Benthos_greater_North_Sea

Peter M.J. Herman

20-6-2020

Introduction

The large databases of EMODNET Biology only store confirmed presences of species. However, when mapping species distribution, it is also important where the species did not occur: there is at least as much information in absences as in presences. Inferring absences from presence-only databases is difficult and always involves some guesswork. In this product we have used as much meta-information as possible to guide us in inferring absences. There is important meta-information at two different levels: the level of the data set, and the level of the species. Datasets can contain implicit information on absences when they have uniformly searched for the same species over a number of sample locations. Normally, if the species would have been present there, it would have been recorded. Other datasets, however, are not informative at all about absences. Typical examples are museum collections. The fact that a specimen is found at a particular place confirms that it lived there, but does not give information on any other species being present or absent in the same spot. A difficulty is that some datasets have searched for a restricted part of the total community, e.g. only sampled shellfish but no worms. In this case, absence of a shellfish species is relevant, but absence of a worm is not. The dataset can only be used to infer absence for the species it has targeted. Here we implicitly assume that a dataset inventoring the endomacrobenthos, is targeting all species belonging to this functional group. Usually, the distinction can be made on the basis of the metadata. It is also helpful to plot the total number of species versus the total number of samples. Incomplete datasets have far less species than expected for their size, compared to 'complete' datasets. At the species level, taxonomic databases such as WoRMS give information on the functional group the species belongs to. This information is present for many species, but it is incomplete. However, we can use it in a step to select the most useful datasets. We could use it to restrict the species to be used in mapping, but since the data are incomplete that would cause some loss of interesting information.

Data selection and retrieval

The retrieval of data goes in three steps.

Step 1. Use functional group information to harvest potentially interesting datasets.

A query is performed for data on species known to be benthic (in WoRMS) and to occur in a number of different sea regions. This yields a large dataset with benthic data, but many of these data come from datasets that are not useful for our purpose. As an example, planktonic datasets contain many benthic animals, because larvae of benthic animals occur in the zooplankton (the so-called meroplankton). We cannot use the plankton datasets to infer anything about absence of benthos in the sea floor.

This procedure is executed in the script "requestData_step1.R".

```r
source("./scripts/requestData_step1.R")
```

Step 2. Get dataset information and select datasets.

From the dataset resulting from step 1 we harvest all potentially interesting datasets, that contain at least one benthic animal in the region of interest. We subsequently use the imis database with meta-information on the datasets, to list the meta-data of all these datasets. This results in the file ./data/derived_data/allDatasets.csv.

```r
source("./scripts/requestData_step2.R")
```

In this file we perform the (manual) selection of datasets to be used. We also qualify the datasets as either 'complete' or 'incomplete'. The result of this manual procedure is the file ./data/derived_data/allDatasets_selection.csv.

Step 3. Download by dataset.

In this step, we download the part of all these useful datasets that occur in the region of interest. For practical reasons this region is subdivided in many subregions - in that way the downloaded files are not too big and there is less risk of interruptions of the process. After download, all these files are recombined into one big datafile.

```
source("./scripts/requestData_step3.R")
```

Production of maps

For the production of maps, we define 'sampling events' as the ensembles of records that share time and place. We consider such events as one sample. For the incomplete datasets, we inventory what species they have targeted. Finally, for every species we determine whether or not it was present in all sampling events of all relevant datasets. This presence/absence information is written to the output file, together with the spatial location and the sampling date. The intention is to use this information to produce interpolated maps covering also the non-sampled space. As a first step in visualisation, we rasterize this information and show it in a map per species.

```
source("./scripts/produce_maps.R")
```