1 *naturaList*: a package to classify occurrence records in levels of confidence in species
 2 identification

3

4 *naturaList* R package

5

6 Arthur Vinicius Rodrigues^{1*}, Gabriel Nakamura¹ and Leandro Duarte¹

7

8 ¹Departamento de Ecologia, Universidade Federal do Rio Grande do Sul, CP 15007,
 9 Porto Alegre, 91501-970, Brazil.

10 * Corresponding author: rodrigues.arthur.v@gmail.com

11

12

13 Abstract

14 **1.** There is a big volume of occurrence records available in biodiversity databases, but
 15 researchers should guarantee its quality before use it in scientific studies. A problem
 16 that might compromise the quality of occurrence data is species misidentification. We
 17 address this issue by presenting *naturaList*, a R package designed to classify species
 18 occurrence data according to identification reliability.

19 **2.** *naturaList* allows to classify species occurrences up to six levels of confidence in
 20 species identification, and to filter occurrence data accordingly. The highest level of
 21 confidence is assigned to records identified by a specialist, whose name must be
 22 provided by the user. The other five levels of confidence are derived from the
 23 occurrence data. We demonstrate *naturaList* functions using occurrences of *Alsophila*
 24 *setosa*, a tree fern species from Atlantic Forest, as example. We classified and filtered
 25 data in grid cells in order to maintain only the highest-level records in each cell. Then
 26 we selected only those records classified in the two highest levels of confidence.

27 **3.** From 323 occurrences of *Alsophila setosa* displaying geographic coordinates, 69
 28 (21%) were identified by a specialist. After filtering the highest-level records inside grid
 29 cells, 102 records remained. From these grid cell filtered data, 38 occurrences (37%)
 30 were classified into the highest confidence level. Three records were removed using an
 31 interactive map module, due to falling in sea sites or outside the native range size of the
 32 species. Since we selected only records classified in the two highest levels of
 33 confidence, the final dataset contained 94 occurrence records.

34 **4.** *naturaList* guarantees the reproducibility of occurrence data processing and cleaning.
 35 Macroecologists, biogeographers and taxonomists might benefit from using *naturaList*
 36 package to evaluate the quality of species identification in occurrence data and by
 37 identify sites that need evaluation of taxonomic classification of species.

38

39 **Keywords:** biodiversity databases, confidence, filtering of occurrence records,
40 identification reliability, identification errors, misidentification, occurrence records,
41 specialists.

42

43

44 **Introduction**

45 Open biodiversity databases (e.g. Global Biodiversity Information Facility – GBIF)
46 provide data essential to studies of global biodiversity patterns. However, their use
47 imposes big challenges regarding the quality of occurrence records data (Yesson et al.,
48 2007; Anderson, 2012; Serra-Diaz, Enquist, Maitner, Merow, & Svenning, 2017).
49 Solutions to problems related to georeferencing of occurrence records (Yesson et al.,
50 2007) have gained attention and some reproducible solutions (Serra-Diaz et al., 2017;
51 Zizka et al., 2019) have been presented. Undoubtedly, georeferencing quality is an
52 important issue in species occurrence records; nonetheless, the reliability of species
53 identification itself is paramount to guarantee data quality (Anderson, 2012; Araújo et
54 al., 2019; Feng et al., 2019). This problem lacks automated processes to cope with,
55 which is ideal to deal with big data. To assure for the quality of species identification
56 remains being a subjective task, usually depending on expert taxonomists, or worst,
57 being ignored.

58 There are some challenges in detect species identification errors from
59 biodiversity databases. First, the user must have taxonomic skills to perform manual
60 checking, and it is possible only if the occurrence records have images associated to
61 them; nonetheless, even for experts, images are not always enough to confirm species
62 identity. Second, depending on the amount of occurrence records, manual checking

could be very time consuming, or even impossible to conclude. The presence of misidentification in datasets used in Species Distribution Models (SDM) may impact the performance and prediction of the models, enhancing the uncertainty in these SDM (Anderson, 2012; Araújo et al., 2019). Therefore, a method that account for such inaccuracies in the data by identifying and classifying them in levels of quality is of great importance because allows to cope with big data in a fast and reproducible way.

Thus, our aim is to present the *naturaList* R package, which provides tools for classifying occurrence records based on well-defined confidence levels of species identification. Further, *naturaList* enables to filter occurrence records within spatial grid cells based on the highest confidence level of species identification and provides an interactive module that allows to check out occurrence records in space and select them. We demonstrate package tools by applying *naturaList* functions to classify occurrence records of a tree fern species from South America.

Materials and Methods

We propose a method for classification of the confidence in species determination based on the information provided by occurrence records databases. Our rationale is that the most reliable identification of a species is made by an expert taxonomist in the taxon (Araújo et al., 2019). To check if a species was determined by a specialist, *naturaList* user must provide a list of specialist names to be compared with the taxonomist name that determined the species. Moreover, other information is extracted from the occurrence dataset to produce subsequent confidence levels, detailed in the next section.

naturaList R package

87 The *naturaList* package can be installed from github using devtools package (Wickham,
88 Hester, & Chang 2019) with the code
89 `devtools::install_github("avrodrigues/naturaList", built_vignette =`
90 `TRUE)`. The package has functions that allow the user to classify, filter and check
91 species occurrences data, as described in the Table 1.

92 **Table 1.** Names and general description of the functions of the *naturaList* package.

Function	Description
	Classifies the occurrence records based on confidence in species
<i>classify_occ</i>	determination.
<i>create_spec_df</i>	Creates a specialist data frame from character vector.
	Facilitates the search for non-taxonomist strings in the 'determined.by'
<i>get_det_names</i>	column of occurrence records dataset.
	Filters the occurrence with most confidence in species identification
<i>grid_filter</i>	inside grid cells.
<i>map_module</i>	Checks out the occurrence records in an interactive map module.

93

94 The classification in confidence levels of species identification is done with
95 *classify_occ* function. This function demands a data frame with the occurrence data and
96 a data frame with specialist names. Each occurrence record is compared with the criteria
97 from lower to higher levels of confidence and it is flagged with the highest criterion it
98 met (Fig. 1). By default, *classify_occ* function uses six confidence levels (see Table 2),
99 which, except for level 1, can be reordered according to the adequacy of the study. The
100 output is a data frame containing the occurrence records with all the original
101 information presented in the input dataset plus a column named 'naturaList_levels',

102 indicating the confidence level of the occurrence record (codes are presented in Table
103 2).
104

105
106 **Figure 1.** The scheme shows the classification procedure of occurrence records made
107 by *classify_occ* function. Based on data frames of occurrence records and specialist
108 names, *classify_occ* function checks for each occurrence record with the criteria of
109 levels of confidence from lower to higher level and flags the occurrence record when
110 the highest criterion it met.

111

112 **Table 2.** Levels of confidence in species determination used by *classify_occ* function.

113 The level codes identify the criteria used and the numbers identify the order chosen by

114 the user. The ordering presented in this table is the default order used in the function,

115 which may be altered by the user. Level 1 is the highest level of confidence.

Level code	Description	Default level order
det_by_spec	Species name is determined by a specialist in the biological group, if not;	1
taxonomist	Taxonomist's name is provided, if not;	2
image	Occurrence has an associated image, if not;	3
sci_colection	Specimen is part of a scientific collection, if not;	4
field_obs	Occurrence is determined in a field observation, if not;	5
no_criteria_met	No previous criteria are met.	6

116

117 The occurrence dataset demanded by *classify_occ* must have at least the

118 columns with the information showed in Table 3. By default, these column names

119 follow the Darwin Core format (Wieczorek et al., 2012; <https://dwc.tdwg.org/>), thus

120 data from databases that uses this format, e.g. GBIF, could be readily used in this

121 function. Furthermore, other formats can be used by informing the corresponding

122 column name in the function arguments. The specialist data frame demanded by

123 *classify_occ* may be easily created from a character vector with *create_spec_df*

124 function, which is ready to use in *classify_occ*.

125

126 **Table 3.** General format for occurrence records data frame used in *classify_occ*
 127 function.

	Default name for		
Argument in	columns (Darwin		
<i>classify_occ</i>	Core format)	Class	Description
			Code that identifies the occurrence
occ.id	occurrenceID	character	record in the source database
scientific.name	species	character	Binomial scientific name
longitude	decimalLongitude	numeric	Longitude in decimal degrees
latitude	decimalLatitude	numeric	Latitude in decimal degrees
year.event	year	numeric	Year of record event
determined.by	identifiedBy	character	Taxonomist name
date.identified	dateIdentified	character	Date of identification
			Name of the institution that provided
institution.source	institutionCode	character	the occurrence record
collection.code	collectionCode	character	Code for the institution name
			Code for occurrence record in the
catalog.number	catalogNumber	character	institution that holds the record
			Type of media associated to the
media.type	mediaType	character	record ("STILLIMAGE")
			Type of record
			("HUMAN_OBSERVATION" or
basis.of.rec	basisOfRecord	character	"PRESERVED_SPECIMEN")

128

129 To classify an occurrence record as identified by a specialist, the algorithm in
130 *classify_occ* searches first if the surname of one or more specialist is present in the
131 ‘determined.by’ column of the occurrence dataset. For those records that specialist
132 surname matches to the names presented in ‘determined.by’ column the algorithm
133 proceed in two ways: if the name in ‘determined.by’ column perfectly matches with
134 specialist name, then the occurrence is classified at level 1. Otherwise, the function lists
135 all strings that need verification by the user to assure it corresponds to a specialist name.
136 In the manual verification, the user is asked by the function if the string printed in the R
137 console corresponds (y), or do not (n) to a specialist’ name.

138 The classification process depends both on the quality of specialist dataset
139 provided and the strings in the column indicated in the “determined.by” argument of
140 *classify_occ* function. In some datasets, some strings may do not correspond to a valid
141 name of a taxonomist (e.g. the string “unknown”). To overcome this problem, we
142 suggest the use of *get_det_names* function to find strings that must be ignored as a
143 taxonomist name. Based on this check out, the user should provide a character vector to
144 *classify_occ* function in “ignore.det.names” argument, in order to ignore such strings
145 that are not taxonomists’ names.

146 The *grid_filter* function selects the occurrence with highest level of confidence
147 inside a spatial extent, given by a cell size defined by the user or from a raster layer.
148 Thus, this function returns a data frame only with the best classified occurrence record
149 of each cell. If two or more occurrences have the same confidence level in a cell, the
150 function gives priority for occurrence that was 1) more recently determined; 2) more
151 recently observed. If the equality persists or if a date was not provided, one of the
152 occurrences is randomly chosen.

153 The *map_module* function is an interactive application that allows the user to
154 manually filter records in the map and easily access the information of each record. This
155 function could operate with output from *classify_occ* or *grid_filter* functions. In this
156 map, the occurrence points are coloured according to the confidence levels
157 classification. By clicking over a point, the user can see some information about that
158 occurrence. When setting on the option ‘delete points with click’, clicking over a point
159 works to delete that occurrence record. Furthermore, the user may select the
160 classification levels to filter points or yet may draw a polygon to select occurrence in a
161 region. These functionalities are helpful to find for spatial outliers and/or occurrences
162 out of the native range (if the native range is known).

163 **Example**

164 To demonstrate the use of *naturaList* package, we conducted a classification of the
165 occurrence records of *Alsophila setosa*, a tree fern species from Atlantic Forest. For
166 this, we downloaded all the occurrence records of *Alsophila setosa* species from GBIF
167 (GBIF, 2019). We assumed that occurrence records had no issues regarding
168 geographical coordinates and did not conduct any procedure to remove records before
169 performing the classification procedure. To produce a list of specialists, we took the
170 names of the authors of a checklist of ferns and lycophytes species of Brazil (Prado et
171 al., 2015). Both datasets are available in the *naturaList* package.

172 We first conducted the classification of occurrences according to the six
173 confidence levels, using default settings in *classify_occ*. Then we filtered the occurrence
174 records in a grid cell with size of 0.5° with *grid_filter* function. The size of the grid cell
175 was arbitrarily chosen only to show how to use the function. Finally, we checked if the
176 occurrence points were inside the native range using *map_module* function. We
177 considered information available in Flora do Brasil (<http://floradobrasil.jbrj.gov.br>),

178 which inform the Brazilian states where species occur, and Plants of The World Online
179 (<http://www.plantsoftheworldonline.org/>), which provides political regions where the
180 species occur in South America, to define the native distribution range of *A. setosa*. The
181 code used to produce this example is provided in the supporting information.

182 Additionally, an introduction to the package can be found with the code:

183 `vignette("naturalList_vignette").`

184

185 **Results**

186 The dataset of *Alsophila setosa* has 508 occurrence records. After using *classify_occ* the
187 number of records decreased to 323 due to records without coordinates in the raw GBIF
188 dataset, which were automatically removed by the default setting of *classify_occ*. The
189 *classify_occ* function asked for manual checking of four strings for specialist names,
190 which were further confirmed as true ones. From the 323 occurrences returned, 69
191 occurrences were classified as level 1, 218 as level 2, 12 as level 3 and 24 as level 4,
192 which represented 21.4%, 67.5%, 3.7% and 7.4% of the total classified occurrence
193 records, respectively.

194 The dataset with 323 classified occurrence records was then filtered in grid cells
195 with 0.5° of size. This filtering reduced the dataset to 102 occurrences, in which 38
196 (37.3%), 58 (56.9%), 3 (2.9%), 3 (2.9%) were classified, respectively, in levels 1 to 4.
197 The grid cell filtering enhanced the representativeness of occurrences classified in the
198 level 1 (from 21.1 to 37.3%), by removing occurrences classified in lower levels that are
199 placed in the same grid cell.

200 Finally, we used this dataset with 102 occurrences in the *map_module* to
201 visually check for potential errors and to select only occurrence records in levels 1 and 2
202 (Fig. 2). First, we deleted three occurrence records, by setting on the button 'delete

203 points with click' (Fig. 2b-II) and clicking on them; two of them were deleted because
 204 were placed in the sea and one because was out of the native range recognized by Flora
 205 do Brasil (the 'x' in Fig. 2b). Then, we selected the levels 1 and 2 to be maintained in
 206 the output dataset (Fig. 2b-I). The current selected records may be visualized above the
 207 'Done!' button (Fig. 2b-III). Note that selection made in the Fig. 2b-IV only serves to
 208 display the points in the map. Polygons should also serve as spatial selection that the
 209 user can draw with tools in the left side of the map (Fig 2b-V). Finally, we clicked in
 210 'Done!' button to assign the selected points to an R object. At the end of these
 211 procedure 94 occurrence records were maintained in the occurrence dataset.

212

213 **Figure 2.** Features of *map_module* function used in the example with *Alsophila setosa*

214 occurrence data. a) the initial display of occurrence records data; b) highlights the

215 features of *map_module*. I – Select levels to be maintained in the output dataset; II –

216 enable to delete point with click over occurrences; III – show how many points have

217 been selected from the initial dataset; IV – display in the map only occurrences that

218 belong to the selected levels, and turn visualization of the map between OpenStreetMap
219 and Satellite; V – options to draw and edit polygons to select occurrence records that
220 fall within polygon; x – highlights the deleted occurrence points.

221

222 **Discussion**

223 Our example demonstrated that *naturaList* enables researchers to classify and quantify,
224 based on well-defined criteria, the reliability of species identification in occurrence
225 records from biodiversity databases. This may assist researchers to decide which levels
226 of confidence is suitable to their study objective. Also, our example showed that the
227 proportion of records with high quality can be improved after the use of *grid_filter*
228 function, therefore it might enhance data quality to be used in SDMs. This increase,
229 however, is dependent on the size of grid cell used in the filtering and the spatial
230 distribution of records identified by specialists. The usefulness of such grid cell filtering
231 is to evaluate the reliability of identification in the same scale of the environmental
232 layer used to fit SDM. To this end, in the *grid_filter* the user may provide a raster layer,
233 from which the function will use the cells to conduct the filtering.

234 Although SDM is the most popular application aiming to produce accurate
235 estimates of species distribution, occupancy models have been also used for this
236 purpose (Kéry, Gardner, & Monnerat, 2010). The classification of occurrence records
237 regarding confidence in the identification might be useful for occupancy models that
238 incorporate false positive probabilities (Royle & Link, 2006; Miller et al., 2011).

239 The tools described here could be also useful for taxonomists, that might
240 identify areas in which there are occurrence records for a given taxon, but with
241 uncertain reliability. Based on that information, the specialist could confirm (or not) the

242 occurrence of the species – by visiting the areas or a scientific institution with preserved
243 material – and thereby improving the quality of identification.

244 On one hand, the tools in *naturaList* enable researchers to account for the quality
245 of species identification in big datasets as well as to report the steps used in the cleaning
246 of data, which enhance the reproducibility of such studies. On the other hand, those
247 tools might enhance the quality of species identification by guiding specialists in
248 taxonomy to revise specimens with low determination confidence.

249

250 **Acknowledgments**

251 We are thankful to the 2019 GBIF Ebbe Nielsen Challenge by recognize the importance
252 of the tools described here. AVR is thankful to PhD grant from Coordenação de
253 Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.
254 LD research activities have been supported by CNPq Productivity Fellowship (grant
255 307527/2018-2). LD and GN research have been conducted in the context of the
256 National Institutes for Science and Technology (INCT) in Ecology, Evolution and
257 Biodiversity Conservation, supported by MCTIC/CNPq (proc. 465610/2014-5) and
258 FAPEG (proc. 201810267000023).

259

260 **Author contributions:** A.V.R. conceived the idea. A.V.R and G.N implement the
261 package. All authors contributed in discussions, suggestions and in the writing of the
262 manuscript.

263

264 **Conflict of Interest**

265 The authors declare no conflict of interest.

266

267 **Data availability**

268 The data used in the example and the package are available at

269 <https://github.com/avrodrigues/naturalList>

270

271 **References**

272 Anderson, R. P. (2012). Harnessing the world's biodiversity data: Promise and peril in

273 ecological niche modeling of species distributions. *Annals of the New York*

274 *Academy of Sciences*, 1260(1), 66–80. doi:10.1111/j.1749-6632.2011.06440.x

275 Araújo, M. B., Anderson, R. P., Barbosa, A. M., Beale, C. M., Dormann, C. F., Early,

276 R., ... Rahbek, C. (2019). Standards for distribution models in biodiversity

277 assessments. *Science Advances*, 5(1). doi:10.1126/sciadv.aat4858

278 Feng, X., Park, D. S., Walker, C., Peterson, A. T., Merow, C., & Papeş, M. (2019). A

279 checklist for maximizing reproducibility of ecological niche models. *Nature*

280 *Ecology & Evolution*, 3(October), 1–15. doi:10.1038/s41559-019-0972-5

281 GBIF.org (08 July 2019) GBIF Occurrence Download.

282 <https://doi.org/10.15468/dl.6jesg0>

283 Kéry, M., Gardner, B., & Monnerat, C. (2010). Predicting species distributions from

284 checklist data using site-occupancy models. *Journal of Biogeography*, 37(10),

285 1851–1862. doi:10.1111/j.1365-2699.2010.02345.x

286 Miller, D. A., Nichols, J. D., Mcclintock, B. T., Campbell, E. H., Bailey, L. L., Weir, L.

287 A., & Grant, E. H. C. (2011). Improving occupancy estimation when two types of

288 observational error occur: non- detection and species misidentification Stable

289 URL: <http://www.jstor.org/stable/23035095> REFERENCES Linked references

290 are available on JSTOR for this article: Improving oc. *Ecology*, 92(7), 1422–

291 1428.

292 Prado, J., Sylvestre, L. D. S., Labiak, P. H., Windisch, P. G., Salino, A., Barros, I. C. L.,
 293 ... Matos, F. B. (2015). Diversity of ferns and lycophytes in Brazil. *Rodriguesia*,
 294 66(4), 1073–1083. doi:10.1590/2175-7860201566410

295 Royle, J. A., & Link, W. A. (2006). Generalized site occupancy models allowing for
 296 false positive and false negative errors, 87(4), 1–7. doi:10.1890/0012-9658

297 Serra-Diaz, J. M., Enquist, B. J., Maitner, B., Merow, C., & Svenning, J. C. (2017). Big
 298 data of tree species distributions: how big and how good? *Forest Ecosystems*, 4(1),
 299 0–12. doi:10.1186/s40663-017-0120-0

300 Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Döring, M., Giovanni, R., ...
 301 Vieglais, D. (2012). Darwin core: An evolving community-developed biodiversity
 302 data standard. *PLoS ONE*, 7(1). doi:10.1371/journal.pone.0029715

303 Wickham, H., Hester, J. and Chang, W.(2019). devtools: Tools to make developing R
 304 packages easier. R package version 2.2.1. [https://CRAN.R-](https://CRAN.R-project.org/package=devtools)
 305 [project.org/package=devtools](https://CRAN.R-project.org/package=devtools)

306 Yesson, C., Brewer, P. W., Sutton, T., Caithness, N., Pahwa, J. S., Burgess, M., ...
 307 Culham, A. (2007). How global is the global biodiversity information facility?
 308 *PLoS ONE*, 2(11). doi:10.1371/journal.pone.0001124

309 Zizka, A., Silvestro, D., Andermann, T., Azevedo, J., Duarte Ritter, C., Edler, D., ...
 310 Antonelli, A. (2019). CoordinateCleaner: Standardized cleaning of occurrence
 311 records from biological collection databases. *Methods in Ecology and Evolution*,
 312 10(5), 744–751. doi:10.1111/2041-210X.13152

313