). This threshold of ϕ_{low} represents the *lowest practicable aggregation* of the original variables, and fixes G_{min}, that is the *minimum* number of surrogates sufficient to detect species-level patterns consistently (see text).
(B) Conceptual framework for surrogate selection in BestAgg. Once ϕ_{low}, and therefore the minimum number of surrogates (G_{min}) is fixed, surrogate selection in BestAgg aims to aggregate species (or taxa, groups, etc.) into surrogates that maximize ecological information (see text). In summary, one or more species may be selected to form a surrogate following the logic of three unifying macrocriteria: *relevance* (general, context-specific, and/or study-specific ecological importance), *easiness* (the distinctiveness of a given species, taxon, or group of organisms leading to be easily identified from a taxonomic, morphological, or functional point of view), and *resemblance* (shared characteristics among organisms, from common ancestry to functional similarity that allow meaningful groupings). High priority for selection should be given to relevant species that are also easy to identify (REL+E). Relevant species whose identification is difficult should be aggregated, if possible, in easy-to-identify but still relevant surrogates (REL+E→RES), and intermediate priority should be given to these surrogates, because their easiness is achieved through resemblance. Finally, not relevant species have low priority and may be grouped to form surrogates following any appropriate aggregation criterion to facilitate their identification (E→RES).


A)
1
1
ρmin
Probability of failure
Information (ρ)
α
ϕlow
0
0
0
1
Aggregation ratio (ϕ)

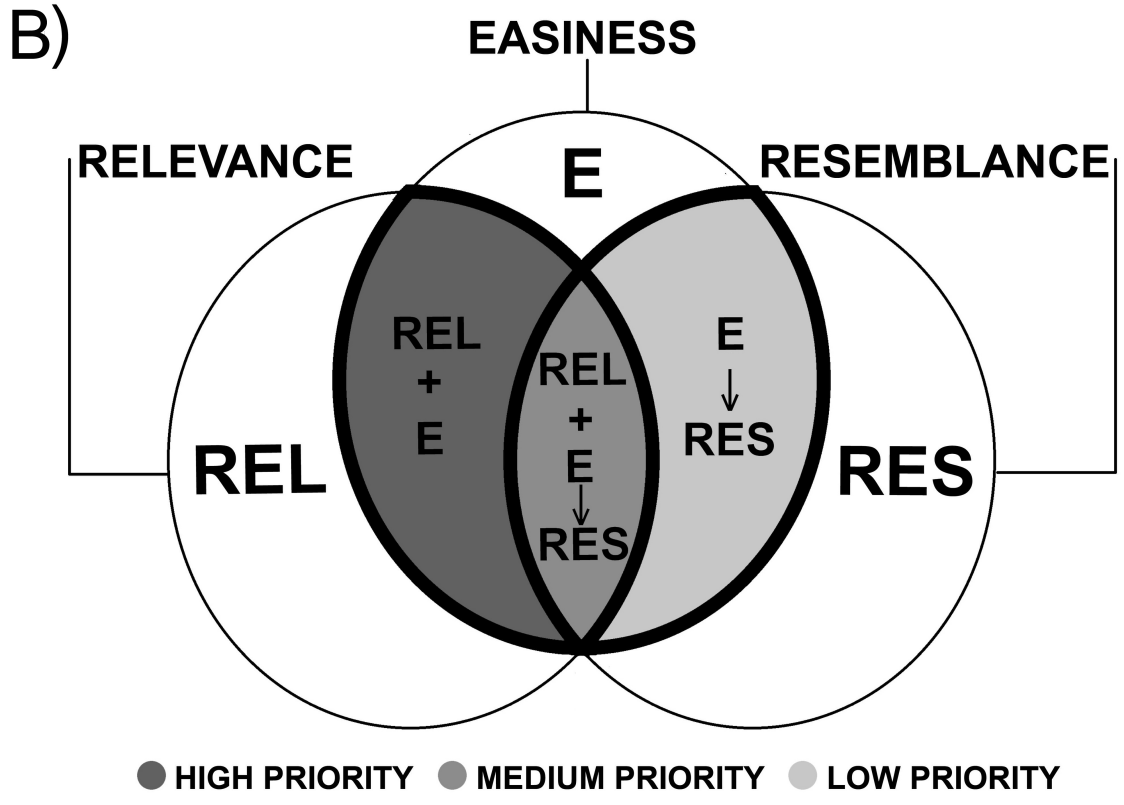

B)
EASINESS
RELEVANCE
E
RESEMBLANCE
REL + E
REL + E ↓ RES
E ↓ RES
REL
RES
HIGH PRIORITY
MEDIUM PRIORITY
LOW PRIORITY

Table S1. Taxonomic list of taxa recorded.

Species	Genus	Family	Order	Class	Phylum
Aiptasia mutabilis	*Aiptasia*	Aiptasiidae	Actiniaria	Anthozoa	Cnidaria
Amphiroa spp.	*Amphiroa*	Corallinaceae	Corallinales	Florideophyceae	Rhodophyta
Ascidia mentula	*Ascidia*	Ascidiidae	Phlebobranchia	Ascidiacea	Tunicata
Ascidia sp.	*Ascidia*	Ascidiidae	Phlebobranchia	Ascidiacea	Tunicata
Balanophyllia europaea	*Balanophyllia*	Dendrophylliidae	Scleractinia	Anthozoa	Cnidaria
Balanus perforatus	*Balanus*	Balanidae	Cirripedia	Maxillopoda	Crustacea
Cereus pedunculatus	*Cereus*	Sagartiidae	Actiniaria	Anthozoa	Cnidaria
Chondrilla nucula	*Chondrilla*	Chondrillidae	Chondrosida	Demospongiae	Porifera
Chondrosia reniformis	*Chondrosia*	Chondrillidae	Chondrosida	Demospongiae	Porifera
Ciona intestinalis	*Ciona*	Cionidae	Phlebobranchia	Ascidiacea	Tunicata
Cladocora caespitosa	*Cladocora*	Caryophylliidae	Scleractinia	Anthozoa	Cnidaria
Cliona celata	*Cliona*	Clionaidae	Hadromerida	Demospongiae	Porifera
Cliona viridis	*Cliona*	Clionaidae	Hadromerida	Demospongiae	Porifera
Codium bursa	*Codium*	Codiaceae	Bryopsidales	Ulvophyceae	Chlorophyta
Codium coralloides	*Codium*	Codiaceae	Bryopsidales	Ulvophyceae	Chlorophyta
Codium effusum	*Codium*	Codiaceae	Bryopsidales	Ulvophyceae	Chlorophyta
Codium vermilara	*Codium*	Codiaceae	Bryopsidales	Ulvophyceae	Chlorophyta
Colpomenia sinuosa	*Colpomenia*	Scytosiphonaceae	Ectocarpales	Phaeophyceae	Heterokontophyta
Corallina officinalis	*Corallina*	Corallinaceae	Corallinales	Florideophyceae	Rhodophyta
Crambe crambe	*Crambe*	Crambeidae	Poecilosclerida	Demospongiae	Porifera
Cutleria adspersa	*Cutleria*	Cutleriaceae	Cutleriales	Phaeophyceae	Heterokontophyta
Cystodytes dellechiajei	*Cystodytes*	Polycitoridae	Aplousobranchia	Ascidiacea	Tunicata
Dictyota spp.	*Dictyota*	Dictyotaceae	Dictyotales	Phaeophyceae	Heterokontophyta
Didemnum spp.	*Didemnum*	Didemnidae	Aplousobranchia	Ascidiacea	Tunicata
Diplosoma listerianum	*Diplosoma*	Didemnidae	Aplousobranchia	Ascidiacea	Tunicata
Encrusting Bryozoans	Encrusting Bryozoans	Encrusting Bryozoans	Encrusting Bryozoans	Encrusting Bryozoans	Bryozoa
Encrusting Coralline Algae	Encrusting Coralline Algae	Corallinaceae	Corallinales	Florideophyceae	Rhodophyta
Filamentous Algae	Filamentous Algae	Filamentous Algae	Filamentous Algae	Filamentous Algae	Filamentous Algae
Filamentous Green Algae	Filamentous Green Algae	Filamentous Green Algae	Filamentous Green Algae	Ulvophyceae	Chlorophyta
Flabellia petiolata	*Flabellia*	Udoteaceae	Bryopsidales	Ulvophyceae	Chlorophyta
Halimeda tuna	*Halimeda*	Halimedaceae	Bryopsidales	Ulvophyceae	Chlorophyta
Halocynthia papillosa	*Halocynthia*	Pyuridae	Stolidobranchia	Ascidiacea	Tunicata
Hydrozoa	Hydrozoa	Hydrozoa	Hydrozoa	Hydrozoa	Cnidaria
Ircinia variabilis	*Ircinia*	Irciniidae	Dictyoceratida	Demospongiae	Porifera
Jania rubens	*Jania*	Corallinaceae	Corallinales	Florideophyceae	Rhodophyta
Laurencia complex	*Laurencia* complex	Rhodomelaceae	Ceramiales	Florideophyceae	Rhodophyta
Microcosmus sabatieri	*Microcosmus*	Pyuridae	Stolidobranchia	Ascidiacea	Tunicata

Padina pavonica	*Padina*	Dictyotaceae	Dictyotales	Phaeophyceae	Heterokontophyta
Palmophyllum crassum	*Palmophyllum*	Palmophyllaceae	Palmophyllales	Chlorophyta i.s.	Chlorophyta
Petrosia ficiformis	*Petrosia*	Petrosiidae	Haplosclerida	Demospongiae	Porifera
Peyssonnelia spp.	*Peyssonnelia*	Peyssonneliaceae	Peyssonneliales	Florideophyceae	Rhodophyta
Phorbas fictitius	*Phorbas*	Hymedesmiidae	Poecilosclerida	Demospongiae	Porifera
Rocellaria dubia	*Rocellaria*	Gastrochaenidae	Gastrochaenoidea	Bivalvia	Mollusca
Sarcotragus spinosulus	*Sarcotragus*	Irciniidae	Dictyoceratida	Demospongiae	Porifera
Sargassum vulgare	*Sargassum*	Sargassaceae	Fucales	Phaeophyceae	Heterokontophyta
Serpulids	Serpulids	Serpulidae	Sabellida	Polychaeta	Annelida
Soft Branched Algae	Soft Branched Algae	Soft Branched Algae	Soft Branched Algae	Soft Branched Algae	Soft Branched Algae
Sphaerococcus coronopifolius	*Sphaerococcus*	Sphaerococcaceae	Gigartinales	Florideophyceae	Rhodophyta
Stypocaulon scoparium	*Stypocaulon*	Stypocaulaceae	Sphacelariales	Phaeophyceae	Heterokontophyta
Thin tubular sheet-like Algae	Thin tubular sheet-like Algae	Thin tubular sheet-like Algae	Thin tubular sheet-like Algae	Thin tubular sheet-like Algae	Rhodophyta
Tricleocarpa fragilis	*Tricleocarpa*	Galaxauraceae	Nemaliales	Florideophyceae	Rhodophyta
Valonia macrophysa	*Valonia*	Valoniaceae	Siphonocladales	Siphonocladophyceae	Chlorophyta
Wrangelia penicillata	*Wrangelia*	Wrangeliaceae	Ceramiales	Florideophyceae	Rhodophyta

Table S2. Selected surrogates for BestAgg based on pilot data. *Relevance* is reported according to evidence from literature and SIMPER analysis (see Method section, see also Table S3). For *Easiness*: E = easy identification, D = difficult identification. In the *Resemblance* column are reported aggregation criteria if applied (NA = not applied). Priority has been assigned following the procedure described in Bevilacqua et al. 2013 (see also Figure S1B). Numbers in brackets are the number of species included in taxa and surrogates.

Phylum	Species (or taxon, group)	Relevance	Easiness	Resemblance	Priority	BestAgg surrogate
Algae (27)	*Jania rubens*	SIMPER	D			
	Corallina officinalis	SIMPER	D	Difficult to distinguish among erect coralline species, aggregated in: Articulated Corallines	Low	Articulated Corallines (3)
	Amphiroa spp.	Not relevant	E			
	Stypocaulon scoparium	Not relevant	D			
	Wrangelia penicillata	Not relevant	D	Some species could be difficult to identify especially in intricate turf complex, aggregated in: Turf-forming Algae	Low	Turf-forming Algae (3)
	Tricleocarpa fragilis	Not relevant	E			
	Colpomenia sinuosa	Indicator (environmental stress)[3,4]	E	NA	High	*Colpomenia sinuosa* (1)
	Codium bursa	Not relevant	E	NA	Low	*Codium bursa* (1)

Codium coralloides	Not relevant	D			
Codium effusum	Not relevant	D	Difficult to distinguish at species level, except for some easy-to-identify species (e.g. *Codium bursa*), aggregated in Unbranched Green Algae	Low	Unbranched Green Algae (4)
Codium vermilara	Not relevant	D			
Palmophyllum crassum	Not relevant	D	Potential confusion with *Codium* spp., aggregated in Unbranched Green Algae		
Valonia macrophysa	Not relevant	D	Difficult to distinguish from congeneric, aggregated in: *Valonia* spp.	Low	*Valonia* spp. (1)
Filamentous Algae	SIMPER, Indicator[1]	E	NA	High	Filamentous Algae (1)
Dictyota spp.	SIMPER	E	NA	High	*Dictyota* spp. (1)
Encrusting Coralline Algae	SIMPER	E	NA	High	Encrusting Coralline Algae (1)
Filamentous Green Algae	SIMPER, Indicator[1]	E	NA	High	Filamentous Green Algae (1)
Halimeda tuna	SIMPER	E	NA	High	*Halimeda tuna* (1)

	Laurencia complex	SIMPER, ecological role as canopy-forming algae	E	NA	High	*Laurencia* complex (1)
	Sargassum vulgare	Ecological role as canopy-forming algae	D	Possible confusion with other canopy algae, aggregated in: Canopy-forming Algae	Medium	Canopy-forming Algae (2)
	Sphaerococcus coronopifolius	SIMPER, Ecological role as canopy-forming algae	D			
	Peyssonnelia spp.	SIMPER	E	NA	High	*Peyssonnelia* spp. (1)
	Cutleria adspersa	Not relevant	D	Mostly not relevant, similar morphology	Low	Coarsely branched Algae (4)
	Flabellia petiolata	SIMPER	E			
	Padina pavonica	Not relevant	E			
	Soft Branched Algae	Not relevant	E			
	Thin tubular sheet-like Algae	Not relevant	E	NA	Low	Thin tubular sheet-like Algae (1)
Annelida (1)	Serpulids	Indicator (harbor fouling)	E	NA	High	Serpulids (1)

Arthropoda (1)	*Balanus perforatus*	Not relevant	D	Difficult to distinguish at species level, aggregated in: Barnacles	Low	Barnacles (1)
Bryozoa (1)	Encrusting Bryozoans	Not relevant	E	NA	Low	Encrusting Bryozoans (1)
Cnidaria (5)	*Aiptasia mutabilis*	Not relevant	E			
	Balanophyllia europaea	Not relevant	E	Not relevant, possible confusion with other anthozoan species, aggregated in: Solitary anthozoans	Low	Solitary Anthozoans (3)
	Cereus pedunculatus	Not relevant	D			
	Cladocora caespitosa	Ecological role as biocostructor, madreporarian	E	NA	High	*Cladocora caespitosa* (1)
	Hydrozoa	SIMPER, Indicator (harbor fouling)	E	NA	High	Hydrozoa (1)
Mollusca (1)	*Rocellaria dubia*	Not relevant	D	Difficult to distinguish from other species, aggregated in: Boring Bivalves	Low	Boring Bivalves (1)
Porifera (9)	*Chondrilla nucula*	Not relevant	E		Low	*Chondrilla nucula* (1)
	Chondrosia reniformis	Not relevant	D	Potential confusion among some species, some species could be difficult to identify, ecological role as main suspension-filter	Low	Massive Sponges (4)

				feeders, aggregated in: Massive Sponges		
	Ircinia variabilis	Not relevant	D			
	Sarcotragus spinosulus	Not relevant	D			
	Petrosia ficiformis	Not relevant	E			
	Cliona celata	Not relevant	D	Difficult to distinguish at species level, aggregated in: Boring Sponges	Low	Boring Sponges (2)
	Cliona viridis	Not relevant	D			
	Crambe crambe	Not relevant	D	Potential confusion among species, and with other species potentially present (e.g. *Spirastrella* sp.) aggregated in Encrusting Sponges	Low	Encrusting Sponges (2)
	Phorbas fictitius	Not relevant	D			
Tunicata (8)	*Ascidia mentula*	Indicator[2] (harbor fouling, environmental stress)	D			
	Ascidia sp.	Not relevant	D	Difficult to distinguish at species level, aggregated in: Solitary ascidians	Medium	Solitary Ascidians (3)
	Ciona intestinalis	Indicator[2] (harbor fouling, environmental stress)	D			

Cystodites dellechiaje	Not relevant	D	Difficult to distinguish at species level, aggregated in: Colonial Ascidians	Low	Colonial Ascidians (2)
Diplosoma listerianum	Not relevant	D			
Didemnum spp.	Not relevant	E	NA	Low	*Didemnum* spp. (1)
Halocynthia papillosa	Potential sensitive[2]	E	NA	Low	*Halocynthia papillosa* (1)
Microcosmus sabatieri	Not relevant	D	Difficult to distinguish at species level, aggregated in: *Microcosmus* spp.	Low	*Microcosmus* spp. (1)

[1]Gray, J.S., 1992. Eutrophication in the Sea. In *Marine Eutrophication and Population Dynamics*, G. Colombo, I. Ferrari, V.U. Ceccherelli, and R. Rossi, pp. 3-15, Olsen & Olsen, Fredensborg, Denmark.

[2]Naranjo, S.A., Carballo, J.C., García-Gómez, J.C., 1996. Effects of environmental stress on ascidian populations in Algericas Bay (southern Spain). Possible marine bioindicators. *Marine Ecology Progress Series*, 144, 119-131.

[3]Chryssovergis, F., Panayotidis, P., 1995. Évolution des peuplements macrophytobentiques le long d'un gradient d'eutrophisation (Golfe de Maliakos, Mer Égée, Grèce). *Oceanologica Acta*, 18, 649-658.

[4]Terlizzi, A., Benedetti-Cecchi, L., Bevilacqua, S., Fraschetti, S., Guidetti, P., Anderson, MJ., 2005. Multivariate and univariate asymmetrical analyses in environmental impact assessment: a case study of Mediterranean subtidal sessile assemblages. *Marine Ecology Progress Series*, 289, 27-42.

Table S3. Results of SIMPER analysis separated for each time of sampling reporting the % contribution of species (or taxa, groups) to assemblage dissimilarities between the impacted (*I*) and Control (*Cs*) locations. Only species with contributions higher than 3% were reported.

Time 1 (T1)

Taxon	Average abundance *Cs*	Average abundance *I*	Contrib. %	Cum. %
Encrusting Coralline Algae	18.3	14.6	18.7	18.7
Filamentous Algae	6.2	17.5	16.35	35.0
Halimeda tuna	15.4	5.8	16.1	51.1
Filamentous Green Algae	8.9	6.9	9.8	60.9
Peyssonnelia spp.	3.6	4.2	6.2	67.1
Dictyota spp.	1.6	4.5	4.9	72.0
Flabellia petiolata	0.6	3.6	4.6	76.6
Laurencia complex	3.3	0.1	3.9	80.5

Average dissimilarity *I*-vs-*Cs* = 64.2

Time 2 (T2)

Taxon	Average abundance *Cs*	Average abundance *I*	Contrib. %	Cum. %
Encrusting Coralline Algae	24.8	13.6	19.7	19.7
Peyssonnelia spp.	3.6	13.9	14.2	33.9
Filamentous Green Algae	6.1	8.7	8.6	42.5
Dictyota spp.	0.5	6.3	7.1	49.6
Halimeda tuna	4.3	4.4	5.9	55.5
Filamentous Algae	5.9	3.5	5.8	61.2
Hydrozoa	4.4	3.2	3.7	65.9
Jania rubens	1.5	4.0	3.9	69.8
Sphaerococcus coronopifolius	0.2	3.0	3.5	73.3
Corallina officinalis	0.2	3.1	3.4	76.7
Laurencia complex	1.7	0.4	3.3	80.0

Average dissimilarity *I*-vs-*Cs* = 66.3

Time 3 (T3)

Taxon	Average abundance *Cs*	Average abundance *I*	Contrib. %	Cum. %
Encrusting Coralline Algae	25.7	18.5	17.3	17.3
Halimeda tuna	13.9	10.2	13.5	30.8
Filamentous Green Algae	12.9	3.4	9.7	40.5
Filamentous Algae	3.8	12.6	9.3	49.8
Peyssonnelia spp.	4.0	9.1	7.9	57.6
Sphaerococcus coronopifolius	0.2	8.7	7.8	65.4
Corallina officinalis	0.1	5.9	5.3	70.7
Flabellia petiolata	1.1	4.0	4.1	74.8
Hydrozoa	4.5	1.1	3.8	78.5
Laurencia complex	3.6	0.0	3.1	81.7

Average dissimilarity *I*-vs-*Cs* = 68.7

Table S4. Percentage of tests (n = 1,000) consistent from with those from species-level analyses based on null models at decreasing levels of aggregation (ϕ). The corresponding number of surrogates (G) is also provided. The lowest practicable aggregation ϕ_{low} (% of tests ≥ 95%) and the corresponding minimum number of surrogates G_{min} for T1 and T2 are given in bold (T3 not analysed due to the lack of significant effects of the impact in this sampling time). The overall sufficient ϕ_{low} (and G_{min}) is underlined.

Number of surrogates (G)	Aggregation ratio (ϕ)	% of significant tests for *I*-vs-*Cs*	
		T1	T2
50	0.94	100%	100%
47	0.89	99%	99%
44	0.83	99%	99%
41	0.77	97%	98%
38	0.72	96%	99%
35	0.66	97%	96%
32	**0.60**	**95%**	97%
29	0.55	93%	97%
26	0.49	92%	97%
23	**0.43**	90%	**95%**
20	0.38	86%	93%
17	0.32	84%	90%
14	0.26	79%	87%
11	0.21	71%	85%
8	0.15	62%	77%
5	0.09	50%	66%
2	0.04	33%	46%

Supplement 2.
R code for BestAgg analyses with asymmetrical designs (2 NESTED factors) and user guide.

This R code applies to asymmetrical experimental designs involving **2 NESTED factors**. Specifically, it has been implemented to extend the BestAgg approach (see Bevilacqua et al. 2013) to the analysis of **asymmetrical designs**. For example, the R code allows analyzing a typical After/Control-Impact (see Glasby, 1997) asymmetrical design in which there are multiple Locations (L) with a single impacted location (*I*) and multiple control locations (*Cs*) and multiple Sites nested in Location [S(L)]. As the appropriate denominator term for the main test of interest in asymmetrical designs may vary (see Terlizzi et al. 2005 for details), the appropriate code (see Step 2.3. below) should be used accordingly. When the design involves multiple time of sampling, therefore including also the factor Time (T), the analysis to obtain G_{min} can be carried out separately for each time of sampling. Then, the most conservative G_{min} (i.e. the highest G_{min} obtained among those derived from each time of sampling) should be adopted.
The R code is articulated in 2 main procedures:

1. *Creating the database format from Species × Sample matrix*
   Implementation of a database, which is necessary to perform analyses in procedure 2, starting from the original species × sample data matrix.
2. *Calculating simulated correlation and PERMANOVA results for each group of random surrogates*
   This procedure allows identifying the minimum number of surrogate groups G_{min} sufficient to obtain results consistent with those obtained at species level, quantifying the information retained in the BestAgg aggregated matrix, and defining the probability of Type I error when using $G_{BestAgg}$ as the effective number of surrogates (see Methods for further details).

Note that **procedure 2 requires necessarily the database format provided by procedure 1**. Note also that **procedure 2 is provided only for asymmetrical designs with 2 nested factors,** whereas more complex designs require some modifications of the R code. Multivariate statistical tests in the R code are based by default on **PERMANOVA** (Anderson 2001).

1. Create the database format from Species × Sample matrix
This first procedure allows constructing the required data format, which **is necessary for subsequent analyses** (procedure 2). **Two input files (.csv)** are needed for this step. **A first file contains the Species × Sample data matrix**, in which **species are rows and sample are columns**. For each species, the matrix should specify also the full taxonomic tree, from species to phylum (see the example data "Data.csv" in Sample File 1). Note that the procedure can be applied also when the original matrix involves operational units other than species (e.g. higher taxa, morphological groups, or mixed). **The second file contains factors for PERMANOVA analyses** (see the example file "Factors.csv" in Sample File 2).
The input file for data must be named "**Data.csv**". Note that **columns involving taxonomic information in "Data.csv" must be named as "Species", "Genus", Family", "Order", "Class", "Phylum" and provided in this precise order**. Note that the full taxonomic hierarchy must be provided if interest lies also in checking for the information retained at higher taxonomic resolution. The R code for this procedure is provided in Bevilacqua *et al.* (2013). If this is not in the aims of the investigator, columns for the taxonomic hierarchy could be filled as follows. For instance:

Species	Genus	Family	Order	Class	Phylum
C. ionica	*C. ionica*	*C. ionica*	*C. ionica*	*C. ionica*	*C. ionica*
Paraonidae	Paraonidae	Paraonidae	Paraonidae	Paraonidae	Paraonidae
Erect sponges	Erect sponges	Erect sponges	Erect sponges	Erect sponge	Erect sponges

The input file for factors must be named "**Factors.csv**". **Columns in "Factors.csv" must be named "FactorA", "FactorB", and "FactorC"**. **FactorA indicates the asymmetry** in the design, whereas **FactorB and FactorC are the two nested factors, with FactorC nested in FactorB**. **Labels for the single level (asymmetrical) of FactorA** (e.g. *I*) **must be set always at the end of the sequence**. For instance, if the design involves 3 locations, one impacted and 2 control locations, with 2 sites in each location and 3 replicates in each site, then the input file for factors should be as follow (see also the example file "Factors.csv" in Sample File 2):

```
FactorA;FactorB;FactorC
C1;1;1
C1;1;1
C1;1;1
C1;1;2
C1;1;2
C1;1;2
```

C2;2;1
C2;2;1
C2;2;1
C2;2;2
C2;2;2
C2;2;2
I;3;1
I;3;1
I;3;1
I;3;2
I;3;2
I;3;2

Note that **the sequence of factor levels in "Factors.csv" has to correspond to the sequence of samples in "Data.csv"**. Check carefully input files for errors, symbols not recognized by R, and so on. Input files must be exactly in the form of example files (see Sample File 1, Sample File 2), with names and numbers separated by semicolons.
The output of the analysis is a .csv file named by default "Data.o.csv", containing the database required for subsequent analyses.

2. Calculating simulated correlation and PERMANOVA results for each group of random surrogates
This part of the R code contains the script calculating (a) correlation ρ values between the species-level data matrix and randomly aggregated matrices for each *G* obtained from the step-wise reduction of fixed detriments *d* (see Methods, see Bevilacqua et al. 2013), and (b) PERMANOVA results for each randomly aggregated matrix. This procedure also allows calculating the correlation ρ values between the species-level matrix and matrices in which species are randomly aggregated in the $G_{BestAgg}$ groups (see Methods), in order to check the amount of information on species-level community patterns retained and the probability of Type I error when using $G_{BestAgg}$ surrogates.
The first step here (*Step 2.1 Build sub-groups table – Determine number of sub-groups for each dataset*) serves to define the set of *G* groups in which species will be aggregated at random. *G* groups come from the step-wise reduction of the original number of species *S*. *S* is progressively reduced by fixed detriments $d = 10\%\ S$, or by $d = 5\%\ S$ if there is the need to reduce gaps between different aggregation levels (e.g., for very speciose assemblages). The line of the script:

(1) *Data.groups=c(50,47,44,41,38,35,32,29,26,23,20,17,14,11,8,5,2)*

contains, at the present, the *G* groups defined for the example data provided in Sample File 1, which refers to the to the first time of sampling (Time 1) in the case study presented. In this case $S = 53$ and a fixed detriment of 3 species (corresponding approximately to $d = 5\%$ of *S*) was applied. Considering 53 species and fixed detriments, the original species were randomly aggregated in 50, 47, 44, 41,…,11, 8, 5, 2 groups. Therefore, **in this line of the R code, the sequence of *G* groups in which species have to be aggregated based on chosen detriments should be inserted in brackets**. The R code performs **1,000 random aggregations for each *G***.
The next step (*Step 2.2 Loading required functions*) involves **loading libraries and/or functions for analyses**.
Subsequent steps, instead, include scripts **calculating ρ values and performing the correct PERMANOVA test for the main term of interest of the asymmetrical design** (see Terlizzi et al. 2005 for details on variance partitioning in asymmetrical designs and selection of correct denominators in multivariate analysis). For example, if the design involve a single impacted location (*I*) and multiple control locations (*Cs*), with multiple sites nested in locations [S(L)], then to test Impact versus Control (*I*-vs-*Cs*):

1) *Step 2.3a Calculating correlations and PERMANOVA [*Cs *as DENOMINATOR]*. It applies when **the correct denominator for the *F* test of *I*-vs-*Cs* is the mean square of the *Cs* term**.
2) *Step 2.3b Calculating correlations and PERMANOVA [S(L) as DENOMINATOR]*. It applies when **the correct denominator for the *F* test of *I*-vs-*Cs* is the mean square of the S(L) term**.
3) *Step 2.3c Calculating correlations and PERMANOVA [Residuals as DENOMINATOR]*. It applies when all other terms may be excluded from the analysis and **the test for *I*-vs-*Cs* is done on the residuals**.

Note that the R code, as it is reported, is set for designs involving 2 control locations. A simple modification of the code, however, allows analyzing designs involving more than 2 locations. Lines of the R code that need to be modified are highlighted in light grey:

(1) `FirstPermanova=adonis(tabAggSimPerm~C(FactorA,c(-1,-1,2),how.many=1)+…`

and, if present,

(2) `SecondPermanova=myAdonis(tabAggSimPerm~C(FactorA,c(-1,-1,2),how.many=1)+…`

The parts of the script highlighted in light grey are currently set for the analysis of designs involving 2 control locations, but require a modification if the number of control locations is higher. For instance, in the case of 3 control locations, the parts of the script above (in light grey) have to be changed as follows:

```
c(-1,-1,-1,3)
```

Note that **PERMANOVA analyses are set, by default, on Bray-Curtis dissimilarities of untransformed data, with 999 permutations**. However, PERMANOVA analyses may be done using any distance metric after modifying the R code appropriately. If data transformation is required, data can be transformed before the analysis and used in the input file Data.csv (see above). Note that, depending on the size of the dataset, the number of step-wise reductions, i.e. numbers in line (1) (see above), and number of permutations used for PERMANOVA, analyses may have different duration, from hours to days. In this view, the number of permutations for PERMANOVA has been fixed to 999, allowing testing terms of interest even with α = 0.001. More conservative significance levels may be obtained increasing the number of permutations, although the time required for analyses may rise consequently.
The final step of this procedure (*Step 2.4 Writing result table*) **provides a .csv file**, named by default as "**Data.res.csv**" **containing results** of previous analyses. In the result table, **columns reporting correlation ρ values from 1,000 random aggregations for each *G* are indicated with the number of the specific *G* groups followed by ".cor"** (e.g. "50.cor"), whereas the remaining **columns reporting *P*-values of the test of interest from the corresponding 1,000 PERMANOVA analyses are indicated with the number of the specific *G* groups followed by ".aov"** (e.g. "50.aov").
Based on such results, the % of analyses for each *G* giving consistent results with those obtained analyzing species-level data can be easily calculated, and regression analyses of ρ against ln(ϕ) (where $\phi = G/S$, see Methods) can be then performed using any statistical package. **The lowest *G* allowing at least 95% of PERMANOVA analyses to give a *P*-value equal or lower than the significance level obtained analyzing species-level data (or than the significant level defined a priori according with the aim of the study) represents the sufficient number of surrogate groups G_{min}** (see Bevilacqua et al. 2013 for details). Note that the procedure allows identifying G_{min} directly, whereas **the corresponding sufficient level of aggregation is given as $\phi_{low} = G_{min}/S$**.
Once G_{min} has been defined, surrogates for BestAgg can be selected, and the final number of BestAgg surrogates, namely $G_{BestAgg}$, obtained (see text, and se also Bevilacqua et al. 2013 for details). The above procedure also allows **checking for the amount of information on species-level community patterns retained when using BestAgg surrogates**, **expressed as $\rho_{BestAgg}$** (i.e. the correlation between the original species-level matrix and the matrix in which species have been aggregated in the BestAgg surrogates, which is not provided by the R code but can be calculated using any statistical computer program), and **defining the probability of Type I error for $G_{BestAgg}$**. The whole procedure, in this case, can be performed **inserting the number of $G_{BestAgg}$ in the line (1) of the R code**. For instance, for the example data provided in Sample File 1 (referred to T1 of the case study presented), G_{min} = 32 and the selection procedure led to $G_{BestAgg}$ = 32 (see Table S1). Thus, the line (1) will be:

(2) *Data.groups=c(32)*

At end of the analysis, **in this case, the resulting file "Data.res.csv" will report correlation ρ values and *P*-values specific for $G_{BestAgg}$**. The **ρ values obtained in this way can be used to construct the 95%CI or frequency distribution against which testing $\rho_{BestAgg}$** (e.g. see Figure 3). As previously described, **the % of PERMANOVA analyses showing *P*-values equal or lower than the significance level of species-level analyses (or than the significant level defined a priori according with the aim of the study) represents the probability of Type I error specific for $G_{BestAgg}$** (under the null hypothesis that BestAgg surrogates are random subsets of the original pool of species).
In this framework, any set of surrogates (e.g. using families as surrogates of species) can be checked. Just to insert in line (1) of the R code the number of surrogates of the set to obtain ρ values and *P*-values from 1,000 randomizations, which can be then used, as explained above, to the check the amount of information retained and define the related probability of Type I error.
A full explanation of the approach and further details on data entry, factors, and a user guide to the BestAgg approach, along with R codes for balanced designs are provided in Bevilacqua et al. (2013). Full theoretical framework underlying the approach is provided in Bevilacqua et al. (2012).

References

Anderson MJ (2001) A new method for non-parametric multivariate analysis of variance. Aust Ecol 26:32–46.
Bevilacqua S, Terlizzi A, Fraschetti S, Claudet J, Boero F (2012) Taxonomic relatedness does not matter for species surrogacy in the assessment of community responses to environmental drivers. J App Ecol 49:357–366.
Bevilacqua S, Terlizzi A, Claudet J (2013) Best Practicable Aggregation of Species: a step forward for species surrogacy in environmental assessment and monitoring. Ecol Evol 3:3780–3793.

Terlizzi A, Benedetti-Cecchi L, Bevilacqua S, Fraschetti S, Guidetti P, Anderson MJ (2005) Multivariate and univariate asymmetrical analyses in environmental impact assessment: a case study of Mediterranean subtidal sessile assemblages. Mar Ecol Prog Ser 289:27–42.

Sample File 1. Example data: species level matrix of sampling Time 1 (T1). Provided as separate .csv file (see Data.csv).

Sample File 2. Factors for example data. Provided as separate .csv file (see Factors.csv).