Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2014, **5,** 708–713 doi: 10.1111/2041-210X.12209

APPLICATION

Using MODESTR to download, import and clean species distribution records

Emilio García-Roselló[1], Cástor Guisande[2]*, Juergen Heine[1], Patricia Pelayo-Villamil[3], Ana Manjarrés-Hernández[4], Luis González Vilas[2], Jacinto González-Dacosta[1], Antonio Vaamonde[5] and Carlos Granado-Lorencio[6]

[1]Department of Computer Science, Universidad de Vigo, Campus Lagoas-Marcosende s/n, Vigo 36310, Spain; [2]Facultad de Ciencias del Mar, Universidad de Vigo, Campus Lagoas-Marcosende s/n, Vigo 36310, Spain; [3]Grupo de Ictiología, Universidad de Antioquia, Medellín, A.A. 1226, Colombia; [4]Instituto Amazónico de Investigaciones (IMANI), Universidad Nacional de Colombia, Km 2 vía Tarapacá, Leticia, Colombia; [5]Departamento de Estadística e Investigación Operativa Facultad de CCEE y Empresariales, Universidad de Vigo, Torrecedeira 105, Vigo 36208, Spain; and [6]Departamento de Biología Vegetal y Ecología, Facultad de Biología, Universidad de Sevilla, Avenida de Reina Mercedes s/n, Sevilla 41012, Spain

Summary

1. Data quality is one of the highest priorities for species distribution data warehouses, as well as one of the main concerns of data users. There is the need, however, for computational procedures with the facility to automatically or semi-automatically identify and correct errors and to seamlessly integrate expert knowledge and automated processes.

2. New version MODESTR 2.0 (http://www.ipez.es/ModestR) makes it easy to download occurrence records from the Global Biodiversity Information Facility (GBIF), to import shape files with species range maps such as those available at the website of the International Union for Conservation of Nature (IUCN), to import KML files, to import CSV files with records of the users, to import ESRI ASCII grid probability files generated by distribution modelling software and show the resulting records on a map.

3. MODESTR supports five different methods for cleaning the data: (i) data filtering when downloading records from GBIF, (ii) habitat data filtering, (iii) taxonomic disambiguation filtering, (iv) automatic spatial dispersion and environmental layer filters and (v) custom data filtering.

Key-words: data cleaning, geographic records, GBIF, IUCN

Introduction

Global Biodiversity Information Facility (GBIF, 2013) and the International Union for Conservation of Nature (IUCN, 2013) give access to millions of current digitized biodiversity records. The primary search tool in GBIF and the IUCN is to utilize names of species. If the name is entered incorrectly, then access to the information by users will be difficult, if not impossible. This is probably the area that causes the most angst and loss of confidence among users in primary species data bases, due to inexperience among users in the necessity for taxonomic and nomenclatural changes (Berendsohn 1997). This issue is commonly referred to as taxonomic disambiguation and one of the proposed solutions is to provide globally unique identifiers (GUID) using the Life Science Identifier (LSID) system to uniquely label species (Jones, White & Orme 2011; Roskov *et al.* 2013). Therefore, the possibility of basing the search on a data base with updated taxonomy or globally unique identifiers would clearly facilitate the downloading of data from the GBIF.

In addition to facilitating the download or importation of geographic records, data quality is also an important issue that remains a pervasive and thorny problem, but which has received little attention. Data quality is particularly important for data sourced from species distribution data warehouses such as GBIF and national nodes, as well as one of the main concerns of data users (Yesson *et al.* 2007; Otegui *et al.* 2013).

Data cleaning is a process used to determine inaccurate, incomplete or unreasonable data and then to improve the quality through correction of detected errors and omissions (Chapman 2005). Several on-line and stand-alone tools have been developed to assist users, for example Diva-GIS (Hijmans *et al.* 2001, 2005b) and the CRIA Data Cleaning tool (CRIA 2005), among others (Chapman 2005).

Among other aspects, data cleaning usually includes removing all duplicates, correcting errors in geocoding (latitude and longitude) and sampling bias, as well as the identification of

*Correspondence author. E-mail: castor@uvigo.es false records. The possibility of discriminating between habitats could be useful for data cleaning prior to undertaking species distribution mapping and modelling. Indeed, this could constitute another potential source of error and should be included in data-cleaning programs.

Data quality and the ability to clean and correct data are the responsibility of the community and cannot be assigned to any one agent in the process (Mesibov 2013; Belbin *et al.* 2013), but there is the need for computational procedures capable of automatically or semi-automatically identifying and correcting errors. The aim of this study is to show the potential of the package MODESTR (http://www.ipez.es/ModestR) for downloading, importing and cleaning data from GBIF and the IUCN, or any other source.

ModestR

MODESTR is an autonomous package developed in C# that consists so far of the applications DATAMANAGER, MAPMAKER and MRFINDER (García-Roselló *et al.* 2013; see video demonstration on the web www.ipez.es/ModestR), and RWIZARD that is an interface designed to interact with R environment and allows the running of scripts specifically designed to use the outputs of MODESTR. Geographic data about country borders and the freshwater habitats were obtained from the web page http//www.openstreetmap.org/. RWIZARD is open code and an easy to use graphical user interface for R environment that it will available soon at the website http://www.ipez.es/ModestR. The main strengths of RWIZARD not available in other applications are: (i) to search for any of the functions available in R environment by the name or the content of the function and to include the function into a script in a very easy way, (ii) to see all arguments of the functions and their details, (iii) to modify the functions of a script with a very friendly menu and, finally, (iv) to have available RWIZARD applications that are a series of plug-ins which extend the range of application. MODESTR is so far only available for Windows, but future versions will be available for other operating systems.

With MODESTR, there is the possibility of using a data set with the list of valid species that may be developed by experts, thereby solving a problem for users unfamiliar with the taxonomy of the group in question and also solving the problem of entering incorrect names. There is an example of a data base called Elasmobranchii.DB (Pelayo-Villamil *et al.* 2012; Guisande *et al.* 2013), also available on the quoted web page, with the geographical distribution of all valid marine species of elasmobranchs currently recognised by systematists (Eschmeyer & Fricke 2013) and available in IPez (http://www.ipez.es/index%20ingles.html, Guisande *et al.* 2010), which may be used as an example of a MODESTR data set.

Collecting the information

DOWNLOADING DATA FROM GBIF

One of the different ways of creating a map with MODESTR is by importing data from the GBIF data portal, by merely introducing the name of the species or by selecting the species from a MODESTR data base. It is possible to select among the different kinds of records supported by GBIF: specimen, observation, living, fossil, germplasm and/or unknown. In addition to searching all records for the selected species, another advantage of MODESTR is that it also includes the synonyms and vernacular names. There is also the possibility of using LSIDs when downloading records from GBIF in MAPMAKER. The user may enter a LSID, and MAPMAKER will try to resolve it and then use the corresponding species name identified by this LSID. Finally, we also implemented a temporal filter when downloading data from GBIF in both DATAMANAGER (many species at the same time) and in MAPMAKER (just for one species). Start date returns only records occurring on or after the supplied date, and end date returns only records occurring on or before the supplied date. This temporal filter may be useful to match temporally referenced species occurrences data with the time period used by the environmental layer data and also for updating a DATAMANAGER data set with new records uploaded recently to GBIF.

IMPORTING DATA FROM SHAPE FILES AVAILABLE AT THE IUCN

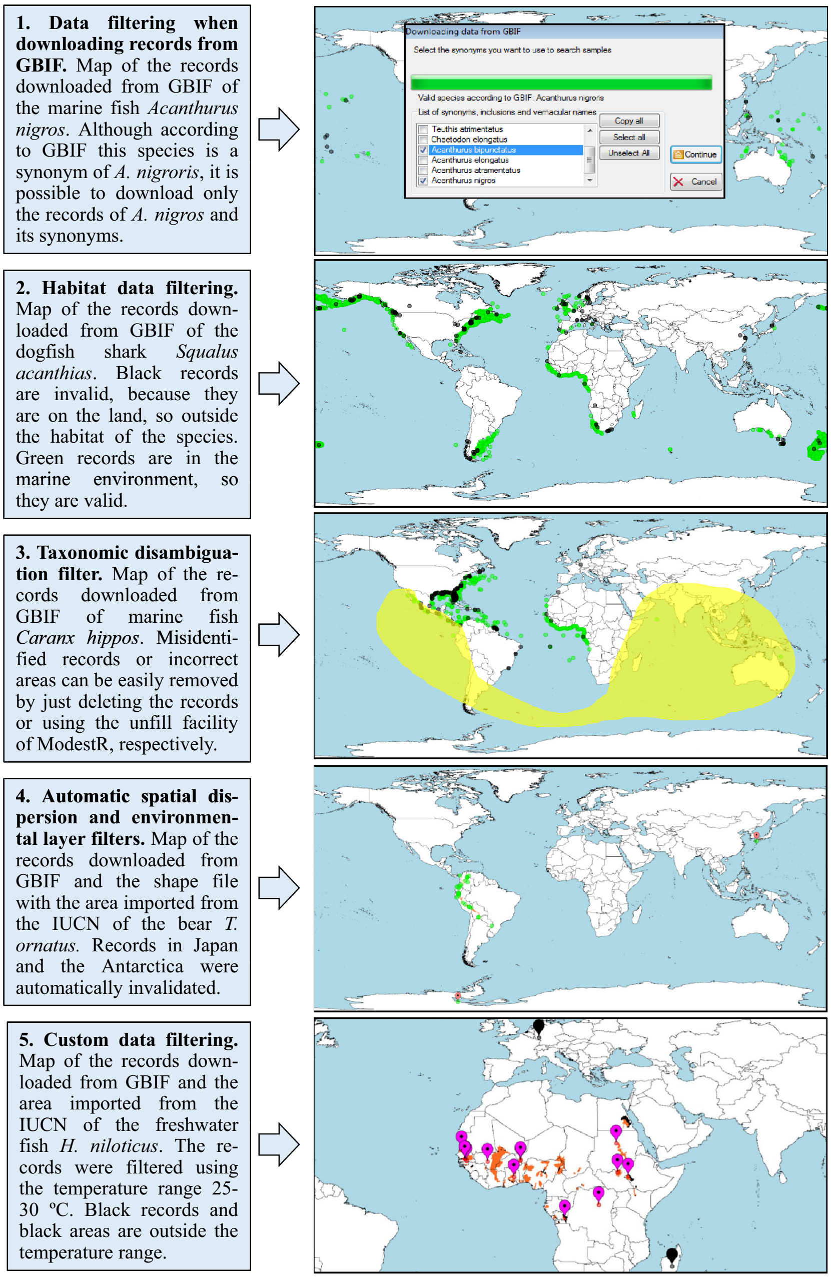
With MODESTR, it is easy to import shape files such as those available at the website of IUCN with the range maps of species. It is possible to filter according to any of the different fields contained in the shape file. For example, when importing the shape file of the spectacled bear (*Tremarctos ornatus*), also known as the Andean bear, MODESTR allows the selection of 'Extant (resident)' and not 'Probably Extant (resident)'. Therefore, the range map obtained includes only the areas where the presence of the species is confirmed and, hence, excludes areas where the presence of the species is uncertain.

Environmental variables for the filtering process

MODESTR allows the organization of a set of different environmental variables (García-Roselló *et al.* 2013). These data can be easily imported from files in CSV or ESRI ASC formats, for example the terrestrial data set WorldClim (Hijmans *et al.* 2005a), the marine data set BIO-ORACLE (Tyberghein *et al.* 2012). One of the purposes of this feature is to integrate exportation of geographic records of species distributions and environmental variables. These environmental variables, however, may also be used in MODESTR to filter the geographic records according to a range or threshold of one or several environmental variables.

Filtering data

MODESTR supports five different methods for filtering the data: (i) data filtering when downloading records from GBIF, (ii) habitat data filtering, (iii) taxonomic disambiguation filtering, (iv) automatic spatial dispersion and environmental layer filters and, finally, (v) custom data filtering.



1. Data filtering when downloading records from GBIF. Map of the records downloaded from GBIF of the marine fish *Acanthurus nigros*. Although according to GBIF this species is a synonym of *A. nigroris*, it is possible to download only the records of *A. nigros* and its synonyms.

Downloading data from GBIF

Select the synonyms you want to use to search samples

Valid species according to GBIF: Acanthurus nigroris

List of synonyms, inclusions and vernacular names

- □ Teuthis atrimentatus
- □ Chaetodon elongatus
- ☑ Acanthurus bipunctatus
- □ Acanthurus elongatus
- □ Acanthurus atramentatus
- ☑ Acanthurus nigros

2. Habitat data filtering. Map of the records downloaded from GBIF of the dogfish shark *Squalus acanthias*. Black records are invalid, because they are on the land, so outside the habitat of the species. Green records are in the marine environment, so they are valid.

3. Taxonomic disambiguation filter. Map of the records downloaded from GBIF of marine fish *Caranx hippos*. Misidentified records or incorrect areas can be easily removed by just deleting the records or using the unfill facility of ModestR, respectively.

4. Automatic spatial dispersion and environmental layer filters. Map of the records downloaded from GBIF and the shape file with the area imported from the IUCN of the bear *T. ornatus*. Records in Japan and the Antarctica were automatically invalidated.

5. Custom data filtering. Map of the records downloaded from GBIF and the area imported from the IUCN of the freshwater fish *H. niloticus*. The records were filtered using the temperature range 25-30 °C. Black records and black areas are outside the temperature range.

Fig. 1. Data-cleaning techniques available in MODESTR.

FILTERING DATA WHEN DOWNLOADING RECORDS FROM GBIF

This cleaning process has the following filter options from which the user can select: (i) to remove duplicates (it is possible to discriminate between real duplicates – records of the same specimen sent to different collections – and multiple records of a species from a particular location taken by different collectors); (ii) to remove records with the geographic coordinates 0° longitude and 0° latitude, which frequently appear in GBIF data sets as an error; (iii) to eliminate records in which the longitude and latitude values are identical and likely represent erroneous repetitive data entry, also a common error in the GBIF data sets and, finally, (iv) correct erroneous synonyms.

In the last filter, MODESTR has the capacity to handle synonyms very efficiently. Synonyms and vernacular names can all be included in the search, thereby allowing the detection of potential errors in the association of synonyms to a specific species. As an example, according to GBIF, the marine fish *Acanthurus nigros* is a synonym of *Acanthurus nigroris*, but *A. nigros* is considered a valid species (Eschmeyer & Fricke 2013), and using MODESTR, it is possible to download separately from the GBIF the records for *A. nigros* and its synonyms *A. bipunctatus* and *Hepatus atramentatus* (Fig. 1-1).

HABITAT FILTERING

MODESTR offers the possibility of discriminating between environments, a particularly useful option when using GBIF data sets in which there are frequently records of occurrence of a species in areas outside the habitat of the species, for example records in the ocean for a terrestrial species and vice versa

(Fig. 1-2). With MODESTR, the user may specify the habitat of the species which may be one or a combination of the following habitat options: land, marine environment, lentic freshwater habitats (lakes, ponds, swamps, etc.), wetlands, large rivers and small rivers (creeks, streams, etc.), reservoirs, canals, ditches and drains.

Those records outside the habitat selected by the user are considered invalid, but are not eliminated. An invalid record will be omitted when exporting the information downloaded from the GBIF to a file with the geographic coordinates of the species, to a presence/absence matrix or to a species richness matrix of any order, family or genus. The invalid samples are not eliminated so as to permit reconsideration of their validity in further manual reviews of the records, or as a result of a modification in the reference world map. For example, a missing lake on the map can subsequently be added, thereby validating a sample previously considered invalid. In this context, it is worth pointing out that, when the reference world map is modified, MODESTR will automatically recheck the validity of the records of a map (or of an entire MODESTR data base).

Taxonomic disambiguation filtering

An important advantage of MODESTR is that it integrates expert knowledge and automated processes. Taxonomic accuracy is very important in any study involving species distribution. However, misidentification is sometimes observed in GBIF data sets, although this can easily be corrected with MODESTR. For example, in the spatial distribution provided by the GBIF for *Caranx hippos* (Fig. 1-3), the reports from the Pacific Ocean refer to *C. caninus* and the Indian Ocean records are probably misidentifications of *C. ignobilis* (Froese & Pauly 2013). These records can be selected and deleted using the option ‘Delete Samples’. Therefore, it is easy to remove potential false records and to obtain the hypothetical distribution of *C. hippos*.

In the case of areas obtained from a shape file, incorrect areas are easily removed by just using the unfill facility of MODESTR.

Automatic spatial dispersion and environmental layer filters

There is an automatic data-filtering facility of MODESTR that guides the user step by step through the data-filtering process making it considerably easier. In the first step, the user can select the variables to be used to clean presence data. Besides environmental variables, three specifically calculated variables related to dispersal capacity are also available: mean distance, latitudinal dispersion and longitudinal dispersion. Mean distance for each occurrence record is calculated as the mean distance from this record to all the other records. Latitudinal dispersion and longitudinal dispersion are merely the latitude or longitude value of each occurrence record. These variables allow the validation of records based on their relative geographical dispersion, so records excessively distant from the others will be cleaned. It must be pointed out that those variables can only be used on maps that contain occurrence records, but not on range maps.

Once the environmental variables are selected, MODESTR will calculate and display the coverage percentage of each one; that is, the percentage of occurrence records where a variable has a not null value. This can be useful to detect and remove variables that are not suitable to be used in data filtering. For example, if we select a variable that contains altimetry values to clean data from a marine species, the coverage percentage will almost be of 0%, showing that this variable is not appropriate to clean presence data from this species.

MODESTR will also display the minimum and maximum values that each selected variable takes in the zones where the species is present. The user can then select the method to be used to detect outliers for each variable. Currently, MODESTR supports four methods:

1. Quartiles: this validation method is based on the values of the Q1 and Q3 quartiles of the values taken by a variable. This option will consider any value of this variable as an outlier if it is outside the range Q1 − X(Q3 − Q1), Q3 + X(Q3 − Q1) where X can be set by the user (typically $X = 1{\cdot}5$ is used to consider both mild and extreme outliers as invalid; or $X = 3$ to consider only extreme outliers as invalid).

2. Jackknife: this validation method is described in Chapman (2005), and it is considered quite reliable. MODESTR supports two variants of this method: the one proposed by CRIA and the one proposed in DIVA-GIS (both are described in Chapman 2005).

3. Autoselect best: this is the default method. It consists of selecting the most conservative method among the ones described above. That is, this option selects the method that returns lower number of outliers. The effectively selected method as the best for each variable can be consulted by the user just by clicking on the variable.

4. Custom: in this case, the user can manually set the range of valid values for a variable. Any value falling outside this range will be considered an outlier.

Once selected the method to be used to detect outliers for each variable (or just left on auto select mode), MODESTR will filter data applying the selected settings. For example, Fig. 1-4 shows the records downloaded from GBIF and the area imported from the shape file available at the IUCN of the spectacled bear (*Tremarctos ornatus*). There are two erroneous records: one in the Ibaraki Nature Museum (Japan) because the record was catalogued as specimen instead of as preserved specimen and another in Antarctica due to an error with the latitude. The data were filtered using the annual mean temperature from 1950 to 2000 of the WorldClim data set (Hijmans *et al.* 2005a), the altitude and the mean distance. This spatial dispersion and environmental layer filters resulted in the invalidation of these two erroneous records (Fig. 1-4).

Custom data filtering

Custom data filtering allows the user to freely enter a validation rule that can combine any environmental variables. The validation rule has to be boolean (logically true or false). All occurrence records or presence areas of the species where the rule is false according to the values of the involved environmental variables will be filtered by MODESTR.

For example, according to Fishbase (Froese & Pauly 2013), the temperature range of the freshwater fish *Heterotis niloticus* (African bonytongue) is between 25°C and 30°C. We elaborated the range map of *H. niloticus* with the areas imported from the shape file available for this species at the website of IUCN, and the geographic records downloaded from GBIF (Fig. 1-5). The data were filtered using the annual mean temperature from 1950 to 2000 of the WorldClim data set (Hijmans *et al.* 2005a) entering a simple rule in MODESTR custom filtering feature:

$$[\text{Annual mean temperature}] \geq 25 \text{ AND}$$
$$[\text{Annual mean temperature}] \leq 30$$

The resulting map is shown in Fig. 1-5. The geographic records available in GBIF of Germany (8·5°C) and Madagascar (18·6°C) are clearly outside the hypothetical temperature range of this species. Some areas in southern east of Egypt, where *H. niloticus* is present according to the information available in the IUCN website, are also outside the temperature range of this species, with temperatures ranged between 21·8°C and 24°C. Therefore, probably the temperature ranging of *H. niloticus* is between 21·5°C and 30°C, if the record in Madagascar is incorrect.

Minimizing the error of working with low spatial resolution records

There is a major source of species occurrence data error resulting from the imprecision of spatial coordinates. In many species, the records are the centroid of a grid cell or a reference or midpoint position within a region. These low spatial resolution records sometimes are expressed with a high spatial precision (5 or more decimal places), but it is incorrect to work with these records at such high resolution. In DataManager, we implemented the option of exporting the data, for use in other applications as MaxEnt of RWizard, with the format of pseudosamples. If quality resolution of the data is low, the proper way to use these data is to create a raster with a grid cell for instance of 5′ × 5′, 30′ × 30′, 1° × 1°, etc. Therefore, the output file of MODESTR is a list of species within each of the grid cells with the size defined by the user and, furthermore, it is also possible to obtain the number of records for each species within the grid cell and the mean value of some environmental variables selected by the user. Instead of the list of species, it is also possible to obtain the species richness within each grid cell.

Strengths

The main strengths of MODESTR compared with other applications are the following five facilities, which are not available in any other software.

1. It is possible to apply five different automatic and semi-automatic data-cleaning techniques to the geographic data, thereby facilitating the process of identification and correction of errors and therefore improving data quality.

2. It is possible to apply to a data set with many species, without the need to proceed species by species, the data filtering when downloading records from GBIF and also the habitat, spatial dispersion and environmental layer filtering techniques.

3. It facilitates the process of downloading information from the GBIF and importing shape files available at the IUCN, with the option of selecting among different kinds of records, and by providing the possibility of using a data set with the list of valid species that may be developed by experts, thereby solving a problem for users unfamiliar with the taxonomy of the group in question. This is especially important as the taxonomy of many groups changes constantly and is a consequence of the description of new species and/or taxa being revised. This data set with the list of valid species for any taxon can easily be updated using MODESTR, and this file can easily be shared with other researchers.

4. It is possible to download all the records of a class, order, family or genus without the need to proceed species by species. Our findings show, however, that it is necessary to correct erroneous synonyms and misidentifications. Misidentification is especially difficult to detect and rarely discussed in the literature (Scott & Hallam 2003), but once detected, it can easily be corrected with MODESTR.

5. One of the most important contributions of MODESTR to the cleaning process of data tested in this study was that it provided the possibility of discriminating between habitats. These erroneous records may introduce an important source of error, for example, when estimating species distribution models. Care needs to be exercised because it is possible that those records outside the habitat zone are not in fact errors, as a freshwater species may occur in streams too small for mapping or in a habitat for which the geographic data are not yet available in MODESTR.

Acknowledgements

We thank ENDESA for financial support.

Data accessibility

This manuscript does not use data.

References

Belbin, L., Daly, J., Hirsch, T., Hobern, D. & La Salle, J. (2013) A specialist's audit of aggregated occurrence records: an 'aggregator's' perspective. *ZooKeys*, **305**, 67–76.

Berendsohn, W.G. (1997) A taxonomic information model for botanical databases: the IOPI model. *Taxon*, **46**, 283–309.

Chapman, A.D. (2005) *Principles and Methods of Data Cleaning – Primary Species and Species-Occurrence Data, version 1.0*. Report for the Global Biodiversity Information Facility, Copenhagen.

CRIA (2005). *speciesLink*. *Dados e ferramentos*. *Data Cleaning*. Centro de Referência em Informação Ambiental. Available at http://splink.cria.org.br/dc/

Eschmeyer, W.N. & Fricke, R. (eds). (2013). *Catalog of Fishes electronic*. Available at http://researchcalacademyorg/ichthyology/catalog/fishcatmainasp.

Froese, R. & Pauly, D. (2013) *FishBase World Wide Web electronic publication*. www.fishbase.org version (02/2013).

García-Roselló, E., Guisande, C., González-Dacosta, J., Heine, J., Pelayo-Villamil, P., Manjarrés-Hernández, A., Vaamonde, A. & Granado-Lorencio, C. (2013) ModestR: a software tool for managing and analyzing species distribution map databases. *Ecography*, **36**, 102–1207.

GBIF (2013) *GBIF data portal*. Available at http://datagbiforg (accessed September 2013).

Guisande, C., Manjarrés-Hernández, A., Pelayo-Villamil, P., Granado-Lorencio, C., Riveiro, I., Acuña, A. *et al.* (2010) IPez: an expert system for the taxonomic identification of fishes based on machine learning techniques. *Fisheries Research*, **102**, 240–247.

Guisande, C., Patti, B., Vaamonde, A., Manjarrés-Hernández, A., Pelayo-Villamil, P., García-Roselló, E., González-Dacosta, J., Heine, J. & Granado-Lorencio, C. (2013) Factors affecting species richness of marine elasmobranchs. *Biodiversity and Conservation*, **22**, 1703–1714.

Hijmans, R., Guarino, L., Cruz, M. & Rojas, E. (2001) Computer tools for spatial analysis of plant genetic resources data: 1. DIVA-GIS. *Plant Genetic Resources Newsletter*, **27**, 15–19.

Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G. & Jarvis, A. (2005a) Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, **25**, 1965–1978.

Hijmans, R.J., Guarino, L., Bussink, C., Mathur, P., Cruz, M., Barrentes, I. & Rojas, E. (2005b) *DIVA-GIS Version 5. A geographic information system for the analysis of biodiversity data*. Available at http://www.diva-gis.org.

IUCN (2013) *The IUCN Red List of Threatened Species. Version 2012.2*. Available at http://www.iucnredlist.org. Downloaded on Septembre 2013.

Jones, A.C., White, R.J. & Orme, E.R. (2011) Identifying and relating biological concepts in the Catalogue of Life. *Journal of Biomedical Semantics*, **2**, 7.

Mesibov, R. (2013) A specialist's audit of aggregated occurrence records. *ZooKeys*, **293**, 1–18.

Otegui, J., Ariño, A.H., Encinas, M.A. & Pando, F. (2013) Assessing the primary data hosted by the Spanish node of the Global Biodiversity Information Facility (GBIF). *PLoS One*, **8**, e55144.

Pelayo-Villamil, P., Guisande, C., González-Vilas, L., Carvajal-Quintero, J.D., Jiménez-Segura, L.F., García-Roselló, E. *et al.* (2012) ModestR: Una herramienta infromática para el estudio de los ecosistemas acuáticos de Colombia. *Actualidades Biológicas*, **34**, 225–239.

Roskov, Y., Kunze, T., Paglinawan, L., Orrell, T., Nicolson, D., Culham, A. *et al.*, eds (2013). *Species 2000 & ITIS Catalogue of Life, 2013 Annual Checklist*. Digital resource at www.catalogueoflife.org/annual-checklist/2013/. Species 2000, Reading, UK.

Scott, W.A. & Hallam, C.J. (2003) Assessing species misidentification rates through quality assurance of vegetation monitoring. *Plant Ecology*, **165**, 101–115.

Tyberghein, L., Verbruggen, H., Pauly, K., Troupin, C., Mineur, F. & de Clerck, O. (2012) Bio-ORACLE: a global environmental dataset for marine species distribution modelling. *Global Ecolology and Biogeography*, **21**, 272–281.

Yesson, C., Brewer, P.W., Sutton, T., Caithness, N., Pahwa, J.S., Burgess, M. *et al.* (2007) How global is the global biodiversity information facility? *PLoS One*, **2**, e1124.

Received 2 March 2013; accepted 16 May 2014
Handling Editor: David Orme