

1 Classifying degrees of species commonness: North Sea  
2 fish as a case study

3 Gianpaolo Coro<sup>a,1,2,\*</sup>, Thomas J. Webb<sup>b</sup>, Ward Appeltans<sup>c</sup>, Nicolas Bailly<sup>d</sup>,  
4 André Cattrijsse<sup>e</sup>, Pasquale Pagano<sup>a</sup>

5 <sup>a</sup>*Istituto di Scienza e Tecnologie dell'Informazione "Alessandro Faedo" – CNR, Pisa,*  
6 *Italy*

7 <sup>b</sup>*Department of Animal & Plant Sciences, University of Sheffield, Sheffield S10 2TN, UK*

8 <sup>c</sup>*Intergovernmental Oceanographic Commission (IOC) of UNESCO, Oostende, Belgium*

9 <sup>d</sup>*WorldFish, Penang, Malaysia*

10 <sup>e</sup>*Vlaams Instituut voor de Zee (VLIZ), Oostende, Belgium*

---

11 **Abstract**

Species commonness is often related to abundance and species conservation status. Intuitively, a “common species” is a species that is abundant in a certain area, widespread and at low risk of extinction. Analysing and classifying species commonness can help discovering indicators of ecosystem status and can prevent sudden changes in biodiversity. However, it is challenging to quantitatively define this concept. This paper presents a procedure to automatically characterize species commonness from biological surveys. Our approach uses clustering analysis techniques and is based on a number of numerical parameters extracted from an authoritative source of biodiversity data, i.e. the Ocean Biogeographic Information System. The analysis takes into account abundance, geographical and temporal aspects of species distributions. We apply our model to North Sea fish species and show that the classification agrees with independent expert opinion although sampling

---

\*Corresponding author,  
*Preprint submitted to Ecological Modelling*  
Email addresses: [coro@isti.cnr.it](mailto:coro@isti.cnr.it) (Gianpaolo Coro),

May 25, 2015

[t.j.webb@sheffield.ac.uk](mailto:t.j.webb@sheffield.ac.uk) (Thomas J. Webb), [w.appeltans@unesco.org](mailto:w.appeltans@unesco.org) (Ward Appeltans), [n.bailly@cgiar.org](mailto:n.bailly@cgiar.org) (Nicolas Bailly), [andre.cattrijsse@vliz.be](mailto:andre.cattrijsse@vliz.be) (André Cattrijsse), [pagano@isti.cnr.it](mailto:pagano@isti.cnr.it) (Pasquale Pagano)

<sup>1</sup>Telephone Number: +39 050 315 2978

<sup>2</sup>Fax Number: +39 050 621 3464

biases affect the data. Furthermore, we show that our approach is robust to noise in the data and is promising in classifying new species. Our method can be used in conservation biology, especially to reduce the effects of the sampling biases which affect large biodiversity collections.

12 *Keywords:* Species Commonness, OBIS, Conservation biology, North Sea,  
13 Clustering, D4Science

---

## 14 **1. Introduction**

15 The term “common species” refers intuitively to a species that is abun-  
16 dant in a certain area, widespread and at low risk of extinction. By con-  
17 sequence, “rare species” are less abundant and possibly threatened. Auto-  
18 matically detecting common and rare species, and how their status changes  
19 through time, is an important step in understanding the consequences of en-  
20 vironmental change for ecosystem functioning. In particular, the abundance  
21 of a species in a community or ecosystem is a key indicator of its ecological  
22 role and ecosystem function therefore depends on the identities and relative  
23 numbers of common and rare species [1]. For instance, rare species may have  
24 unique functional traits [2] and make particular contributions to diversity  
25 [3]. On the other hand, common species may underpin ecosystem function  
26 where they dominate in terms of biomass [4, 5, 6]. Both human activity  
27 and natural environmental change typically affect the relative abundances  
28 of species [7]. Monitoring changes in the relative abundance of species is  
29 straightforward when working on individual, well-monitored systems. How-

30 ever, anthropogenic-driven environmental change is affecting entire ecosys-  
31 tems, requiring large-scale ecological efforts [8]. One approach to monitor  
32 species commonness at large scale and in a certain time frame, is to perform  
33 meta-analyses on studies of multiple individual communities. This is useful  
34 for extracting general trends across multiple taxa [9]. An alternative is to  
35 take advantage of the increasing availability of large-scale compilations of  
36 biodiversity data, such as the UK's National Biodiversity Network (NBN)  
37 [10], the Global Biodiversity Information Facility (GBIF) [11], or the Ocean  
38 Biogeographic Information System (OBIS) [12]. These compilations include  
39 millions of opportunistic records of the distributions of very large numbers  
40 of species, often across multiple decades. This temporal dimension offers  
41 significant potential to track the relative commonness of species through  
42 time. However, it is difficult to extract robust estimates that are insensitive  
43 to changes and biases in sampling effort, from those heterogeneous and un-  
44 structured data sources [13]. The major issue is that it is hard to separate the  
45 signal of the actual relative commonness of a species in the system from the  
46 noise of sampling effort that varies in time and space, and in its taxonomic  
47 focus. For instance, a species may appear common across a given decade in  
48 a large dataset because there was at that time an intensive sampling pro-  
49 gramme targeting it. Its subsequent reduction in apparent abundance may  
50 simply reflect the end of that programme, rather than anything of ecological  
51 significance.

52 In this paper, we present a method to classify the degree of commonness

53 of marine fish species in a certain area and time frame, using a large data  
54 collection of biodiversity data. In particular, we rely on the OBIS data col-  
55 lection and, for the purposes of methodological development, we focus on  
56 fish from the North Sea, a subset of 70 well-studied but unevenly-sampled  
57 species. We use clustering analysis to automatically extract commonness  
58 classes from unstructured data and compare these classes with expert opin-  
59 ion. Reliable concordance between our method and experts, suggests that  
60 classifying commonness for less well-studied taxa or regions from data col-  
61 lections such as OBIS may be possible. We also assess the performance of  
62 our method in terms of (i) accuracy (using cross-validation), (ii) robustness  
63 to random noise in the data, (iii) dependency on the variables we chose to  
64 represent species commonness and (iv) dependency on our definition of these  
65 variables.

66 The paper is structured as follows: section 2 gives an overview on tech-  
67 niques for identifying species commonness. Section 3 describes the survey  
68 data we used. Section 4 reports the variables we defined to model the prob-  
69 lem and describes our modelling approach. Section 5 reports an evaluation  
70 of the robustness of our method. It includes a comparison between our auto-  
71 matic classification and the classifications produced by two experts. Section  
72 6 discusses the results, suggests possible usages of our technique and includes  
73 conclusive remarks.

## 74 2. Overview

75 Species commonness and rarity have been investigated in several scientific  
76 works. Most approaches derive species commonness from species abundance  
77 distributions (SADs) [14, 15]. The intimate connection between abundance  
78 and commonness (or rarity) is widely recognized, even if an explicit defini-  
79 tion of this dependency is unknown [5]. Approaches to model such depen-  
80 dency and to discover new correlated parameters, range from machine learn-  
81 ing based approaches to explicit modelling. In this last case, models specify  
82 the role that each parameter has in defining species commonness. Search-  
83 ing for these parameters usually requires analyses by domain experts. For  
84 example, Preston [16] analyses how abundance is distributed among species.  
85 He recognises the importance of characteristics like (i) the total number of  
86 living individuals, (ii) the total number of individuals living at any instant  
87 on a given area, (iii) the ratio of the number of individuals with respect to  
88 another species, (iv) the number of observed individuals in different data  
89 collections. Some authors suggest that common species tend to be common  
90 everywhere, as reflected in a general positive relationship between local pop-  
91 ulation density and regional distribution [17, 18, 19, 20]. These species also  
92 tend to remain common through time [21, 22], with major changes in the  
93 rank-order of species commonness rather rare. In other studies, common  
94 species have been identified with species widely distributed on a territory,  
95 whereas rare species have been indicated as those in the Red List for the  
96 same territory. For example, using these definitions, Pearman et al. [23]

97 detect spatial patterns for common species in Switzerland. In order to ac-  
98 count for this heterogeneity of parameters, other works have promoted using  
99 standard measures and data to compare common and rare species [24].

100 Unfortunately, no single satisfactory formal definition of species common-  
101 ness and rarity has been found, especially using explicit modelling. Clustering  
102 analysis is a promising approach coming from machine learning techniques  
103 that may help to address this. This technique has been widely used for  
104 identifying classes of species characteristics. For example, clustering envi-  
105 ronmental properties has proven to be useful in detecting vegetation types  
106 [25], in modelling the coexistence of plants in agro-ecosystems [26] and in  
107 detecting new agro-ecosystems [27]. Clustering analysis can also account for  
108 the lack of sampling uniformity in data collections, for example to group  
109 several species together when few data are available [28].

### 110 **3. Data**

111 Our model needs to be trained on species observation data. In order  
112 to identify the best training data, we searched for a dataset which was (i)  
113 sufficiently large and complex that relative commonness was not straight-  
114 forward to ascertain but where (ii) the number of species was not too large  
115 and (iii) independent estimates of relative commonness were available from  
116 expert opinion. Points (ii) and (iii) restricted us to well-known species, with  
117 officially accepted scientific names available from the authoritative World  
118 Register of Marine Species (WoRMS) [29, 30]. In order to extract data, we

119 consulted the Ocean Biogeographic Information System (OBIS) [31]. OBIS is  
120 the world's largest database on the diversity, distribution and abundance of  
121 all marine life. OBIS was initiated in 2000 by the Census of Marine Life and  
122 now runs under the auspices of UNESCO's Intergovernmental Oceanographic  
123 Commission. It currently provides free access to 40 million observations of  
124 115,000 marine species, integrated from more than 1,600 datasets provided by  
125 nearly 500 institutions worldwide. OBIS is an amalgam of many individual  
126 datasets from research projects, national monitoring programmes, museum  
127 collections and so on, targeting different taxa in different areas, often using  
128 different methods over different years. We limited our analysis on North Sea  
129 fish, because fish (Pisces<sup>3</sup>) represents 50% of all data in OBIS and the North  
130 Sea has relatively the highest amount of observations of all areas in the world.  
131 Thus, we extracted observation records from OBIS and defined the spatial  
132 boundaries of North Sea according to the International Hydrographic Orga-  
133 nization (IHO) indications. Furthermore, we selected only species observed  
134 between 2000-2009, as OBIS is particularly rich of datasets and occurrence  
135 records for the North Sea in this period. This selection produced a list of  
136 247 scientific species names, 70 of which had distinct and accepted species  
137 names according to WoRMS. We used this subset of 70 species from OBIS  
138 as a benchmark to develop and evaluate our method.

---

<sup>3</sup>LSID: urn:lsid:marinespecies.org:taxname:11676

## 139 4. Method

140 Starting from the dataset described in the section 3, we used clustering  
141 analysis to automatically derive classes of commonness. The aim was also  
142 to search for a classification robust enough to account for sampling biases.  
143 Clustering analysis requires defining variables on the data. This section re-  
144 ports the steps of our analysis from the definition of these variables to the  
145 selection and application of the clustering model.

### 146 4.1. Variables definition

147 The choice of the variables to use in a data mining experiment is very  
148 difficult when there is no formal definition of the phenomenon to model.  
149 Clustering analysis requires that each element to cluster is associated with a  
150 numeric vector. Thus, in our case we had to associate a vector of real numbers  
151 to each species, where the numbers were correlated with species commonness.  
152 Furthermore, such numbers had to be as independent as possible from each  
153 other. This was necessary to reduce noise during the clustering process.

154 The works reported in section 2, suggest that factors related to abundance  
155 and extent are correlated with species commonness. On the other hand, we  
156 know that collections of observations can contain biases. In particular, non-  
157 uniform sampling in time of the observations affects the estimation of species  
158 extents. We decided to classify the degree of commonness of each species in  
159 our benchmark dataset on the time frame of one decade (2000-2009), and  
160 to produce one classification per species for the decade. The main reason



161 is that we wanted to explore the robustness of the classification rather than  
162 producing an analysis of commonness trends. Thus, we took into account the  
163 rate of species observations in the decade. In particular, we considered the  
164 monthly observations of the species. This rate depends also on the datasets  
165 contained in the OBIS collection. A species that is contained in several  
166 datasets (each with a different survey scope) is likely to be often encountered  
167 in that area.

168 This process resulted in the following variables, whose definition was  
169 guided by a cycle of interactions with domain experts. They refer only to  
170 records from the North Sea, extracted with proper geo-spatial queries:

171 *Abundance (A)*: average number of reported individuals per observation.  
172 This quantity takes into account the number of individuals reported each  
173 time a species is observed:

$$A = \frac{n. \text{ of individuals reported in the record}}{n. \text{ of observation records}}$$

174 *Intra-Dataset Observations (IntraDO)*: average number of observations per  
175 dataset. These datasets come from different OBIS contributors, e.g. Fish-  
176 Base and NOAA. This parameter accounts for the frequency of presence of  
177 a species in each dataset. If the quantity is high, then the species is often  
178 reported by the OBIS contributors:

$$IntraDO = \frac{\sum_D n. \text{ of observations in dataset } D}{n. \text{ of datasets in OBIS}}$$

179 *Inter-Dataset Observations (**InterDO**)*: fraction of datasets containing ob-  
180 servation records for a species. This parameter accounts for the observation  
181 frequency of a species among the OBIS contributors:

$$InterDO = \frac{n. \text{ of datasets with at least one observation for the species}}{n. \text{ of datasets in OBIS}}$$

182 *Extension (**E**)*: fraction of 0.1 degree cells in the North Sea, for which at least  
183 one observation was reported. This measure accounts for the distributional  
184 extent of the species:

$$E = \frac{n. \text{ of 0.1 degree cells containing observations for the species in North Sea}}{n. \text{ of 0.1 degree cells in North Sea}}$$

185 *Time Rate (**TR**)*: fraction of months containing at least one observation  
186 record. This measure accounts for the time rate of the species observations:

$$TR = \frac{n. \text{ of months containing species observations between 2000 and 2009}}{n. \text{ of months between 2000 and 2009}}$$

187 *Time Rate of Many Observations (**TRMO**)*: fraction of months containing  
188 a significant number of observations. This is an alternative measure of the  
189 observation rate, which accounts for the months in which it was frequent to  
190 observe the species. Based on the values of species known to be common or  
191 rare, we calculated that 10 observations were a significant threshold in the

192 2000-2009 decade.

$$TRMO = \frac{n \text{ months containing at least 10 species observations}}{n. \text{ of months between 2000 and 2009}}$$

193     Extracting the values of these variables from our benchmark generated  
194 a set of 70 vectors of 6 Real numbers, each referring to one species between  
195 2000 and 2009. The values of the variables would need to be recalculated  
196 if the focus area and time range change. Applying the same calculations  
197 to other data collectors than OBIS, would require finding correspondence in  
198 the new collection for the elements constituting the above formulae. These  
199 elements can be reconstructed from (i) geo-localized observation records, (ii)  
200 the number of individuals per observation, (iii) the identity of the datasets  
201 containing the observations, (iv) observation dates. Most data collectors (e.g.  
202 GBIF and FishBase) support such information, which reassures us of the  
203 potential generality of this approach. Nevertheless, the OBIS Postgres-based  
204 collection provides very easy and fast access to retrieve the above values.

#### 205 *4.2. Clustering*

206     Clustering analysis is a data mining technique which is able to group  
207 together numeric vectors, according to a certain similarity criterion. In the  
208 case of real valued vectors, similarity is usually measured in terms either of  
209 density or of euclidean distances. In our case, we wanted to verify if clustering  
210 could extract classes of similarity related to species commonness. To this end,  
211 we selected two alternative clustering techniques, named X-Means [32] and

212 DBScan [33]. The former uses a distance based approach, while the latter  
213 uses a density-based approach. We selected such algorithms because they  
214 automatically find the best number of clusters from the data.

215 DBScan is a density-based clustering algorithm. It searches for an optimal  
216 number of clusters on the basis of two parameters: *epsilon* and *min points*.  
217 The former is a distance threshold that defines the neighbourhood of a point  
218 (epsilon-neighbourhood), while the latter is the minimum number of points  
219 required to form a dense region. The DBSCAN algorithm starts selecting an  
220 arbitrary point. Then it takes the epsilon-neighbourhood of the point and,  
221 if this contains at least *min points* elements, it aggregates the points into a  
222 cluster. Otherwise, it assumes that this point could be later found in the  
223 epsilon-neighbourhood of another point (and thus added to the cluster of  
224 that point), and moves to another point. The process analyses all the points  
225 and creates density-connected clusters. For further details see Ester et al.  
226 [33].

227 X-Means is a variant of the popular K-Means algorithm [34], which intro-  
228 duces several efficiency enhancements. An important difference with respect  
229 to K-Means is that the number of optimal clusters to search for is not speci-  
230 fied *a priori*. Instead, it requires to set a minimum and a maximum number  
231 of clusters ( $K_{min}$  and  $K_{max}$ ) to search for. The X-Means algorithm starts  
232 from  $K_{min}$  and adds centroids as far as  $K_{max}$  is reached. At each step, the  
233 K-Means algorithm is run, which finds the best assignment of the vectors to  
234 the indicated number of clusters. K-Means indicates a score for this assign-

235 ment, based on the distortion measure, i.e. the average squared distance of  
236 the points to their clusters centroids. The X-Means algorithm outputs the  
237 result of the K-Means that gave the best score, and consequently the best  
238 number of clusters. X-Means also adds efficiency enhancements to K-Means,  
239 using *kd*-trees [35] and *blacklisting* to support processing. Furthermore, at  
240 each step of the computation, the location of the centroids of the additional  
241 clusters is decided using the Bayesian Information Criterion (BIC) [36]. For  
242 further details see Pelleg and Moore [32].

243 We applied clustering analysis to our North Sea species benchmark. In our  
244 experiment, we searched for the clustering analysis detecting the lowest num-  
245 ber of clusters and presenting a uniform distribution of the vectors in these  
246 clusters. We used the implementations running on the D4Science Statistical  
247 Manager Service [37, 38], which hosts such procedures as-a-Service. We used  
248 several configurations for both the algorithms. Eventually, the best configu-  
249 ration for DBScan was obtained by setting  $\epsilon = 100$  and  $\minpoints = 2$ .  
250 Unfortunately, this ended in 38 clusters and was not practical to use. On  
251 the other hand, the X-Means algorithm was executed by asking to search  
252 for a number of clusters between 1 and 50. Although the interval was large,  
253 the algorithm ended in only four clusters. The algorithm found an optimal  
254 separation of the vectors according to their relative euclidean distance. Fur-  
255 thermore, we noticed that such clusters could be given an interpretation.  
256 The dataset and the results are available as supplementary material of this  
257 paper.

258 The normalized distribution of the mean values of the variables is re-  
259 ported in Table 1 for each X-Means cluster. Table 2 reports examples of  
260 vectors associated to the clusters and Figure 1 displays the distribution of  
261 the values of the clustering variables over the clusters. Table 3 reports the  
262 interpretation we gave to these clusters, based on the distributions of their  
263 centroids and of the variables values. Cluster number 1, interpreted as the  
264 class of “Common” species, contains 12 vectors (corresponding to 12 species),  
265 and is characterized by very high values of almost each variable. This means  
266 that the species in this cluster are frequent, widespread and with high in-  
267 dividual density. Cluster 2 (“Moderate Commonness”) contains 21 vectors  
268 with lower individual density with respect to cluster 1. The most evident  
269 characteristics are moderate distributional extent and moderate frequency of  
270 observation. Cluster 3 (“Moderate-Low Commonness”) contains 23 vectors  
271 presenting a low individual density and only moderate reporting frequency  
272 by several datasets. Finally, cluster 4 (“Low Commonness”, which includes  
273 rare species) contains 14 species which are very localized and with low indi-  
274 vidual density. In this case, we use the term *widespread* to indicate that the  
275 species has a large geographical range, in which it is likely to be observed.  
276 The term *localized* means that the species lives in highly localized zones, but  
277 there could be a certain distance between such zones. Finally, individual  
278 density is defined *high* if a large number of individuals are encountered each  
279 time the species is observed.

## 280 **5. Evaluation**

### 281 *5.1. Agreement with experts*

282 In this section, we evaluate the performance of the classification produced  
283 by X-Means with respect to expert opinion. In order to create a comparison  
284 reference, two of us (Bailly and Cattrijsse) performed independent classifi-  
285 cation assignments on the 70 benchmark species of North Sea fish, based  
286 on expert opinion. Each expert separately assigned the appropriate cluster  
287 to each species, selecting among those in Table 3. The experts did not be-  
288 long to the same institute: Expert 1 (Cattrijsse) is a researcher in Coastal  
289 Marine Biology working for the Vlaams Instituut voor de Zee (VLIZ), while  
290 Expert 2 (Bailly) is a biologist working in the biodiversity informatics field  
291 for the World Fish Center. The result of this classification is available as  
292 supplementary material attached to this paper.

293 We estimated the agreement between all the classifications using the ab-  
294 solute percentage of agreement, defined as the percentage of matching assign-  
295 ments. Furthermore, we also calculated Cohen’s Kappa [39], which estimates  
296 the agreement between two evaluators with respect to purely random assign-  
297 ments. Cohen’s Kappa allows comparing complex classification tasks (e.g.  
298 with many classes) with simpler ones (e.g. dichotomous scenarios) where  
299 high agreement could have occurred by chance. Table 4 reports the Cohen’s  
300 Kappa values of the agreements, along with two different interpretations  
301 commonly used in literature [40, 41]. It is notable that in this experiment  
302 the absolute percentage agreement reflects the Kappa values. The values are

303 symmetric, thus we report them once per pair of evaluators.

304 In order to give insight about the differences between the classifications  
305 assignments, we report the example of the lesser pipefish *Syngnathus rostellatus*<sup>4</sup>,  
306 which Expert 2 and X-Means assign to *Moderate-Commonness*, and  
307 Expert 1 to *Common*. This species presents an *Abundance* (A) parameter  
308 value equal to 17.16, quite far from the 325.27 of the common dab *Limanda*  
309 *limanda*<sup>5</sup>, which is “Common” according to all the assignments. A signifi-  
310 cant difference is recorded also for the *IntraDO* values, which is 101.75 for  
311 the lesser pipefish and 24521.14 for the common dab. Indeed, *Syngnathus*  
312 *rostellatus* has a lower number of observation records for (407 records) with  
313 respect to *Limanda limanda* (171648 records). This influences the behaviour  
314 of X-Means, but its classification can be still considered viable because it  
315 agrees with one of the two experts. Figure 2 depicts the distribution of the  
316 observation records of the above species, aggregated at 0.5 degrees resolution.

317 One interesting consideration is that, even if the classification classes were  
318 automatically detected by the X-Means algorithm, the overall agreement  
319 with both the experts is good. On the other hand, the agreement between  
320 the two experts is poor. This indicates that the problem is objectively hard,  
321 but clustering seems able to reconcile the divergent opinions in some way.

322 The disagreement between experts could be due to their different inter-  
323 pretation of the clusters descriptions. Thus, we investigated this aspect by

---

<sup>4</sup>LSID: urn:lsid:marinespecies.org:taxname:127389

<sup>5</sup>LSID: urn:lsid:marinespecies.org:taxname:127139



324 aggregating the not *Common* clusters into a generic *Non-Common* cluster.  
325 Table 5 reports the evaluation in this case. The agreement between Expert  
326 2 and clustering is excellent, while the aggregation introduces misalignment  
327 between Expert 1 and clustering. This is due to a general tendency by Expert  
328 1 to classify more in the *Moderate-Commonness* class.

329 We repeated the same evaluation aggregating the *Common* and the *Moderate-*  
330 *Commonness* clusters into one cluster, and the *Moderate-Low* and *Low-*  
331 *Commonness* clusters into another cluster. Table 6 reports the agreement  
332 in this case. With this aggregation, the agreement by both the experts with  
333 the clustering analysis is good, and highest agreement is still with Expert 2.

334 These experiments highlight that even changing the definition of the clus-  
335 ters, there is a sensible agreement between experts and clustering. This  
336 indicates reliability of the automatic classification. It is notable that the  
337 variables used by the clustering analysis are likely to be affected by biases,  
338 especially when the species is poorly reported in time and is rarely reported  
339 by the OBIS contributors. Clustering accounts for the lack of information of  
340 some variables, because it compensates with information from the other vari-  
341 ables. This comes out from the variables combination made by the euclidean  
342 distances and by the subsequent optimization process. Furthermore, produc-  
343 ing classes of commonness (instead of commonness scores) hides fine-grain  
344 differences between the vectors.

345 *5.2. Performance evaluation*

346 We measured the robustness of our method in terms of (i) classifying new  
347 species, (ii) dependency on noise, (iii) dependency on the clustering variables  
348 and (iv) on their definitions. In particular, we calculated the performance on  
349 classifying species that were not included in the training set. To this aim, we  
350 used cross-validation. We randomly selected 90% of the species to produce  
351 clusters. We checked if the clusters coincided with the ones extracted using  
352 100% of the species (complete set), and then we used the other 10% of the  
353 species to check if their associated vectors were assigned to the same clusters  
354 as in the complete set. We used only 10% of the species as test set because  
355 our benchmark dataset had small size. In each experiment, we calculated the  
356 *accuracy* of the classification as the ratio between correct assignments and  
357 overall assignments. In the end, we averaged the accuracies of ten executions.  
358 In all the experiments the clusters coincided with the ones of the complete set.  
359 The overall (averaged) accuracy was 98.57%. This means that for the North  
360 Sea case our clusters are stable and the model is promising in classifying new  
361 species.

362 As further step, we checked the robustness of our classification to noise.  
363 As explained before, the data we extracted from OBIS contain sampling  
364 biases. The good agreement of our method with expert opinion already sug-  
365 gests that our approach can manage these biases. Nevertheless, we explored  
366 this aspect further by adding an increasing amount of white noise to our  
367 data and checking if the clusters remained stable, i.e. if the newly identified

368 clusters were still the ones of Table 3. We added white noise directly to our  
369 variables and Table 7 reports the results: a 10% noise level means that we  
370 randomly added or subtracted up to the 10% of a variable value. Referring  
371 to Table 7, up to 1% of noise there is no change in the clustering and even  
372 at 5% the clusters are very similar to the ones without noise, because most  
373 of the species in the original (“clean” data) clusters are found in the corre-  
374 sponding newly found clusters. The number of clusters changes when 10%  
375 of noise is reached, but at this level the newly found clusters have still corre-  
376 spondence with the original clusters. For example, the species belonging to  
377 the original cluster 1 are largely included in the newly found cluster 1. The  
378 original cluster 2 corresponds to both the new clusters 1 and 2, whereas the  
379 original cluster 3 and 4 correspond to the new clusters 2 and 3 respectively.  
380 Over 10% of noise the original clusters are no more recognizable. It is our  
381 opinion that this limit is a reasonable indicator of robustness to noise. It  
382 is remarkable, in fact, that our data are already biased and the white noise  
383 only adds more bias.

384 As additional step, we evaluated the influence of each variable on the  
385 clustering analysis. Table 8 reports the results of the clustering analysis  
386 when we exclude one variable at time. The number of clusters changes and  
387 the identity of the original clusters is lost in most of the cases. It is notable  
388 that when *InterDO* is missing, the number of clusters is overestimated. In the  
389 other cases, the clustering is very simplistic and does not allow easy semantic  
390 interpretations. In particular, clusters 1, 3 and 4 are merged together, which

391 means that common and uncommon species are mixed up. These changes  
392 indicate that all the variables have an important role (i.e. carry a remarkable  
393 amount of information) in the definition of the clusters of Table 3. Our  
394 definitions are related to indicators taken from other studies and come from  
395 expert opinion (see section 4.1). This analysis confirms that they all have  
396 a key role in producing species commonness classes that agree with expert  
397 opinion.

398 As final step, we checked if the commonness classes depend on our defini-  
399 tions of the variables (see section 4.1). Table 9 reports how the results of the  
400 clustering analysis change when the variables definitions are slightly altered.  
401 The new definitions in Table 9 still include information that is correlated to  
402 the original definitions. For example, in one of the experiments we redefined  
403  $A$  as the number of recorded individuals, without dividing for the number of  
404 observations. In another case, we defined one time variable as the ratio be-  
405 tween the two time variables  $TRMO$  and  $TR$ . The last row of Table 9 reports  
406 the case in which all the variables definitions are altered. In all the cases, the  
407 clustering analysis identifies four clusters. Furthermore, the original clusters  
408 are recognizable in all the cases and sometimes the output coincides with the  
409 one of the original model. This means that the clustering analysis is flexible  
410 enough to exploit the information associated to the variables, even when the  
411 variables definitions change.

## 412 **6. Discussion and conclusions**

413 In this paper we have presented an approach to classify species common-  
414 ness. We have trained our models on a dataset extracted from the OBIS data  
415 collection and focusing on North Sea fishes. The performance has been eval-  
416 uated by comparing automatic assessments with the opinions of two experts.  
417 We have demonstrated that our process has good agreement with expert  
418 opinion although our analysed dataset contains sampling biases. We have  
419 further explored this robustness, by evaluating the effects that random noise  
420 in the data has on the classification. The results indicate that the model  
421 is reasonably robust in managing noise. Furthermore, we have used cross-  
422 validation to calculate the performance of our model in classifying species  
423 that had not been included in the training set. The performance indicates  
424 that the identified clusters are stable for the North Sea species. This gives  
425 suggestions about the possible generalisation of our method. In fact, our  
426 clustering analysis is also applicable to other areas and large biodiversity  
427 data collections. Applying our method to other regions than North Sea re-  
428 quires the model to be trained on new data. Indeed, we conducted the same  
429 analysis on 222 species from OBIS at global scale. Also in this case, we  
430 found an optimal separation into four clusters<sup>6</sup> having the same percentage  
431 distributions as in Table 1. This result indicates that our classification could  
432 be valid for other areas too, but validating this hypothesis requires further

---

<sup>6</sup>The complete classification is available on the D4Science e-Infrastructure for consul-  
tation: <http://goo.gl/FYuD6P>

433 investigation and much more effort in terms of experts' reviews. We will  
434 address this issue in future experiments.

435 We have demonstrated that our process is more dependent on the in-  
436 formation included in the variables than to their definition. This is useful  
437 when applying our analysis to other biodiversity data collections that report  
438 information in a different way from OBIS.

439 Finally, we have demonstrated also that our set of variables contains a  
440 sufficient amount of information to identify four reliable commonness clas-  
441 sifications. Using a lower number of variables would produce less refined  
442 classifications and less clusters (see Table 8). This is a remarkable property,  
443 since we defined the variables based on interactions with ecology and data  
444 experts (i.e. not using automatic data selection [42]). This may suggest that  
445 our variables are ecologically meaningful, i.e. they are really correlated to  
446 species commonness.

447 From our analysis, new biodiversity and ecosystem indicators could be  
448 identified and this will be part of our future investigations. For example,  
449 using our method a species could be found, today, to be “less common” in  
450 a certain area with respect to a previous time period. This could indicate a  
451 change of the ecosystem in that area or that the species has been overfished.  
452 Our method could be also a way to reconcile the opinions of different experts  
453 about the commonness of a set of species. For example, it could be used as a  
454 supporting tool for biologists, who would rely on an “external” opinion when  
455 discussing about species commonness. Furthermore, classifying commonness

456 for fishes in a well-studied region is a first step towards working on less known  
457 taxa in other regions.

458 Our experiments highlight the intrinsic difficulty of the problem, but the  
459 proposed technique represents a step forward in classifying species common-  
460 ness and in understanding which factors are related to this concept. A data  
461 provider like OBIS could embed such method to alert a user about the pos-  
462 sible commonness of a species in a certain area. In this context, we are  
463 planning to build an interface allowing a user to select an IHO area and  
464 a time range, and to retrieve the species possibly classified as *Common* or  
465 *Moderately-Common*. Currently, our clustering technique is released as soft-  
466 ware [43, 44] inside the i-Marine e-infrastructure [45], which grants free access  
467 to statistics about the OBIS database and allows sharing datasets, biological  
468 analyses and experimental results.

## 469 **Acknowledgments**

470 The reported work has been partially supported by the i-Marine project  
471 (FP7 of the European Commission, INFRASTRUCTURES-2011-2, Contract  
472 No. 283644). Thomas J. Webb is a Royal Society University Research Fellow.

## 473 **References**

- 474 [1] A. E. Magurran, Biodiversity in the context of ecosystem function, Ma-  
475 rine biodiversity & ecosystem functioning-frameworks, methodologies  
476 and integration (2012) 16–23.

- 477 [2] D. Mouillot, D. R. Bellwood, C. Baraloto, J. Chave, R. Galzin,  
478 M. Harmelin-Vivien, M. Kulbicki, S. Lavergne, S. Lavorel, N. Mou-  
479 quet, et al., Rare species support vulnerable functions in high-diversity  
480 ecosystems, *PLoS biology* 11 (5) (2013) e1001569.
- 481 [3] X. Mi, N. G. Swenson, R. Valencia, W. J. Kress, D. L. Erickson, A. J.  
482 Pérez, H. Ren, S.-H. Su, N. Gunatilleke, S. Gunatilleke, et al., The  
483 contribution of rare species to community phylogenetic diversity across  
484 a global network of forest plots, *The American Naturalist* 180 (1) (2012)  
485 E17–E30.
- 486 [4] K. J. Gaston, R. A. Fuller, Commonness, population depletion and con-  
487 servation biology, *Trends in Ecology & Evolution* 23 (1) (2008) 14–19.
- 488 [5] K. J. Gaston, Valuing Common Species, *Science* 327 (5962) (2010) 154–  
489 155. doi:10.1126/science.1182818.  
490 URL <http://dx.doi.org/10.1126/science.1182818>
- 491 [6] K. J. Gaston, Common ecology, *Bioscience* 61 (5) (2011) 354–362.
- 492 [7] F. S. Chapin III, E. S. Zavaleta, V. T. Eviner, R. L. Naylor, P. M.  
493 Vitousek, H. L. Reynolds, D. U. Hooper, S. Lavorel, O. E. Sala, S. E.  
494 Hobbie, et al., Consequences of changing biodiversity, *Nature* 405 (6783)  
495 (2000) 234–242.
- 496 [8] J. T. Kerr, H. M. Kharouba, D. J. Currie, The macroecological contri-  
497 bution to global change solutions, *Science* 316 (5831) (2007) 1581–1584.



- 498 [9] M. Dornelas, N. J. Gotelli, B. McGill, H. Shimadzu, F. Moyes, C. Siev-  
499 ers, A. E. Magurran, Assemblage time series reveal biodiversity change  
500 but not systematic loss, *Science* 344 (6181) (2014) 296–299.
- 501 [10] National Biodiversity Network (NBN)., [nbn.org.uk](http://nbn.org.uk) (2014).
- 502 [11] Global Biodiversity Information Facility (GBIF)., [gbif.org](http://gbif.org) (2014).
- 503 [12] Intergovernmental Oceanographic Commission (IOC) of UNESCO.  
504 The Ocean Biogeographic Information System., <http://www.iobis.org>  
505 (2014).
- 506 [13] N. J. Isaac, A. J. Strien, T. A. August, M. P. Zeeuw, D. B. Roy, Statistics  
507 for citizen science: extracting signals of change from noisy ecological  
508 data, *Methods in Ecology and Evolution*.
- 509 [14] S. R. Connolly, M. A. MacNeil, M. J. Caley, N. Knowlton, E. Cripps,  
510 M. Hisano, L. M. Thibaut, B. D. Bhattacharya, L. Benedetti-Cecchi,  
511 R. E. Brainard, et al., Commonness and rarity in the marine biosphere,  
512 *Proceedings of the National Academy of Sciences* (2014) 201406664.
- 513 [15] B. J. McGill, R. S. Etienne, J. S. Gray, D. Alonso, M. J. Anderson, H. K.  
514 Benecha, M. Dornelas, B. J. Enquist, J. L. Green, F. He, et al., Species  
515 abundance distributions: moving beyond single prediction theories to in-  
516 tegration within an ecological framework, *Ecology letters* 10 (10) (2007)  
517 995–1015.

- 518 [16] F. W. Preston, The commonness, and rarity, of species, *Ecology* 29 (3)  
519 (1948) 254–283.
- 520 [17] K. J. Gaston, T. M. Blackburn, J. J. Greenwood, R. D. Gregory, R. M.  
521 Quinn, J. H. Lawton, Abundance–occupancy relationships, *Journal of*  
522 *Applied Ecology* 37 (s1) (2000) 39–59.
- 523 [18] T. M. Blackburn, P. Cassey, K. J. Gaston, Variations on a theme:  
524 sources of heterogeneity in the form of the interspecific relationship be-  
525 tween abundance and distribution, *Journal of Animal Ecology* 75 (6)  
526 (2006) 1426–1439.
- 527 [19] T. J. Webb, R. P. Freckleton, K. J. Gaston, Characterizing abundance–  
528 occupancy relationships: there is no artefact, *Global Ecology and Bio-*  
529 *geography* 21 (9) (2012) 952–957.
- 530 [20] T. Hughes, D. Bellwood, S. Connolly, H. Cornell, R. Karl-  
531 son, Double jeopardy and global extinction risk in corals  
532 and reef fishes, *Current Biology* 24 (24) (2014) 2946 – 2951.  
533 doi:<http://dx.doi.org/10.1016/j.cub.2014.10.037>.  
534 URL [http://www.sciencedirect.com/science/article/pii/](http://www.sciencedirect.com/science/article/pii/S0960982214013463)  
535 [S0960982214013463](http://www.sciencedirect.com/science/article/pii/S0960982214013463)
- 536 [21] T. J. Webb, D. Noble, R. P. Freckleton, Abundance–occupancy dynam-  
537 ics in a human dominated environment: linking interspecific and in-

- 538       traspecific trends in british farmland and woodland birds, *Journal of*  
539       *Animal Ecology* 76 (1) (2007) 123–134.
- 540 [22] T. J. Webb, Marine and terrestrial ecology: unifying concepts, revealing  
541       differences, *Trends in ecology & evolution* 27 (10) (2012) 535–541.
- 542 [23] P. B. Pearman, D. Weber, Common species determine richness patterns  
543       in biodiversity indicator taxa, *Biological Conservation* 138 (1) (2007)  
544       109–119.
- 545 [24] R. Bevill, S. Louda, Comparisons of related rare and common species in  
546       the study of plant rarity, *Conservation Biology* 13 (3) (1999) 493–498.
- 547 [25] M. B. Dale, P. Dale, P. Tan, Supervised clustering using decision trees  
548       and decision graphs: An ecological comparison, *Ecological modelling*  
549       204 (1) (2007) 70–78.
- 550 [26] M. Debeljak, G. R. Squire, D. Kocev, C. Hawes, M. W. Young,  
551       S. Džeroski, Analysis of time series data on agroecosystem vegetation  
552       using predictive clustering trees, *Ecological Modelling* 222 (14) (2011)  
553       2524–2529.
- 554 [27] M. Liu, A. Samal, A fuzzy clustering approach to delineate agroecozones,  
555       *Ecological Modelling* 149 (3) (2002) 215–228.
- 556 [28] N. Picard, F. Mortier, V. Rossi, S. Gourlet-Fleury, Clustering species  
557       using a model of population dynamics and aggregation theory, *Ecological*  
558       *modelling* 221 (2) (2010) 152–160.

- 559 [29] W. Appeltans, P. Bouchet, G. Boxshall, K. Fauchald, D. Gordon,  
560 B. Hoeksema, G. Poore, R. Van Soest, S. Stöhr, T. Walter, et al., World  
561 register of marine species, <http://www.marinespecies.org> (2011).
- 562 [30] V. Leen, B. Vanhoorne, W. Decock, A. Trias-Verbeek, S. Dekeyzer,  
563 S. Colpaert, F. Hernandez, World register of marine species, Book of.
- 564 [31] J. Grassle, The ocean biogeographic information system (obis):  
565 an on-line, worldwide atlas for accessing, modeling and map-  
566 ping marine biological data in a multidimensional geographic con-  
567 text, OCEANOGRAPHY-WASHINGTON DC-OCEANOGRAPHY  
568 SOCIETY- 13 (3) (2000) 5–7.
- 569 [32] D. Pelleg, A. W. Moore, X-means: Extending k-means with efficient  
570 estimation of the number of clusters., in: ICML, 2000, pp. 727–734.
- 571 [33] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, A density-based algorithm  
572 for discovering clusters in large spatial databases with noise., in: Kdd,  
573 Vol. 96, 1996, pp. 226–231.
- 574 [34] J. MacQueen, et al., Some methods for classification and analysis of mul-  
575 tivariate observations, in: Proceedings of the fifth Berkeley symposium  
576 on mathematical statistics and probability, Vol. 14, California, USA,  
577 1967, pp. 281–297.
- 578 [35] J. L. Bentley, Multidimensional binary search trees used for associative  
579 searching, Communications of the ACM 18 (9) (1975) 509–517.

- 580 [36] G. Schwarz, et al., Estimating the dimension of a model, *The annals of*  
581 *statistics* 6 (2) (1978) 461–464.
- 582 [37] G. Coro, A. Gioia, P. Pagano, L. Candela, A Service for Statistical  
583 Analysis of Marine Data in a Distributed e-Infrastructure, *Bollettino di*  
584 *Geofisica Teorica e Applicata* 54 (1) (2013) 68–70.
- 585 [38] G. Coro, L. Candela, P. Pagano, A. Italiano, L. Liccardo, Parallelizing  
586 the execution of native data mining algorithms for computational bi-  
587 ology, *Concurrency and Computation: Practice and Experience* (2014)  
588 n/a–n/doi:10.1002/cpe.3435.  
589 URL <http://dx.doi.org/10.1002/cpe.3435>
- 590 [39] J. Cohen, et al., A coefficient of agreement for nominal scales, *Educa-*  
591 *tional and psychological measurement* 20 (1) (1960) 37–46.
- 592 [40] J. L. Fleiss, Measuring nominal scale agreement among many raters.,  
593 *Psychological bulletin* 76 (5) (1971) 378.
- 594 [41] J. R. Landis, G. G. Koch, The measurement of observer agreement for  
595 categorical data, *biometrics* (1977) 159–174.
- 596 [42] I. Jolliffe, *Principal component analysis*, Wiley Online Library, 2002.
- 597 [43] G. Coro, L. Candela, gcube statistical manager: the algorithms, Tech-  
598 nical report, ISTI–CNR, technical report, 2014. (2014).

- 599 [44] G. Coro, gCube clustering analysis, algorithms code,  
600 <http://svn.research-infrastructures.eu/public/d4science/gcube/trunk/data->  
601 [analysis/EcologicalEngine/src/main/java/org/gcube/dataanalysis/ecoengine/clustering/](http://svn.research-infrastructures.eu/public/d4science/gcube/trunk/data-analysis/EcologicalEngine/src/main/java/org/gcube/dataanalysis/ecoengine/clustering/)  
602 (2014).
- 603 [45] i-Marine, i-Marine European Project, <http://www.i-marine.eu> (2011).

	A	IntraDO	InterDO	E	TR	TRMO
Cluster 1	85.3%	85.4%	33.9%	64.3%	35.4%	47.1%
Cluster 2	9.5%	12.4%	26.6%	26.4%	31.5%	37.5%
Cluster 3	4.8%	2.1%	21.4%	8.3%	23.4%	14.7%
Cluster 4	0.4%	0.1%	18.1%	1.0%	9.6%	0.6%

Table 1: Normalized distributions of the mean values of the variables in the X-Means clusters.

Sp. scientific name	A	IntraDO	InterDO	E	TR	TRMO	Cluster
<i>Sprattus sprattus</i>	7921.81	2779.67	0.44	0.031	0.44	0.39	1
<i>Trisopterus esmarkii</i>	5477.46	2502.11	0.44	0.027	0.45	0.44	1
<i>Gadus aeglefinus</i>	1680.20	8869.78	0.67	0.039	0.49	0.48	1
<i>Trachurus trachurus</i>	2067.49	1294.33	0.56	0.035	0.45	0.42	2
<i>Pollachius virens</i>	250.39	1433	0.44	0.013	0.43	0.37	2
<i>Platichthys flesus</i>	11.02	647.89	0.56	0.013	0.59	0.5	2
<i>Ammodytes lancea</i>	663.20	49.22	0.67	0.0036	0.26	0.1	3
<i>Mustelus asterias</i>	16.52	96.89	0.33	0.0046	0.38	0.21	3
<i>Scophthalmus rhombus</i>	2.58	82.33	0.56	0.010	0.4	0.17	3
<i>Pomatoschistus pictus</i>	38.17	2.67	0.33	0.00032	0.083	0	4
<i>Ciliata septentrionalis</i>	5.75	6.22	0.33	0.00076	0.1	0.0083	4
<i>Labrus bergylta</i>	0.07	6.56	0.33	0.00044	0.13	0.017	4

Table 2: Examples of vectors of parameters (with related clusters) for some of the species included in our benchmark dataset.



Cluster Number	Label	Definition
Cluster 1	Common	Frequent, widespread, high individual density
Cluster 2	Moderate Commonness	Moderately frequent, moderately widespread, medium individual density
Cluster 3	Moderate-Low Commonness	Poorly widespread, poorly-moderately frequent, low individual density
Cluster 4	Low Commonness	Localized, not frequent, very low individual density

Table 3: Interpretation of the X-Means clusters as classes of species commonness.

<b>Kappa values on 4 Clusters</b>		
	Expert 2	Clustering
Expert 1	0.24	<b>0.57</b>
Expert 2		0.48

<b>Kappa interpretation Fleiss/Landis–Koch</b>		
	Expert 2	Clustering
Expert 1	Poor/Slight	<b>Good/Moderate</b>
Expert 2		Good/Moderate

<b>Absolute Percentage of Agreement</b>		
	Expert 2	Clustering
Expert 1	46.5%	<b>67.4%</b>
Expert 2		61.4%

Table 4: Agreement with Kappa statistic and absolute percentage of agreement on the classification of species in four clusters: *Common*, *Moderate–Commonness*, *Moderate–Low Commonness*, *Low–Commonness*. The table in the middle reports interpretations for the Kappa values.

<b>Kappa values on Comm./Non-Comm. classes</b>		
	Expert 2	Clustering
Expert 1	0.34	0.39
Expert 2		<b>0.78</b>

<b>Kappa interpretation Fleiss/Landis–Koch</b>		
	Expert 2	Clustering
Expert 1	Marginal/Fair	Marginal/Fair
Expert 2		<b>Excellent/ Substantial</b>

<b>Absolute Percentage of Agreement</b>		
	Expert 2	Clustering
Expert 1	67.4%	69.8%
Expert 2		<b>92.9%</b>

Table 5: Agreement with Kappa statistic and absolute percentage of agreement on the classification of species in two clusters: *Common*, *Non-Common*. The table in the middle reports interpretations for the Kappa values.

<b>Kappa values on 2 aggregated Clusters</b>		
	Expert 2	Clustering
Expert 1	0.26	<b>0.67</b>
Expert 2		0.52

<b>Kappa interpretation Fleiss/Landis–Koch</b>		
	Expert 2	Clustering
Expert 1	Marginal/Fair	<b>Good/Substantial</b>
Expert 2		Good/Moderate

<b>Absolute Percentage of Agreement</b>		
	Expert 2	Clustering
Expert 1	67.4%	<b>83.7%</b>
Expert 2		75.7%

Table 6: Agreement with Kappa statistic and absolute percentage of agreement on the classification of species in two aggregated clusters: *Common and Moderate–Common vs. Moderate–Low and Low–Commonness*. The table in the middle reports interpretations for the Kappa values.

Response to Noise					
		Distribution of the original clusters on the newly found clusters			
Added noise	Found Clusters (C1, C2,...,Cn)	Cluster 1	Cluster 2	Cluster 3	Cluster 4
0.1%	4	100% C1	0% C1	0% C1	0% C1
		0% C2	100% C2	0% C2	0% C2
		0% C3	0% C3	100% C3	0% C3
		0% C4	0% C4	0% C4	100% C4
1%	4	100% C1	0% C1	0% C1	0% C1
		0% C2	100% C2	0% C2	0% C2
		0% C3	0% C3	100% C3	0% C3
		0% C4	0% C4	0% C4	100% C4
5%	4	100% C1	4% C1	0% C1	0% C1
		0% C2	96% C2	0% C2	0% C2
		0% C3	0% C3	91% C3	0% C3
		0% C4	0% C4	9% C4	100% C4
10%	3	70% C1	43% C1	17% C1	0% C1
		30% C2	48% C2	66% C2	14% C2
		0% C3	9% C3	18% C3	86% C3
50%	1	100% C1	100% C1	100% C1	100% C1

Table 7: Output of our clustering analysis in response to random noise added to the data. The results are reported with respect to an increasing percentage of added noise. The percentages indicate the distribution of the clusters associated to the clean data over the clusters found for the noisy data.

Variables influence on the clustering analysis					
Distribution of the original clusters on the newly found clusters					
Excluded variable	Found Clusters (C1, C2,...,Cn)	Cluster 1	Cluster 2	Cluster 3	Cluster 4
A	2	100% C1 0% C2	78% C1 22% C2	100% C1 0% C2	100% C1 0% C2
IntraDO	2	100% C1 0% C2	78% C1 22% C2	100% C1 0% C2	100% C1 0% C2
InterDO	5	100% C1	13% C1	0% C1	0% C1
		0% C2	87% C2	0% C2	0% C2
		0% C3	0% C3	61% C3	0% C3
		0% C4	0% C4	39% C4	29% C4
		0% C5	0% C5	0% C5	71% C5
E	1	100% C1	100% C1	100% C1	100% C1
TR	2	100% C1	70% C1	100% C1	100% C1
		0% C2	30% C2	0% C2	0% C2
TRMO	2	100% C1	30% C1	100% C1	100% C1
		0% C2	70% C2	0% C2	0% C2

Table 8: Modifications in the species clustering when one variable at time is excluded. The percentages indicate the distribution of the original clusters over the newly calculated clusters.

Influence of variables redefinitions on the clustering analysis					
Redefined variable	Found Clusters (C1, C2,...,Cn)	Distribution of the original clusters on the newly found clusters			
		Cluster 1	Cluster 2	Cluster 3	Cluster 4
$A'$ =n. of individuals	4	100% C1	0% C1	0% C1	0% C1
		0% C2	100% C2	0% C2	0% C2
		0% C3	0% C3	100% C3	0% C3
		0% C4	0% C4	0% C4	100% C4
$A''$ =n. of obs.	4	100% C1	0% C1	0% C1	0% C1
		0% C2	96% C2	0% C2	0% C2
		0% C3	4% C3	91% C3	0% C3
		0% C4	0% C4	9% C4	100% C4
$IntraDO'$ =avg. n. of obs. in datasets containing species obs.	4	100% C1	9% C1	0% C1	0% C1
		0% C2	91% C2	0% C2	0% C2
		0% C3	0% C3	100% C3	0% C3
		0% C4	0% C4	0% C4	100% C4
$InterDO'$ =n. of datasets containing species obs.	4	100% C1	0% C1	0% C1	0% C1
		0% C2	100% C2	0% C2	0% C2
		0% C3	0% C3	100% C3	0% C3
		0% C4	0% C4	0% C4	100% C4
$TR'$ =n. of months with obs.	4	100% C1	30% C1	0% C1	0% C1
		0% C2	70% C2	40% C2	0% C2
		0% C3	0% C3	60% C3	0% C3
		0% C4	0% C4	0% C4	100% C4
$TRMO'$ =n. of months with at least 10 obs.	4	100% C1	35% C1	0% C1	0% C1
		0% C2	65% C2	0% C2	0% C2
		0% C3	0% C3	100% C3	0% C3
		0% C4	0% C4	0% C4	100% C4
T=TRMO/TR (subst. to TR and TRMO)	4	100% C1	30% C1	0% C1	0% C1
		0% C2	70% C2	0% C2	0% C2
		0% C3	0% C3	61% C3	0% C3
		0% C4	0% C4	39% C4	100% C4
$A'$ , $IntraDO'$ , $InterDO'$ , $TR'$ , $TRMO'$	4	100% C1	8% C1	0% C1	0% C1
		0% C2	92% C2	0% C2	0% C2
		0% C3	0% C3	100% C3	0% C3
		0% C4	0% C4	0% C4	100% C4

Table 9: Modifications in the species clustering when variables are redefined in a slightly different way from our default definitions. The percentages indicate the distribution of the original clusters over the newly calculated clusters.

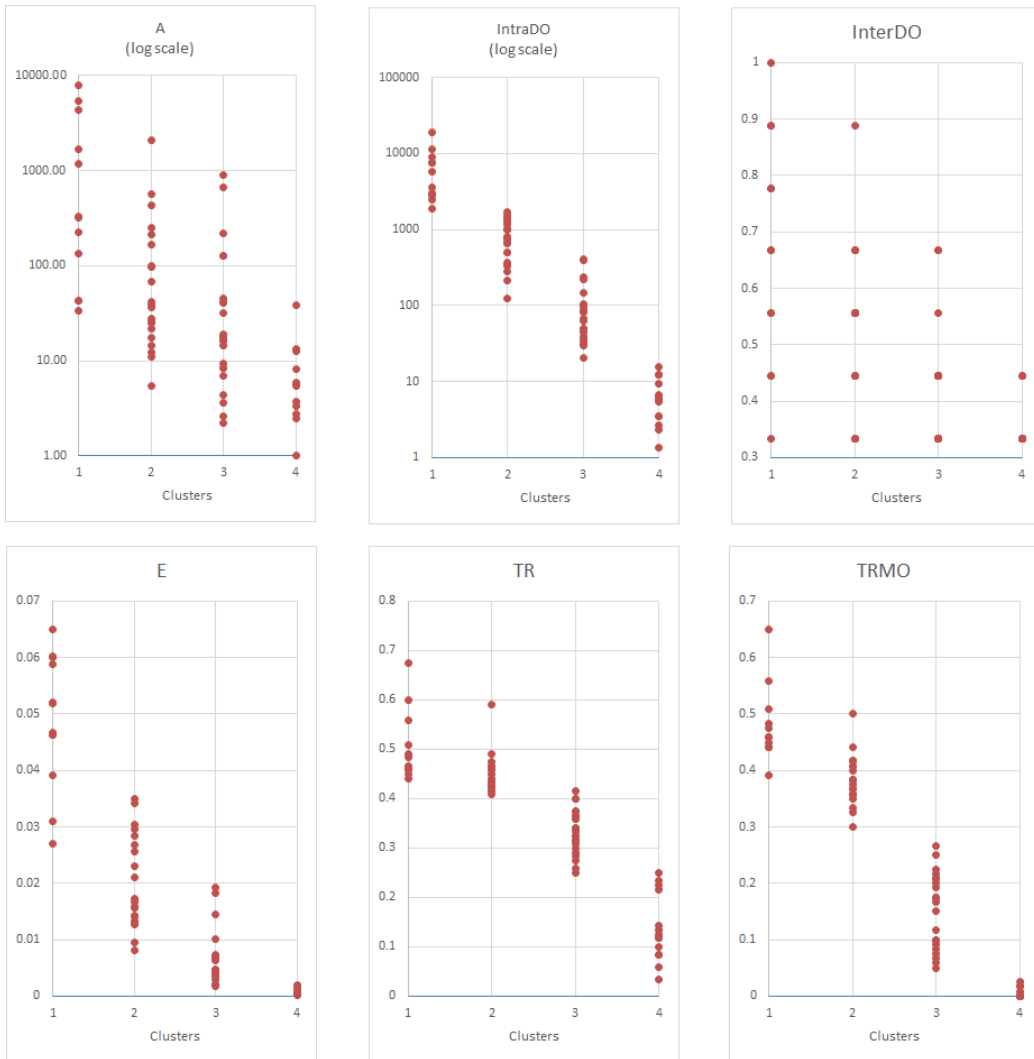


Figure 1: Distribution of the values of our variables over the four clusters identified by our model.



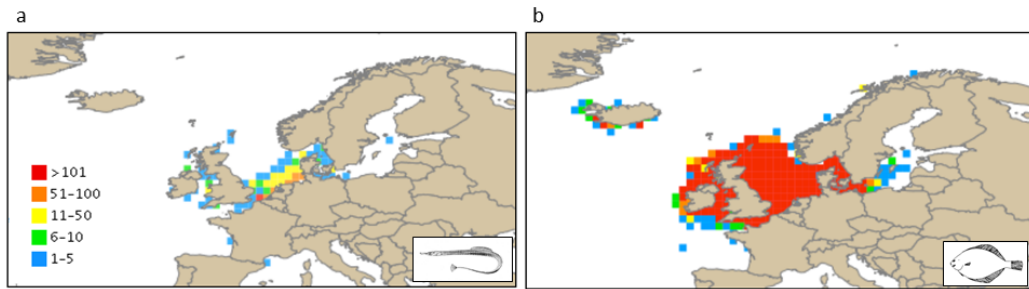


Figure 2: a. Representation of observation records from OBIS for *Syngnathus rostellatus*, aggregated at 0.5 degrees b. Representation of observation records from OBIS for *Limanda limanda*, aggregated at 0.5 degrees.