Imperial College London

IMPERIAL COLLEGE LONDON

DEPARTMENT OF LIFE SCIENCES

Establishing Baselines: Are our conservation baselines accurate?

Author:

Jake Curry

Supervisor:

Dr. Natalie Cooper

Prof. Andy Purvis

A thesis submitted in partial fulfilment of the requirements for the degree of Master of Research at Imperial College London

Formatted in the journal style of Methods in Ecology and Evolution

Submitted for the MRes in Computational Methods in Ecology and Evolution

August 15, 2019

Declaration

I'd like to thank and acknowledge those who contributed to my project; Dr Natalie Cooper for her excellent supervision and guidance throughout, and help choosing the appropriate analysis. Emily Buckingham for the use of the data she collected throughout her Masters project. And finally Rachel Bates and Hannah O'Sullivan, for keeping me sane throughout. Cleaning of data was conducted by myself and no mathematical models were developed as part of this project.

Abstract

Conservation efforts often require the use of a pre-impact baseline to establish if they are having an effect. However, these baselines are almost all relatively recently established, making them post-impact in almost all cases. Pre-impact baselines can be constructed using locality data from historical natural history collections, such as that at the Natural History Museum, London. There is error inherent in georeferencing historical specimens, but many studies ignore this as it is complicated and computationally demanding to incorporate this into analyses. Here, I created tools in R, bundled into a package (sfe), that make it easy to incorporate this error, as well as generating convex hulls for each species in an analysis and functionality to find the kilometer distance between centroids of present-day range maps and those convex hulls. I carried out a case study on the pangolins (Manidae) and have implemented version control of data up to 2GB in size in an accessible manner. In pangolins there was a correlation between continent of collection, and the percentage overlap among present-day range maps and ranges generated from historical specimen data. In general, African species had lower percentage overlaps than Asian species. I found no significant correlation among percentage overlap and decade of collection. These results are likely due to complex factors, including socio-economic drivers such as wealth disparity between China and the home ranges of the most exploited pangolin species. Future geospatial analysis of historical specimen data can and should include georeferencing error, as sfe makes it easy to do so in a computationally efficient manner. The tools I have created better informs us of baseline pre-anthropocene distributions of species, which can greatly aid conservation efforts. Future work should look to increase the number of taxa investigated, and to tease apart the drivers for the continental difference observed in pangolins.

Contents

1 Introduction

Global biodiversity is undoubtedly under threat. Between 200 and 100,000 species are becoming extinct every year (Pimm et al. 2014), representing a rate of species loss between 1000 and 10,000 times higher than the natural background rate (De Vos et al. 2015). This unprecedented loss of species has led to the current era being defined as the sixth mass extinction (Barnosky et al. 2011).

What drives the loss of a species is complex, and depends on the species in question, however the primary culprit in the majority of species extinctions is now thought to be habitat loss (Brooks et al. 2002). Habitat loss is itself generally driven by human activities and the effects of this are now being further compounded by Global Climate Change (Thomas et al. 2004). Human activities here is taken to mean any activities undertaken by humans which have an impact on the ecosystem in which they are taking place.

Loss of biodiversity (or species extinction) has accelerated during the Anthropocene (De Vos et al. 2015), and is a pressing issue, as humanity depends on biodiversity for all the ecosystem functions which keep us alive. Costanza et al. (1997) made a conservative estimate of the monetary value of these ecosystem services to humanity of up to $54trillion United States Dollars per year, which was roughly twice the global domestic product (GDP) at the time. This huge cost makes it infeasible from an economic standpoint to replace, even if the requisite technology was ready and available now, which in many cases it is not. As such, we should be deeply concerned about the loss of the biodiversity that allows ecosystems to continue to function.

To limit the loss of biodiversity we must conserve species, to do this it is necessary to know where a species is. Expert range maps are an excellent way to keep this information. Range changes are one of the key metrics used in determining the risk of species extinction (Mace et al. 2008), and how well a species is responding to conservation efforts. However, herein lies the problem; expert range maps with the purview for conservation have only been created within the last 30 years or so, well into the Anthropocene and therefore potentially after a great deal of human impact on species ranges. This is problematic, as using incorrect conservation baselines leads

to erroneous conclusions as to the effectiveness of conservation strategies (Froyd & Willis 2008, Willis et al. 2005, Willis & Birks 2006), and how at risk a species is.

Fortunately there is a solution. We can retroactively create historical range maps using the wealth of location data stored in museums, in the metadata of specimens in collections, such as that at the Natural History Museum, London (NHM). It is possible to use these locality descriptions to georeference a specimen, however, location metadata varies in the accuracy of the locality description given, which leads to uncertainty about the accuracy of the data.

Various projects are creating historical range maps, however they often encounter problems with the accuracy of their data because of this variation in locality descriptions. For example, previous work that involves georeferencing of specimens often does not incorporate the uncertainty of measurement. Of the first 100 relevant papers taken from Google Scholar search (see supplementary materials) of "georeferencing of specimens" 46 spoke about error or uncertainty of measurement. Of those that did, the most frequently used method was Wieczorek et al. (2004) method of incorporating error (19 papers). The other 27 papers used a mix of Bayesian regression, machine learning approaches such as MAXENT, simple data filtering and manually double checking gazetteer coordinates as methods of accounting for error, or simply do not state how error was calculated. The majority of papers (54 papers) make no mention of error. In my thesis I will find simple methods of measuring and incorporating error, so that future studies can stop ignoring georeferencing error.

Incorporating error is a challenge, and I suspect that this is why many papers do not consider it. Either that, or the papers are using digitised archives which do not contain error, and the authors are unfamiliar with historic geospatial data and the uncertainties involved in generating it. Using modern geospatial tools such as the R R Core Team (2019) package *sf* (Pebesma 2018), it is relatively easy to incorporate error into geospatial analysis. The protocol outlined in Wieczorek et al. (2004) can be used to add an error radius to point data, which in turn can be used to create spatial polygons. These polygons are analogous to a primitive range map, and can be used for comparison against IUCN range map data across a wide variety of taxa.

Researchers at NHM have been georeferencing historical specimens (Fig 1) to increase the availability of this data, to expedite research across fields such as taxonomy and conservation (Edwards 2004) in line with the goals of the Global Biodiversity Information Facility (GBIF), as well as for research projects occurring at the NHM (Cooper et al. 2012). With this data it is possible to compare IUCN maps and modern ranges and observe if and how a species range has changed throughout the Anthropocene. Once we have determined how a range has changed it is possible to link it to drivers, such as changes in land use and climate change, by comparing how these have changed for the same area (Mace et al. 2008).

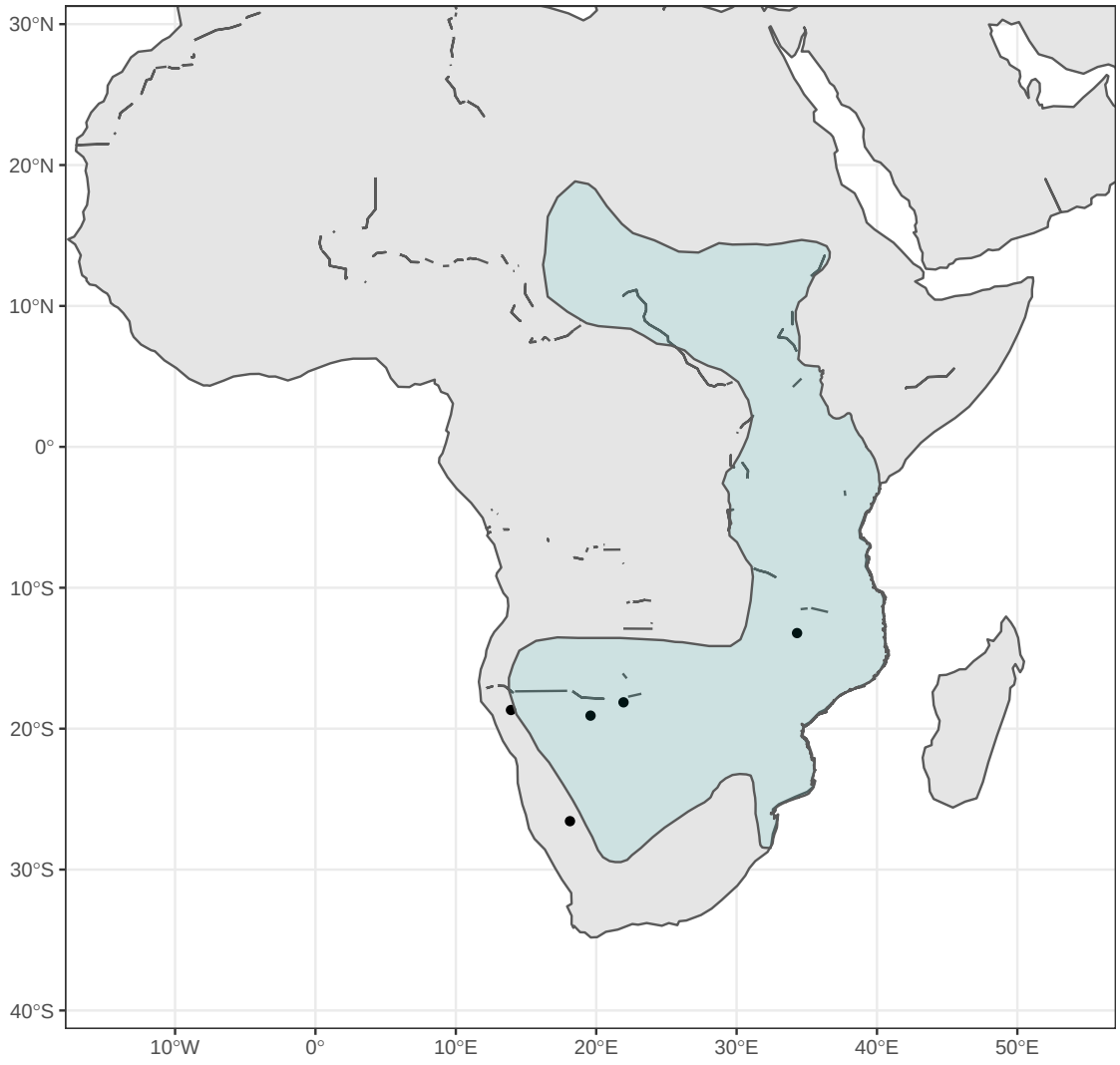

Figure 1: Previous work done at the NHM has georeferenced points and looked at whether they fall within the IUCN range (shaded blue here). This is an example using *Smutsia temminckii*, where each point represents an individual specimen which has been georeferenced.

Unfortunately the utility of this work is limited by a variety of computational issues, many of which are common to all of the literature on the subject. Below I outline the objectives of this thesis, with key areas highlighted in Fig 2.

Above I briefly discussed the first major issue, and that is the lack of incorporating error in georeferencing into the analysis, despite the knowledge that retroactive georeferencing generates uncertainty.

The second major issue is that workflows are rarely reproducible. This is due in part to the collaborative nature of modern scientific work, in that a project like this usually has several contributors, all working on their own machines, creating their own code which is unlikely to integrate into a seamless workflow, if such a thing was even considered in the first place. The usual scenario is that the code an author generates remains solely on that authors computer, and only the final results of said code ever make it to publication. This is partly why ecology is one of the "at-risk" disciplines for reproducibility (Fidler et al. 2017). Creating an R package, or otherwise bundling the entire workflow so that it can be shipped to another machine and rerun producing the same results effectively combats the issue of reproducibility. I will therefore bundle the tools described below into an R package as part of this project.

The third major issue is one of computational power. Creating range polygons from map data and then computing overlaps among specimen localities and current ranges is a computationally intensive process. Visualising range maps and their overlap is equally computationally demanding. The recently released R package *sf* (Pebesma 2018) is more computationally efficient than the previous R methods. Implementing *sf* in this pipeline will greatly reduce analysis times, which is imperative, as the end goal of this work is to implement geospatial analysis on more than two million vertebrate specimens.

The fourth major issue relates to the data used in the creation of historic range maps. Currently multiple individuals from a variety of knowledge backgrounds (experts through to volunteers with no prior knowledge of the taxa) are involved in the georeferencing process. Each individual will georeference slightly differently, despite following the same protocol,leading to a potential source of error. At the extreme,

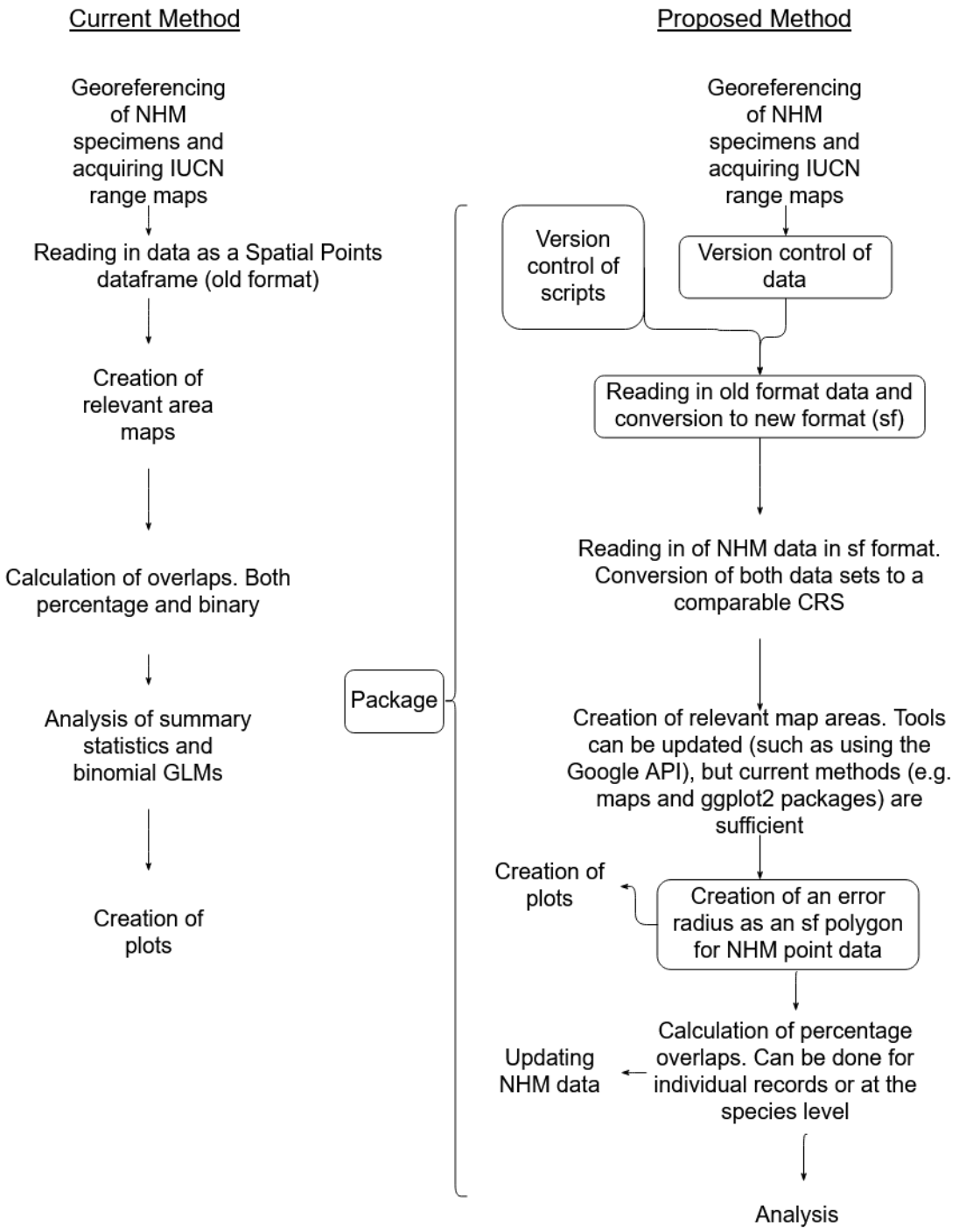

Figure 2: Comparison of current and proposed pipelines for analysing historical geospatial data. I have highlighted key differences where I believe there is scope for the greatest improvement with boxes.

there may be occasions where the project leads need to remove or modify all records input by an individual, or on a certain date, due to a systematic error creeping in . At present, everyone contributes to the same dataset, and there is no version control of this dataset. This means that if one person makes a mistake and overwrites previous work erroneously, or adds poorly georeferenced data, there is no way to roll back the data to a previous state. Version control of data is hard because the geospatial data generated is “big data”, and as such difficult to store, particularly in multiple copies. In addition, GitHub (GitHub Inc. 2019) is not designed to store such large datasets, and tools that exist to help with this require specialist knowledge to use effectively. To address this issue I will implement version control of the datasets I use in my thesis with the R package datastorr (Falster et al. 2019), and provide documentation for this to be used effectively within this project.

In summary, in this project I aim to:

1. Incorporate error into geospatial analysis of historical range maps.

2. Use the most up-to-date geospatial tools in R to improve computational efficiency.

3. Create an R package to make workflows entirely reproducible.

4. Implement version control of the “big data” geospatial records already generated.

5. Use 84 georeferenced pangolin (Manidae) records as a test case to understand the feasibility of implementing these features, as well as answering the question “Are our baselines accurate for pangolins?”

These aims will address a key gap in the literature, and practical concerns of how to routinely incorporate the wealth of archive data into modern conservation, while achieving a high standard of reproducibility.

2 Methods

2.1 Incorporating error into geospatial analysis of historical range maps

In this section I will briefly describe how the data were originally generated and how I used the most up-to-date geospatial tools in R to incorporate error into an analysis of historical range maps. I will also explain several methods of calculating overlap between range polygons, from the most basic to some more advanced functions. I will address both aims 1 and 2 of this project here. I am combining the first two aims into one section as they go hand in hand - to incorporate error (aim 1) I need to use the most up-to-date geospatial tools in R (aim 2).

2.1.1 Georeferencing protocol

For a full explanation of georeferencing please see the NHM georeferencing protocol attached in the supplementary materials. In brief, museum specimens have labels that contain locality data of varying accuracy, ranging from just a country to exact coordinates. Most cases will be somewhere between these extremes. In georeferencing the area on the label is found using Google Maps and given an appropriate set of coordinates in longitude and latitude. These coordinates and a description of how you arrived at that conclusion are recorded, as are other metadata, including how certain you are that the longitude and latitude are correct, what type of area it is (e.g. a country, named place, river, mountain etc.) and the uncertainty measured in kilometers. The uncertainty depends on what type of area is being georeferenced (see NHM protocol in supplementary materials), but for a named place (e.g. London) you use the 'Measure' tool in Google Maps to measure the straight line distance between the latitude and longitude point and the furthest border of what would be considered to lie within the named place. Following Wieczorek et al. (2004) point-radius method you then draw a circle around the point with a radius of that straight line distance, to fully encompass the whole possible area the specimen could have been found in. This circle represents the potential error in georeferencing.

2.1.2 Creating error polygons

Previous work using the NHM georeferenced localities did not incorporate the extent measurements made in the initial generation of the data (these projects were completed by Masters students with limited time to collect data and learn new geospatial tools in R). Fortunately the recently developed R geospatial analysis package sf Pebesma (2018) has functionality that can be used to this end.

A single command allowed me to transform the point geometry at the latitude/longitude coordinates into a circle of a specified radius (Snippet 1.):

```
Snippet 1.
df <- st_buffer(df, df$Extent_km)
```

The output from this is a geometry collection, where each data point is now a circle of radius extent, centred around its specified longitude/latitude point. While creating these is simple, actually using the now incorporated error is more tricky, and required me to build several functions which used sf and the tidyverse packages (see below).

2.1.3 Calculating overlaps between georeferenced localities and their extent, and species current ranges

The next step in the pipeline (see Figure 2.), is calculating whether georeferenced localities with error included overlapped with current species ranges, and if so what the percentage overlap of the two was.

The R package sf (Pebesma 2018) greatly speeds geospatial operations in R, and brings GIS tools to the tidyverse, however, this also means sf sacrifices some ease of use, namely that performing operations that are not native to this package can be quite difficult. This makes calculating the percentage overlaps between two convex hulls a non-trivial task.

The long code snippet (Snippet 2.) below shows that three functions using sf native operations are needed to calculate percentage overlap.

```
Snippet 2.
# two input function for calculating the percentage overlap
calcOverlaps <- function(df1, df2) {
  df1 <- lwgeom::st_make_valid(df1)
  # gives percentage overlap between NHM and IUCN
  overlap <- st_intersection(df1, df2) %>%
             st_area() / st_area(df1, df2) * 100
  # at this point the output is of class "units" which don't play nice
  overlap <- units::drop_units(overlap)
  # allows for handling of cases of zero overlap
  if (purrr::is_empty(overlap) == T) {
    # as it otherwise returns a list of length zero,
    # which cannot be appended to a df
    overlap <- c(0)
  }
  overlap <- as.list(overlap)
  # returns the result, so can be passed to another fun
  return(overlap)
}

hullOverFun <- function(df1, df2) {
  # adds a column of na's
  df1$Percent_overlap <- NA
  # for each row in first df's geometry col
  for (row in 1:nrow(df1)) {
    # extract the geometry
    geom <- df1$geometry[row]
    geom <- st_transform(geom, 2163)
    # use previous fun to calculate overlaps
    x <- calcOverlaps(geom, df2$geometry)
    df1$Percent_overlap[row] <- x
  }
  # return the modified df for use in another fun
  return(df1)
}
```

```r
calculateOverlaps <- function(x, y) {
  # create an empty list to store results
  output <- c()
  # find all entries in both dfs which match var
  for (var in unique(x$binomial)) {
    IUCN_var <- y[y$binomial == var,]
    NHM_var <- x[x$binomial == var,]
    # ensure planar crs is in use
    NHM_var <- st_transform(NHM_var, 2163)
    IUCN_var <- st_transform(IUCN_var, 2163)
    # then pass to the over_function
    tmp <- hullOverFun(NHM_var, IUCN_var)
    # rebuilding the input df with a new col
    output <- rbind(tmp, output)
  }
  output$Percent_overlap <- as.numeric(output$Percent_overlap)
  return(output)
}
```

2.1.4 More advanced measures of overlap

The methods above provide a basic insight into overlap, but the reality may be more complex, which requires more complex tools to explore. At its core this means investigating the overlap of two convex hulls (i.e. the simplest shape containing all the points of a collection of records). To achieve this I have included functions to create convex hulls for each species (Snippet 2.) and calculation of centroid-centroid distances (Snippet 4..; Fig 3.) for use in more complex models. Only a single convex hull can be calculated for each species, as by definition it includes all the possible localities in the simplest possible shape. Convex hulls are used in the calculation of area of occupancy and extent of occurrence (IUCN methods of calculation, IUCN Standards and Petitions Subcommittee (2017)), which are required for assessing species risk of extinction. *Guidelines for Using the IUCN Red List Categories and Criteria* (IUCN Standards and Petitions Subcommittee 2017) gives a full methodology for this, so I will not expand upon that here.

```
Snippet 3.
makeLandClippedHulls <- function(x) {
  # transform crs to make st_intersect happy
  x <- st_transform(x, 2163)
  # makes a base for hulls to be clipped to
  landMap <- rnaturalearth::ne_countries(returnclass = 'sf') %>%
    st_union()
  # transform crs to make st_intersect happy
  landMap <- st_transform(landMap, 2163)
  # empty list to rebuild the df from
  output <- c()
  for (var in unique(x$binomial)) {
    # splits data into species groups
    subsetOfDf <- x[x$binomial == var,]
    subsetOfDf$geometry <- st_convex_hull(st_combine(subsetOfDf$geometry))
    # sets to correct (long/lat) crs for comparison
    clippedHull <- suppressMessages(st_intersection(
            lwgeom::st_make_valid(subsetOfDf$geometry),
                      lwgeom::st_make_valid(landMap)))
    if (purrr::is_empty(clippedHull)) {
      # this function is ONLY for terrestrial convex hulls
      print('Hull is entirely in the ocean')
    } else {
      subsetOfDf$geometry <- clippedHull
    }
    subsetOfDf <- st_transform(subsetOfDf, 4326)
    output <- rbind(output, subsetOfDf)
    output <- st_transform(output, 4326)
  }
  return(output)
}
```

```
Snippet 4.
calculateCentroidDistance <- function(x, y) {
  output <- c()
  x$distance <- NA
  x$distance2 <- NA
  for (var in unique(x$binomial)) {
    subsetOfDf <- x[x$binomial == var,]
    subsetOfIUCN <- y[y$binomial == var,]
    subsetOfDf$geometry <- st_transform(subsetOfDf$geometry, 2163)
    subsetOfIUCN <- st_transform(subsetOfIUCN, 2163)
    # finds the centroid (point geom of convex hull)
    centroid <- st_centroid(subsetOfDf$geometry)
    IUCNCentroid <- st_centroid(subsetOfIUCN$geometry)
    edgeDist <- st_distance(centroid, IUCNCentroid)
    edgeDist <- units::drop_units(edgeDist)
    edgeDist <- edgeDist/1000
    # allows for handling of cases of zero overlap
    if (purrr::is_empty(edgeDist) == T) {
      # as it otherwise returns a list of length zero, which can-
not be appended to a df
      edgeDist <- c(0)
    }
    matrixSize <- ncol(edgeDist)
    if (ncol(edgeDist) == 1) {
      subsetOfDf$distance <- edgeDist
    } else if (ncol(edgeDist) == 2) {
      subsetOfDf$distance <- edgeDist[, 1]
      subsetOfDf$distance2 <- edgeDist[, 2]
    } else {
      print('IUCN data contains more than two polygons,
      please reduce to areas of interest and try again')
    }
    output <- rbind(output, subsetOfDf)
  }
  return(output)
```

```
}
```

Centroid-centroid distance is the straight line distance between the centroids (or centre points) of two convex hulls (Fig 3). Centroid-Centroid distance can be used to inform range shift (Lyons 2003). In this case I will be using the kilometer distance difference between the IUCN range centroid and the NHM convex hull centroid. This, along with other measures, can be important for modelling species extinction risk (Midgley et al. 2003), but in this case I am more interested in any absolute change.

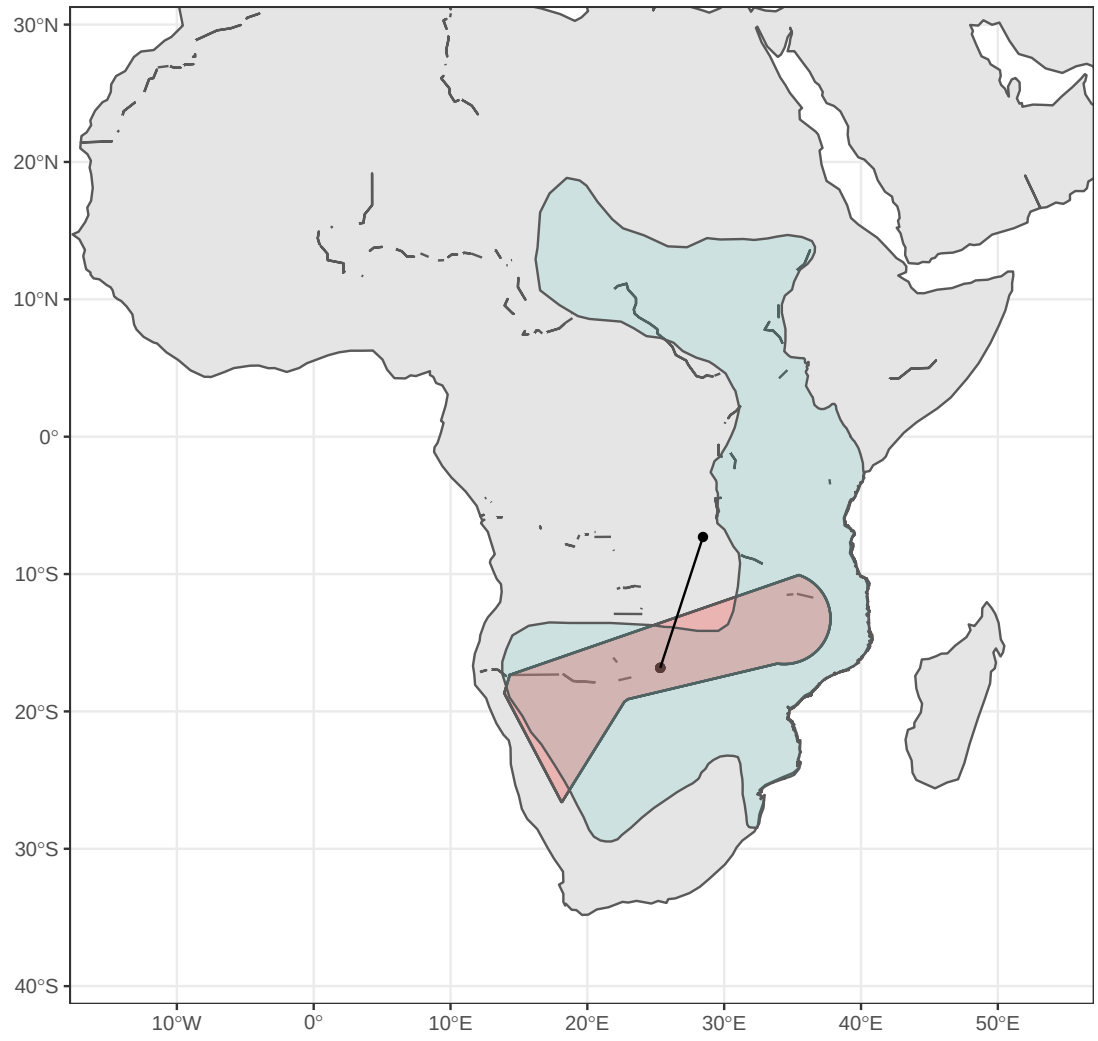

Figure 3: Centroid-centroid distance is the straight line measurement between the two centre points (shown above) of two separate convex hulls. In this case the centroids are from the NHM convex hull (red) and the IUCN range map (blue) for *Smutsia temminckii*.

2.2 Creating an R package to make workflows entirely reproducible

To make the methods described above truly reproducible, and so that others could use the tools I created, I bundled all the functions into an R package called sfe. I used RStudio (RStudio Team 2018) and the R package roxygen2 (Wickham et al. 2018) to create the package and its documentation following the guidelines set out by Wickham (2015).The sfe package is accessible on my GitHub page here https://github.com/JCur96/sfe. The landing page is the vignette, which gives a detailed walkthrough of the package.

2.3 Implementing Version Control

Whilst version control of the data is the second stage in the workflow (Fig 2) I implemented it late in the development of the package as this was a "stretch goal" for this project. Implementing version control was relatively simple, the package Datastorr (Falster et al. 2019) provides a set of easy to use tools to do this. The vignette on the landing page of Datastorr (Falster et al. 2019) which can be found here (https://github.com/ropenscilabs/datastorr), provides instructions on both setting up version control of data and how to use versioning as an end user. I will therefore only give a brief description here. First I ran the following code (Snippet 5.)

```
Snippet 5.
datastorr:::autogenerate(repo="JCur96/sfe", read="read.csv")
```

This generates a set of functions that are fully documented and are used for version control. These are included as functions in sfe and built at the same time as the rest of the package. Data is initially released using this function (Snippet 6.)

```
Snippet 6.
sfe::mydata_release("First data release", "../data/PangolinData.csv")
```

Versions are changed using the same function, but the version number must be increased, to automate this I created an *updateVersion* function which when run prompts the user to enter a new version number. Users can then choose which version of the data to download using the *mydata_versions* and *mydata* functions to view version numbers and then download the data respectively. If a user wants to upload data with a new version they can run the *mydata_release* function and the following

bash code (Snippet 7.) if working on Linux, or use GitKraken (Axosoft 2019), or other Git management software on their preferred operating system.

```
Snippet 7.
git add "DataFileName"
git add "DescriptionFileName"
git commit -a
# A text editor will open and you must
# enter a description of the changes made, then
git push
```

I have also implemented a test case for “Big Data”, as a proof of concept of Datastorr (Falster et al. 2019) using the IUCN range data for Amphibians. I used this data as it is much closer to the requirements of “Big Data” and is pushing the limit of what Datastorr (Falster et al. 2019) can do at approximately 2GB in size.

2.4 Pangolins case study

To demonstrate proof of concept I conducted a case study using the Manidae family (pangolins). Data were collected by Buckingham (2019, unpublished Masters Thesis) from the NHM collections, and consisted of 215 georeferenced specimens from seven of the eight species of pangolin. There are eight species of pangolin in three genera, distributed over two continents (Africa and Asia). The NHM collections are missing specimens of *Manis culionensis* as this species was only formally recognised as a distinct species in 2005 (Gaubert & Antunes 2005, Wilson & Reeder 2005), prior to which *M. javanica* and *M. culionensis* were generally considered to be the same species. *M. culionensis* is only found in the Philippines and no NHM records of *M. javanica* exist in the Philippines, leading us to believe we have no *M. culionensis* in the collections.

The three factors I investigate below (decade of collection, region of collection and species) are likely to be correlated with the number of specimens found within their current IUCN ranges as follows.

First, I would expect to see a species level effect due to the life history differences among the species. For example, *Manis javanica* is arboreal, so harder to harvest (Newton et al. 2008), whereas *Manis pentadactyla* is terrestrial (Newton et al. 2008, Challender et al. 2014), and therefore easier to harvest, so would have fewer overlaps between IUCN range and collection locale than *M. javanica*.

Second, the region where specimens were collected is likely to correlate with percentage overlap as there is a growing demand for pangolin material (Challender & MacMillan 2014), which is being fueled by African pangolins (Ingram et al. 2018, Challender & MacMillan 2014), so I would expect to see fewer overlaps between IUCN range and collection locale in Africa as opposed to Asia as a result. This is due in part to the overexploitation of Asian pangolin species, leading to a lack of supply in Asia (Heinrich et al. 2016).

Finally, the NHM collections have specimens which were collected both before and after the start of the Anthropocene (as defined by the Anthropocene working group, Zalasiewicz et al. 2015; 2017, Steffen et al. 2015) and as such I would expect to see fewer overlaps between IUCN ranges and collection localities for those specimens collected before the start of the Anthropocene as these would have experienced the least human impact.

I first cleaned the data in R so it contained only species for which there were also corresponding IUCN range records and data on decade of collection. This reduced the data to 84 records across seven species, meaning there were 84 locality points and seven convex hulls.

I then used functions I created as part of sfe to add error radii to the point data, following the methods of Wieczorek et al. (2004). From these I first calculated percentage overlap between point-radius polygons and IUCN range polygons for each pangolin specimen (see Fig 4a). Next, I used the point-radius polygons to calculate convex hulls for each species and then calculated the percentage overlap between NHM range polygons and IUCN range polygons for each of the seven pangolin species I had data for (see Fig 4b).

To determine whether overlap differed by species, by the region the species came from (Asia or Africa), or by the decade the specimen was collected in, I then used generalized linear models (GLMs) with binomial errors to analyse the point-radius data. I calculated the number of successes and failures for each species, i.e. the number of specimen localities that overlapped with the current IUCN range (successes), and the number of specimen localities that did not overlap with the current IUCN range (failures). I used these numbers as the response variable. I fitted one model with species binomial name as the explanatory variable, another with region as the explanatory variable, and another with collection decade. With only seven species there were too few data to fit more complex models.

For the convex hull percentage overlaps data, there is only one value for each species, hence there were not enough data points to statically analyse so I just report the percentage overlaps between IUCN range and the convex hull generated.

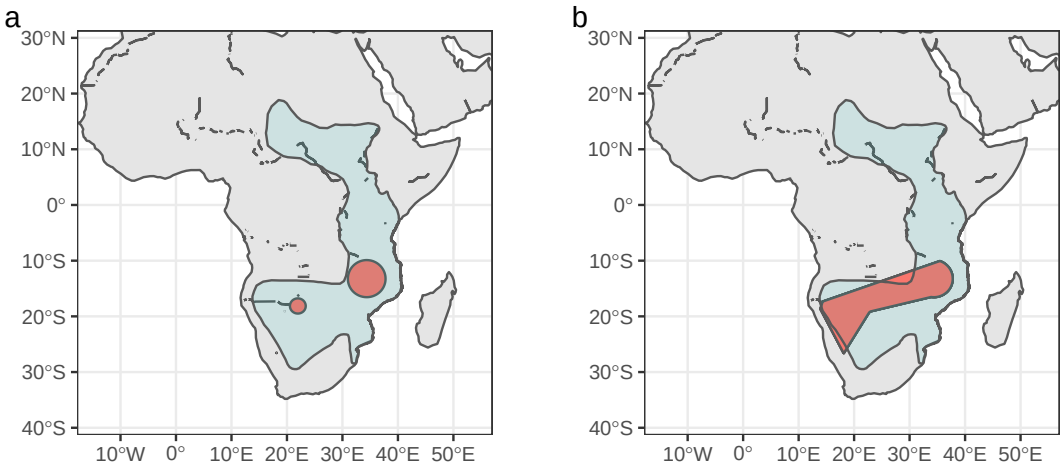

Figure 4: Point-radius data for *Smutsia temminckii*, where red circles represent NHM records, which are longitude-latitude point data that has had an uncertainty radius added according to Wieczorek et al. (2004) method. The red polygon in the second panel is a convex hull created from those point-radius records. The Light blue polygon is an IUCN range polygon.

3 Results

3.1 Incorporating error into geospatial analysis of historical range maps

I have successfully incorporated error into geospatial analysis using the *sf* package. Please see the ReadMe file on the landing page of the GitHub repository for *sfe* https://github.com/JCur96/sfe for more details.

3.2 Creating an R package to make workflows entirely reproducible.

I created the *sfe* R package. This is accessible on my GitHub page here https://github.com/JCur96/sfe. The landing page is the vignette, which gives a detailed walkthrough of the package.

3.3 Implementing version control of geospatial records

See my GitHub page (https://github.com/JCur96/sfe) for a working example using the Manidae (pangolin) family as a test case. I also provide "Big data" test cases involve the use of the IUCN Amphibian data as this is much closer to the requirements of "big data", which can be found in the following git repo, https://github.com/JCur96/bigDataTest this data is not for redistribution, and I am only using this as a proof of concept.

3.4 Pangolins case study

The seven species show a large range in their overlap values across specimens, with *Phataginus tricuspis* specimens showing the most consistently high percentage overlaps of IUCN range and possible collection locale (Fig 5a, mean percent overlap $\pm$ standard error, with raw data shown in grey.) For *Manis crassicaudata* and *Smutsia temminckii*, percent overlap was only at the extremes (0% or 100% overlap; Fig 5a) For *P. tricuspis* percentage overlap was mostly 100%. Percent overlap showed no clear trend with decade of collection (Fig 5b), but Asian species show higher levels of overlap than African species.

Centroid-centroid distance also varies across species (Fig 5c; Tab 1), with *M. crassicaudata* and *S. temminckii* showing the greatest displacement. I also found

that the overlap between convex hulls created from NHM data and the current IUCN ranges varied across the seven species (Table 1).

In the models run for the point-radius data; percentage overlap of IUCN range and collection locale as a function of species did not converge. There was no significant relationship between the number of IUCN range/collection locale overlaps and collection decade (binomial GLM: devicance = 193.0, df = 82, p = 0.358). There was a significant relationship between region and the number of IUCN range/collection locale overlaps (binomial GLM: deivance = 173.0, df = 82, $p < 0.005$), (Fig 5d, mean percent overlap $\pm$ standard error, with raw data shown in grey).

Species Binomial	**Percent Overlap**	**Centroid-Centroid distance (km)**	**Continent**
Manis crassicuadata	45.7	2908.0	Asia
Manis javanica	74.7	290.7	Asia
Manis pentadactyla	75.4	253.9	Asia
Phataginus tetradactyla	55.7	1260.7	Africa
Phataginus tricuspis	64.4	725.9	Africa
Smutsia gigantea	23.3	667.8	Africa
Smutsia temminckii	67.6	1622.6	Africa

Table 1: Raw data results from the generation of convex hulls for the Manidae family. Convex hulls were generated using data from Natrual History Museum, London, collections to function as a primitive range map. Percent overlap is the percentage area overlap of the convex hull with the corresponding IUCN range map for each species. Centroid-Centroid distance is the straight line distance in kilometers from the calculated geometric center of both the IUCN range map and NHM convex hull.

4 Discussion

I have successfully incorporated error into the geospatial analysis of historical range maps using the methods described by Wieczorek et al. (2004), using the most up-to-date geospatial tools R has to offer to do so, and bundled my methods into an R package for maximum reproducibility. I have also brought the data I used under version control, and as a proof of concept have conducted geospatial analysis on the records of the Manidae family. I have therefore fulfilled all five objectives of my thesis,

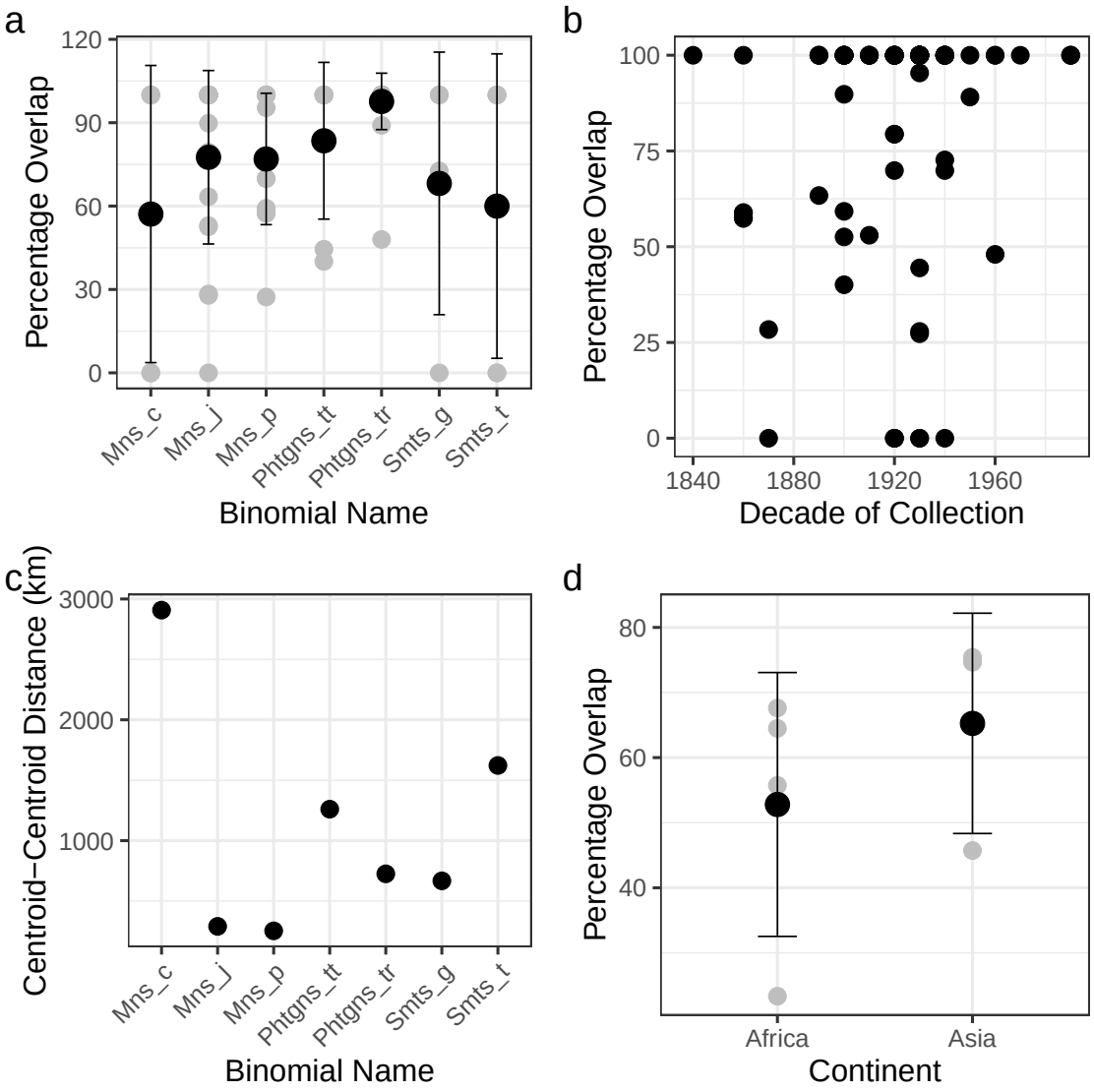

Figure 5: Geospatial analysis of pangolin (family Manidae) specimens. (a) Percent overlap of IUCN range and NHM collection localities; (b) the relationship between the decade specimens were collected and percentage overlap of localities and IUCN range maps; (c) centroid to centroid distance between IUCN ranges and NHM convex hulls in kilometers; (d) the relationship between continent of collection and percentage overlap of IUCN ranges and NHM convex hulls, there is a significant effect.

addressing a so far mostly ignored part of historical specimen use (that is georeferencing error), and creating easy to use methods of incorporating georeferencing error into any geospatial analysis.

Creation conservation baselines is something that ecologists have been trying to do since the early 20th century (Alagona et al. 2012) using a variety of historical records (e.g. oral histories, photographs, land surveys and catch logs, McClenachan et al. 2012), however many of these are subjective, and require records to have been kept. For commercially important species this is not much of an issue, (see Rick & Lockwood (2013) for an example), but for more cryptic or rarer species creating a baseline using anything other than georeferenced specimens is near impossible. Museum collections will also suffer from having few specimens from exceptionally rare species but early collectors generally were more interested in collecting the exotic than the mundane, so museums do generally overcome this barrier. More recently ecologists have been using computationally intense methods of determining species baselines, such as incorporating life history and energetics into ecologically explicit distribution models (Rodhouse et al. 2015). Models such as this provide an unparalleled example of creating a statistical basis for conservation, however they require a great deal of knowledge of each individual species, including metabolic ecology, which is not feasible across millions of vertebrate taxa, some of which are highly cryptic. Below I will first discuss the results of the pangolin case study, and then discuss more general limitations and future directions.

4.1 Pangolins case study

All seven species of pangolin in this study showed a range of percentage IUCN range/collection locale overlaps, with no clear correlation with decade of collection, but with more IUCN range/collection locale overlaps in Asian than African species. The most heavily exploited species seem to have the fewest range/collection locale overlaps, showing that these species used to inhabit quite different ranges. This is an expected result, as the most at risk species are likely to have historically experienced the greatest anthropogenic pressures, which will have caused range changes. This is problematic if we use the IUCN ranges as baselines, as that would be incorrect, leading to a belief that a population is healthy when it is in fact degraded.

These species (*M. crassicaudata* and *S. temminckii*) are currently some of the most heavily exploited because they are the most easily harvested and there is the greatest motivation in the form of financial compensation for doing so for these species (Newton et al. 2008, Challender & Hywood 2012, MacMillan & Nguyen 2014), relative to the other species of pangolin. Both *M. crassicaudata* and *S. temminckii* are terrestrial, sharing a similar life history (Mahmood et al. 2014, Baillie et al. 2014, Pietersen et al. 2015), which gives a potential cause for this correlation. *Phataginus tetradactyla* also shows a large displacement, but is arboreal. In terms of conservation baselines my results suggest there is a need to reassess conservation goals for Manidae, particularly those found in Africa. These species appear to have been affected by anthropogenic pressures to the greatest extent, so any future conservation work should aim to restore species to their historic extent if at all possible.

I did not see any effect of collection decade, possibly because I had too few specimens to observe this effect, or possibly because many of the specimens were collected during and after 'The Great Acceleration' (Steffen et al. 2015), particularly post-1950. The term 'Great Acceleration' refers to the marked increase in both human impacts on earth systems and socio-economic systems (e.g. atmospheric carbon dioxide levels and primary energy use, Steffen et al. 2015). If a majority of the specimens were collected during or after this time, then I would not expect to see a correlation as this is a demarking point at which humanity's impact on the natural world increased greatly. This means the population those specimens were collected from are post-impact populations, just as the populations used for the IUCN range maps are. This can only be addressed by using a larger number of specimens, which would involve accessing collections at other institutions.

Both continents have typically exploited pangolins for traditional medicine, however there is a growing trade primarily fueled by the Chinese market (Challender & Hywood 2012). This is being fed by a marked increase in pangolin harvesting in Africa (Ingram et al. 2018). Why Africa is experiencing this drastic increase is possibly due to greater enforcement of legal protections in Asia (Shepherd 2009, Soewu & Seodinde 2016), but more likely due to factors of relative scarcity of pangolins in Asia (such as in China, Challender & MacMillan 2014, Nash et al. 2016), and the increase in relative poverty between areas of supply and demand - China has had an average increase of Gross

National Product (GDP) of 9.5% over the last 39 years, and gone from the lowest GDP per capita to the highest by a large margin when compared to Senegal, Angola, South Sudan and Sudan (International Monetary Fund 2019).

The conclusions I can draw from this study may not necessarily be widely generalised to other taxa as they are specific to the anthropogenic pressures and life histories of pangolins, however as an indication of whether or not our baselines are correct this study suggests that in many cases they are not, as all species face significant anthropogenic pressures (De Vos et al. 2015, Rapacciuolo et al. 2017) and have done so for some decades (Steffen et al. 2015, Zalasiewicz et al. 2015; 2017). Future work should therefore expand the number of taxa investigated, to confirm whether or not our baselines are correct (such as corroborating studies such as Rick & Lockwood 2013), and if possible to modify conservation targets in response. At a smaller scale, I would like to include *Manis culionensis* (the eighth species of pangolin) in future analysis of Manidae. Similarly, I would like to include specimens from a wider source of collections if at all possible, to help overcome some of the issues of having limited data points for each species.

4.2 Limitations of *sfe* and future directions

There are some limitations and difficulties with *sf* and *sfe*; as *sf* (Pebesma 2018) is written in C and compiled for use in R creating functions that perform operations on subsets of a dataframe and rebuild that dataframe is not as intuitive as most operations in R. In particular the creation of convex hulls on a each species basis was very error prone, but this is likely due to the difficulties in computing convex hulls (*sf* does not make it clear which algorithm is used for computation, however Barber et al. (1996) present a well known algorithm and discuss some of the difficulties, such as the use of the algorithm when floating point arithmetic is conducted leads to rounding errors and therefore incorrect geometries).

One issue common to all functions, was that in subsetting the dataframe in such a way that both R and *sf* (Pebesma 2018) would correctly interpret the input and perform whichever *sf* (Pebesma 2018) function it was that was required. For the most complex case, calculating the percentage overlap, the simplest way to overcome this error was

to break the operation into three nested functions, each of which internally subsets the data, using respectively smaller subsets. For example, the first function subsets the data by species binomial, this subset is passed to the next function, which subsets the data by row, which is then passed to the final function where percent overlap is calculated on a geometry by geometry basis. This demonstrates that this error type was surmountable, however it took a great deal of time to find the solution, and how the data were subset was slightly different depending on function, so a universal solution could not be applied.

Due to time constraints I could not include all the functionality in *sfe* that I wanted to. I would like to add functionality to clip convex hulls to aquatic environments, as currently *sfe* can only clip hulls to landmasses. This would allow for better exploration of data derived from the numerous aquatic taxa that are held in the NHM's collections. I also did not have time to fully unit test *sfe*. Unit testing is usually a standard part of program development, however as I am new to the subject I did not have time to fully integrate it into my workflow, *sfe* is however informally tested on two separate and quite different computer systems, so should work well on the majority of systems. I would also like to integrate *sfe* with a species distribution modelling package, as I feel refinement of the convex hulls *sfe* can generate into distribution models which account for biologically important factors (such as temperature, altitude and rainfall), would help further inform any questions of how far our baselines have shifted.

In conclusion, I have created tools to incorporate percentage overlap into geospatial analysis of ecological data, from both collection locality and convex hulls (which I have also added tools for creating for each species as a whole). When combined with the wealth of museum collections which are being georeferenced this allows for a vast array of species to be given pre-anthropocene baselines, overcoming the key issue most taxa have when ascribing historical baselines, which is a lack of data (Rick & Lockwood 2013). Creating baselines is difficult (Alagona et al. 2012), however museum specimens provides an objective method of doing so, overcoming issues of baseline shift due to intergenerational knowledge loss (Pauly 1995), and the inevitable inaccuracies of oral records. There are more accurate methods of creating baselines (Rodhouse et al. 2015), but those require a much more detailed knowledge of each species, which is not feasible for many taxa. My methodology therefore provides

a middle ground, where many taxa can have their baselines assessed without the need for detailed knowledge of life history, which is key to conserving species as the pressures of the Anthropocene grow.

References

Alagona, P., Sandlos, J. & Wiersma, Y. (2012), 'Past Imperfect: Using Historical Ecology and Baseline Data for Conservation and Restoration Projects in North America', *Environmental Philosophy* **9**(1), 49–70.

Axosoft (2019), 'GitKraken'.
URL: *https://www.gitkraken.com/*

Baillie, J., Challender, D., Kaspal, P., Khatiwada, A., Mohapatra, R. & Nash, H. (2014), 'Manis crassicaudata. The IUCN Red List of Threatened Species', **8235**.

Barber, C. B., Dobkin, D. P., Huhdanpaa, H. & Huhdanpaa, H. (1996), 'The quickhull algorithm for convex hulls', *ACM Transactions on Mathematical Software* **22**(4), 469–483.

Barnosky, A. D., Matzke, N., Tomiya, S., Wogan, G. O. U., Swartz, B., Quental, T. B., Marshall, C., McGuire, J. L., Lindsey, E. L., Maguire, K. C., Mersey, B. & Ferrer, E. A. (2011), 'Has the Earth's sixth mass extinction already arrived?', *Nature* **471**(7336), 51–57.

Brooks, T. M., Mittermeier, R. A., Mittermeier, C. G., da Fonseca, G. A. B., Rylands, A. B., Konstant, W. R., Flick, P., Pilgrim, J., Oldfield, S., Magin, G. & Hilton-Taylor, C. (2002), 'Habitat Loss and Extinction in the Hotspots of Biodiversity', *Conservation Biology* **16**(4), 909–923.

Challender, D., Baillie, J., Ades, G., Kaspal, P., Chan, B., Khatiwada, A., Xu, L., Chin, S., KC, R., Nash, H. & Hsieh, H. (2014), 'Manis pentadactyla (Chinese Pangolin)', *The IUCN Red List of Threatened Species* **8235**.
URL: *http://www.iucnredlist.org/details/12764/0*

Challender, D. W. S. & Hywood, L. (2012), 'African pangolins under increased pressure from poaching and intercontinental trade', *TRAFFIC Bulletin* **24**(3), 53–55.

Challender, D. W. S. & MacMillan, D. C. (2014), 'Poaching is more than an Enforcement Problem', *Conservation Letters* **7**(5), 484–494.

Cooper, N., Griffin, R., Franz, M., Omotayo, M. & Nunn, C. L. (2012), 'Phylogenetic host specificity and understanding parasite sharing in primates', *Ecology Letters* **15**(12), 1370–1377.

Costanza, R., D'Arge, R., de Groot, R., Farber, S., Grasso, M., Hannon, B., Limburg, K., Naeem, S., O'Neill, R. V., Paruelo, J., Raskin, R. G., Sutton, P. & van den Belt, M. (1997), 'The value of the world's ecosystem services and natural capital', *Nature* **387**(6630), 253–260.

De Vos, J. M., Joppa, L. N., Gittleman, J. L., Stephens, P. R. & Pimm, S. L. (2015), 'Estimating the normal background rate of species extinction', *Conservation Biology* **29**(2), 452–462.

Edwards, J. L. (2004), 'Research and Societal Benefits of the Global Biodiversity Information Facility', *BioScience* **54**(6), 485–486.

Falster, D. S., FitzJohn, R. G., Pennell, M. W. & Cornwell, W. K. (2019), 'Datastorr: a workflow and package for delivering successive versions of 'evolving data' directly into R', *GigaScience* **8**(5).

Fidler, F., Chee, Y. E., Wintle, B. C., Burgman, M. A., McCarthy, M. A. & Gordon, A. (2017), 'Metaresearch for Evaluating Reproducibility in Ecology and Evolution', *BioScience* **67**(3), biw159.

Froyd, C. & Willis, K. (2008), 'Emerging issues in biodiversity conservation management: The need for a palaeoecological perspective', *Quaternary Science Reviews* **27**(17-18), 1723–1732.

Gaubert, P. & Antunes, A. (2005), 'Assessing the Taxonomic Status of the Palawan Pangolin Manis culionensis (Pholidota) Using Discrete Morphological Characters', *Journal of Mammalogy* **86**(6), 1068–1074.

GitHub Inc. (2019), 'GitHub'.
URL: *https://github.com/*

Heinrich, S., Wittmann, T. A., Prowse, T. A., Ross, J. V., Delean, S., Shepherd, C. R. & Cassey, P. (2016), 'Where did all the pangolins go? International CITES trade in pangolin species', *Global Ecology and Conservation* **8**, 241–253.

Ingram, D. J., Coad, L., Abernethy, K. A., Maisels, F., Stokes, E. J., Bobo, K. S., Breuer, T., Gandiwa, E., Ghiurghi, A., Greengrass, E., Holmern, T., Kamgaing, T. O. W., Ndong Obiang, A.-M., Poulsen, J. R., Schleicher, J., Nielsen, M. R., Solly, H., Vath, C. L., Waltert, M., Whitham, C. E. L., Wilkie, D. S. & Scharlemann, J. P. (2018),

'Assessing Africa-Wide Pangolin Exploitation by Scaling Local Data', *Conservation Letters* **11**(2), e12389.

International Monetary Fund (2019), 'World Economic Outlook Database April 2019'. **URL:** *https://www.imf.org/external/pubs/ft/weo/2019/01/weodata/index.aspx*

IUCN Standards and Petitions Subcommittee (2017), 'Guidelines for Using the IUCN Red List Categories and Criteria. Version 13.', *Iucn* **13**(March), 60.

Lyons, S. K. (2003), 'A Quantitative Assessment of the Range Shifts of Pleistocene Mammals', *Journal of Mammalogy* **84**(2), 385–402.

Mace, G., Collar, N., Gaston, K., Hilton-Taylor, C., Akçakaya, H., Leader-Williams, N., Milner-Gulland, E. & Stuart, S. (2008), 'Quantification of Extinction Risk: IUCN's System for Classifying Threatened Species', *Conservation Biology* **22**(6), 1424–1442.

MacMillan, D. C. & Nguyen, Q. A. (2014), 'Factors influencing the illegal harvest of wildlife by trapping and snaring among the Katu ethnic group in Vietnam', *Oryx* **48**(2), 304–312.

Mahmood, T., Irshad, N. & Hussain, R. (2014), 'Habitat preference and population estimates of Indian pangolin (Manis crassicaudata) in district Chakwal of Potohar Plateau, Pakistan', *Russian Journal of Ecology* **45**(1), 70–75.

McClenachan, L., Ferretti, F. & Baum, J. K. (2012), 'From archives to conservation: why historical data are needed to set baselines for marine animals and ecosystems', *Conservation Letters* **5**(5), 349–359.

Midgley, G., Hannah, L., Millar, D., Thuiller, W. & Booth, A. (2003), 'Developing regional and species-level assessments of climate change impacts on biodiversity in the Cape Floristic Region', *Biological Conservation* **112**(1-2), 87–97.

Nash, H. C., Wong, M. H. & Turvey, S. T. (2016), 'Using local ecological knowledge to determine status and threats of the Critically Endangered Chinese pangolin (Manis pentadactyla) in Hainan, China', *Biological Conservation* **196**, 189–195.

Newton, P., Nguyen, T., Roberton, S. & Bell, D. (2008), 'Pangolins in peril: using local hunters' knowledge to conserve elusive species in Vietnam', *Endangered Species Research* **6**(1), 41–53.

Pauly, D. (1995), 'Pauly 1995 Anecdotes of the shifting baseline syndrome of fisheries', *Trends in ecology & evolution* **10**(10), 430.

Pebesma, E. (2018), 'Simple Features for R: Standardized Support for Spatial Vector Data', *The R Journal* **10**(1), 439.

Pietersen, D., Waterman, C., Hywood, L., Rankin, P. & Soewu, D. (2015), 'Smutsia temminckii. The IUCN Red List of Threatened Species 2014', **8235**.
URL: *https://www.iucnredlist.org/species/12765/45222717*

Pimm, S. L., Jenkins, C. N., Abell, R., Brooks, T. M., Gittleman, J. L., Joppa, L. N., Raven, P. H., Roberts, C. M. & Sexton, J. O. (2014), 'The biodiversity of species and their rates of extinction, distribution, and protection', *Science* **344**(6187).

R Core Team (2019), 'R: A language and environment for statistical computing'.
URL: *https://www.r-project.org/*

Rapacciuolo, G., Marin, J., Costa, G. C., Helmus, M. R., Behm, J. E., Brooks, T. M., Hedges, S. B., Radeloff, V. C., Young, B. E. & Graham, C. H. (2017), 'The signature of human pressure history on the biogeography of body mass in tetrapods', *Global Ecology and Biogeography* **26**(9), 1022–1034.

Rick, T. C. & Lockwood, R. (2013), 'Integrating Paleobiology, Archeology, and History to Inform Biological Conservation', *Conservation Biology* **27**(1), 45–54.

Rodhouse, T. J., Ormsbee, P. C., Irvine, K. M., Vierling, L. A., Szewczak, J. M. & Vierling, K. T. (2015), 'Establishing conservation baselines with dynamic distribution models for bat populations facing imminent decline', *Diversity and Distributions* **21**(12), 1401–1413.

RStudio Team (2018), 'RStudio: Integrated Development for R'.
URL: *http://www.rstudio.com/*

Shepherd, C. (2009), Overview of pangolin trade in Southeast Asia, *in* 'Proccedings of the Workshop on Trade and Conservation of Pangolins Native to South and Southeast Asia', Vol. 30, pp. 6–9.

Soewu, D. A. & Seodinde, O. A. (2016), 'Utilization of pangolins in Africa: Fuelling factors, diversity of uses and sustainability', *International Journal of Biodiversity and Conservation* **7**(1), 1–10.

Steffen, W., Broadgate, W., Deutsch, L., Gaffney, O. & Ludwig, C. (2015), 'The trajectory of the Anthropocene: The Great Acceleration', *The Anthropocene Review* **2**(1), 81–98.

Thomas, C. D., Cameron, A., Green, R. E., Bakkenes, M., Beaumont, L. J., Collingham, Y. C., Erasmus, B. F. N., de Siqueira, M. F., Grainger, A., Hannah, L., Hughes, L., Huntley, B., van Jaarsveld, A. S., Midgley, G. F., Miles, L., Ortega-Huerta, M. A., Townsend Peterson, A., Phillips, O. L. & Williams, S. E. (2004), 'Extinction risk from climate change', *Nature* **427**(6970), 145–148.

Wickham, H. (2015), *R packages*, O'Reilly Media, Inc.
URL: *http://r-pkgs.had.co.nz/*

Wickham, H., Danenberg, P. & Eugster, M. (2018), *roxygen2: In-Line Documentation for R*. R package version 6.1.1.
URL: *https://CRAN.R-project.org/package=roxygen2*

Wieczorek, J., Guo, Q. & Hijmans, R. (2004), 'The point-radius method for georeferencing locality descriptions and calculating associated uncertainty', *International Journal of Geographical Information Science* **18**(8), 745–767.

Willis, K. & Birks, H. (2006), 'What Is Natural? The Need for a Long-Term Perspective', *Science* **314**(5803), 1261–1266.

Willis, K. J., Gillson, L., Brncic, T. M. & Figueroa-Rangel, B. L. (2005), 'Providing baselines for biodiversity measurement.', *Trends in ecology & evolution* **20**(3), 107–8.

Wilson, D. E. & Reeder, D. M. (2005), *Mammal species of the world : a taxonomic and geographic reference*, Johns Hopkins University Press.

Zalasiewicz, J., Waters, C. N., Summerhayes, C. P., Wolfe, A. P., Barnosky, A. D., Cearreta, A., Crutzen, P., Ellis, E., Fairchild, I. J., Gałuszka, A., Haff, P., Hajdas, I., Head, M. J., Ivar do Sul, J. A., Jeandel, C., Leinfelder, R., McNeill, J. R., Neal, C., Odada, E., Oreskes, N., Steffen, W., Syvitski, J., Vidas, D., Wagreich, M. & Williams, M. (2017), 'The Working Group on the Anthropocene: Summary of evidence and interim recommendations', *Anthropocene* **19**, 55–60.

Zalasiewicz, J., Waters, C. N., Williams, M., Barnosky, A. D., Cearreta, A., Crutzen, P., Ellis, E., Ellis, M. A., Fairchild, I. J., Grinevald, J., Haff, P. K., Hajdas, I., Leinfelder, R., McNeill, J., Odada, E. O., Poirier, C., Richter, D., Steffen, W., Summerhayes, C., Syvitski, J. P., Vidas, D., Wagreich, M., Wing, S. L., Wolfe, A. P., An, Z. & Oreskes, N. (2015), 'When did the Anthropocene begin? A mid-twentieth century boundary level is stratigraphically optimal', *Quaternary International* **383**, 196–203.

A Appendix

NHM GEOREFERENCING

Updated Guidelines based on MaNIS/HerpNET/ORNIS By Malcolm Penn

Note: The most important thing to remember that assumptions should be noted.

With modifications in italics by Andrea Sartorius 2017.

LOCALITY TYPE	GEOREFERENCING PROCEDURE	DETERMINING EXTENT
Named Places **Named Place or Urban Area:** locality consists of a reference to a geographic feature having a spatial extent e.g.: "Reading"	It is best to use the geographic centre (the centroid/midpoint of both the latitude and longitude extremes) for the coordinates of named places. Use the designated centre from Google maps or Google Earth to at least ensure source consistency.	Use the radius/distance from the coordinates of the named place to the furthest point within that named place.
Named Places **Remote Named Place:** locality does not have a clear boundary shown on the map e.g.: "Dirty Socks": The extent is 0.4 mi because the nearest named feature, Hot Springs, is 0.8 mi away.	Localities that do not have a shaded boundary or a topographic symbol for buildings shown on the map, place coordinates at the dot for that locale on the map.	The extent is half the distance to the nearest named place. (Make a note of this in comment field). *The nearest named place should be of comparable size/type*
Named Places **Near a Named Place:** e.g.: "Near(by) Swindon", "vicinity of Killin" "close to Killin" “above Killin” etc.	Use the geographic centre of the named place for the geographic coordinates. Remember word Near Nr, or above are appended after the place name Worcester (Nr)	The extent will be the distance from the geographic centre of the named place to the halfway point between that geographic centre and the centre of the next nearest named place. (The named place entered into the spreadsheet should include "near", or "vicinity of", or another such modifier). *The nearest named place should be of comparable size/type*

Named Places **Street Address:** e.g. "1 Lovington lane, Lower Broadheath, Worcester"	Locate the address using GeoRef interface with Google maps	The extent is the smallest area possible that cannot be mistaken for any other address.
Named Places **Ranch/Farm, Golf course,Quarry,Mine Estates, or Parks** e.g. "Apple tree Farm"	Treat all as named places. If you are not able to locate them with GeoRef interface, use search engines such as Google to locate them in relation to nearby geographic entities. If farms or Golf course appear on a map, usually only the building will be shown. Take coordinates from the buildings themselves, unless you can identify the exact border of the farm.	If you can find exact boundaries, then treat the ranch or farm as you would a named place. Otherwise, the extent is half the distance between the buildings and the next nearest named place. *The nearest named place should be of comparable size/type*
Named Places **Junction** e.g. "junction of Elsham Rd. and Russell Rd.", "junction of Rio Claro and Rio La Hondura"	Locate the two roads or rivers in Georef interface or a map and obtain the coordinates of the point of intersection. Use Streetmap/Google Earth or Google Maps to help locate the road names, as roads may not be labelled on the map you are using. Treat the Road junction as a Precise location and enter the full junction description. Treat the river junction as a Feature and enter the full description.	Measure the extent of the junction as if it were a named place. If the extent or a road junction cannot be measured on the maps available, use the following standards: The extent is 10 m for two-lane city streets and two-lane highways. The extent is 20 m for four-lane highways. The extent is 30 m for large highways with medians.
Named Places **Cave:** e.g. "Las Cuevas Cave", "Chiquibul Caverns"	Georeference the entrance to the cave.	The extent is usually the surface length of the cave.
Named Places ***Mountain – specific named mountain***	*Attempt to find coordinates for the mountain in question*	*Use terrain view to see how large the mountain is – extent is the furthest low point from the centre coordinates*
Named Places ***Mountain range***	*Use the centre point of the mountain range*	*Use terrain view to determine the full extent of the mountain range – make extent include the furthest point where there are continuous mountains*

Named Places **River, Mouth of River, or Head of River:** e.g. "River Thames", "Mouth of Severn River"	River: Make a straight line from the mouth of the river to the head of the river. Calculate the centre of this line, and place the coordinates closest to the centre of the line on the river itself. Do not use the coordinates given by gazetteers, as these points usually correspond to the mouths of the river, not the geographic centres. *Do not make a straight line from the mouth to the head of the river. Instead, trace the river using the Google maps measure distance tool from mouth to head. Divide the length of the river in half and place the coordinates at the midpoint.* River Mouth: Georeference where the river meets a larger body of water; this is usually the point of the river with the lowest elevation. River Head: Georeference where the river starts (usually in mountains, canyons, or lakes); this should be the point of the river that has the highest elevation	The extent is half the length of the line drawn. Make sure to only include the portion of the river that is within the specified higher geography. The extent is half the distance across the river mouth or head (this is usually rather small). *The smallest possible extent is 10 meters.*
Named Places **In between two Places** e.g. "Between Bristol and Bath, Uk."	Georeference the midpoint between the centres of both named places.	The extent is half the distance between the centres of both named places.
Named Places **Names States and Parishes** First order Admin area, State/Province e.g. Florida e.g. Settlement has the same place name as a parish/county/commune/	Use the geographic centre of the State/ Province for the geographic coordinates use. Use **Province** Make assumption that the collector means the settlement, unless the label states parish/county/commune municipality. **Georeference the settlement.**	The extent will be the distance from the geographic centre of the State/Province to furthest point of the Province shape using the radius tool. Use the radius/distance from the coordinates of the named place to the furthest point within that named place.

municipality. e.g. If a label clearly states 2nd or 3rd Admin area (Department/ Parish/County/Commune Municipality) e.g. If label shows a place name which can only be a parish/county	*If a settlement has the same name as a larger area, use the larger area (unless there is clear indication that this should not be done) since it will also include the settlement.* Use the geographic centre of the parish/county/commune/municipality/ county for the geographic coordinates use. **Precise Locality**. Follow above and maps as parish/county/ commune/municipality use **Precise Locality**.	 The extent will be the distance from the geographic centre of the county/parish to furthest point of the shape using the radius tool. The extent will be the distance from the geographic centre of the county/parish to furthest point of the shape using the radius tool.
Offsets **Offset Only**: **locality consists of an offset from a named place without any direction specified** e.g.: "5 km outside Brisbane"	Record the geographic coordinates of the centre of the named place, just as you would for a "normal" named place. (Note that the precise locality is 5km outside Brisbane, so not Brisbane, but still use Brisbane for the coordinates)	Use the extent of the named place + distance.
Offsets **Direction Only:** **locality consists of a direction from or within a named place without any distance specified** e.g. "N Reading", "N of Reading"	If only a direction is given, such as "N Reading" and there is no town named "North Reading", then there is no way of knowing if the collector meant "northern portion of Reading" or "North of Reading." Find the distance from the centre of the named place (Reading) to the centre of the next nearest named place to the north. Place the coordinate at one half of the distance to the centre of the next nearest named place in the direction specified. **Remember the Direction should be appended after the place name, e.g. Reading (North or East etc.,)**	For such localities, the extent is one half of the distance between the centre of the named place in question and the centre of the next nearest named place in the specified direction. *The nearest named place should be of comparable size/type.*

Offsets **Offset at a Heading: locality contains a distance in a given direction** e.g. "50 miles E of Lima"	Assume the collector measured the distance "by air." unless stated otherwise. Use the GeoRef interface to measure 50 miles in an easterly direction from the centroid of the named place. *Measure the distance using the measure distance tool in Google maps*	Calculate the extent to the next nearest place name. *Extent is the offset distance (ex: 50 miles).*
Offsets **Offset Along a Path, in One Direction: locality describes a route from a named place** e.g. "7.9 mi N Beatty, on US 95"	If "by road" is specified in the locality description, Use the line tool to follow the route. Begin at the centre of the starting point and use the measuring tool to follow the road until you have travelled the distance given. The coordinates come from this ending point.	Use the extent of the starting point.
Offsets **Offset Along a River, in One Direction** e.g. "3 miles above Worcester on River Severn on left bank"	Treat the stream as you would a road. Above refers to upstream and below refers to downstream. Left and right sides of a river are determined from the perspective of facing downstream.	Use the extent of the starting point.
Coordinates **GPS (Global Positioning System)**	When georeferencing GPS coordinates, make sure to note whether the accuracy and the datum where reported. Always record coordinates in decimal degrees and make sure we distinguish the master records by using Reading (North) or Reading (North of)	The accuracy of the GPS at the time the coordinates were recorded. If none was recorded, assume 30m.
Coordinates **Latitude and Longitude Coordinates: coordinates from unknown source, given in locality description**	Always record coordinates in decimal degrees. Enter these coordinates as the Precise locality.	Extent is 30m

How georeferencing error is dealt with in the literature

Key words used: georeferencing of specimens

Reference	Method of dealing with geospatial error
Aedo & Pando 2017	1-minute accuracy
Anacker & Strauss 2014	None used - applied a 10km buffer as range estimation, but no error added to geographical point
Andrew et al. 2011	Mention sampling resolution of 1km2 at best and median of 101km2, georef error not explicitly mentioned
Andrew et al. 2012	None mentioned - used cbif (canadian biodiversity information facility) so may have used data that had point-radius error, but if so no mention of incorporating it into the study is made
Arrigo et al. 2013	None used - discarded imprecise georefs (or gave them new coords, no mention of error)
Barros et al. 2012	Arbitrary precision of 5 arc minutes given to all localities
Beaman & Conn 2003	None used, but state error analysis is needed
Beentje et al. 2006	None mentioned
Bendiksby et al. 2014	None mentioned

Boakes et al. 2010	A bewildering variety of methods used accuracy to 1degree, or 10minutes or if description matched two or more places a midpoint was taken, so long as it was accurate to 1degree. Data was then required to meet arbitrary requirements such as that it was within a reasonable distance of the species known range (what this constitutes is not specified)
Boedeker et al. 2010	None mentioned
Bontrager & Angert 2016	Error distance calculated, but how is not said. Potentially point raidius.
Boumans 2011	None used - they simply say they georef'd the specimens as best they could to a coord.
Brummitt et al. 2008	None specified, alludes to error being computed for georeferences
Buckley 2008	None mentioned, however data was from GBIF and herpnet (GBIF definitely uses point-radius for its data, but as the author makes no mention, I assume they haven't included error)
Burgio et al. 2018	Point radius
Campbell et al. 2011	None used - some data was accepted to have error as was georef'd using older mapping systems, but still only used point data
Campbell et al. 2012	None mentioned - translational errors between map projections are talked about, but they do not mention anything else. Projection errors are deemed acceptable 20-90m long and 292-300m lat.
Carlson et al. 2017	Probably point-radius (Wieczorek & Chapman, 2006)

Cason et al. 2016	Point-radius (by the sound of it, assigned coords and error radii)
?	Yes, seems like point-radius but a little difficult to tell
Chatfield-Taylor & Cole 2017	None mentioned – say they use GEOLocate which can calculate polygonal error
Christenhusz & Toivonen 2008	None mentioned
Cook et al. 2014	None mentioned
Couvreur et al. 2011	None mentioned
Craven & Vorster 2006	Quarter degree square system of Edwards and Leistner (1971)
Crawford & Hoagland 2009	Georef'd to township (93.3km2) resolution (present or absent essentially)
Damerval et al. 2018	None mentioned
Davenport et al. 2010	None mentioned - modern GPS loggers used for georefing so little error on the points I think
De Giovanni et al. 2012	Point-radius
de la Torre et al. 2012	None mentioned
DeWalt et al. 2009	None used
Dodd et al. 2015	None mentioned
Donoso et al. 2009	Point radius, follows methods outlined in Wieczorek et al., (2004) categorises description data into nine bins depending on certainty/quality of description

Droissart et al. 2011	None used - imprecise data was discarded, what counted as imprecise isn't mentioned, however they do mention that spp were grouped into classes based on distance from the ocean at 1degree, more than 2-3degree and more than 3degree classes. In this case 1deg corresponds to 111km
Droissart et al. 2012	Filtered data to precise only (accurate to 10km)
Duursma et al. 2013	None mentioned. GBIF used, but also georeferenced themselves, do not mention error/uncertainty
Erb et al. 2011	Point radius
Escudero et al. 2012	None mentioned - went for high precision of lat/long coords but didn't discuss error of measurement
Feeley & Silman 2010	Data filtering (removing data before modelling if it does not meet minimum requirements (which can be quite strict) of quality.
Foley et al. 2007	None used. Precision was either of lat long or 1km - 100m depending on if MGRS coords were used or specimens were re-georef'd
Funk et al. 1999	None used
Gómez-Mendoza & Arriaga 2007	None mentioned - as with many, if the author isn't doing the georefing they don't seem to think about uncertainty
Garcia-Milagros & Funk 2010	Gazetteer coordinates as a measure of uncertainty, Point-radius
GBIF.org n.d.	Uses point-radius method
Gotelli et al. 2012	None mentioned

Graham et al. 2007	Demonstrate that MaxEnt and boosted regression trees are both robust to moderate geographical error, interestingly
Graham et al. 2013	Alludes to standard georeferencing techniques and states anything with error greater than 5km was dropped from modelling, however no indication of what standard technique used actually is. MAXENT is also used, however I do not believe this in of itself deals with georef errors
Guralnick & Van Cleve 2005	Points to methods of data prep in supplementary materials, which are not available (broken web link)
Guralnick et al. 2006	Point-radius method
Gutiérrez et al. 2014	Error is acknowledged, but how it was calculated is not specified. I think, based on the supplementary material and figures, point-radius or an equivalent was used.
Henebry et al. 2001	Error of one quarter section (65 ha) if specimens had to be georef'd. Point data was used for the other previously ref'd specimens
Hopkins 2007	No method described for georef'd points, but say that uncertainty was +-50km (this was used as a reason for not using a particular model)
Kozak & Wiens 2007	None used - say all georefecences were from systematic studies with ref being taken from original authors (so probably lat long coords)
Kozak et al. 2008	None mentioned
Lash et al. 2012	Point-radius

?	None mentioned, GBIF was used for part of the dataset so some point radius possibly
Lozier et al. 2009	Georef'd to place name only (presumably Yellowstone has equal meaning here to Medstead (a small village), so resolution varies wildly).
Magwé-Tindo et al. 2016	Point radius (also using handheld gps logger, so less error in modern collection)
Martellos et al. 2014	Difficult to tell, mention wiezoreck 2004 but also say georef'd to 1km2 grids
Martin & Omland 2011	Point-radius
Matthews & Mazer 2016	None mentioned
McAllister et al. 2019	None mentioned - climatic data associated with locality was at the 30-arc second resolution, so might be a proxy in some cases
McCormack et al. 2010	None mentioned. Some data was direct observation recorded with a GPS logger, but museum specimens were used as well
McElwain 2004	None mentioned
McGowan & Kiessling 2013	Resolution of 10km - I'm guessing point radius as BioGeoMancer used. Justification used of 10km is broadly considered acceptable in ecological studues, and 100km in paleontology
Miller et al. 2013	None mentioned
Miller et al. 2009	Coords assigned through 10km locality names
Molgo et al. 2017	None mentioned
Nemitz et al. 2012	Mentions error distances, but not how they were calculated

Neufeld et al. 2003	None used
Nuelle et al. 2018	None mentioned
Phillips & Dudík 2008	A machine learning method of distribution probability, does not seem to deal with initial error in georef's, but many papers use this method
Rajbhandary et al. 2011	None mentioned
Ralston & Kirchman 2012	MAXENT used, but no error of coords mentioned (visual checks for obvs discrepancies)
Ralston & Kirchman 2013	Point-radius (not explicitly mentioned, but talk about uncertainty radius and the georefs coming from GBIF)
Riordan & Rundel 2009	Data filtering (I think, they basically eyeballed it to remove any obvious errors in data entry)
Rissler & Apodaca 2007v	None mentioned
Rivers et al. 2011	None used
Rivers et al. 2010	None mentioned, assuming that as with many of these if the authors are not conducting the georeferencing someone else has done so probably using point-radius as that is the most widely used method
Roberts et al. 2016	None used. Data was filtered
Rowe 2005	Post-hoc 3-dimensional georeferencing (point radius but including a z parameter as far as I can tell).
Särkinen et al. 2011	Data filtering (they allude to georef errors but don't incorporate into study)
Sérgio et al. 2012	1kmx1km scale of georef

Sandall & Deans 2018	Point radius (resolution to 30arcseconds)
Schmidt et al. 2005	5-10km if older record made using gazetteer, precise location from newer GPS ref'd specimens
Sidlauskas & Vari 2012	None mentioned
Snyder et al. 2016	Point -raidus
Soberón et al. 2000	No formal method, localities were assigned to be 1 min arc (translating to pixels of 1.1km a side)
Stein & Wieczorek 2004	Point-radius method
Stigall et al. 2014	Point radius method
Stockwell et al. 2006	None mentioned - pulls georef'd records from a wide variety of sources it seems. Error not thought about in this context
Syfert et al. 2017	Error calculated, don't say how. Give a median value of 8km and range of 500m to over 100km
Syfert et al. 2016	None mentioned - georeferenced to a "high standard" no quantative measure of what this is given
Tobler et al. 2007	Manually checked coords to reduce error (error between 1-100km reported in this paper, no formal method of adding uncertainty mentioned)
Velásquez-Tibatá et al. 2016	Bayesian logistic regression with measurement error
Wehr et al. 2013	None mentioned
Wieringa & Sosef 2011	None mentioned - do talk about species being included if they were within a 10km buffer of parks, but this isn't the same as error radius really

Wilkin et al. 2009	None mentioned - I think GPS handloggers were used here
Zeilinger et al. 2017	Point-radius method
Zhang et al. 2013	Ref'd to five arc minutes, as this was resolution of climate data

References

Aedo, C. & Pando, F. (2017), 'A distribution and taxonomic reference dataset of Geranium in the New World', *Scientific Data* **4**(1), 170049.

Anacker, B. L. & Strauss, S. Y. (2014), 'The geography and ecology of plant speciation: range overlap and niche divergence in sister species', *Proceedings of the Royal Society B: Biological Sciences* **281**(1778), 20132980.

Andrew, M. E., Wulder, M. A. & Coops, N. C. (2011), 'How do butterflies define ecosystems? A comparison of ecological regionalization schemes', *Biological Conservation* **144**(5), 1409–1418.

Andrew, M. E., Wulder, M. A., Coops, N. C. & Baillargeon, G. (2012), 'Beta-diversity gradients of butterflies along productivity axes', *Global Ecology and Biogeography* **21**(3), 352–364.

Arrigo, N., Therrien, J., Anderson, C. L., Windham, M. D., Haufler, C. H. & Barker, M. S. (2013), 'A total evidence approach to understanding phylogenetic relationships and ecological diversity in Selaginella subg. Tetragonostachys', *American Journal of Botany* **100**(8), 1672–1682.

Barros, F. S. M., de Siqueira, M. F. & da Costa, D. P. (2012), 'Modeling the potential geographic distribution of five species of Metzgeria Raddi in Brazil, aiming at their conservation'.

Beaman, R. & Conn, B. (2003), 'Automated geoparsing and georeferencing of Malesian collection locality data', *Telopea* **10**(1), 43–52.

Beentje, H., Luke, W., S.A., G. & J., M. (2006), 'Restricted range endemism in East African plants', *Taxonomy and ecology of African plants, their conservation and sustainable use. Proceedings of the 17th AETFAT Congress* pp. 229–245.

Bendiksby, M., Mazzoni, S., Jørgensen, M. H., Halvorsen, R. & Holien, H. (2014), 'Combining genetic analyses of archived specimens with distribution

modelling to explain the anomalous distribution of the rare lichen Staurolemma omphalarioides: long-distance dispersal or vicariance?', *Journal of Biogeography* **41**(11), 2020–2031.

Boakes, E. H., McGowan, P. J. K., Fuller, R. A., Chang-qing, D., Clark, N. E., O'Connor, K. & Mace, G. M. (2010), 'Distorted Views of Biodiversity: Spatial and Temporal Bias in Species Occurrence Data', *PLoS Biology* **8**(6), e1000385.

Boedeker, C., Eggert, A., Immers, A. & Wakana, I. (2010), 'Biogeography of Aegagropila linnaei (Cladophorophyceae, Chlorophyta): a widespread freshwater alga with low effective dispersal potential shows a glacial imprint in its distribution', *Journal of Biogeography* **37**(8), no–no.

Bontrager, M. & Angert, A. L. (2016), 'Effects of range-wide variation in climate and isolation on floral traits and reproductive output of Clarkia pulchella', *American Journal of Botany* **103**(1), 10–21.

Boumans, L. (2011), 'The Plecoptera Collection At The Natural History Museum In Oslo', *Illiesia* **7**(25), 280–290.

Brummitt, N., Bachman, S. & Moat, J. (2008), 'Applications of the IUCN Red List: towards a global barometer for plant diversity', *Endangered Species Research* **6**(2), 127–135.

Buckley, L. B. (2008), 'Linking traits to energetics and population dynamics to predict lizard ranges in changing environments.', *The American naturalist* **171**(1), E1–E19.

Burgio, K. R., Carlson, C. J. & Bond, A. L. (2018), 'Georeferenced sighting and specimen occurrence data of the extinct Carolina Parakeet (Conuropsis carolinensis) from 1564 - 1944.', *Biodiversity data journal* (6), e25280.

Campbell, T. L., Lewis, P. J., Thies, M. L. & Williams, J. K. (2012), 'A Geographic Information Systems (GIS)-based analysis of modern South African rodent distributions, habitat use, and environmental tolerances', *Ecology and Evolution* **2**(11), 2881–2894.

Campbell, T. L., Lewis, P. J. & Williams, J. K. (2011), 'Analysis of the modern distribution of South African Gerbilliscus (Rodentia: Gerbillinae) with implications for Plio-Pleistocene palaeoenvironmental reconstruction', *South African Journal of Science* **107**(1/2), 1–7.

Carlson, C. J., Burgio, K. R., Dougherty, E. R., Phillips, A. J., Bueno, V. M., Clements, C. F., Castaldo, G., Dallas, T. A., Cizauskas, C. A., Cumming, G. S., Doña, J., Harris, N. C., Jovani, R., Mironov, S., Muellerklein, O. C., Proctor, H. C.

& Getz, W. M. (2017), 'Parasite biodiversity faces extinction and redistribution in a changing climate', *Science Advances* **3**(9), e1602422.

Cason, M. M., Baltensperger, A. P., Booms, T. L., Burns, J. J. & Olson, L. E. (2016), 'Revised distribution of an Alaskan endemic, the Alaska Hare (Lepus othus), with implications for taxonomy, biogeography, and climate change', *Arctic Science* **2**(2), 50–66.

Chatfield-Taylor, W. & Cole, J. A. (2017), 'Living rain gauges: cumulative precipitation explains the emergence schedules of California protoperiodical cicadas', *Ecology* **98**(10), 2521–2527.

Christenhusz, M. J. M. & Toivonen, T. K. (2008), 'Giants invading the tropics: the oriental vessel fern, Angiopteris evecta (Marattiaceae)', *Biological Invasions* **10**(8), 1215–1228.

Cook, D., Lee, S. T., Taylor, C. M., Bassüner, B., Riet-Correa, F., Pfister, J. A. & Gardner, D. R. (2014), 'Detection of toxic monofluoroacetate in Palicourea species', *Toxicon* **80**, 9–16.

Couvreur, T. L., Porter-Morgan, H., Wieringa, J. J. & Chatrou, L. W. (2011), 'Little ecological divergence associated with speciation in two African rain forest tree genera', *BMC Evolutionary Biology* **11**(1), 296.

Craven, P. & Vorster, P. (2006), 'Patterns of plant diversity and endemism in Namibia', *Bothalia* **36**(2), 175–189.

Crawford, P. H. C. & Hoagland, B. W. (2009), 'Can herbarium records be used to map alien species invasion and native species expansion over the past 100years?', *Journal of Biogeography* **36**(4), 651–661.

Damerval, C., Ben Othman, W., Manicacci, D. & Jabbour, F. (2018), 'Distribution area of the two floral morphs of Nigella damascena L. (Ranunculaceae): a diachronic study using herbarium specimens collected in France', *Botany Letters* **165**(3-4), 396–403.

Davenport, T. R. B., De Luca, D. W., Bracebridge, C. E., Machaga, S. J., Mpunga, N. E., Kibure, O. & Abeid, Y. S. (2010), 'Diet and feeding patterns in the kipunji (Rungwecebus kipunji) in Tanzania's Southern Highlands: a first analysis', *Primates* **51**(3), 213–220.

De Giovanni, R., Bernacci, L. C., De Siqueira, M. F. & Rocha, F. S. (2012), 'The real task of selecting records for ecological niche modelling', *Natureza a Conservacao* **10**(2), 139–144.

de la Torre, L., Cerón, C. E., Balslev, H. & Borchsenius, F. (2012), 'A biodiversity informatics approach to ethnobotany: Meta-analysis of plant use patterns in Ecuador', *Ecology and Society* **17**(1).

DeWalt, R. E., Cao, Y., Hinz, L. & Tweddale, T. (2009), 'Modelling of historical stonefly distributions using museum specimens', *Aquatic Insects* **31**(sup1), 253–267.

Dodd, A. J., Burgman, M. A., McCarthy, M. A. & Ainsworth, N. (2015), 'The changing patterns of plant naturalization in Australia', *Diversity and Distributions* **21**(9), 1038–1050.

Donoso, D. A., Salazar, F., Maza, F., Cárdenas, R. E. & Dangles, O. (2009), 'Diversity and distribution of type specimens deposited in the Invertebrate section of the Museum of Zoology QCAZ, Quito, Ecuador', *Annales de la Société entomologique de France (N.S.)* **45**(4), 437–454.

Droissart, V., Hardy, O. J., Sonké, B., Dahdouh-Guebas, F. & Stévart, T. (2012), 'Subsampling Herbarium Collections to Assess Geographic Diversity Gradients: A Case Study with Endemic Orchidaceae and Rubiaceae in Cameroon', *Biotropica* **44**(1), 44–52.

Droissart, V., Sonké, B., Hardy, O. J., Simo, M., Taedoumg, H., Nguembou, C. K. & Stévart, T. (2011), 'Do plant families with contrasting functional traits show similar patterns of endemism? A case study with Central African Orchidaceae and Rubiaceae', *Biodiversity and Conservation* **20**(7), 1507–1531.

Duursma, D. E., Gallagher, R. V., Roger, E., Hughes, L., Downey, P. O. & Leishman, M. R. (2013), 'Next-Generation Invaders? Hotspots for Naturalised Sleeper Weeds in Australia under Future Climates', *PLoS ONE* **8**(12), e84222.

Erb, L. P., Ray, C. & Guralnick, R. (2011), 'On the generality of a climate-mediated shift in the distribution of the American pika (Ochotona princeps)', *Ecology* **92**(9), 1730–1735.

Escudero, M., Hipp, A. L., Hansen, T. F., Voje, K. L. & Luceño, M. (2012), 'Selection and inertia in the evolution of holocentric chromosomes in sedges (Carex, Cyperaceae)', *New Phytologist* **195**(1), 237–247.

Feeley, K. J. & 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', *Journal of Biogeography* **37**(4), 733–740.

Foley, D. H., Weitzman, A. L., Miller, S. E., Faran, M. E., Rueda, L. M. & Wilkerson, R. C. (2007), 'The value of georeferenced collection records for predicting pat-

terns of mosquito species richness and endemism in the Neotropics', *Ecological Entomology* **0**(0), 071203162814003–???

Funk, V., Zermoglio, M. F. & Nasir, N. (1999), 'Testing the use of specimen collection data and GIS in biodiversity exploration and conservation decision making in Guyana', *Biodiversity and Conservation* **8**(6), 727–751.

Garcia-Milagros, E. & Funk, V. A. (2010), 'data: Improving the use of information from museum specimens: Using Google Earth© to georeference Guiana Shield specimens in the US National Herbarium', *Frontiers of Biogeography* **2**(3).

GBIF.org (n.d.), 'GBIF Home Page'.
URL: *https://www.gbif.org/*

Gómez-Mendoza, L. & Arriaga, L. (2007), 'Modeling the Effect of Climate Change on the Distribution of Oak and Pine Species of Mexico', *Conservation Biology* **21**(6), 1545–1555.

Gotelli, N. J., Chao, A., Colwell, R. K., Hwang, W.-h. & Graves, G. R. (2012), 'Specimen-Based Modeling, Stopping Rules, and the Extinction of the Ivory-Billed Woodpecker', *Conservation Biology* **26**(1), 47–56.

Graham, C. H., Elith, J., Hijmans, R. J., Guisan, A., Townsend Peterson, A. & Loiselle, B. A. (2007), 'The influence of spatial errors in species occurrence data used in distribution models', *Journal of Applied Ecology* **45**(1), 239–247.

Graham, M. R., Jaeger, J. R., Prendini, L. & Riddle, B. R. (2013), 'Phylogeography of the Arizona hairy scorpion (Hadrurus arizonensis) supports a model of biotic assembly in the Mojave Desert and adds a new Pleistocene refugium', *Journal of Biogeography* **40**(7), 1298–1312.

Guralnick, R. P., Wieczorek, J., Beaman, R., Hijmans, R. J. & Group, t. B. W. (2006), 'BioGeomancer: Automated Georeferencing to Map the World's Biodiversity Data', *PLoS Biology* **4**(11), e381.

Guralnick, R. & Van Cleve, J. (2005), 'Strengths and weaknesses of museum and national survey data sets for predicting regional species richness: comparative and combined approaches', *Diversity and Distributions* **11**(4), 349–359.

Gutiérrez, E. E., Boria, R. A. & Anderson, R. P. (2014), 'Can biotic interactions cause allopatry? Niche models, competition, and distributions of South American mouse opossums', *Ecography* **37**(8), 741–753.

Henebry, G., Putz, B. C. & Merchant, J. W. (2001), 'Modeling Reptile and Amphibian Range Distributions from Species Occurrences and Landscape Variables', *GAP Analysis Bulletin No. 10* **10**, 22–26.

Hopkins, M. J. G. (2007), 'Modelling the known and unknown plant biodiversity of the Amazon Basin', *Journal of Biogeography* **34**(8), 1400–1411.

Kozak, K. H., Graham, C. H. & Wiens, J. J. (2008), 'Integrating GIS-based environmental data into evolutionary biology', *Trends in Ecology & Evolution* **23**(3), 141–148.

Kozak, K. H. & Wiens, J. J. (2007), 'Climatic zonation drives latitudinal variation in speciation mechanisms', *Proceedings of the Royal Society B: Biological Sciences* **274**(1628), 2995–3003.

Lash, R., Carroll, D. S., Hughes, C. M., Nakazawa, Y., Karem, K., Damon, I. K. & Peterson, A. (2012), 'Effects of georeferencing effort on mapping monkeypox case distributions and transmission risk', *International Journal of Health Geographics* **11**(1), 23.

Lozier, J. D., Aniello, P. & Hickerson, M. J. (2009), 'Predicting the distribution of Sasquatch in western North America: anything goes with ecological niche modelling', *Journal of Biogeography* **36**(9), 1623–1627.

Magwé-Tindo, J., Zapfack, L. & Sonké, B. (2016), 'Diversity of wild yams (Dioscorea spp., Dioscoreaceae) collected in continental Africa', *Biodiversity and Conservation* **25**(1), 77–91.

Martellos, S., Attorre, F., Farcomeni, A., Francesconi, F., Pittao, E. & Tretiach, M. (2014), 'Species distribution models backing taxa delimitation: the case of the lichen Squamarina cartilaginea in Italy', *Flora - Morphology, Distribution, Functional Ecology of Plants* **209**(12), 698–703.

Martin, M. D. & Omland, K. E. (2011), 'Environmental Niche Modeling Reveals Climatic Differences among Breeding Ranges of Orchard Oriole Subspecies', *The American Midland Naturalist* **166**(2), 404–414.

Matthews, E. R. & Mazer, S. J. (2016), 'Historical changes in flowering phenology are governed by temperature precipitation interactions in a widespread perennial herb in western North America', *New Phytologist* **210**(1), 157–167.

McAllister, C. A., McKain, M. R., Li, M., Bookout, B. & Kellogg, E. A. (2019), 'Specimen-based analysis of morphology and the environment in ecologically dominant grasses: the power of the herbarium', *Philosophical Transactions of the Royal Society B: Biological Sciences* **374**(1763), 20170403.

McCormack, J. E., Zellmer, A. J. & Knowles, L. L. (2010), 'Does niche divergence accompany allopatric divergence in Aphelocoma Jays as predicted under ecological speciation?: Insights from tests with niche models', *Evolution* **64**(5), 1231–1244.

McElwain, J. C. (2004), 'Climate-independent paleoaltimetry using stomatal density in fossil leaves as a proxy for CO2 partial pressure', *Geology* **32**(12), 1017.

McGowan, A. & Kiessling, W. (2013), 'Using abundance data to assess the relative role of sampling biases and evolutionary 2 radiations in Upper Muschelkalk ammonoids', *Acta Palaeontologica Polonica* **58**(3), 561–572.

Miller, J. S., Krupnick, G. A., Stevens, H., Porter-Morgan, H., Boom, B., Acevedo-Rodríguez, P., Ackerman, J., Kolterman, D., Santiago, E., Torres, C. & Velez, J. (2013), 'Toward Target 2 of the Global Strategy for Plant Conservation: An Expert Analysis of the Puerto Rican Flora to Validate New Streamlined Methods for Assessing Conservation Status', *Annals of the Missouri Botanical Garden* **99**(2), 199–205.

Miller, R. J., Carroll, A. D., Wilson, T. P. & Shaw, J. (2009), 'Spatiotemporal Analysis of Three Common Wetland Invasive Plant Species Using Herbarium Specimens and Geographic Information Systems', *Castanea* **74**(2), 133–145.

Molgo, I. E., Soltis, D. E. & Soltis, P. S. (2017), 'Cytogeography of Callisia section Cuthbertia (Commelinaceae).', *Comparative cytogenetics* **11**(4), 553–577.

Nemitz, D., Huettmann, F., Spehn, E. M. & Dickoré, W. B. (2012), Mining the Himalayan Uplands Plant Database for a Conservation Baseline Using the Public GMBA Webportal*, *in* 'Protection of the Three Poles', Springer Japan, Tokyo, pp. 135–158.

Neufeld, D. L., Guralnick, R. P., Glaubitz, R. & Allen, J. R. (2003), 'Museum Collections Data and Online Mapping Applications', *https://doi.org/10.1659/0276-4741(2003)023[0334:MCDAOM]2.0.CO;2* **23**(4), 334–337.

Nuelle, R. J. J., Aicezs, K. K., Nuelle, R. J. I. & Whitbeck, M. (2018), 'Automeris louisiana (Lepidoptera: Saturniidae) populations in the chenier plain habitat of coastal Texas, with new distributional and larval host plant records', *Journal of Entomology and Zoology Studies* **6**(2), 1182–1188.

Phillips, S. J. & Dudík, M. (2008), 'Modeling of species distributions with Maxent: new extensions and a comprehensive evaluation', *Ecography* **31**(2), 161–175.

Rajbhandary, S., Hughes, M., Phutthai, T., Thomas, D. & Shrestha, K. K. (2011), 'Asian Begonia: out of Africa via the Himalayas', *Gard Bull Singapore* **63**, 277–286.

Ralston, J. & Kirchman, J. J. (2012), 'Continent-scale genetic structure in a boreal forest migrant, the Blackpoll Warbler (Setophaga striata)', *The Auk* **129**(3), 467–478.

Ralston, J. & Kirchman, J. J. (2013), 'Predicted range shifts in North American boreal forest birds and the effect of climate change on genetic diversity in blackpoll warblers (Setophaga striata)', *Conservation Genetics* **14**(2), 543–555.

Riordan, E. C. & Rundel, P. W. (2009), 'Modelling the distribution of a threatened habitat: the California sage scrub', *Journal of Biogeography* **36**(11), 2176–2188.

Rissler, L. J. & Apodaca, J. J. (2007), 'Adding More Ecology into Species Delimitation: Ecological Niche Models and Phylogeography Help Define Cryptic Species in the Black Salamander (Aneides flavipunctatus)', *Systematic Biology* **56**(6), 924–942.

Rivers, M. C., Bachman, S. P., Meagher, T. R., Nic Lughadha, E. & Brummitt, N. A. (2010), 'Subpopulations, locations and fragmentation: applying IUCN red list criteria to herbarium specimen data', *Biodiversity and Conservation* **19**(7), 2071–2085.

Rivers, M. C., Taylor, L., Brummitt, N. A., Meagher, T. R., Roberts, D. L. & Lughadha, E. N. (2011), 'How many herbarium specimens are needed to detect threatened species?', *Biological Conservation* **144**(10), 2541–2547.

Roberts, D. L., Taylor, L. & Joppa, L. N. (2016), 'Threatened or Data Deficient: assessing the conservation status of poorly known species', *Diversity and Distributions* **22**(5), 558–565.

Rowe, R. J. (2005), 'Elevational gradient analyses and the use of historical museum specimens: a cautionary tale', *Journal of Biogeography* **32**(11), 1883–1897.

Sandall, E. & Deans, A. (2018), 'Temporal differentiation in environmental niche modeling of Nearctic narrow-winged damselflies (Odonata: Coenagrionidae)', *PeerJ Preprints* pp. 0–14.

Särkinen, T., Iganci, J. R., Linares-Palomino, R., Simon, M. F. & Prado, D. E. (2011), 'Forgotten forests - issues and prospects in biome mapping using Seasonally Dry Tropical Forests as a case study', *BMC Ecology* **11**(1), 27.

Schmidt, M., Kreft, H., Thiombiano, A. & Zizka, G. (2005), 'Herbarium collections and field data-based plant diversity maps for Burkina Faso', *Diversity and Distributions* **11**(6), 509–516.

Sérgio, C., Garcia, C. A., Hespanhol, H., Vieira, C., Stow, S. & Long, D. (2012), 'Bryophyte diversity in the peneda-Gerês National Park (Portugal): Selecting important plant areas (IPA) based on a new survey and past records', *Botanica Complutensis* **36**, 39–50.

Sidlauskas, B. L. & Vari, R. P. (2012), 'Diversity and distribution of anostomoid fishes (Teleostei: Characiformes) throughout the Guianas', *Cybium* **36**(1), 71–103.

Snyder, J. L., Powell, G. S., Behring, R. S., Alford, A. M., Mccarty, M. E. & Zaspel, J. M. (2016), 'Distribution, Phenology, and Notes on the Life History of Calyptra canadensis (Bethune) (Erebidae: Calpinae)', *Journal of the Lepidopterists' Society* **70**(4), 253–259.

Soberón, J. M., Llorente, J. B. & Oñate, L. (2000), 'The use of specimen-label databases for conservation purposes: an example using Mexican Papilionid and Pierid butterflies', *Biodiversity and Conservation* **9**(10), 1441–1466.

Stein, B. R. & Wieczorek, J. R. (2004), 'Mammals of the World: MaNIS as an example of data integration in a distributed network environment', *Biodiversity Informatics* **1**(0).

Stigall, A. L., Bauer, J. E. & Brame, H. M. R. (2014), 'The digital Atlas of Ordovician life: Digitizing and mobilizing data for paleontologists and the public', *Estonian Journal of Earth Sciences* **63**(4), 312–316.

Stockwell, D. R., Beach, J. H., Stewart, A., Vorontsov, G., Vieglais, D. & Pereira, R. S. (2006), 'The use of the GARP genetic algorithm and Internet grid computing in the Lifemapper world atlas of species biodiversity', *Ecological Modelling* **195**(1-2), 139–145.

Syfert, M. M., Castañeda-Álvarez, N. P., Khoury, C. K., Särkinen, T., Sosa, C. C., Achicanoy, H. A., Bernau, V., Prohens, J., Daunay, M.-C. & Knapp, S. (2016), 'Crop wild relatives of the brinjal eggplant (Solanum melongena): Poorly represented in genebanks and many species at risk of extinction', *American Journal of Botany* **103**(4), 635–651.

Syfert, M. M., Serbina, L., Burckhardt, D., Knapp, S. & Percy, D. M. (2017), 'Emerging New Crop Pests: Ecological Modelling and Analysis of the South American Potato Psyllid Russelliana solanicola (Hemiptera: Psylloidea) and Its Wild Relatives', *PLOS ONE* **12**(1), e0167764.

Tobler, M., Honorio, E., Janovec, J. & Reynel, C. (2007), 'Implications of collection patterns of botanical specimens on their usefulness for conservation planning: an example of two neotropical plant families (Moraceae and Myristicaceae) in Peru', *Biodiversity and Conservation* **16**(3), 659–677.

Velásquez-Tibatá, J., Graham, C. H. & Munch, S. B. (2016), 'Using measurement error models to account for georeferencing error in species distribution models', *Ecography* **39**(3), 305–316.

Wehr, J. D., Stancheva, R., Truhn, K. & Sheath, R. G. (2013), 'Discovery of the Rare Freshwater Brown Alga Pleurocladia lacustris (Ectocarpales, Phaeophyceae) in California Streams', *Western North American Naturalist* **73**(2), 148–157.

Wieringa, J. J. & Sosef, M. S. (2011), 'The applicability of relative floristic resemblance to evaluate the conservation value of protected areas', *Plant Ecology and Evolution* **144**(3), 242–248.

Wilkin, P., Hladik, A., Weber, O., Marcel Hladik, C. & Jeannoda, V. (2009), 'Dioscorea orangeana (Dioscoreaceae), a new and threatened species of edible yam from northern Madagascar', *Kew Bulletin* **64**(3), 461–468.

Zeilinger, A. R., Rapacciuolo, G., Turek, D., Oboyski, P. T., Almeida, R. P. P. & Roderick, G. K. (2017), 'Museum specimen data reveal emergence of a plant disease may be linked to increases in the insect vector population', *Ecological Applications* **27**(6), 1827–1837.

Zhang, M.-G., Zhou, Z.-K., Chen, W.-Y., Cannon, C. H., Raes, N. & Slik, J. W. F. (2013), 'Major declines of woody plant species ranges under climate change in Yunnan, China', *Diversity and Distributions* **20**(4), 405–415.