Title: GridDER: Grid Detection and Evaluation in R

Authors: Xiao Feng[1*], Tainá Rocha[1,2], Hanna T Thammavong[3], Rima Tulaiha[3], Xin Chen[1], Yingying Xie[3], Daniel Park[3,4]

[1]Department of Geography, Florida State University, Tallahassee, FL 32306, USA

[2]Botanical Garden Research Institute of Rio de Janeiro, Rua Jardim Botânico 1008, 22470- 19 180 Rio de Janeiro, Brazil

[3]Department of Biological Sciences, Purdue University, West Lafayette, IN 47907, USA

[4]Purdue Center for Plant Biology, Purdue University, West Lafayette, IN 47906, USA

*Corresponding author: Xiao Feng, fengxiao.sci@gmail.com

Conflict of Interest The authors declare no conflicts of interest.

Biosketch The authors share a common research interest in biodiversity informatics, with special interest in developing tools to integrate and analyze biodiversity data, in order to advance the discovery, study, and preservation of biodiversity.

Title: GridDER: Grid Detection and Evaluation in R

Running title: GridDER

Abstract

Aim Observations and collections of organisms form the basis of our understanding of Earth's biodiversity and are an indispensable resource for global change studies. Geographic information is key, serving as the link between organisms and the environments they reside in. However, the geographic information associated with these records is often inaccurate, thus limiting their efficacy for research. Along these lines multiple solutions for identifying erroneous coordinate data and records georeferenced to centroids or landmarks have been developed. Another prominent, but less discussed and documented source of inaccuracies arises due to the use of gridded survey systems in many regions of the world.

Innovation Here we present GridDER, a tool for identifying biodiversity records that have been designated locations on widely used grid systems. Our tool also estimates the degree of environmental heterogeneity associated with grid systems, allowing users to make informed decisions about how to use such occurrence data in global change studies. We show that a significant proportion (~13.5%; 261 million) of records on GBIF, largest aggregator of natural history collection data, are potentially gridded data, and demonstrate that our tool can reliably identify such records and quantify the associated uncertainties.

Main conclusions GridDER can serve as a tool to not only screen for gridded points, but to quantify the geographic and environmental uncertainties associated with these records, which can be used to inform models and analyses that utilize these data, including those pertaining to global change.

Keywords: biogeography, biodiversity, GBIF, GIS, occurrences, spatial uncertainty

Introduction

The past few decades have witnessed an exponential increase of biodiversity data worldwide, which come from rapid digitization and aggregation of existing biodiversity collections and ongoing specimen collections, field surveys, and human observations (Feng *et al.*, 2022). For example, the Global Biodiversity Information Facility (GBIF) – the world's leading repository of biodiversity observations – recently reached ~2 billion records (accessed January 2022). Collections and observations of organisms form the basis of our understanding of Earth's biodiversity, and are an indispensable resource for studying the response of biodiversity under global change studies (Tingley *et al.*, 2009; Hedrick *et al.*, 2020).

The geographic information associated with primary biodiversity data (e.g., preserved specimens, field observations) serves as the key link between organisms and their environment, and such information has been used broadly in studies of ecology, biogeography, and beyond. For example, one of the most basic uses of the geographic information associated with biodiversity collections is studying the historical and current richness of species at various spatial scales. The locality or coordinates of species' occurrences can be associated with the environmental conditions to study their ecological niche and model their geographic ranges, an approach termed ecological niche modeling (or species distribution modeling) that got tremendous attention in the recent literature (Peterson *et al.*, 2011). For instance, such geographic information can also be used to study the changes of species' geographic distributions, such as range shifts under climate change (Blowes *et al.*, 2019), range reductions due to habitat loss and anthropogenic disturbance (Doughty *et al.*, 2016; Feng *et al.*, 2021), biological invasions (Feng

& Papeş, 2015; Park & Potter, 2015a,b), and the conservation of rare or endangered species (Hannah *et al.*, 2020).

However, aggregated biodiversity information is often prone to issues of data quality. Typical issues include missing values of certain fields, or fields filled with erroneous or non-standardized values. A notable example is the non-standardized scientific names in biodiversity databases that were aggregated from many sources; commonly biodiversity databases rely on software to automatically correct or standardize taxonomic information based on a reference checklist (e.g. (Boyle *et al.*, 2013)). A series of issues also exist for fields about the geographic information, such as inaccurate or unfilled coordinates and inconsistency between locality descriptions and coordinates. Since the aggregated data could be originally collected for various purposes using different protocols, they could be associated with varying degrees of uncertainties. The underlying spatial uncertainties are particularly relevant for studies that use aggregated geographic information to extract the environmental conditions (Park & Davis, 2017). As a response, many tools have been developed to solve the underlying issues in geographic information. For records with more complete geographic information, the locality description and coordinates can be used to cross-validate each other (Buitrago, 2020). For records with missing coordinates, georeferencing tools can be used to infer the coordinates based on locality descriptions (Guralnick *et al.*, 2006; GeoLocate, https://www.geo-locate.org). When precise locality information is not available, records are often georeferenced to the centroid of political divisions (e.g. country or state/province). Records georeferenced in this fashion are common in aggregated biodiversity databases, but the method of georeferencing (e.g. country centroid) is not always properly recorded. This is concerning because of the huge potential spatial uncertainty,

and thus multiple tools have been developed to flag out such records (e.g. CoordinateCleaner (Zizka *et al.*, 2019), Geocoordinate Validation Service, https://github.com/ojalaquellueva/gvs).

Another potentially prominent but less discussed source of inaccuracies arises from the use of grid systems to aggregate biodiversity data, where species occurrences are assigned to the center coordinates of predefined grids (Fig. 1). A major source of such gridded coordinates is grid-based surveys of species' geographic distribution, where the focal landscape is divided into equal-area grids (Franklin, 2010). The spatial resolution of the grid could be meters for local studies (Tinkham *et al.*, 2018; DeSisto *et al.*, 2020) or at much coarse resolution (e.g. 1km-500km) for regional, national, or global grid systems (e.g. Ordnance Survey National Grid or British National Grid). When gridded coordinates are aggregated in biodiversity databases, the metadata of the grid system (e.g. spatial resolution) are not always preserved, as those metadata are not supported by current biodiversity metadata standard (e.g. (Wieczorek *et al.*, 2012)). Our review of GBIF data revealed that approximately 261 million records (or 13.5% of GBIF records) (assessed on January 1, 2022) could be associated with grid systems of varying spatial resolution and extent (see section 2.2 and (Park *et al.* in review)). It is concerning that such large volumes of gridded coordinates could be applied to biodiversity studies without accounting for the potentially large degrees of spatial uncertainty (Franklin, 2010). It is actually difficult to detect the presence of gridded coordinates in individual studies that focus on one or few species or a relatively small geographic extent, since the pattern of regular spacing of gridded coordinates is usually more visible at aggregated scale across large breadths of space and taxa.

Here we present *GridDER*, a potential solution for detecting and evaluating gridded coordinates and grid systems. We first synthesized grid systems that were broadly used and developed a metadata standard to characterize grid systems. We also developed functions to infer the spatial attributes of grid systems, generate grid systems based on spatial attributes, and detect the origin of gridded coordinates.

2. Methods

2.1 Overview of a grid system

A grid system is a set of grids (squares/rectangles) originating from predefined spatial reference systems, which can be defined by a set of spatial attributes, including coordinate reference system (CRS), spatial resolution of horizontal and vertical dimensions, spatial extent, and spatial origin (minimum coordinate of all the left-bottom corners of all grids). The grid systems could have varied spatial attributes. For example, the CRS could be a geographic reference system (e.g. WGS84), then the unit of the spatial resolution will be arc-degrees/minutes/seconds; the CRS can also be a projected reference system (e.g. Universal Transverse Mercator; UTM), in which case the unit of the spatial resolution could be kilometers. The spatial extent of a grid system could be local (e.g. a country), regional (e.g. a continent), or global. For two grid systems that have the same CRS, extent and resolution, they could have slightly different origins, which can be caused by unintentional uncertainty in implementing the same non-digitized map, or caused by multiple independent implementations of the same grid design but without coordination for consistency. It is also possible to have a series of grid systems with a gradient of resolution (e.g. 1,2,10,100 km) while other attributes are the same.

2.2 Compile metadata of a grid system

We mainly focused on grid systems that have been applied broadly, in the sense that many biodiversity observations are associated with such grid systems. We started from the most duplicated coordinates from GBIF. We virtually checked the coordinates from QGIS and compiled a list of coordinates that show regular spacings among each other (here termed gridded occurrences) (Park *et al.* in review). We further checked the metadata (e.g. dataset key, publishing institution in Darwin Core), based on which we queried the original dataset that contains the gridded occurrences. Those datasets were used as the baseline for compiling the metadata of grid systems.

We compiled the following metadata of grid systems: name of a grid system (if available), country name/s and ISO-3166 code/s (if a grid system is applicable to one or multiple country), the European Petroleum Survey Group (EPSG) code of the CRS (EPSG hosts a database of CRSs), the resolution of the grid system on horizontal and vertical dimensions (size and unit), and the spatial extent (Table 1). We compiled the metadata from online documentations of a dataset or project, and/or acquired from the author or manager of the dataset. In situations where spatial data (vector or raster) of a grid system is available, we extracted metadata from the spatial data to supplement metadata information.

2.3 Simulating grid systems based on metadata

With the compiled metadata of grid systems, we simulated grid systems as spatial vectors. In brief, we developed *grid_generation* (Table 2; Fig. 2) to simulate regularly spaced points representing the four vertices of grids in a grid system. The vertices were then transformed as

spatial polygons (square or rectangles), crop to a predefined extent, or further masked to predefined polygons (e.g. country boundaries). The simulated grid systems were also validated with the gridded occurrences, and some of the metadata used during simulation were adjusted accordingly to better fit the existing occurrences. The simulated grid systems were exported as spatial vectors in shapefile format.

2.4 Simulating grid systems based on empirical gridded occurrences

We developed functions to simulate grid systems based on empirical gridded occurrences (Table 2; Fig. 2). This applies to situations when no metadata is available or when the metadata of a grid system is incomplete.

2.4.1 Inferring CRS

Among the metadata used for simulating a grid, the CRS is the most important attribute; it is almost always true for a spatial dataset. The CRS with which a grid system and gridded occurrences are originally defined could get lost during the data aggregation, where occurrences from different sources with different CRSs are commonly unified to the same CRS (e.g. WGS84). We developed a function (*infer_crs*) to infer the CRS of a grid system, using coordinates that are associated with an unknown grid system as input. In a nutshell, this inference builds upon a simple assumption -- the true CRS will lead to the most regular layout (distance and angle) of the gridded coordinates. The workflow of this inference is as following: 1) download gridded occurrences from GBIF and simplify the dataset to be unique spatial points (in CRS of WGS84) using latitude and longitude; 2) project the spatial points to ~6000 different registered CRSs that were compiled by GDAL (GDAL/OGR contributors, 2020); 3) for each

projected point, find its nearest neighbor along four directions (up, down, left, right); 4) calculate the distance between a focal point and its nearest neighbors (from previous step), along x-axis, y-axis, and both axes (in which case Euclidean distance is calculated), respectively; 5) calculate the angle between a focal point and its nearest neighbors; 6) the distances calculated from the up-side neighbors were pooled to find the mode and its frequency, the latter of which is used to quantify the regularity of spacing of projected points along y-axis; similarly, the distances calculated from left-side neighbors were used to quantify the regularity along x-axis; 7) the angles calculated from the up-side and left-side neighbors were divided by 90 degrees, and the reminders were polled to further calculate the mean, which is used to quantify the deviation from vertical or horizontal directions (the regularity of layout of projected points); 8) the ~6000 CRSs were ranked based on the frequency calculate from step 6 and the deviations calculate from step 7; ideally, the best CRS will lead to highest frequency in step 6 and close to 0 deviations in step 7.

2.4.2 Inferring resolution

In situations when a CRS, a unit of spatial distance, is known, and the spatial resolution is unknown, we developed a function (*infer_resolution*) to infer the spatial resolution based on input occurrences. The workflow of this inference is as following: 1) project the input points to the CRS where the grid system was originally defined; 2) for each projected point, find its nearest neighbor along four directions (up, down, left, right); 3) calculate the distance between a focal point and its nearest neighbors (from previous step), along x-axis and y-axis, respectively; 4) the distances calculated from the up-side neighbors were pooled to find the value with highest

frequency, which is considered as (or the approximation of) the resolution along y-axis; similarly, the distances calculated from left-side neighbors are used to quantify the resolution along x-axis.

2.4.3 Inferring origin of the grid

We developed a function (*infer_origin*) to calculate the origin of the grids, when the CRS and spatial resolutions are known. The origin is calculated as the minimum x and y among the input spatial points minus half of the resolution along x and y-axis, respectively.

2.4.4 Defining spatial extent

The spatial extent of a grid system can be calculated by the *infer_extent* function. The extent could be simply as the spatial extent where a CRS is defined, the bounding box of a country boundary, or the bounding box of the gridded occurrences. The spatial extent could also be further masked to be along the boundary of a country. Additional buffers, in the unit of grid resolution, are also possible to be added along the grid system, using parameter *flag_buffer* of function *grid_generation*.

2.5 Spatial & environmental uncertainty

We calculated the spatial and environmental uncertainty for each grid system (Table 2; Fig. 2). The spatial uncertainty is calculated as the half of the diagonal of a grid ($\frac{\sqrt{res_x^2+res_y^2}}{2}$); this represents the maximum distance toward the center of a grid within that grid. The spatial uncertainty is the same for all grids within a grid system. The environmental uncertainty can vary by the environmental variable being considered, as well as the location of a grid. Here we used NASA DEM elevation data (Crippen *et al.*, 2016) with fine spatial resolution (30 meters) at

global scale. The environmental uncertainty is calculated as the standard deviation of elevation for each grid of the grid system. The spatial and environmental uncertainty were all pre-calculated for the grid systems we compiled or simulated. We also developed a function (*assess_env_uncertainty*) to calculate such uncertainty; this function is based on Google Earth Engine API accessed through *R*.

2.6 Identify the grid system for unknown occurrences

We developed a function (*grid_matching*) to help identify whether an occurrence is from a grid system, and which grid system is most likely to be the source of the occurrence (Table 2; Fig. 2). The workflow is as following: 1) subset the grid systems to be checked based on the country information provided by the user; 2) project the occurrences to the CRS of a grid system to be checked; 3) calculate the pairwise distance between occurrences and centers of all grids of a grid system; 4) for each occurrence, find the minimum distance toward a grid center and divided it by the spatial uncertainty of the grid system; 5) loop through all grid systems, for each occurrence find the grid system that is associated with smallest relative distance; 6) if the relative distance is larger than a predefined threshold (e.g. 5% of the diagonal line of a grid or the maximum distance between known gridded coordinates and grid centroids), then reach the conclusion that the occurrence is not from a grid system; otherwise, conclude that the grid system with shortest relative distance is the origin of an occurrence (Fig. 3).

3. Application

```
Case 1. Infer the metadata of a grid system
#1.1 Infer coordinate reference system
> remotes::install_github("BiogeographyLab/gridder")
> result_crs = infer_crs(occ_path=occs_unique,  #demo coordinates
+                        truth_crs_num = "2154",#use true crs as a reference
+                        cup_num =15)            #use more cpus
```

```
> result_crs$selected[c("code","note")]
     code                              note
1    2154                             truth
1336 2154                  RGF93 / Lambert-93
6350 5698 RGF93 / Lambert-93 + NGF-IGN69 height
6351 5699 RGF93 / Lambert-93 + NGF-IGN78 height
4355 3947                       RGF93 / CC47
4354 3946                       RGF93 / CC46
2944 30731       Nord Sahara 1959 / UTM zone 31N
3515 32431             WGS 72BE / UTM zone 31N
4353 3945                       RGF93 / CC45
4352 3944                       RGF93 / CC44

#1.2 Infer the spatial resolution
> input_occ = load_occ(occs_unique)
> input_occ_prj = sp::spTransform(input_occ,crs(paste0("+init=epsg:","2154")))
> result_res = infer_resolution(input_coord=input_occ_prj@coords,flag_unit="meter")
> print(result_res$res_x)
10000
> print(result_res$res_y)
10000

# 1.3 Infer the spatial extent
> result_ext = infer_extent(method = "crs_extent",
+             crs_grid = result_crs$selected$code[1],
+             flag_adjust_by_res = TRUE,
+             res_x = result_res$res_x,
+             res_y = result_res$res_y)


Case 2: Generation of a grid system based on metadata
> simulated_grid = grid_generation(res_x = result_res$res_x,
+                         res_y = result_res$res_y,
+                         unit = "m",
+                         flag_crs=TRUE,
+                         country = "France",
+                         extent_unit="empirical_occ_extent",
+                         input_extent=result_ext,
+                         flag_offset=c(0,-result_res$res_y*10,
+                                       result_res$res_x*10,0),
+                         crs_num = result_crs$selected$code[1],
+                         flag_maskByCountry = TRUE)

Case 3. Match occ dataset against known grid systems
> point_grid = input_occ[sample(1:length(input_occ),100),]
> point_nongrid = point_grid
> point_nongrid@coords = point_nongrid@coords + runif(100)
> point_all = rbind(point_grid,point_nongrid)
```

```
>
> grid_metadata = data.frame(grid_ID=c("88"),
+                       resolution_x=c(10000),
+                       resolution_y=c(10000),
+                       resolution_unit=c("m")  )
>
> temp_result = grid_matching(input_occ = point_all,
+                        input_grid = simulated_grid,
+                        grid_metadata=grid_metadata,
+                        flag_relativeTHD = 0.1,
+                        flag_absoluteTHD = 10)
input 200 occ
checking 1 grid out of 1[1] 1
> table(temp_result@data$grid_closest_both)
grid_ID_88   notFound
       100        100
```

4. Discussion

The software (*GridDER*) we developed here provides a potential solution for detecting and evaluating gridded coordinates and grid systems that are broadly used in ecological and biodiversity studies. We built upon a synthesis of grid systems used in biodiversity records and developed a metadata standard to characterize grid systems with different spatial attributes. When the metadata are partly or completely unknown, *GridDER* can be used to infer the spatial attributes of a grid system, such as CRS and resolution, based on a dataset of coordinates. With the metadata available, *GridDER* can be used to generate grid systems. The software (*GridDER*) can also be used to assess the grid system that coordinates stem from, as well as to assess the spatial and environmental uncertainty associated with gridded coordinates.

We envision this software can be used in a variety of scenarios. First, *GridDER* can be used by dataset managers to simulate the original grid system of gridded coordinates, when the metadata of the grid system is either partly or fully unknown or when the grid system is known but not available in digitized format. During our review of duplicated coordinates from GBIF, we found

multiple cases that documented metadata of a grid system (e.g. resolution) does not match the spatial patterns inferred from the corresponding coordinates. This would be a case where *GridDER* could be used by dataset managers to further validate the relevant grid systems. Second, *GridDER* can be utilized at larger scales by biodiversity databases for data curation purposes. Biodiversity databases can integrate digitized biodiversity collections with coordinates of localities. We demonstrated that such aggregated information commonly includes gridded coordinates, but the origin of the grid systems are not necessarily properly recorded. For example, our review of duplicated coordinates from GBIF identified 107 datasets (Table S2) that are enriched with gridded coordinates; those together correspond to over 261 million records (or 13.5% of GBIF records) (assessed on January 1, 2022). *GridDER* can be used to identify coordinates that represent the centroids of cells in different grid systems. The identified grid system can also be used to infer the potential spatial and environmental uncertainty associated with the gridded coordinates. The inferred spatial uncertainty can be used to fill *coordinateUncertaintyInMetersProperty* (https://dwc.tdwg.org/terms/#dwc:coordinateUncertaintyInMeters), a field in Darwin Core (a widely adopted metadata standard for biodiversity data; (Wieczorek *et al.*, 2012)). Information on the uncertainty of coordinates can also be reported in the specifics and/or metadata of species distribution models and ecological niche models that make use of these data to ensure reproducibility and accurate interpretation of results (Feng *et al.*, 2019; Zurell *et al.*, 2020). The inferred spatial and environmental uncertainties can have important implications for studies that utilize digitized biodiversity records. Third, in addition to the functions for grid simulation and matching, another major component of *GridDER* is the data of known grid systems. *GridDER* can be used to match coordinates that come from multiple datasets (or multiple grid systems) to

known gridded systems. This more likely represents a use case for individual users or individual species, where the compiled coordinates may stem from multiple sources.

6. References

Blowes, S.A., Supp, S.R., Antão, L.H., Bates, A., Bruelheide, H., Chase, J.M., Moyes, F., Magurran, A., McGill, B., Myers-Smith, I.H., Winter, M., Bjorkman, A.D., Bowler, D.E., Byrnes, J.E.K., Gonzalez, A., Hines, J., Isbell, F., Jones, H.P., Navarro, L.M., Thompson, P.L., Vellend, M., Waldock, C. & Dornelas, M. (2019) The geography of biodiversity change in marine and terrestrial assemblages. *Science*, **366**, 339–345.

Boyle, B., Hopkins, N., Lu, Z., Raygoza Garay, J.A., Mozzherin, D., Rees, T., Matasci, N., Narro, M.L., Piel, W.H., McKay, S.J., Lowry, S., Freeland, C., Peet, R.K. & Enquist, B.J. (2013) The taxonomic name resolution service: an online tool for automated standardization of plant names. *BMC Bioinformatics*, **14**, 16.

Buitrago, L. (2020) GBIF issues & flags.

Crippen, R., Buckley, S., Agram, P., Belz, E., Gurrola, E., Hensley, S., Kobrick, M., Lavalle, M., Martin, J., Neumann, M., Nguyen, Q., Rosen, P., Shimada, J., Simard, M. & Tung, W. (2016) Nasadem global elevation model: Methods and progress. *ISPRS - International Archives of the Photogrammetry Remote Sensing and Spatial Information Sciences*, **XLI-B4**, 125–128.

DeSisto, C.M.M., Park, D.S., Davis, C.C., Ramananjato, V., Tonos, J.L. & Razafindratsima, O.H. (2020) An invasive species spread by threatened diurnal lemurs impacts rainforest structure in Madagascar. *Biological Invasions*, **22**, 2845–2858.

Doughty, C.E., Wolf, A., Morueta‐Holme, N., Jørgensen, P.M., Sandel, B., Violle, C., Boyle, B., Kraft, N.J.B., Peet, R.K., Enquist, B.J., Svenning, J., Blake, S. & Galetti, M. (2016) Megafauna extinction, tree species range reduction, and carbon storage in Amazonian forests. *Ecography*, **39**, 194–203.

Feng, X., Enquist, B.J., Park, D.S., Boyle, B., Breshears, D.D., Gallagher, R., Lien, A., Newman, E., Burger, J.R., Maitner, B., Merow, C., Li, Y., Huynh, K.M.-L., Ernst, K., Baldwin, E., Foden, W., Hannah, L., Jørgensen, P.M., Kraft, N.J.B., Lovett, J.C., Marquet, P., McGill, B.J., Morueta-Holme, N., Neves, D.M., Nunez-Regueiro, M., Oliveira-Filho, A.T., Peet, R.K., Pillet, M., Roehrdanz, P.R., Sandel, B., Serra-Diaz, J.M., Šímová, I., Svenning, J.-C., Violle, C., Weitemier, T.D., Wiser, S. & López-Hoffman, L. (2022) A review of the heterogeneous landscape of biodiversity databases: opportunities and challenges for a synthesized biodiversity knowledge base. *Global Ecology and Biogeography.* DOI:10.1111/geb.13497.

Feng, X., Merow, C., Liu, Z., Park, D.S., Roehrdanz, P.R., Maitner, B., Newman, E.A., Boyle, B.L., Lien, A., Burger, J.R., Pires, M.M., Brando, P.M., Bush, M.B., McMichael, C.N.H., Neves, D.M., Nikolopoulos, E.I., Saleska, S.R., Hannah, L., Breshears, D.D., Evans, T.P., Soto, J.R., Ernst, K.C. & Enquist, B.J. (2021) How deregulation, drought and increasing fire impact Amazonian biodiversity. *Nature*, **597**, 516–521.

Feng, X. & Papeş, M. (2015) Ecological niche modelling confirms potential north-east range expansion of the nine-banded armadillo (*Dasypus novemcinctus*) in the USA. *Journal of*

Biogeography, **42**, 803–807.
Feng, X., Park, D.S., Walker, C., Peterson, A.T., Merow, C. & Papeş, M. (2019) A checklist for maximizing reproducibility of ecological niche models. *Nature Ecology & Evolution*, **3**, 1382–1395.
Franklin, J. (2010) *Mapping Species Distributions: Spatial Inference and Prediction*, Cambridge University Press.
GDAL/OGR contributors (2020) GDAL/OGR geospatial data abstraction software library. *Open Source Geospatial Foundation*.
Guralnick, R.P., Wieczorek, J., Beaman, R., Hijmans, R.J. & BioGeomancer Working Group (2006) BioGeomancer: automated georeferencing to map the world's biodiversity data. *PLoS Biology*, **4**, e381.
Hannah, L., Roehrdanz, P.R., Marquet, P.A., Enquist, B.J., Midgley, G., Foden, W., Lovett, J.C., Corlett, R.T., Corcoran, D., Butchart, S.H.M., Boyle, B., Feng, X., Maitner, B., Fajardo, J., McGill, B.J., Merow, C., Morueta-Holme, N., Newman, E.A., Park, D.S., Raes, N. & Svenning, J.-C. (2020) 30% land conservation and climate action reduces tropical extinction risk by more than 50%. *Ecography*, **43**, 943–953.
Hedrick, B.P., Mason Heberling, J., Meineke, E.K., Turner, K.G., Grassa, C.J., Park, D.S., Kennedy, J., Clarke, J.A., Cook, J.A., Blackburn, D.C., Edwards, S.V. & Davis, C.C. (2020) Digitization and the Future of Natural History Collections. *BioScience*, **70**, 243–251.
Park, D.S. & Davis, C.C. (2017) Implications and alternatives of assigning climate data to geographical centroids. *Journal of Biogeography*, **44**, 2188–2198.
Park, D.S. & Potter, D. (2015a) A reciprocal test of Darwin's naturalization hypothesis in two mediterranean-climate regions. *Global Ecology and Biogeography*. **24**, 1049–1058.
Park, D.S. & Potter, D. (2015b) Why close relatives make bad neighbours: phylogenetic conservatism in niche preferences and dispersal disproves Darwin's naturalization hypothesis in the thistle tribe. *Molecular Ecology*, **24**, 3181–3193.
Park, D., Xie, Y., Thammavong, H., Tulaiha, R. & Feng, X. ABHOR: Artificial Biodiversity Hotspot Occurrence Registry, in review.
Peterson, A.T., Soberón, J., Pearson, R.G., Anderson, R.P., Martínez-Meyer, E., Nakamura, M. & Araújo, M.B. (2011) *Ecological Niches and Geographic Distributions (MPB-49)*, Princeton University Press.
Tingley, M.W., Monahan, W.B., Beissinger, S.R. & Moritz, C. (2009) Birds track their Grinnellian niche through a century of climate change. *Proceedings of the National Academy of Sciences of the United States of America*, **106 Suppl 2**, 19637–19643.
Tinkham, W.T., Mahoney, P.R., Hudak, A.T., Domke, G.M., Falkowski, M.J., Woodall, C.W. & Smith, A.M.S. (2018) Applications of the United States Forest Inventory and Analysis dataset: a review and future directions. *Canadian Journal of Forest Research*. **48**, 1251–1268.
Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Döring, M., Giovanni, R., Robertson, T. & Vieglais, D. (2012) Darwin Core: an evolving community-developed biodiversity data standard. *PloS ONE*, **7**, e29715.
Zizka, A., Silvestro, D., Andermann, T., Azevedo, J., Duarte Ritter, C., Edler, D., Farooq, H., Herdean, A., Ariza, M., Scharn, R., Svantesson, S., Wengström, N., Zizka, V. & Antonelli, A. (2019) CoordinateCleaner : Standardized cleaning of occurrence records from biological collection databases. *Methods in Ecology and Evolution*, **289**, 110.
Zurell, D., Franklin, J., König, C., Bouchet, P.J., Dormann, C.F., Elith, J., Fandos, G., Feng, X.,

Guillera-Arroita, G., Guisan, A., Lahoz-Monfort, J.J., Leitão, P.J., Park, D.S., Peterson, A.T., Rapacciuolo, G., Schmatz, D.R., Schröder, B., Serra-Diaz, J.M., Thuiller, W., Yates, K.L., Zimmermann, N.E. & Merow, C. (2020) A standard protocol for reporting species distribution models. *Ecography*, **43**, 1261–1277.

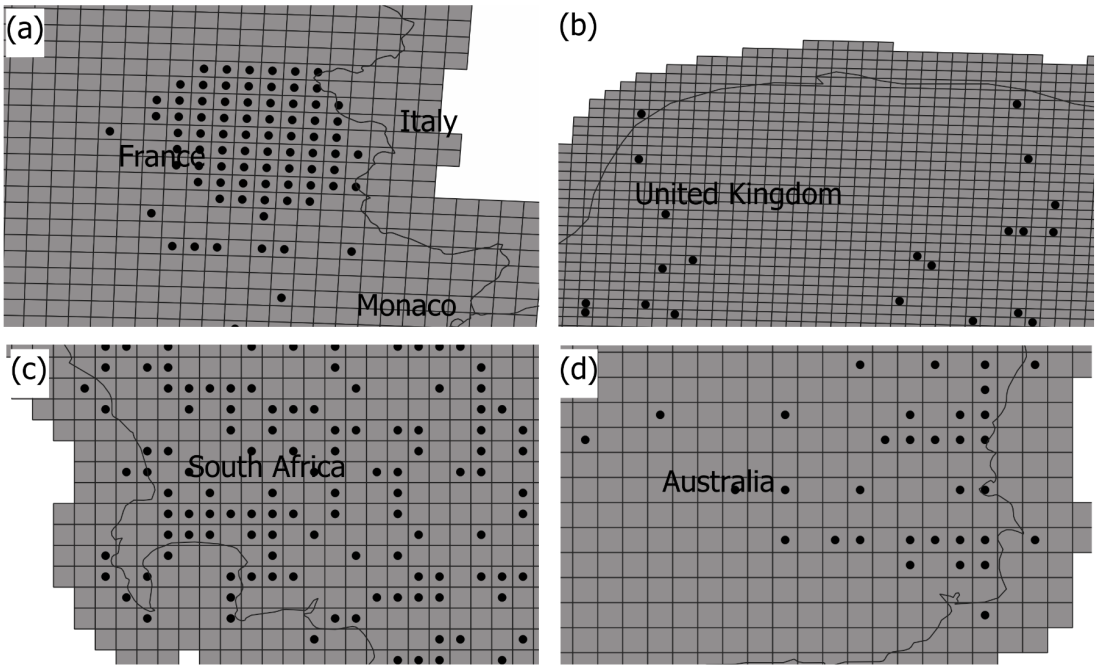

Figure 1. Examples of four grid systems used in France (a), United Kingdom (b), South Africa (c), and Australia (d). The four grid systems have different spatial solutions – 10km for (a), 1km for (b), 5 arc-minute for (c), and 6 arc-minute for (d). The black points represent the biological collections assigned to the centroid of the corresponding grid systems. All maps constructed using WGS84.

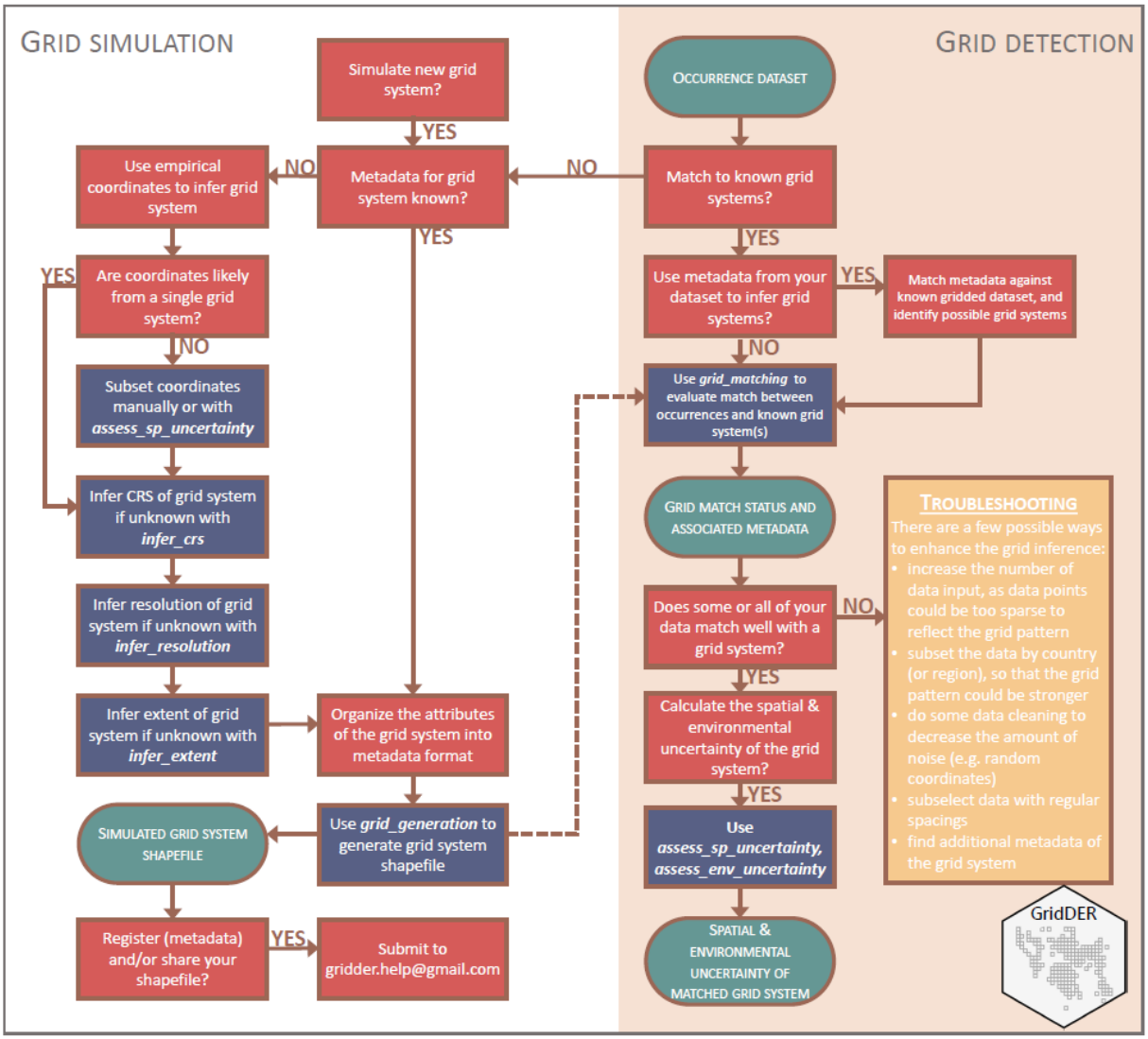

Figure 2. Overview of the workflow of grid detection and evaluation.

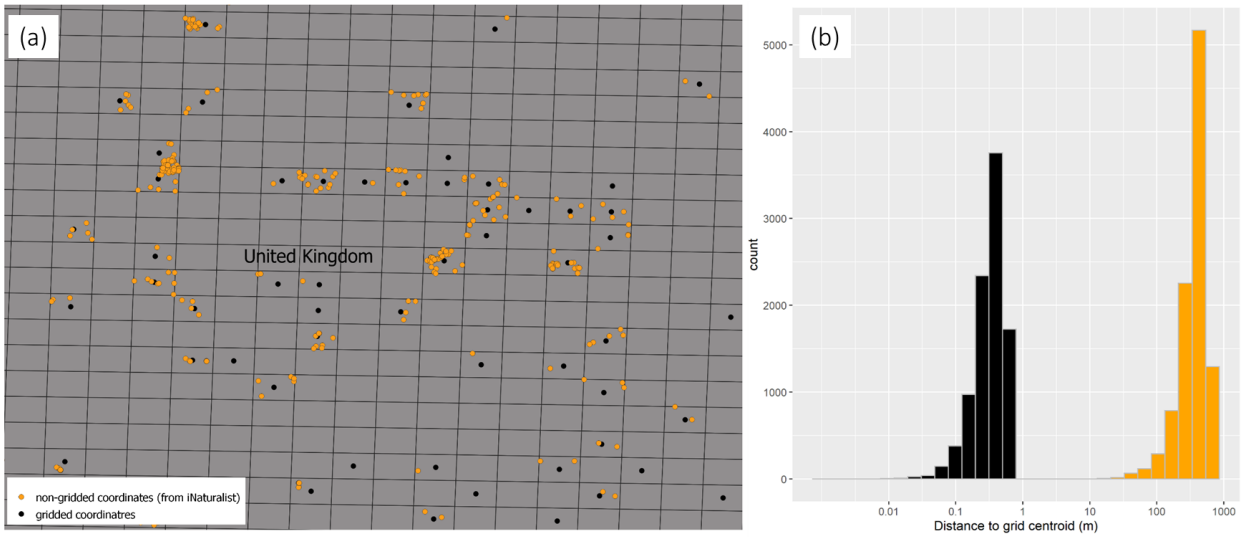

Figure 3. Comparison between gridded and non-gridded coordinates. Panel a shows an example of a grid system (1km resolution) used by the United Kingdom, where the gridded coordinates are shown in black and non-gridded coordinates are shown in orange. Panel b shows the histogram of the distances between gridded or non-gridded coordinates (~10,000) and nearest grid centroids. The x-axis of panel b is log10 transformed.

Table 1. Metadata format of grid systems

File name	Note	Example
grid_ID	the ID of a registered grid system	2
ISO-3166	country name in ISO-3166 code, separated by comma	FR
country_name	country name	France
crs_type (mandatory)	either “EPSG” or “from_shapefile”	EPSG
crs_code (mandatory)	an EPSG code	2154
grid_name	the name of a grid system	Grille nationale (10km x 10km)
resolution_x (mandatory)	the resolution of the grid system along the horizontal axis	10
resolution_y (mandatory)	the resolution of the grid system along the vertical axis	10
resolution_unit (mandatory)	km, degree, minute, or second	km
extent	the spatial extent of a grid system	"20000,6040000,1310000,7130000"
extent_unit	either “empirical_occ_extent”, “crs_web_extent”, or “crs_countryPolygon”	empirical_occ_extent
path_demo_occ	path of point dataset that is associated with a grid system	"data/france_datasetID10/0068190-210914110416597.csv"
spatial_uncertainty_mean_grid	the spatial uncertainty in meters calculated from empirical gridded coordinates	4.349
spatial_uncertainty_mean_nongrid	the spatial uncertainty in meters calculated from empirical non-gridded coordinates (e.g. iNaturalist)	3745.848

Table 2. Overview of functions in gridder

Purpose	New Name	Functionality
internal function	find_angle	calculate angle between two points
internal function	find_crs_extent	automatically find the spatial extent of a coordinate reference system from https://epsg.io
data prep	load_occ	transform a txt file (e.g. a dataset from GBIF in DarwinCore format) that has decimalLongitude and decimalLatitude into simplified spatially points
metadata inference	infer_crs	infer the coordinate reference system in which a set of gridded coordinates were originally defined
metadata inference	infer_resolution	estimate the resolution (or regular spacing of input coordinates) along horizontal and vertical axes, based on distance to nearest neighbors along four directions (i.e. up, down, left, right)
metadata inference	infer_origin	infer the origin a grid system while considering the small uncertainty of the coordinates
metadata inference	infer_extent	infer the spatial extent of a grid system, based on country polygons, coordinate reference system, or input spatial point data
grid generation	grid_generation	generate a grid system based on user defined attributes
grid matching	grid_matching	infer whether a coordinate is from a known grid system, based on absolute and relative distance
grid adjustment	grid_adjustment	adjust the origin of a grid system to minimize the mismatch between grid centroids and input coordinates
uncertainty assessment	assess_sp_uncertainty	assess the spatial uncertainty of a grid system, by calculating the distance between grid centroids toward gridded coordinates and non-gridded coordinates (e.g. iNaturalist data)
uncertainty assessment	assess_env_uncertainty	assess the environmental uncertainty of a grid system, by calculating the variation of environmental conditions (e.g. 30m elevation) within each grid

Data availability statement:

The software is available from Github: https://github.com/BiogeographyLab/gridder. The compiled metadata of grid systems are available in Table S1. The DOIs of the coordinate datasets assessed are available in Table S2.

Table S1. Metadata of grid systems

Table S2. Datasets from GBIF with gridded coordinates