**Cleaning the data**

1. Duplicate occurrences need to be removed

* scrubr: function dedup (need to give a tolerance level between 0 and 1, possibly we use tolerance level of 1 to exclude duplicates which are 100% the same?)

OR

* biogeo: function duplicatesexclude (need to give resolution as input?)

OR

* CoordinateCleaner: function cc\_dupl (seems to work the easiest) + function cc\_equ (to check for identical longitude/latitude) = extra safety

1. Remove impossible (fe. Longitudes higher than 180°), incomplete (rows only consisting of NA’s), unlikely coordinates (0°, indicators for errors and leading to high concentrations around equator or prime meridian)

* scrubr, all three of them

1. Centroid detection

* CoordinateCleaner has a function cc\_cap which flags occurrences near the capital of each country
* There is also the cc\_cen function which flags occurrences in vicinity of country and province centroids. Don’t know if this is better than cc\_cap. Maybe both can be used?

1. Occurrences with a resolution higher than 10 km are removed (This is a criteria from the paper Scott sent me but I don’t understand this one quite well or how to fix for this. I thought the occurrences where point data, so I don’t see how their resolution can be larger than 10 km.)

* Boolean operator on column of “coordinate uncertainty”, but what threshold? Larger than 1 km if we use the climatic data of Worldclim with 1km resolution?

1. Occurrences based on fossil material, germplasm and literature are removed (these ones are the occurrences located at the centre of gravity of each country if I understand correctly?)

* I have found the CleanCoordinates function for this one. It cleans the data automatically and retains the coordinates which are spatially valid. Don’t know if this function really is the relevant one to solve for this criteria.
* Maybe also through the “issue column”?

1. Occurrences located in sea/lakes need to be accounted for

* biogeo: function nearestcell (need input of specific dataframe and new raster)
* after this we have to do remove again duplicates, because it moves again (paper)

1. Originally global species occurrences will be included which will later be clipped to only occurrences within Europe
2. Missing environment checks
3. Environmental outlier

* Steps 8 and 9 don’t matter to much, because the model will still run with a high cut-off of 10

1. Geographical outlier

* CoordinateCleaner has a function cc\_outl which flags geographical outliers.

1. Potential botanical garden checks + hyper-anthropogenic environment

* Maybe this can be done by checking an issue column for words like garden or something in that way and then again using a logical expression on that column
* Coordinate cleaner also has a function for this one: cc\_inst (botanical gardens). It flags records assigned to the location of zoos, botanical gardens, herbaria, universities and museums, based on a global database of approximately 10,000 such biodiversity institutions. Coordinates from these locations can be related to data-entry errors, false automated geo-reference or individuals in captivity/horticulture.
* I think this can be incorporated in the botanical garden check by expanding search terms to for example “city” (but first see if it is possible with Boolean operator)
* CoordinateCleaner has a function cc\_urb (hyper-anthropogenic environment), which flags records inside urban areas. The question remains what is considered as an urban area for this function?

1. One occurrence per grid cell (thinning)
2. At species will only be included for analyses when at least 30 records are available for that species (with if-else relationship)