Ecological Modelling 312 (2015) 272–280

Contents lists available at ScienceDirect

Ecological Modelling

journal homepage: www.elsevier.com/locate/ecolmodel

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

Gianpaolo Coro[a,*], Thomas J. Webb[b], Ward Appeltans[c], Nicolas Bailly[d], André Cattrijsse[e], Pasquale Pagano[a]

[a] Istituto di Scienza e Tecnologie dell'Informazione "Alessandro Faedo" – CNR, Pisa, Italy
[b] Department of Animal & Plant Sciences, University of Sheffield, Sheffield S10 2TN, UK
[c] Intergovernmental Oceanographic Commission (IOC) of UNESCO, Oostende, Belgium
[d] WorldFish, Penang, Malaysia
[e] Vlaams Instituut voor de Zee (VLIZ), Oostende, Belgium

ARTICLE INFO

Article history:
Received 23 January 2015
Received in revised form 25 May 2015
Accepted 26 May 2015
Available online 17 June 2015

Keywords:
Species commonness
OBIS
Conservation biology
North Sea
Clustering
D4Science

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 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.

1. Introduction

The term "common species" refers intuitively to a species that is abundant in a certain area, widespread and at low risk of extinction. By consequence, "rare species" are less abundant and possibly threatened. Automatically detecting common and rare species, and how their status changes through time, is an important step in understanding the consequences of environmental change for ecosystem functioning. In particular, the abundance of a species in a community or ecosystem is a key indicator of its ecological role and ecosystem function therefore depends on the identities and relative numbers of common and rare species (Magurran, 2012). For instance, rare species may have unique functional traits (Mouillot et al., 2013) and make particular contributions to diversity (Mi et al., 2012). On the other hand, common species may underpin ecosystem function where they dominate in terms of biomass (Gaston and Fuller, 2008; Gaston, 2010, 2011). Both human activity and natural environmental change typically affect the relative abundances of species (Chapin et al., 2000). Monitoring changes in the relative abundance of species is straightforward when working on individual, well-monitored systems. However, anthropogenic-driven environmental change is affecting entire ecosystems, requiring large-scale ecological efforts (Kerr et al., 2007). One approach to monitor species commonness at large scale and in a certain time frame, is to perform meta-analyses on studies of multiple individual communities. This is useful for extracting general trends across multiple taxa (Dornelas et al., 2014). An alternative is to take advantage of the increasing availability of large-scale compilations of biodiversity data, such as the UK's National Biodiversity Network (NBN) (2014), the Global Biodiversity Information Facility (GBIF) (2014), or the Ocean Biogeographic Information System (OBIS) (IOCUNES, 2014). These compilations include millions of opportunistic records of the distributions of very large numbers of species, often across multiple decades. This temporal dimension

* Corresponding author. Tel.: +39 050 315 2978; fax: +39 050 621 3464.
E-mail addresses: coro@isti.cnr.it (G. Coro), t.j.webb@sheffield.ac.uk (T.J. Webb), w.appeltans@unesco.org (W. Appeltans), n.bailly@cgiar.org (N. Bailly), andre.cattrijsse@vliz.be (A. Cattrijsse), pagano@isti.cnr.it (P. Pagano).

http://dx.doi.org/10.1016/j.ecolmodel.2015.05.033
0304-3800/© 2015 Elsevier B.V. All rights reserved.

offers significant potential to track the relative commonness of species through time. However, it is difficult to extract robust estimates that are insensitive to changes and biases in sampling effort, from those heterogeneous and unstructured data sources (Isaac et al., 2014). The major issue is that it is hard to separate the signal of the actual relative commonness of a species in the system from the noise of sampling effort that varies in time and space, and in its taxonomic focus. For instance, a species may appear common across a given decade in a large dataset because there was at that time an intensive sampling programme targeting it. Its subsequent reduction in apparent abundance may simply reflect the end of that programme, rather than anything of ecological significance.

In this paper, we present a method to classify the degree of commonness of marine fish species in a certain area and time frame, using a large data collection of biodiversity data. In particular, we rely on the OBIS data collection and, for the purposes of methodological development, we focus on fish from the North Sea, a subset of 70 well-studied but unevenly sampled species. We use clustering analysis to automatically extract commonness classes from unstructured data and compare these classes with expert opinion. Reliable concordance between our method and experts, suggests that classifying commonness for less well-studied taxa or regions from data collections such as OBIS may be possible. We also assess the performance of our method in terms of (i) accuracy (using cross-validation), (ii) robustness to random noise in the data, (iii) dependency on the variables we chose to represent species commonness and (iv) dependency on our definition of these variables.

The paper is structured as follows: section 2 gives an overview on techniques for identifying species commonness. Section 3 describes the survey data we used. Section 4 reports the variables we defined to model the problem and describes our modelling approach. Section 5 reports an evaluation of the robustness of our method. It includes a comparison between our automatic classification and the classifications produced by two experts. Section 6 discusses the results, suggests possible usages of our technique and includes conclusive remarks.

2. Overview

Species commonness and rarity have been investigated in several scientific works. Most approaches derive species commonness from species abundance distributions (SADs) (Connolly et al., 2014; McGill et al., 2007). The intimate connection between abundance and commonness (or rarity) is widely recognized, even if an explicit definition of this dependency is unknown (Gaston, 2010). Approaches to model such dependency and to discover new correlated parameters, range from machine learning based approaches to explicit modelling. In this last case, models specify the role that each parameter has in defining species commonness. Searching for these parameters usually requires analyses by domain experts. For example, Preston (1948) analyses how abundance is distributed among species. He recognizes the importance of characteristics like (i) the total number of living individuals, (ii) the total number of individuals living at any instant on a given area, (iii) the ratio of the number of individuals with respect to another species, (iv) the number of observed individuals in different data collections. Some authors suggest that common species tend to be common everywhere, as reflected in a general positive relationship between local population density and regional distribution (Gaston et al., 2000; Blackburn et al., 2006; Webb et al., 2012; Hughes et al., 1346). These species also tend to remain common through time (Webb et al., 2007; Webb, 2012), with major changes in the rank-order of species commonness rather rare. In other studies, common species have been identified with species widely distributed on a territory, whereas rare species have been indicated as those in the Red List for the same territory. For example, using these definitions, Pearman and Weber (2007) detect spatial patterns for common species in Switzerland. In order to account for this heterogeneity of parameters, other works have promoted using standard measures and data to compare common and rare species (Bevill and Louda, 1999).

Unfortunately, no single satisfactory formal definition of species commonness and rarity has been found, especially using explicit modelling. Clustering analysis is a promising approach coming from machine learning techniques that may help to address this. This technique has been widely used for identifying classes of species characteristics. For example, clustering environmental properties has proven to be useful in detecting vegetation types (Dale et al., 2007), in modelling the coexistence of plants in agro-ecosystems (Debeljak et al., 2011) and in detecting new agro-ecosystems (Liu and Samal, 2002). Clustering analysis can also account for the lack of sampling uniformity in data collections, for example to group several species together when few data are available (Picard et al., 2010).

3. Data

Our model needs to be trained on species observation data. In order to identify the best training data, we searched for a dataset which was (i) sufficiently large and complex that relative commonness was not straightforward to ascertain but where (ii) the number of species was not too large and (iii) independent estimates of relative commonness were available from expert opinion. Points (ii) and (iii) restricted us to well-known species, with officially accepted scientific names available from the authoritative World Register of Marine Species (WoRMS) (Appeltans et al., 2011; Leen et al., 2008). In order to extract data, we consulted the Ocean Biogeographic Information System (OBIS) (Grassle, 2000). OBIS is the world's largest database on the diversity, distribution and abundance of all marine life. OBIS was initiated in 2000 by the Census of Marine Life and now runs under the auspices of UNESCO's Intergovernmental Oceanographic Commission. It currently provides free access to 40 million observations of 115,000 marine species, integrated from more than 1600 datasets provided by nearly 500 institutions worldwide. OBIS is an amalgam of many individual datasets from research projects, national monitoring programmes, museum collections and so on, targeting different taxa in different areas, often using different methods over different years. We limited our analysis on North Sea fish, because fish (Pisces[1]) represents 50% of all data in OBIS and the North Sea has relatively the highest amount of observations of all areas in the world. Thus, we extracted observation records from OBIS and defined the spatial boundaries of North Sea according to the International Hydrographic Organization (IHO) indications. Furthermore, we selected only species observed between 2000 and 2009, as OBIS is particularly rich of datasets and occurrence records for the North Sea in this period. This selection produced a list of 247 scientific species names, 70 of which had distinct and accepted species names according to WoRMS. We used this subset of 70 species from OBIS as a benchmark to develop and evaluate our method.

4. Method

Starting from the dataset described in Section 3, we used clustering analysis to automatically derive classes of commonness. The aim was also to search for a classification robust enough to account for sampling biases. Clustering analysis requires defining variables on the data. This section reports the steps of our analysis from the

[1] LSID: urn:lsid:marinespecies.org:taxname:11676.

definition of these variables to the selection and application of the clustering model.

4.1. Variables definition

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

The works reported in Section 2, suggest that factors related to abundance and extent are correlated with species commonness. On the other hand, we know that collections of observations can contain biases. In particular, non-uniform sampling in time of the observations affects the estimation of species extents. We decided to classify the degree of commonness of each species in our benchmark dataset on the time frame of one decade (2000–2009), and to produce one classification per species for the decade. The main reason is that we wanted to explore the robustness of the classification rather than producing an analysis of commonness trends. Thus, we took into account the rate of species observations in the decade. In particular, we considered the monthly observations of the species. This rate depends also on the datasets contained in the OBIS collection. A species that is contained in several datasets (each with a different survey scope) is likely to be often encountered in that area.

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

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

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

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

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

*Inter-dataset observations **(InterDO)***: fraction of datasets containing observation records for a species. This parameter accounts for the observation frequency of a species among the OBIS contributors:

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

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

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

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

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

*Time rate of many observations **(TRMO)***: fraction of months containing a significant number of observations. This is an alternative measure of the observation rate, which accounts for the months in which it was frequent to observe the species. Based on the values of species known to be common or rare, we calculated that 10 observations were a significant threshold in the 2000–2009 decade.

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

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

4.2. Clustering

Clustering analysis is a data mining technique which is able to group together numeric vectors, according to a certain similarity criterion. In the case of real valued vectors, similarity is usually measured in terms either of density or of Euclidean distances. In our case, we wanted to verify if clustering could extract classes of similarity related to species commonness. To this end, we selected two alternative clustering techniques, named X-Means (Pelleg and Moore, 2000) and DBScan (Ester et al., 1996). The former uses a distance-based approach, while the latter uses a density-based approach. We selected such algorithms because they automatically find the best number of clusters from the data.

DBScan is a density-based clustering algorithm. It searches for an optimal number of clusters on the basis of two parameters: *epsilon* and *min points*. The former is a distance threshold that defines the neighbourhood of a point (epsilon-neighbourhood), while the latter is the minimum number of points required to form a dense region. The DBSCAN algorithm starts selecting an arbitrary point. Then it takes the epsilon-neighbourhood of the point and, if this contains at least *min points* elements, it aggregates the points into a cluster. Otherwise, it assumes that this point could be later found in the epsilon-neighbourhood of another point (and thus added to the cluster of that point), and moves to another point. The process analyses all the points and creates density-connected clusters. For further details see Ester et al. (1996).

X-Means is a variant of the popular K-Means algorithm (MacQueen et al., 1967), which introduces several efficiency enhancements. An important difference with respect to K-Means is that the number of optimal clusters to search for is not specified *a priori*. Instead, it requires to set a minimum and a maximum number of clusters ($K_{\min}$ and $K_{\max}$) to search for. The X-Means algorithm starts from $K_{\min}$ and adds centroids as far as $K_{\max}$ is reached. At each step, the K-Means algorithm is run, which finds the best assignment of the vectors to the indicated number of clusters. K-Means indicates a score for this assignment, based on the distortion measure, i.e. the average squared distance of the points to their clusters centroids. The X-Means algorithm outputs the result of the K-Means that gave the best score, and consequently the best number of clusters. X-Means also adds efficiency enhancements to K-Means, using *kd*-trees (Bentley, 1975) and *blacklisting* to

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

	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%

support processing. Furthermore, at each step of the computation, the location of the centroids of the additional clusters is decided using the Bayesian Information Criterion (BIC) (Schwarz et al., 1978). For further details see Pelleg and Moore (2000).

We applied clustering analysis to our North Sea species benchmark. In our experiment, we searched for the clustering analysis detecting the lowest number of clusters and presenting a uniform distribution of the vectors in these clusters. We used the implementations running on the D4Science Statistical Manager Service (Coro et al., 2013, 2014), which hosts such procedures as-a-Service. We used several configurations for both the algorithms. Eventually, the best configuration for DBScan was obtained by setting *epsilon* = 100 and *minpoints* = 2. Unfortunately, this ended in 38 clusters and was not practical to use. On the other hand, the X-Means algorithm was executed by asking to search for a number of clusters between 1 and 50. Although the interval was large, the algorithm ended in only four clusters. The algorithm found an optimal separation of the vectors according to their relative Euclidean distance. Furthermore, we noticed that such clusters could be given an interpretation. The dataset and the results are available as Supplementary material of this paper.

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

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

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

5. Evaluation

5.1. Agreement with experts

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

We estimated the agreement between all the classifications using the absolute percentage of agreement, defined as the percentage of matching assignments. Furthermore, we also calculated Cohen’s Kappa (Cohen et al., 1960), which estimates the agreement between two evaluators with respect to purely random assignments. Cohen’s Kappa allows comparing complex classification tasks (e.g. with many classes) with simpler ones (e.g. dichotomous scenarios) where high agreement could have occurred by chance. Table 4 reports the Cohen’s Kappa values of the agreements, along with two different interpretations commonly used in literature (Fleiss, 1971; Landis and Koch, 1977). It is notable that in this experiment the absolute percentage agreement reflects the Kappa values. The values are symmetric, thus we report them once per pair of evaluators.

In order to give insight about the differences between the classifications assignments, we report the example of the lesser pipefish

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

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

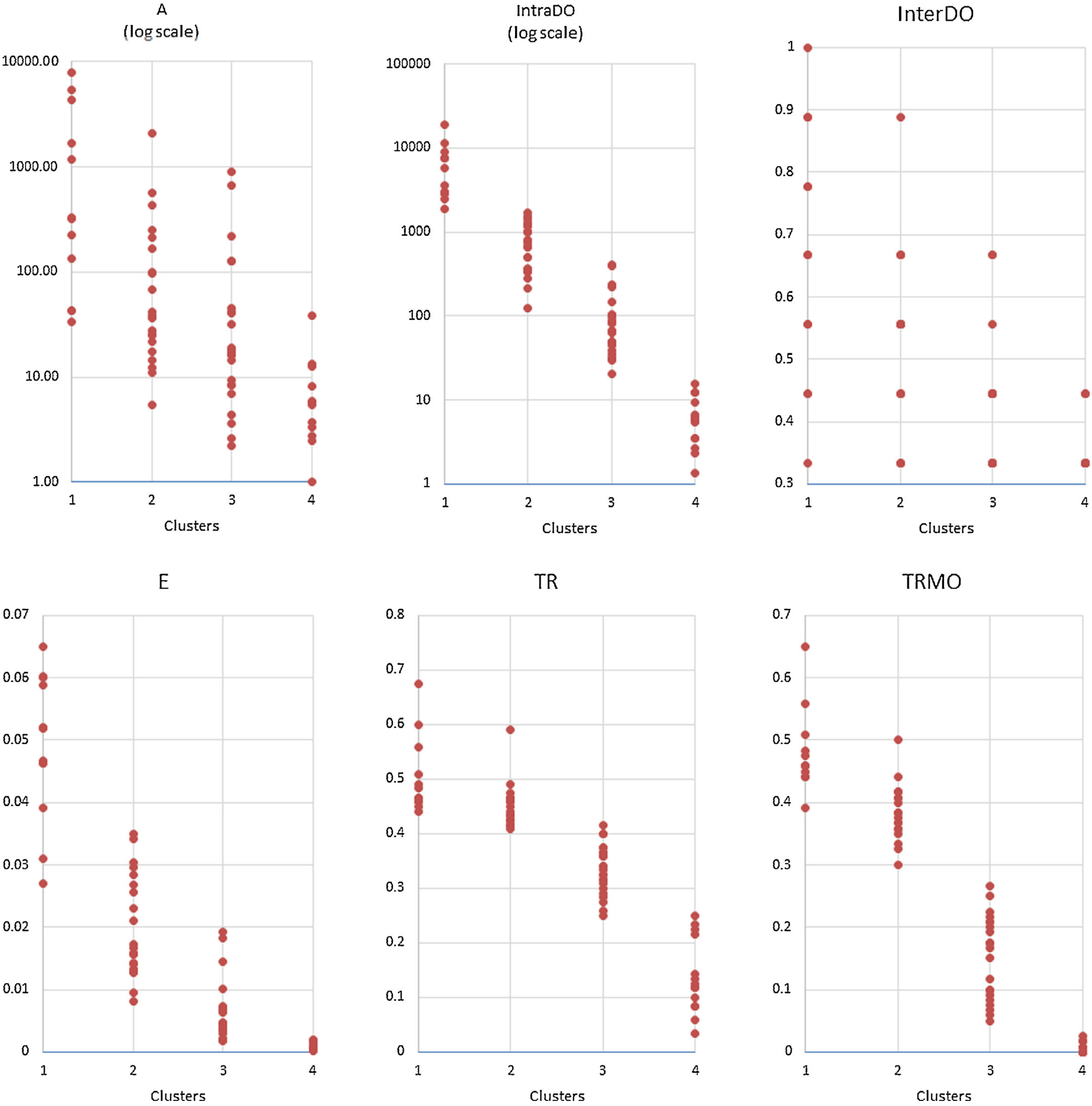

Fig. 1. Distribution of the values of our variables over the four clusters identified by our model.

Syngnathus rostellatus,[2] which Expert 2 and X-Means assign to *Moderate-Commonness*, and Expert 1 to *Common*. This species presents an *Abundance* (A) parameter value equal to 17.16, quite far from the 325.27 of the common dab *Limanda limanda*,[3] which is "Common" according to all the assignments. A significant difference is recorded also for the *IntraDO* values, which is 101.75 for the lesser pipefish and 24521.14 for the common dab. Indeed, *S. rostellatus* has a lower number of observation records for (407 records) with respect to *L. limanda* (171,648 records). This influences the behaviour of X-Means, but its classification can be still considered viable because it agrees with one of the two experts. Fig. 2 depicts the distribution of the observation records of the above species, aggregated at 0.5° resolution.

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

The disagreement between experts could be due to their different interpretation of the clusters descriptions. Thus, we investigated this aspect by aggregating the not *Common* clusters into a

[2] LSID: urn:lsid:marinespecies.org:taxname:127389.

[3] LSID: urn:lsid:marinespecies.org:taxname:127139.

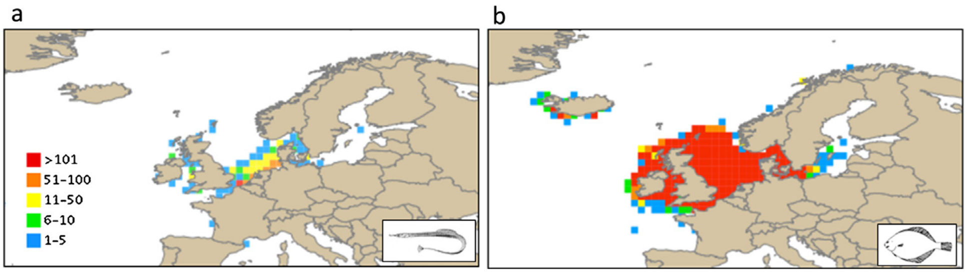

Fig. 2. (a) Representation of observation records from OBIS for *Syngnathus rostellatus*, aggregated at 0.5°. (b) Representation of observation records from OBIS for *Limanda limanda*, aggregated at 0.5°.

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.

	Expert 2	Clustering
Kappa values on four clusters		
Expert 1	0.24	**0.57**
Expert 2		0.48
Kappa interpretation Fleiss/Landis–Koch		
Expert 1	Poor/slight	**Good/moderate**
Expert 2		Good/moderate
Absolute percentage of agreement		
Expert 1	46.5%	**67.4%**
Expert 2		61.4%

The results highlighted in bold indicate the best agreement scores and interpretations in each agreement calculation.

generic *Non-Common* cluster. Table 5 reports the evaluation in this case. The agreement between Expert 2 and clustering is excellent, while the aggregation introduces misalignment between Expert 1 and clustering. This is due to a general tendency by Expert 1 to classify more in the *Moderate-Commonness* class.

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

These experiments highlight that even changing the definition of the clusters, there is a sensible agreement between experts and clustering. This indicates reliability of the automatic classification. It is notable that the variables used by the clustering analysis are likely to be affected by biases, especially when the species is poorly reported in time and is rarely reported by the OBIS contributors. Clustering accounts for the lack of information of some variables, because it compensates with information from the other variables. This comes out from the variables combination made by the Euclidean distances and by the subsequent optimization process. Furthermore, producing classes of commonness (instead of commonness scores) hides fine-grain differences between the vectors.

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.

	Expert 2	Clustering
Kappa values on comm./non-comm. classes		
Expert 1	0.34	0.39
Expert 2		**0.78**
Kappa interpretation Fleiss/Landis–Koch		
Expert 1	Marginal/fair	Marginal/fair
Expert 2		**Excellent/substantial**
Absolute percentage of agreement		
Expert 1	67.4%	69.8%
Expert 2		**92.9%**

The results highlighted in bold indicate the best agreement scores and interpretations in each agreement calculation.

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.

	Expert 2	Clustering
Kappa values on two aggregated clusters		
Expert 1	0.26	**0.67**
Expert 2		0.52
Kappa interpretation Fleiss/Landis–Koch		
Expert 1	Marginal/fair	**Good/substantial**
Expert 2		Good/moderate
Absolute percentage of agreement		
Expert 1	67.4%	**83.7%**
Expert 2		75.7%

The results highlighted in bold indicate the best agreement scores and interpretations in each agreement calculation.

5.2. Performance evaluation

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

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

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.

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

identified clusters were still the ones of Table 3. We added white noise directly to our variables and Table 7 reports the results: a 10% noise level means that we randomly added or subtracted up to the 10% of a variable value. Referring to Table 7, up to 1% of noise there is no change in the clustering and even at 5% the clusters are very similar to the ones without noise, because most of the species in the original ("clean" data) clusters are found in the corresponding newly found clusters. The number of clusters changes when 10% of noise is reached, but at this level the newly found clusters have still correspondence with the original clusters. For example, the species belonging to the original cluster 1 are largely included in the newly found cluster 1. The original cluster 2 corresponds to both the new clusters 1 and 2, whereas the original clusters 3 and 4 correspond to the new clusters 2 and 3, respectively. Over 10% of noise the original clusters are no more recognizable. It is our opinion that this limit is a reasonable indicator of robustness to noise. It is remarkable, in fact, that our data are already biased and the white noise only adds more bias.

As additional step, we evaluated the influence of each variable on the clustering analysis. Table 8 reports the results of the clustering analysis when we exclude one variable at time. The number of clusters changes and the identity of the original clusters is lost in most of the cases. It is notable that when *InterDO* is missing, the number of clusters is overestimated. In the other cases, the clustering is very simplistic and does not allow easy semantic interpretations. In particular, clusters 1, 3 and 4 are merged together, which means that common and uncommon species are mixed up. These changes indicate that all the variables have an important role (i.e. carry a remarkable amount of information) in the definition of the clusters of Table 3. Our definitions are related to indicators taken from other studies and come from expert opinion (see Section 4.1). This analysis confirms that they all have a key role in producing species commonness classes that agree with expert opinion.

As final step, we checked if the commonness classes depend on our definitions of the variables (see Section 4.1). Table 9 reports how the results of the clustering analysis change when the variables definitions are slightly altered. The new definitions in Table 9 still include information that is correlated to the original definitions. For example, in one of the experiments we redefined *A* as

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.

Variables influence on the clustering analysis					
Excluded variable	Found clusters (C1, C2,. . ., C*n*)	Distribution of the original clusters on the newly found clusters			
		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 0% C2 0% C3 0% C4 0% C5	13% C1 87% C2 0% C3 0% C4 0% C5	0% C1 0% C2 61% C3 39% C4 0% C5	0% C1 0% C2 0% C3 29% C4 71% C5
E	1	100% C1	100% C1	100% C1	100% C1
TR	2	100% C1 0% C2	70% C1 30% C2	100% C1 0% C2	100% C1 0% C2
TRMO	2	100% C1 0% C2	30% C1 70% C2	100% C1 0% C2	100% C1 0% C2

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.

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

the number of recorded individuals, without dividing for the number of observations. In another case, we defined one time variable as the ratio between the two time variables *TRMO* and *TR*. The last row of Table 9 reports the case in which all the variables definitions are altered. In all the cases, the clustering analysis identifies four clusters. Furthermore, the original clusters are recognizable in all the cases and sometimes the output coincides with the one of the original model. This means that the clustering analysis is flexible enough to exploit the information associated to the variables, even when the variables definitions change.

6. Discussion and conclusions

In this paper we have presented an approach to classify species commonness. We have trained our models on a dataset extracted from the OBIS data collection and focusing on North Sea fishes. The performance has been evaluated by comparing automatic assessments with the opinions of two experts. We have demonstrated that our process has good agreement with expert opinion although our analysed dataset contains sampling biases. We have further explored this robustness, by evaluating the effects that random noise in the data has on the classification. The results indicate that the model is reasonably robust in managing noise. Furthermore, we have used cross-validation to calculate the performance of our model in classifying species that had not been included in the training set. The performance indicates that the identified clusters are stable for the North Sea species. This gives suggestions about the possible generalization of our method. In fact, our clustering analysis is also applicable to other areas and large biodiversity data collections. Applying our method to other regions than North Sea requires the model to be trained on new data. Indeed, we conducted the same analysis on 222 species from OBIS at global scale. Also in this case, we found an optimal separation into four clusters[4] having the same percentage distributions as in Table 1. This result indicates that our classification could be valid for other areas too, but validating this hypothesis requires further investigation and much more effort in terms of experts' reviews. We will address this issue in future experiments.

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

Finally, we have demonstrated also that our set of variables contains a sufficient amount of information to identify four reliable commonness classifications. Using a lower number of variables would produce less refined classifications and less clusters (see Table 8). This is a remarkable property, since we defined the variables based on interactions with ecology and data experts (i.e. not using automatic data selection Jolliffe, 2002). This may suggest that our variables are ecologically meaningful, i.e. they are really correlated to species commonness.

[4] The complete classification is available on the D4Science e-Infrastructure for consultation: http://goo.gl/TYuD6P

From our analysis, new biodiversity and ecosystem indicators could be identified and this will be part of our future investigations. For example, using our method a species could be found, today, to be "less common" in a certain area with respect to a previous time period. This could indicate a change of the ecosystem in that area or that the species has been overfished. Our method could also be a way to reconcile the opinions of different experts about the commonness of a set of species. For example, it could be used as a supporting tool for biologists, who would rely on an "external" opinion when discussing about species commonness. Furthermore, classifying commonness for fishes in a well-studied region is a first step towards working on less known taxa in other regions.

Our experiments highlight the intrinsic difficulty of the problem, but the proposed technique represents a step forward in classifying species commonness and in understanding which factors are related to this concept. A data provider like OBIS could embed such method to alert a user about the possible commonness of a species in a certain area. In this context, we are planning to build an interface allowing a user to select an IHO area and a time rage, and to retrieve the species possibly classified as *Common* or *Moderately Common*. Currently, our clustering technique is released as software (Coro and Candela, 2014; Coro, 2014) inside the i-Marine e-infrastructure (i-Marine, 2011), which grants free access to statistics about the OBIS database and allows sharing datasets, biological analyses and experimental results.

Acknowledgements

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

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ecolmodel.2015.05.033

References

Appeltans, W., Bouchet, P., Boxshall, G., Fauchald, K., Gordon, D., Hoeksema, B., Poore, G., Van Soest, R., Stöhr, S., Walter, T., et al., 2011. World Register of Marine Species. http://www.marinespecies.org

Bentley, J.L., 1975. Multidimensional binary search trees used for associative searching. Commun. ACM 18 (9), 509–517.

Bevill, R., Louda, S., 1999. Comparisons of related rare and common species in the study of plant rarity. Conserv. Biol. 13 (3), 493–498.

Blackburn, T.M., Cassey, P., Gaston, K.J., 2006. Variations on a theme: sources of heterogeneity in the form of the interspecific relationship between abundance and distribution. J. Anim. Ecol. 75 (6), 1426–1439.

Chapin III, F.S., Zavaleta, E.S., Eviner, V.T., Naylor, R.L., Vitousek, P.M., Reynolds, H.L., Hooper, D.U., Lavorel, S., Sala, O.E., Hobbie, S.E., et al., 2000. Consequences of changing biodiversity. Nature 405 (6783), 234–242.

Cohen, J., et al., 1960. A coefficient of agreement for nominal scales. Educ. Psychol. Meas. 20 (1), 37–46.

Connolly, S.R., MacNeil, M.A., Caley, M.J., Knowlton, N., Cripps, E., Hisano, M., Thibaut, L.M., Bhattacharya, B.D., Benedetti-Cecchi, L., Brainard, R.E., et al., 2014. Commonness and rarity in the marine biosphere. Proc. Natl. Acad. Sci. U. S. A. 111 (23), 8524–8529.

Coro, G., Candela, L., 2014. gCube statistical manager: the algorithms, Technical Report, ISTI-CNR.

Coro, G., Gioia, A., Pagano, P., Candela, L., 2013. A service for statistical analysis of marine data in a distributed e-infrastructure. Boll. Geofis. Teor. Appl. 54 (1), 68–70.

Coro, G., Candela, L., Pagano, P., Italiano, A., Liccardo, L., 2014. Parallelizing the execution of native data mining algorithms for computational biology. Concurr. Comput.: Pract. Exp., http://dx.doi.org/10.1002/cpe.3435

Coro, G., 2014. gCube clustering analysis, algorithms code. http://svn.research-infrastructures.eu/public/d4science/gcube/trunk/data-analysis/EcologicalEngine/src/main/java/org/gcube/dataanalysis/ecoengine/clustering/

Dale, M.B., Dale, P., Tan, P., 2007. Supervised clustering using decision trees and decision graphs: an ecological comparison. Ecol. Model. 204 (1), 70–78.

Debeljak, M., Squire, G.R., Kocev, D., Hawes, C., Young, M.W., Džeroski, S., 2011. Analysis of time series data on agroecosystem vegetation using predictive clustering trees. Ecol. Model. 222 (14), 2524–2529.

Dornelas, M., Gotelli, N.J., McGill, B., Shimadzu, H., Moyes, F., Sievers, C., Magurran, A.E., 2014. Assemblage time series reveal biodiversity change but not systematic loss. Science 344 (6181), 296–299.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In: KDD, vol. 96, pp. 226–231.

Fleiss, J.L., 1971. Measuring nominal scale agreement among many raters. Psychol. Bull. 76 (5), 378.

Gaston, K.J., Fuller, R.A., 2008. Commonness, population depletion and conservation biology. Trends Ecol. Evol. 23 (1), 14–19.

Gaston, K.J., Blackburn, T.M., Greenwood, J.J., Gregory, R.D., Quinn, R.M., Lawton, J.H., 2000. Abundance–occupancy relationships. J. Appl. Ecol. 37 (s1), 39–59.

Gaston, K.J., 2010. Valuing common species. Science 327 (5962), 154–155, http://dx.doi.org/10.1126/science.1182818

Gaston, K.J., 2011. Common ecology. Bioscience 61 (5), 354–362.

Global Biodiversity Information Facility (GBIF), gbif.org, 2014.

Grassle, J., 2000. The ocean biogeographic information system (obis): an on-line, worldwide atlas for accessing, modeling and mapping marine biological data in a multidimensional geographic context. Oceanography 13 (3), 5–7.

Hughes, T., Bellwood, D., Connolly, S., Cornell, H., Karlson, R., 2014. Double jeopardy and global extinction risk in corals and reef fishes. Curr. Biol. 24 (24), 2946–2951, http://dx.doi.org/10.1016/j.cub.2014.10.037, http://www.sciencedirect.com/science/article/pii/S0960982214013463

i-Marine, 2011. i-Marine European Project. http://www.i-marine.eu

Intergovernmental Oceanographic Commission (IOC) of UNESCO, 2014. The Ocean Biogeographic Information System. http://www.iobis.org

Isaac, N.J., Strien, A.J., August, T.A., Zeeuw, M.P., Roy, D.B., 2014. Statistics for citizen science: extracting signals of change from noisy ecological data. Methods Ecol. Evol. 5 (10), 1052–1060.

Jolliffe, I., 2002. Principal Component Analysis. Wiley Online Library.

Kerr, J.T., Kharouba, H.M., Currie, D.J., 2007. The macroecological contribution to global change solutions. Science 316 (5831), 1581–1584.

Landis, J.R., Koch, G.G., 1977. The measurement of observer agreement for categorical data. Biometrics, 159–174.

Leen, V., Vanhoorne, B., Decock, W., Trias-Verbeek, A., Dekeyzer, S., Colpaert, S., Hernandez, F., 2008. World Register of Marine Species, Book of.

Liu, M., Samal, A., 2002. A fuzzy clustering approach to delineate agroecozones. Ecol. Model. 149 (3), 215–228.

MacQueen, J., et al., 1967. Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 14, California, USA, pp. 281–297.

Magurran, A.E., 2012. Biodiversity in the context of ecosystem function. In: Solan, M., Aspden, R.J., Paterson, D.M. (Eds.), Marine Biodiversity & Ecosystem Functioning – Frameworks, Methodologies and Integration. Oxford University Press, Great Clarendon Street, Oxford, pp. 16–23.

McGill, B.J., Etienne, R.S., Gray, J.S., Alonso, D., Anderson, M.J., Benecha, H.K., Dornelas, M., Enquist, B.J., Green, J.L., He, F., et al., 2007. Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework. Ecol. Lett. 10 (10), 995–1015.

Mi, X., Swenson, N.G., Valencia, R., Kress, W.J., Erickson, D.L., Pérez, A.J., Ren, H., Su, S.-H., Gunatilleke, N., Gunatilleke, S., et al., 2012. The contribution of rare species to community phylogenetic diversity across a global network of forest plots. Am. Nat. 180 (1), E17–E30.

Mouillot, D., Bellwood, D.R., Baraloto, C., Chave, J., Galzin, R., Harmelin-Vivien, M., Kulbicki, M., Lavergne, S., Lavorel, S., Mouquet, N., et al., 2013. Rare species support vulnerable functions in high-diversity ecosystems. PLOS Biol. 11 (5), e1001569.

National Biodiversity Network (NBN), nbn.org.uk, 2014.

Pearman, P.B., Weber, D., 2007. Common species determine richness patterns in biodiversity indicator taxa. Biol. Conserv. 138 (1), 109–119.

Pelleg, D., Moore, A.W., 2000. X-means: extending k-means with efficient estimation of the number of clusters. In: ICML, pp. 727–734.

Picard, N., Mortier, F., Rossi, V., Gourlet-Fleury, S., 2010. Clustering species using a model of population dynamics and aggregation theory. Ecol. Model. 221 (2), 152–160.

Preston, F.W., 1948. The commonness, and rarity, of species. Ecology 29 (3), 254–283.

Schwarz, G., et al., 1978. Estimating the dimension of a model. Ann. Stat. 6 (2), 461–464.

Webb, T.J., Noble, D., Freckleton, R.P., 2007. Abundance–occupancy dynamics in a human dominated environment: linking interspecific and intraspecific trends in British farmland and woodland birds. J. Anim. Ecol. 76 (1), 123–134.

Webb, T.J., Freckleton, R.P., Gaston, K.J., 2012. Characterizing abundance–occupancy relationships: there is no artefact. Global Ecol. Biogeogr. 21 (9), 952–957.

Webb, T.J., 2012. Marine and terrestrial ecology: unifying concepts, revealing differences. Trends Ecol. Evol. 27 (10), 535–541.