Biogeo: an R package for assessing and improving data quality of occurrence record datasets

Mark P. Robertson, Vernon Visser and Cang Hui

M. P. Robertson (mrobertson@zoology.up.ac.za), Centre for Invasion Biology, Dept of Zoology and Entomology, Univ. of Pretoria, Hatfield 0028, South Africa.

V. Visser, Centre for Invasion Biology, Dept of Botany and Zoology, Stellenbosch Univ., Matieland, 7602, South Africa, and Invasive Species Programme, South African National Biodiversity Inst., Kirstenbosch Research Centre, Cape Town, South Africa.

C. Hui, Centre for Invasion Biology, Dept of Mathematical Sciences, Stellenbosch Univ., and African Inst. for Mathematical Sciences, Matieland 7602, South Africa.

Abstract

Occurrence data from museum and herbarium collections are valuable for mapping biodiversity patterns in space and time. Unfortunately these collections datasets contain many errors and suffer from several data quality issues that can influence the quality of the products derived from them. It is up to the user to identify these errors and data quality issues when using these data. Despite the large number of potential users of these datasets there are few software tools dedicated to error detection and correction of collections datasets. The R package biogeo was developed for detecting and correcting errors and for assessing of data quality of collections datasets consisting of occurrence records. Features of the package include error detection, such as mismatches between the recorded country and the country where the record is plotted, records of terrestrial species that fall into the sea and outlier detection. A key feature of the package is the ability to identify likely alternative positions for points that represent obvious errors in the dataset and functions to explore records in geographical and environmental space in order to identify possible errors in the dataset. Functions are also available for converting coordinates that are in various text formats into degrees, minutes and seconds and then into decimal degrees.

Vast amounts of biodiversity data are available in museum and herbarium collections (Graham et al. 2004, Suarez and Tsutsui 2004, Boakes et al. 2010, Maldonado et al. 2015). These datasets are based on collections that were assembled for the primary purpose of taxonomy, but are now being used for an array of other analyses and applications (Funk and Richardson 2002, Graham et al. 2004, Chapman 2005a, Newbold 2010). Several recent studies have made use of collections data to investigate various questions in macroecology (Swenson et al. 2012, Lamanna et al. 2014) and invasion biology (Richardson et al. 2011, Novoa et al. 2015). Collections data consist of ad hoc records obtained from specimen labels in museums and herbaria. These specimens were collected at a particular locality and often the coordinates of this locality are given by the collector; alternatively, coordinates can be assigned later if the locality description is sufficiently precise. Many collections datasets have become more easily accessible to users through online databases such as the Global Biodiversity Information Facility (< www.gbif.org >; Edwards 2004). These datasets represent a valuable source of species distribution data and represent a valuable baseline for describing biodiversity patterns (Chapman 2005a, Weiser et al. 2007, Boakes et al. 2010, Swenson et al. 2012, Lamanna et al. 2014, Maldonado et al. 2015). Another valuable source of species distribution data comes from atlas projects (Robertson et al. 2010). Atlas projects are usually initiated to collect data for a particular taxonomic group and have certain minimum data requirements for a record (Robertson et al. 2010). The presence of a species is usually recorded in a grid with a particular spatial resolution e.g. 15 minutes. A key difference between these two data sources is that atlas data tend to be grid-based while collections data are point-based. This has important data quality implications.

These collections datasets are frequently used to develop species distribution models (ecological niche models) that have many applications in biology (Guisan and Zimmermann 2000, Elith and Leathwick 2009). The simplest application of collections datasets is to produce point-based range maps that can be used in field guides or to guide further collection efforts. These range maps are used to calculate range size metrics that are used in IUCN red list assessments (IUCN 2012), which include area of occupancy (AOO) and extent of occurrence (EOO) calculations (Gaston and Fuller 2009). These metrics can be used to calculate changes in range size over time, such as range contractions in the case of threatened species (Joseph and Possingham 2008) or range expansion in the case of invasive species. Range size calculations such as alpha-hulls can be used for investigating macro-ecological questions (Hui et al. 2011a). In addition to knowing the distribution of single species in isolation, it is valuable to document the species assemblage in a given area. Species richness maps are the basis for many macroecological studies and for conservation (Gaston 2000, Weiser et al. 2007, Swenson et al. 2012, Maldonado et al. 2015). Species richness maps can be produced using various approaches, including by converting point data to grids and by combining range maps produced by distribution models (Graham and Hijmans 2006). Point data can be incorporated into grid-based atlas projects, which in turn have a number of applications in biogeography and conservation (Robertson et al. 2010). The quality of the datasets used in these applications has a strong influence on the reliability of the products produced (Freeley and Silman 2010, Maldonado et al. 2015). It is up to the users of these datasets to assess the quality of the data that they obtain and make decisions about the suitability of those data to answer particular questions.

Collections datasets are known to contain errors (Yesson et al. 2007, Robertson 2008, Newbold 2010) and suffer from certain weaknesses, such as sampling bias (Reddy and Davalos 2003, Robertson and Barker 2006, Hortal et al. 2008, Hui et al. 2011b) that can decrease the quality of the products derived from them (Franklin 2009). Several articles have assessed various aspects of data quality of collections datasets (Hijmans et al. 1999, Ponder et al. 2001, Funk and Richardson 2002, Hortal et al. 2007, Yesson et al. 2007, Robertson 2008, Newbold 2010, Maldonado et al. 2015), but fewer have provided advice on how to detect and correct errors (Hijmans et al. 1999, Chapman 2005b, Hortal et al. 2007). Some of the important data cleaning steps are described by Hijmans and Elith (2015) and they show how these can be performed in R using the *dismo* package (Hijmans et al. 2015).

Obvious errors in collection localities can be detected by producing a map of the records for a species and identifying outliers such as points in the sea for terrestrial species (Hijmans et al. 1999). Errors such as these are easily detected and the record can either be corrected or discarded from the dataset. However other errors, such as a record for a species that is geographically close to other records for that species, but that is incorrectly located at the top of a mountain range, may be more difficult to detect and to correct (Newbold 2010). Users need to know which types of errors to look for, identify these errors in records, correct them if possible, or exclude the records from their analyses. Despite the large number of potential users of these datasets there seem to be relatively few software tools dedicated to error detection and correction of point data from collections datasets. To address this need we have developed an R package, *biogeo*, for the detection and correction of errors and for assessment of data quality of collections datasets consisting of occurrence records.

This package has been developed with the primary aim of data cleaning and data quality assessment. Although other software packages can perform some of the data cleaning operations available here, there are none that are as comprehensive or that offer as many different tools. A key feature of the package is that it can cope with a dataset that consists of records that are in a range of different coordinate formats, a common problem with datasets that have been collated from multiple sources. The package has several functions for detecting errors in datasets but also has the functionality to correct these errors instead of simply removing them from the dataset. The package also has functions for detecting various data quality issues, such as low precision coordinates. This package has been written in R, which has become a very popular programming language used by scientists and by biologists in particular. This means that the tools available in this package can be incorporated into user-specific scripts for more experienced R users, to enable quicker and more efficient data cleaning of large datasets. However the functions can also be used by those with limited programming experience as the tutorial demonstrates their application and has been prepared with the inexperienced user in mind.

In order to provide the necessary context for describing the features of the package we first discuss errors and data quality considerations in relation to collections data followed by a section on data preparation and cleaning.

Errors and data quality considerations

The most common type of error in collections datasets is probably locational errors, concerning the geographical position of a given record in space. These errors can often be detected as obvious geographical outliers on a map (Yesson et al. 2007). Locational errors, and geographical outliers in particular, are most problematic for drawing range maps and especially for calculating range size using extent of occurrence and area of occupancy (Gaston and Fuller 2009). These errors can be caused by missing coordinates, substitution of x- and y-coordinates and errors in converting to decimal degrees, which makes them relatively easy to detect (Table 1). Locational errors can be detected if other data such as country names, locality descriptions and elevation are provided as part of the record (see errors e and f in Table 1). Low precision of the coordinates (e.g. when only the degrees have been recorded) is a data quality issue rather than an error but it has important consequences for many applications. This problem can cause records to appear as if they are incorrect e.g. points plotted in the sea for terrestrial species (Yesson et al. 2007).

Table 1. Description of errors detected in point data with explanations of the likely cause of the error. Yesson et al. (2007) described several errors, which we give in brackets in the Error column

Error	Possible cause of problem
a) Point plotted at zero degrees latitude and longitude. ('Lat/Long zero').	No coordinates were available in the original dataset but values of zero assigned to the coordinates.
b) Points in sea for terrestrial species or on land for aquatic species, obvious geographical outliers. ('Lat/Long error', 'Far from valid').	Transposed latitude and longitude coordinates; incorrect sign on the decimal degrees of the latitude or longitude coordinate; degrees and minutes were transposed before the coordinate was converted to decimal degrees; imprecise locality description used to assign coordinates; the specimen was incorrectly identified by the collector or the incorrect name was applied to the species when the data were digitized.
c) Point in sea but close to coast for terrestrial species, or on land but close to coast for marine species. ('Lat/Long error', 'Near Valid').	Low precision coordinates e.g. only degrees were available or the data were originally collected on a coarse scale grid. Imprecise locality description used to assign coordinates.
d) Point plotted along the prime meridian or equator. ('Lat/Long zero').	Missing coordinate for latitude or for longitude that was incorrectly assigned a value of zero.
e) Country name given in the record does not correspond with country where point is plotted.	Likely to be the same errors as for b) above.
f) Elevation given in the record does not correspond with elevation obtained from a digital elevation model where point is plotted.	Likely to be the same errors as for b) above, or the spatial resolution of the digital elevation model is too coarse.

Species distribution models are probably less sensitive to geographical outliers, especially if there are few of these errors in proportion to the remaining records that are correct (but see Freeley and Silman 2010). Environmental data can be extracted from, among others, interpolated climate surfaces, digital elevation models, vegetation and soils maps using the coordinates of the geographical locations of point records. These data are the basis for distribution models and the interplay between geographical and environmental space is important in species distribution modeling (Elith and Leathwick 2009). Environmental outliers are points in environmental space that are far away (not typical) of the rest of the records in the environmental space. Environmental outliers are potentially more serious for species distribution models than geographic outliers (Newbold 2010). A point may be a geographical outlier but have very similar environmental conditions to the remaining records of the species. In contrast, a point may be close geographically to the other points but have a different environment, especially where environmental gradients are steep (Freeley and Silman 2010).

Sampling bias is a known problem in collections datasets (Reddy and Davalos 2003, Robertson and Barker 2006, Hortal et al. 2008), although it is not explicitly addressed in this package since other software are available for correcting sampling bias in datasets e.g. R package spThin (Aiello-Lammens et al. 2015).

Dataset preparation and cleaning

In order to prepare a dataset for analysis data usually have to be collated from a variety of sources e.g. Global Biodiversity Information Facility (GBIF), museum collections and private collections. The *dismo* (Hijmans et al. 2015) and *rgbif* (Chamberlain et al. 2015) packages are especially useful for downloading species occurrence records from GBIF. Several procedures will then usually be followed as part of the data preparation and cleaning process. These data, particularly the coordinates, will be converted into a common format (steps 1 and 2 in Table 2), duplicate records will be removed (step 3 in Table 2), then data quality issues (such as low precision coordinates) will be identified (step 4 in Table 2), error checks and error corrections will be performed (step 5 in Table 2), finally the data will be prepared for particular applications e.g. species richness maps (step 6 in Table 2). The specific steps for dataset preparation and cleaning are described in Table 2 together with the appropriate functions in the *biogeo* package that can be used to assist with the data management or analysis at each step.

Table 2. Description of steps in data preparation and data quality assessment. The names of appropriate functions from the *biogeo* package are given (in italics) that can be used in each step, along with a brief descriptions of what they do

Steps	Function and description
1) Data formatting for compatibility with *biogeo*	*checkdatastr* – ensures that certain required fields are present e.g. the x- and y-coordinates named as ‘x’ and ‘y’ and a unique identifier field called ‘ID’.
	addmainfields – adds required fields to the dataframe.
	keepmainfields – retains user-selected fields from a dataframe.
	renamefields – renames fields in the dataframe.
2) Convert coordinates to decimal degrees and find coordinates for localities that have no coordinates	*dmsparse* – converts all coordinates, regardless of format (e.g. degrees, minutes and seconds; decimal degrees; character; numeric) to a standardized format in decimal degrees.
	dmsabs – separates coordinates that are in text strings into separate fields for degrees, minutes and seconds when there are no delimiters.
	dmsparsefmt – parses coordinate string using a format string.
	uniqueformats – produces a list of unique coordinate formats in the dataset.
	finddecimals – finds coordinates that are in decimal degree format.
	dms2dd – converts coordinates from degrees, minutes and seconds format into decimal degrees.
	missingcoords – finds indices of records in the datasets for which there are no coordinates.
	fromGEarth – obtains coordinates of a point from Google Earth via the clipboard.
3) Identify duplicate records to prevent pseudoreplication	*duplicatesexclude* – flags duplicate point records per species per grid cell.
4) Identify records that may be too imprecise for the analysis	*precisioncheck* – checks the precision of the coordinates.
	precisionenv – checks whether precision of coordinates is less than that of environmental data.
5) Identify records that likely have incorrect coordinates using geographical	*errorcheck* – performs several data quality and error checks (see detailed description below and Table 3).

Steps	Function and description
and environmental information	*nearestcell* – assigns points that fall in the sea to the nearest adjacent terrestrial grid cell, or vice versa.
	pointsworld – plots points on a world map showing countries.
	missingvalsexclude – highlights records which do not have any associated environmental values (depending on the raster used).
	alternatives, alternatives2 – identifies likely alternative positions for points that are known to have positional errors.
	alternativesenv – identifies likely alternative positions for points that are known to have positional errors using geographical and environmental space.
	geo2envid, geo2envpca – error detection using geographical and environmental space.
	elevcheck – identifies records that have a recorded elevation, but this elevation does not match that based on its coordinates and extracting an elevation value from a digital elevation model.
	modifiedtoday – selects records that were modified during the current day.
6) Data summaries and output	*pointsworld* – plots points on a world map showing countries.
	points2shape – converts a dataframe to a point shape file.
	speciescount – counts number of records per species in a dataframe.
	richnessmap – creates a raster map of the number of species or number of records per grid cell.
	quickrich – produces a raster map of species richness values and applies the function *quickclean* to remove records with errors.

Features of the package

The package has been designed to work with a dataset consisting of point records containing x- and y-coordinates for several different species. The *errorcheck* function performs a number of error and data quality checks on a dataset consisting of several records per species. It starts by excluding any records where the x- and y-coordinates are both zero. It then checks for any x-coordinates that are outside the range of −180 to 180 degrees and any y-coordinates that are outside the range of −90 to 90 degrees. It extracts country names for each point record from a user-specified shapefile and compares these to country names in the dataset. If there is a mismatch between these two names for a record then the record is flagged. Records without country names are flagged as being potential errors. Low precision records are flagged by determining whether they occur either at the top left corner or centre of a 10, 15, 20, 25 or 30 minute grid cell. If records have these exact coordinates then it is possible that they were collected at a coarse spatial resolution. A cell identifier is returned for each record based on the grid cell that the record falls into. These identifiers are then used to identify records that have the same cell identifier number. An environmental outlier detection is performed for all species with 10 or more records for each of the user-selected environmental variables. The reverse jackknife algorithm has been used for detecting outliers (Chapman 2005a, 2005b) and has been implemented in DivaGIS (< www.diva-gis.org >). It is considered to be a reliable method of detecting outliers. The second approach to outlier detection is to highlight records that fall a distance of 1.5 times beyond the interquartile range.

The function called *quickclean* performs many of the checks performed by *errorcheck* but instead of indicating records with possible errors it simply removes these records from the dataset. It is intended for the user who wants to rapidly remove any suspect records (e.g. for an analysis including a large number of species). This function performs a country mismatch check if the country field is specified, it performs a check to determine if the records are at the appropriate precision for the spatial resolution, it assigns point records to the nearest cell containing environmental data (using *nearestcell*, explained below) and removes records that are in the wrong environment. It flags duplicate records per species per grid cell but does not remove the duplicates. It does not require environmental data and does not perform the environmental outlier checks as performed in *errorcheck*. The function called *quickrich* produces a species richness map at a selected spatial resolution. It uses *quickclean* to eliminate any records with errors.

A key feature of the package is being able to identify likely alternative positions for points that represent obvious errors in the dataset. These alternative positions are plotted by simulating common errors such as substituting the x- and y-coordinates and changing the signs on one or both coordinates. Using the *alternatives* function, the user can select the correct position for the point on a map based on several alternatives. The *alternativesenv* function is available for exploring the positions of points in geographical and environmental space in order to identify likely alternative positions for points that are known to have positional errors. Similarly, the positions of points in geographical and environmental space can be used to identify possible errors in the dataset using the *geo2envid* function for plotting a two-dimensional environmental space or the *geo2envpca* to use principal components analysis to define the environmental space for several environmental variables. The *nearestcell* function moves points that are in the sea to the nearest grid cell on land (or the converse) if they are within one grid cell of land grid cells. Functions are available for producing species richness maps and maps of numbers of records per species per grid cell.

Another major highlight of the package is the ability to separate (parse) coordinates that are in text format e.g. 23°15’35”S into separate fields for degrees, minutes, seconds and then convert them into decimal degrees. The advantage is that a single function (*dmsparse*) can automatically identify several different coordinate text formats in a single dataset and parse them. Coordinates are often in different formats when datasets are combined from several different sources (e.g. Table 4, second column). There are also several tools for performing coordinate conversions (Table 2). The coordinate management and conversion functions are particularly useful for preparing a dataset and standardizing the data format (Table 2).

Table 3. Output from the *errorcheck* function run for Species U, with certain fields removed. Names of fields are as follows: ID – a unique identifier, x and y are the x- and y-coordinates in decimal degrees, cellid – cell identifiers calculated based on the coordinates; dups – indicates that a particular record represents a duplicate cell identifier for that particular species; country_ext – the country name extracted from the shape file based on the coordinates for the point; CountryMismatch – indicates a mismatch in names for the country in the original record (Country) and the country name extracted (country_ext); lowprec – low precision records. The bioclimatic variables are: bio1 – annual mean temperature; bio5 – maximum temperature of warmest month; bio6 – minimum temperature of coldest month; bio12 – annual precipitation. Bioclimatic variables ending with '_e' indicate outlier detection using boxplot statistics and ending with '_j' indicate outlier detection using the jackknife procedure. The column labeled 'elevMismatch' indicates when there is a mismatch in the elevation in the elevation field and that extracted from a digital elevation model (shown in demElevation). The column labeled as 'error' indicates that there was at least one error for that record. The final column labeled 'spperr' indicates that there was at least one error for that species

ID	x	y	cellid	dups	country_ext	CountryMismatch	wrongEnv	lowprec	bio1_e	bio12_e	bio5_e	bio6_e	bio1_j	bio12_j	bio5_j	bio6_j	elevMismatch	demElevation	error	spperr
1971	28.1	−25.4	1495969	0	South Africa	0	0	0	0	0	0	0	0	0	0	0	0	1127	0	1
1972	32.41666667	−27.03333333	1517595	0	South Africa	0	0	0	0	0	0	0	0	0	0	0	0	78	0	1
1973	32.88333333	−27	1517598	0	NA	1	0	0	0	0	0	0	0	0	0	0	0	25	0	1
1974	32.31	−27.78	1526234	0	South Africa	0	0	0	0	0	0	0	0	0	0	0	0	125	0	1
1975	32.27	−27.65	1524074	0	South Africa	0	0	0	0	0	0	0	0	0	0	0	0	69	0	1
1976	32.27	−27.65	1524074	1	South Africa	0	0	0	0	0	0	0	0	0	0	0	0	69	0	1
1977	32.27	−27.65	1524074	1	South Africa	0	0	0	0	0	0	0	0	0	0	0	0	69	0	1
1978	32.8	−26.96666667	1515437	0	South Africa	0	0	0	0	0	0	0	0	0	0	0	0	51	0	1
1979	32.8	−26.96666667	1515437	1	South Africa	0	0	0	0	0	0	0	0	0	0	0	0	51	0	1
1980	27.38333333	−24.61666667	1485165	0	South Africa	0	0	0	0	0	0	0	0	0	0	0	0	1087	0	1
1981	28.538	−29.752	1552132	0	Lesotho	1	0	0	0	0	1	1	1	0	1	1	0	2716	1	1
1982	30.16666667	−23.83333333	1474382	0	South Africa	0	0	1	0	0	0	0	0	0	0	0	0	720	0	1
1983	19.71666667	−33.26666667	1597439	0	South Africa	0	0	0	0	0	0	0	0	0	0	0	0	934	0	1

Table 4. Output from the *dmsparse* function with certain fields removed. The field x_dms contains the text strings of x-coordinates that were used in the input to *dmsparse* and parsed into degrees (xdeg), minutes (xmin) and seconds (xsec). Coordinates that have been converted into decimal degrees are given in the column labeled 'x'

Place	x_dms	xdeg	xmin	xsec	EW	x
Chimoio	33 28.9 E	33	28	54	E	33.48167
Grahamstown	26d31m59.98 E	26	31	59.98	E	26.53333
Kenton	26°38′59″E	26	38	59	E	26.64972
Ladybrand	27°27′E	27	27	0	E	27.45
Maun	23 25E	23	25	0	E	23.41667
Mwinilunga	E 24 25 59.9880	24	25	59.988	E	24.43333
Pretoria	28°13 45.9840 E	28	13	45.984	E	28.22944
Tsumeb	17 43 0.0120 E	17	43	0.012	E	17.71667
Frostburg	78 55 42.3912 W	78	55	42.3912	W	−78.9284
San Francisco	122 25 9.4116 W	122	25	9.4116	W	−122.419
Seronera	34 49 13.1 E	34	49	13.1	E	34.82031
Paphos	32 25 47.1072 E	32	25	47.1072	E	32.42975
Alumine	070 55 11 W	70	55	11	W	−70.9197
Douala	009°56′41″E	9	56	41	E	9.944722
Mega	038°26′00″E	38	26	0	E	38.43333
Lausanne	006°40′00″E	6	40	0	E	6.666667
Moscow	037°36′56″E	37	36	56	E	37.61556
Harare	31.0 E	31	NA	NA	E	31
Trondheim	10.3999	NA	NA	NA	NA	10.3999
Maputo	32.58	NA	NA	NA	NA	32.58

Some of the functions available in the *biogeo* package are also available in other stand-alone software packages e.g. outlier detection in DivaGIS (< www.diva-gis.org >). Software tools are available for performing certain operations that are not available in *biogeo* e.g. the GBIF name parser for separating species names into component parts (< http://tools/gbif/org/nameparser >) and obtaining coordinates from text descriptions (BioGeomancer, Guralnick et al. 2006). Many other useful tools for managing collections data can be found on the GBIF website (< www.gbif.org/resource-type/tool >).

Example application of *biogeo*

To demonstrate some of the key features of *biogeo* we used a dataset of insect records from southern Africa containing 21 species with several occurrences per species. We renamed the species with letters and included some known errors in order to demonstrate the capability of the package.

The function *alternatives* was applied to the full dataset of records. All records should be plotted in southern Africa, but the point with the identifier 732 in Egypt is clearly an error (Fig. 1). By selecting this point the alternative positions for that point are indicated as purple dots with broken lines leading to them. All other records for that species are indicated as points in black and the records of all other species in the dataset are indicated as blue points. By clicking on the alternative point in southern Africa, the coordinates will automatically be updated to that position and the original incorrect coordinates for the point will be stored.

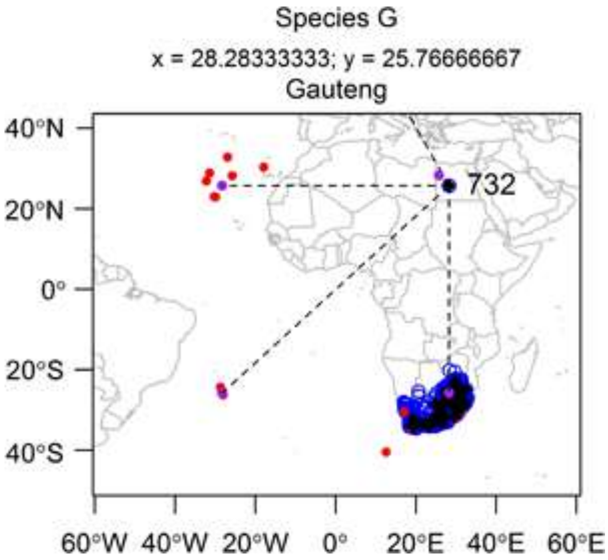

Figure 1. Alternative positions (purple points) for the point with the identifier 732, generated using the function *alternatives*. Records for the species that is found at the selected outlier (Species G) are indicated in black and records for all other species are in blue. Records that fall outside of country boundaries are shown in red.

The function *alternatives2* was applied to a single species (Species A, Fig. 2). This function plots only the points for the selected species. The correct position for the record labelled 39 (Fig. 2) is indicated by the red arrow. The selection of this point instead of the other point in South Africa was based on the locality description for the point (Kosi Bay), which is displayed at the top of the map.

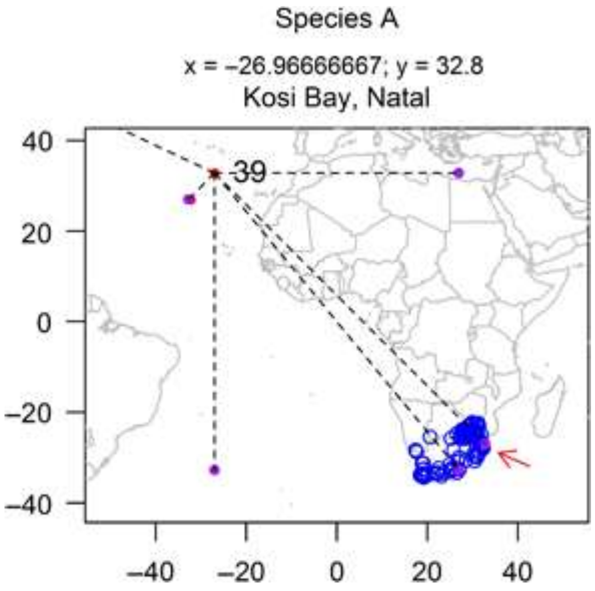

Figure 2. Alternative point records for a selected species (Species A) using the function *alternatives2*. Only the records for Species A are displayed. The red arrow indicates the correct position for the incorrect record labeled as 39.

The interplay between geographical and environmental space is important in species distribution modeling (Elith and Leathwick 2009), but there appear to be few tools to easily examine the distribution of points in both geographical and environmental space at the same time. Several functions make use of geographical and environmental space to detect possible errors and correct errors (e.g. Fig. 3 and 4). The function *geo2envpca* was applied to a single species and demonstrates the use of geographical and environmental space for identifying errors (Fig. 4). The point selected on the map (1981), which occurs in the highlands of Lesotho, is a clear environmental outlier in the environmental space that was defined by performing a principal components analysis on five climatic variables.

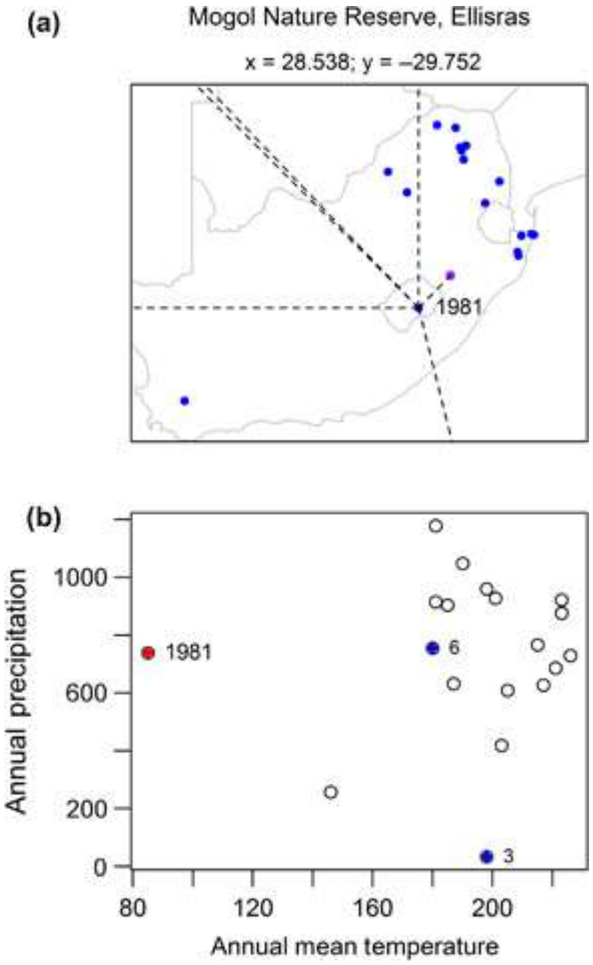

Figure 3. Outputs from the function *alternativesenv*. The alternative points for the record selected (identifier 1981) in the map on the above (a) are displayed in a two-dimensional environmental space below (b) (as blue points numbered 3 and 6). The environmental space is defined in this example by annual precipitation and annual mean temperature.

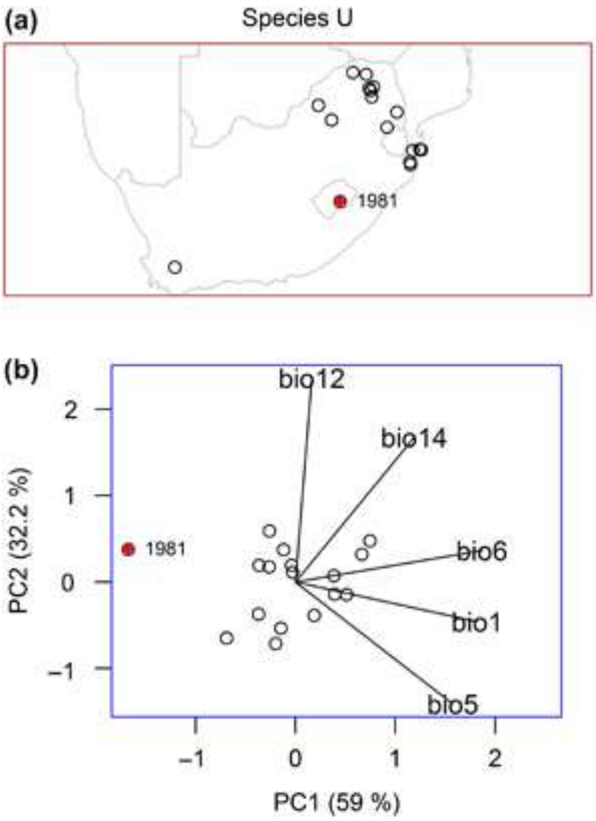

Figure 4. Outputs from the function *geo2envpca*, showing the geographical space above (a) and the environmental space as defined by principal components from a principal components analysis below (b). The environmental variables are: bio 1 – annual mean temperature, bio5 – maximum temperature of warmest month, bio6 – minimum temperature of coldest month, bio12 – annual precipitation, bio14 – precipitation of driest month.

The function *alternativesenv* was applied to a single species and demonstrates the use of alternatives with an environmental space defined by the values of two climatic variables (Fig. 3). The alternative points for the record selected (1981) in the map on the left are displayed in a two-dimensional environmental space on the right, where point 6 appears to be plausible in terms of its proximity in the climatic space to the other records for the species (blue points Fig. 3b).

The function *errorcheck* was run for Species U, the records of which are shown in Fig. 4. A screen shot with some of the fields and records removed is shown in Table 3. For the fourth record (ID 1973) a country mismatch error was recorded because the point was plotted outside the borders of any country, thus returning NA for the country_ext field and a countryMismatch error. For the 12th record (ID 1981) the record was incorrectly plotted in Lesotho (see outlier in Fig. 3), resulting in a country mismatch and being identified as an outlier for several of the environmental variables including bio1 – annual mean temperature; bio5 – maximum temperature of warmest month and bio6 – minimum temperature of coldest month. The 13th record (ID 1982) has low precision coordinates as both the x and y-coordinates were recorded at the top left corner of a 10 minute grid cell.

The *dmsparse* function was applied to a set of coordinates in various text formats for the x-coordinate (x_dms in Table 4) to parse these coordinates into separate fields for degrees, minutes and seconds. The coordinates that are all in different text formats have been successfully parsed into degrees, minutes and seconds. The last two places (Maputo and Trondheim) are recognized as being in decimal degrees and so NA values are assigned to the degrees, minutes and seconds columns.

In summary, this package provides users with a set of functions for easily detecting common errors and data quality issues with occurrence datasets sourced from collections datasets. Most importantly, several of the functions assist the user in correcting the errors in the dataset, rather than simply detecting and excluding them.

To cite *biogeo* or acknowledge its use, cite this Software note as follows, substituting the version of the application that you used for 'version 0':

Robertson, M. P., Visser, V. and Hui, C. 2016. Biogeo: an R package for assessing and improving data quality of occurrence record datasets. – Ecography 39: 394–401 (ver. 0).

Acknowledgements

This research was supported by DST-NRF Centre for Invasion Biology, the National Research Foundation and Univ. of Pretoria.

References

Aiello-Lammens, M. A. et al. 2015. spThin: an R package for spatial thinning of species occurrence records for use in ecological niche models. – Ecography 38: 1–5.

Boakes, E. H. et al. 2010. Distorted views of biodiversity: spatial and temporal bias in species occurrence data. – PLoS Biol. 8: e1000385.

Chamberlain et al. 2015. Rgbif package – interface to the Global Biodiversity Information Facility API v. 0.8.9. – < https://cran.r-project.org/web/packages/rgbif/rgbif.pdf >.

Chapman, A. D. 2005a. Uses of primary species-occurrence data, version 1.0. – Report for the Global Biodiversity Information Facility, Copenhagen.

Chapman, A. D. 2005b. Principles and methods of data cleaning – primary species and species-occurrence data, version 1.0. – Report for the Global Biodiversity Information Facility, Copenhagen.

Edwards, J. L. 2004. Research and societal benefits of the Global Biodiversity Information Facility. – Bioscience 54: 485–486.

Elith, J. and Leathwick, J. R. 2009. Species distribution models: ecological explanation and prediction across space and time. – Annu. Rev. Ecol. Evol. Syst. 40: 677–697.

Franklin, J. 2009. Mapping species distributions: spatial inference and prediction. – Cambridge Univ. Press.

Freeley, K. J. and Silman, M. R. 2010. Modelling the responses of Andean and Amazonian plant species to climate change: the effects of georeferencing errors and the importance of data filtering. – J. Biogeogr. 37: 733–740.

Funk, V. A. and Richardson, K. S. 2002. Systematic data in biodiversity studies: use it or lose it. – Syst. Biol. 51: 303–316.

Gaston, K. J. 2000. Global patterns in biodiversity. – Nature 405: 220–227.

Gaston, K. J. and Fuller, R. A. 2009. The sizes of species' geographic ranges. – J. Appl. Ecol. 46: 1–9.

Graham, C. H. and Hijmans, R. J. 2006. A comparison of methods for mapping species ranges and species richness. – Global Ecol. Biogeogr. 15: 578– 587.

Graham, C. H. et al. 2004. New developments in museum based informatics and applications in biodiversity analysis. – Trends Ecol. Evol. 19: 497–503.

Guisan, A. and Zimmermann, N. E. 2000. Predictive habitat distribution models in ecology. – Ecol. Model. 135: 147–186.

Guralnick, R. P. et al. 2006. BioGeomancer: automated georeferencing to map the world's biodiversity data. – PLoS Biol. 4: e381.Hijmans, R. J. and Elith, J. 2015. Species distribution modelling with R. – < https://cran.r-project.org/web/packages/dismo/vignettes/sdm.pdf >.

Hijmans, R. J. et al. 1999. Using GIS to check coordinates of germplasm accessions. – Genet. Resour. Crop Evol. 46: 291–296.

Hijmans, R. J. et al. 2015. Dismo: species distribution modelling package v. 1.0-12. – < https://cran.r-project.org/web/packages/dismo/index.html >.

Hortal, J. et al. 2007. Limitations of biodiversity databases: case study on seed–plant diversity in Tenerife (Canary Islands). – Conserv. Biol. 21: 853–863.

Hortal, J. et al. 2008. Historical bias in biodiversity inventories affects the observed environmental niche of the species. – Oikos 117: 847–858.

Hui, C. et al. 2011a. Macroecology meets invasion ecology: linking native distribution of Australian acacias to invasiveness. – Divers. Distrib. 17: 872–883.

Hui, C. et al. 2011b. Defining optimal sampling effort for large-scale monitoring of invasive alien plants: a Bayesian method for estimating abundance and distribution. – J. Appl. Ecol. 48: 768–776.

IUCN 2012. IUCN Red List categories and criteria version 3.1. – Gland, Switzerland.

Joseph, L. N. and Possingham, H. P. 2008. Grid-based monitoring methods for detecting population declines: sensitivity to spatial scale and consequences of scale correction. – Biol. Conserv. 141: 1868–1875.

Lamanna, C. et al. 2014. Functional trait space and the latitudinal diversity gradient. – Proc. Natl Acad. Sci. USA 111: 13745–13750.

Maldonado, C. et al. 2015. Estimating species diversity and distribution in the era of Big Data: to what extent can we trust public databases? – Global Ecol. Biogeogr. 24: 973–984.

Newbold, T. 2010. Applications and limitations of museum data for conservation and ecology, with particular attention to species distribution models. – Prog. Phys. Geogr. 34: 3–22.

Novoa, A. et al. 2015. Introduced and invasive cactus species: a global review. – AoB Plants 7: plu078.

Ponder, W. F. et al. 2001. Evaluation of museum collection ata for use in biodiversity assessment. – Conserv. Biol. 15: 648–657.

Reddy, S. and Davalos, M. 2003. Geographical sampling bias and its implications for conservation priorities in Africa. – J. Biogeogr. 30: 1719–1727.

Richardson, D. M. et al. 2011. Human-mediated introductions of Australian acacias – a global experiment in biogeography. – Divers. Distrib. 17: 771–787.

Robertson, D. R. 2008. Global biogeographical data bases n marine fishes: caveat emptor. – Divers. Distrib. 14: 891–892.

Robertson, M. P. and Barker, N. P. 2006. A technique for evaluating species richness maps generated from collections data. – S. Afr. J. Sci. 102: 77–84.

Robertson, M. P. et al. 2010. Getting the most out of atlas data. – Divers. Distrib. 16: 363–375.

Suarez, A. V. and Tsutsui, N. D. 2004. The value of museum collections for research and society. – Bioscience 54: 66–74.

Swenson, N. G. et al. 2012. The biogeography and filtering of woody plant functional diversity in North and South America. – Global Ecol. Biogeogr. 21: 798–808.

Weiser, M. D. et al. 2007. Latitudinal patterns of range size and species richness of New World woody plants. – Global Ecol. Biogeogr. 16: 679–688.

Yesson, C. et al. 2007. How global is the global biodiversity information facility? – PLoS One 2: e1124.