UNEP-WCMC Technical Briefing Note

December, 2015

Absence of evidence is not evidence of absence

Working with data deficiency in global biodiversity spatial datasets

Key messages

More data is unknown than known in both the terrestrial and marine realms, due to numerous technical, logistical and financial challenges in biodiversity data collection at a global scale. Data deficiencies are therefore a common feature within global datasets.

Users of global biodiversity datasets are often faced with uncertainty when interpreting data deficiencies: It is not always possible to infer whether the features themselves are absent, or whether the data required was absent. Cases of confirmed absence of biodiversity features are also rarely highlighted within datasets.

Data deficiency No feature mapped	
Feature Exists	Feature does not exist
False negative	True negative

Interpretative approaches to data deficiency (red: false value, green: true value)

The absence of data may coincide with areas on-the-ground where the feature of interest exists, and areas where it does not exist. This may lead to two interpretative scenarios (see above right):

- **The true negative**: No data is displayed in the area of interest because the biodiversity feature does not occur there.

- **The false negative**: No data is displayed in the area of interest, however the biodiversity feature does occur on site.

It is commonly misinterpreted that data deficiency (or data absence) is evidence that the feature is absent, but this is often not the case. Data deficiencies should be interpreted as "unknowns" with the potential for both presence and absence until further investigation or supplementary information is available to confirm whether a particular feature exists.

This principle also applies to cases of species categorized as *Data Deficient* on the IUCN Red List of Threatened Species[1]. Future research on these species may show that threatened classification is appropriate.

Practical guidance for data users

1. **Consider gaps in global biodiversity spatial datasets to be a lack of information or knowledge rather than an absence of the biodiversity feature on-ground.**
2. **Do not assume an area has been adequately surveyed for all biodiversity features, just because data exists for some features.**
3. **Look for regional, national or local data to supplement global datasets.**
4. **Apply the precautionary principle of assuming "unknown" status rather than "absent" status until ground-truthing demonstrates the feature is absent.**

[1] A taxon is *Data Deficient* if "there is inadequate information to make a direct, or indirect, assessment of its risk of extinction based on its distribution and/or population status. A taxon in this category may be well studied, and its biology well known, but appropriate data on abundance and/or distribution are lacking." (IUCN, 2012)

Introduction

Biodiversity data is collected by an increasing number of organisations for research and conservation purposes. Features of particular interest include the distribution of species of various threat status and the distribution of sensitive marine, coastal and terrestrial habitats.

A range of techniques are used to develop datasets at a global scale, ranging from remote sensing of large areas of the world, to the amalgamation of multiple site-scale datasets. Despite advances in data collection techniques, the mapping of biodiversity features at a global scale is complicated by the large extent to cover, the remoteness and inaccessibility of particular areas of interest such as the deep sea or polar regions, and the mobile nature of species – one of the key units of biodiversity.

The increasing use of spatial datasets in decision-making across sectors has increased the demand for high quality datasets. The ideal dataset is one that is precisely current, perfectly accurate and completely comprehensive.

For lack of such a dataset, users need to be aware of the limitations of the data they use. While most of the data are likely to be accurate, some may be falsely mapped (error), and some may be absent (data deficiency).

This briefing note explores possible causes and issues associated with data deficiency. We encourage data users to consider the possibility of a feature occurring on site, despite not being mapped.

Why is data deficiency a concern?

Biodiversity data is often used by the conservation community to infer areas of high biodiversity value and determine a conservation response.

Global biodiversity spatial datasets are also used by the private sector to assess the operational risks and potential impacts associated with activities. By identifying the distribution of features of high biodiversity value, these datasets support the development of appropriate impact mitigation approaches.

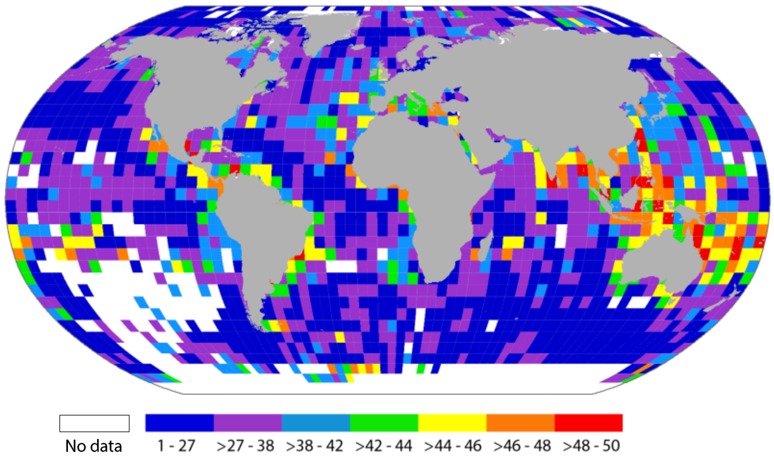


Figure 1: Example of data deficiency in a raster dataset. Figure 1 represents Hurlbert's index of biodiversity in the marine realm (OBIS, 2014). The dataset is derived from a global census of marine life and represents the expected number of species given 50 on-site records. The index was not calculated for grid cells where fewer than 50 records are present in the underlying data. These are recorded as 'No data'.

Data deficiencies add uncertainty to the identification of biodiversity features of interest, and therefore add limitations to these assessments.

Recognising data deficiency

Global biodiversity spatial datasets are most commonly provided in vector and raster formats. Data deficiency appears differently in both formats.

Vector-based datasets represent data as geometrical shapes, such as points, lines and polygons. Examples include polygons outlining the boundaries of protected areas, point features indicating the location of species sightings (e.g. cold-water corals), or line features mapping rivers. Blank spaces between features in a vector dataset are considered data deficient areas, because they neither map a feature, nor confirm its absence.

Raster-based datasets consist of a grid of cells organised in rows and columns where each cell is attributed a value representing a characteristic of that location. Rasters are commonly used to represent the variation of environmental variables measured through remote-sensing, derived from global spatial models, or developed as a biodiversity metric. Common examples are maps of species richness, land cover data, or digital elevation models. In a raster dataset, groups of cells may share a value recording the location of a data deficiency. These are locations where a variable could not be calculated (or measured) with confidence, because the underlying data (or the methodology) was inadequate, unavailable or insufficient (**Figure 1**). The value may state 'No data'[2], 'No data available', 'Insufficient data' or 'Data deficient'.

Interpreting data deficiency

The terminology used in this briefing note to explain data deficiencies is borrowed from accuracy assessment methodologies used to validate classifications in remotely sensed spatial data. These evaluate how often the information provided on a spatial map accurately represents the ground truth.

Four scenarios can be distinguished based on a conceptual model relating the existence of the feature on the ground, to its presence on the map (**Figure 2** and **Figure 3**). Data deficient areas in the dataset may coincide with areas where the feature does not exist on the ground (True negative) or with areas where the feature is actually present on-ground but not recorded in the dataset (False negative).

A high quality dataset aims to minimize false negatives, and maximise true negatives. For example, the spatial information for species listed on the IUCN Red List does not provide information on absence, however species range maps are limited to the habitat

[2] Noting that ArcGIS also uses the default value 'No Data' to identify areas located outside of the extent of the raster dataset (for raster datasets whose extent is not global).

suitable for a species. While the database does not provide evidence of absence, the nature of the methodology for mapping species ranges increases the likelihood of the true negative scenario.

Nonetheless, it is conceptually incorrect to assume that all data deficient areas are evidence that a feature is absent. The precautionary principle warrants treating data deficiencies as unknowns until further investigation or supplementary information is available to confirm the presence or absence of a particular feature.

Figure 2: Possible outcomes when relating features occurring on the ground, with features occurring in a spatial dataset. Absence of data may represent false negatives, as well as true negatives. The terms “Errors of commission” and “Error of omission” are used in accuracy assessments of spatial data, and frequently used in the context of ground-truthing data.

	Feature exists	Feature does not exist
Feature mapped	True positive	False positive (aka error of commission)
No feature mapped	False negative (aka error of omission)	True negative

Red : false values
Green : true values

Data deficiency

The True negative and evidence of absence

No data is displayed in the area of interest because the biodiversity feature does not occur there.

A true negative occurs in situations where there is a known absence of a feature on the ground and this is reflected in the dataset by an empty area or no data points.

In rare cases, evidence of absence is recorded in a dataset (**Figure 3**). The terminology may refer to ‘proven’, ‘identified’ or ‘confirmed’ absence of biodiversity features. Such information classifies as an identified true negative, and is commonly derived from habitat suitability mapping in the case of global biodiversity spatial datasets.

The False negative

No data is displayed in the area of interest, however the biodiversity feature does occur on site.

False negatives are also known as errors of omission. There are several reasons why a false negative could occur, falling into two main categories[3].

A. No research was completed in the area.

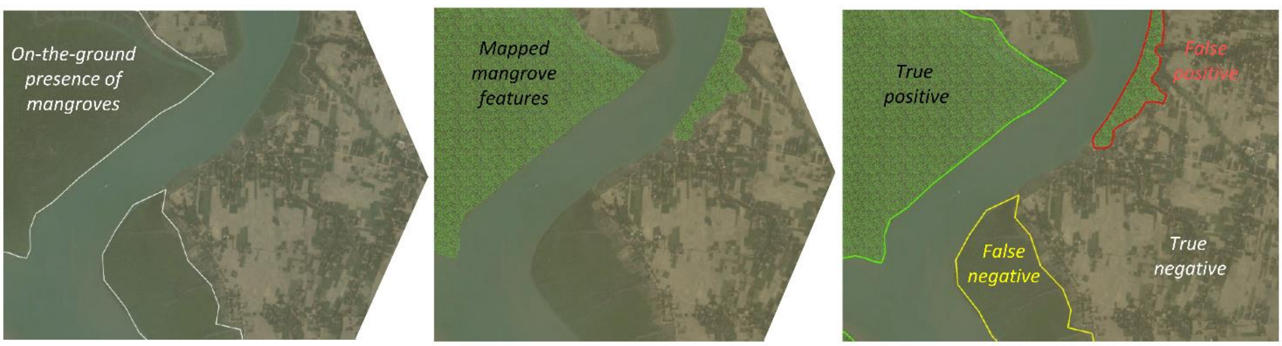
Much of the world remains unexplored, in particular regions which are hard to access and therefore less likely to have been visited by researchers collating data on biodiversity features. Absence of data is common in the deep ocean, remote polar regions and former conflict zones.

B. Data was collected in the area, but the feature is not recorded, although present in reality.

Potential reasons include:

- The feature is rare or elusive. The feature may not have been detected despite best efforts to sample it.
- The feature was seen but not recorded, because it was attributed an incorrect value. Species are commonly mis-identified, especially in cases where there is debate over the taxonomy of a particular species. This can also lead to cases of false positives (a.k.a error of commission).
- The spatial distribution of the feature has changed over time, as a significant time lag may occur between the collection of the data and the publication of the spatial dataset.
- The sampling methodology, or data gathering strategy was inadequate. Common reasons for this include sampling at the wrong time of day (for species showing diurnal or nocturnal movements), at the wrong time of year (for species showing seasonal migrations), or sampling at the wrong location (for species with a patchy distribution).
- The biodiversity feature was not targeted. To date, it is thought that 91% of the estimated 2.2 million species living in the oceans have not yet been described[4]. Researchers might therefore have conducted a survey for one feature, and not recorded the presence of other non-target features. These could include new and unknown species (sometimes referred to as *unknown unknowns*).

Figure 3: Fictional example of a spatial dataset outlining mangrove extent, and illustrating scenarios identified in Figure 2. The true extent of mangroves is visible as dark patches in the underlying satellite imagery, which underwent ground-truthing.

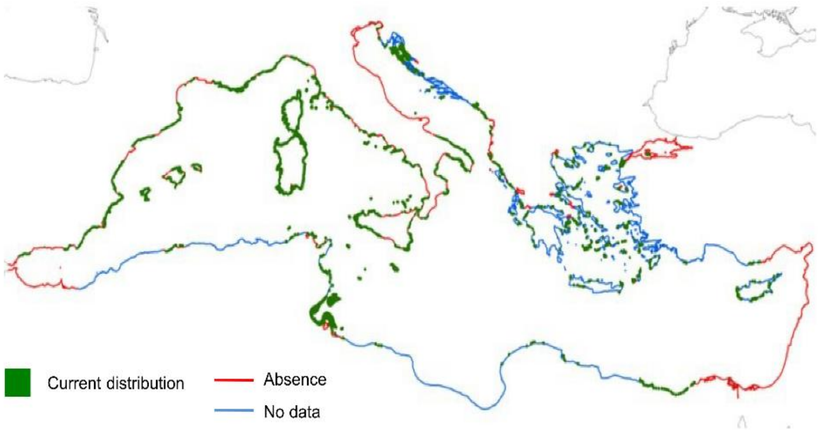


[3] More information in section 4 "Data challenges, gaps and limitations in the marine environment" of the *Manual of Marine and Coastal Datasets of Biodiversity Importance* (Martin et al., 2014) available on the Proteus website.

[4] Mora at al. (2011)

- The collected data was insufficient to confidently attribute a value to the dataset. A threshold of certainty needs to be exceeded when attributing a value within a dataset. Uncertainty can occur when identifying a new species, or when generating data using modelling techniques for which insufficient source data is available. Results presented in **Figure 1** for instance, are not provided for cases where fewer than 50 records are available at a particular location.
- The data are not available to those developing the global dataset. For example the data collected by companies as part of impact assessments or ongoing monitoring are not always made available to the scientific community for inclusion in other projects.

Figure 4: Distribution of *Posidonia oceanica* seagrass across the Mediterranean Sea, where presence ("current distribution"), data gaps ("no data") as well as absence ("absence") are shown. Figure credits: Belluscio et al., 2013.



Working with data deficiency

The precautionary principle warrants treating data deficiencies as unknowns until further investigation or supplementary information is available to determine whether the absence of data represents a true or false negative regarding the presence of the feature on-ground.

Areas of data deficiency may be addressed by accessing complementary regional or national datasets or by gathering new ground-truthing data. Observational data are frequently gathered by local researchers, but may not have been shared on an online platform or published in the grey or peer-reviewed literature. Incorporating such information may pose new challenges due to differences in how the data is collected, collated, verified and recorded. However, such observations may generate valuable evidence of absence in the area of interest.

"Data Deficient" in the IUCN Red List of Threatened Species.

The term Data Deficient has a particular definition in the IUCN Red List, which differs from the explanation of data deficiency described in this briefing note.

The IUCN Red List of Threatened Species is an inventory of the global conservation status of plant and animal species, using a set of scientific criteria to evaluate the extinction risk of species, and allocate one of eight categories of threat to the species assessed. One of these categories is named Data Deficient.

A taxon is categorized as Data Deficient, if "there is inadequate information to make a direct, or indirect, assessment of its risk of extinction based on its distribution and/or population status. A taxon in this category may be well studied, and its biology well known, but appropriate data on abundance and/or distribution are lacking" (IUCN, 2012).

While Data Deficient is a distinct and defined category in the Red List, data deficiency as described in this briefing note is also inherent to the Red List spatial data, as it is to all global biodiversity spatial datasets.

References

Belluscio A, Panayiotidis P, Gristina M, Knittweis L, Pace M, Telesca L, Criscoli A, Apostolaki E, Gerakaris V, Fraschetti S, Spedicato M, Lembo G, Salomidi M, Mifsud R, Fabi G, Badalamenti F, Garofalo G, Alagna A, Ardizzone G, Martin C, Valavanis V (2013) Task 1.1 Seagrass beds distribution along the Mediterranean coasts. In: Giannoulaki M, Belluscio A, Colloca F, Fraschetti S, Scardi M, Smith C, Panayotidis P, Valavanis V, Spedicato M (eds) Mediterranean Sensitive Habitats (MEDISEH), final project report. DG MARE Specific Contract SI2.600741. Hellenic Centre for Marine Research, Heraklion (Greece), p 557 pp.

IUCN (2012) IUCN Red List Categories and Criteria: Version 3.1. Second edition. IUCN, Gland, Switzerland and Cambridge, UK

Ocean Biogeographic Information System (OBIS) (2014). Global map of Hurlbert's index of biodiversity. Oostende (Belgium): Intergovernmental Oceanographic Commission (IOC) of the United Nations Educational, Scientific and Cultural Organization (UNESCO). URLs: www.iobis.org/maps/diversity, data.unep-wcmc.org/datasets/15

Martin CS, Fletcher R, Jones MC, Kaschner K, Sullivan E, Tittensor DP, Mcowen C, Geffert JL, van Bochove JW, Thomas H, Blyth S, Ravilious C, Tolley M, Stanwell-Smith D (2014). Manual of marine and coastal datasets of biodiversity importance. May 2014 release. Cambridge (UK): UNEP World Conservation Monitoring Centre. 28 pp. (+ 4 annexes totalling 174 pp. and one e-supplement)

Mora C, Tittensor DP, Adl S, Simpson AGB, Worm B (2011) How many species are there on Earth and in the ocean? PLoS Biol 9:e1001127

Citation: UNEP-WCMC (2015) Absence of evidence is not evidence of absence: Working with data deficiency in global biodiversity spatial datasets, UNEP-WCMC, Cambridge, UK.

Authors: Kerstin Brauneder, Matt Jones, Melissa Tolley

Available online at: www.proteuspartners.org/

Contact: businessandbiodiversity@unep-wcmc.org

The United Nations Environment Programme World Conservation Monitoring Centre (UNEP-WCMC) is the specialist biodiversity assessment centre of the United Nations Environment Programme (UNEP), the world's foremost intergovernmental environmental organization. The Centre has been in operation for over 30 years, combining scientific research with practical policy advice.

www.unep-wcmc.org/