Questions for - Using Digitized Collections-Based Data in Research: Applications for Ecology, Phylogenetics, and Biogeography
Botany 2020

Please place your question under the relevant heading below; current page numbers for each heading (these are linked to the section) are given but note that these will change as more questions are added.

Workshop Material can be found here
https://github.com/soltislab/BotanyENMWorkshop2020

Workshop Leaders

- Pam Soltis: psoltis@flmnh.ufl.edu
- Doug Soltis: dsoltis@ufl.edu @soltislab
- Maria Cortez: mariacortez@ufl.edu
- Shelly Gaynor: michellegaynor@ufl.edu @ShellyGaynor
- Andre Naranjo: aanaranjo@ufl.edu @dicerandre
- Lauren Whitehurst: laurenwhitehurst@ufl.edu @igotflowerpower

Participant General Information

Add your **name** and **email** here so we can follow up:

- Cara Hastings carahastings@u.boisestate.edu
- jenny_xiang@ncsu.edu
- Joshua Der <jder@fullerton.edu>
- Sheila Lyons-Sobaski - slyons@albion.edu
- Nora Heaphy, nora.heaphy@yale.edu, @noraceae
- Lauren Gardiner, lmg32@cam.ac.uk, @CUHerb
- Sterling Herron, saherron615@gmail.com
- Chung Hyun Cho, cndgus56@gmail.com
- Megan R. King, megan.king@rutgers.edu
- Trista Crook, tristacrook@isu.edu
- Francisco Velasquez, francisco.velasquez_puentes@idiv.de
- Sophie Young, s.young5@lancaster.ac.uk
- Kira Lindelof klindel@ncsu.edu
- Jahnabi Gogoi jangogoilee@gmail.com
- Hyunkyung Oh, ohk92@korea.kr
- Chantal Bussiere, chantal.bussiere@maine.edu
- Chris Tyrrell: tyrrell@mpm.edu
- Andrea Berardi, andrea.berardi@ips.unibe.ch
- Miguel Perez lobelino@gmail.com
- Jing-Xia Liu j.liu@kew.org
- Emily Humphreys ehumphre@oberlin.edu
- Megan Van Etten mlv18@psu.edu
- Angélica Gallego Narbón angelica.gallego@uam.es
- Miao Sun cactusresponsible@gmail.com @Miao_the_Sun
- Ingrid Jordon-Thaden jordonthaden@wisc.edu
- Oluwaseun Akinyemi akinyemiseunfaith@gmail.com
- Ryan Esch - esch.ryla@gmail.com
- Lynn Wellman - Lynn_Wellman@fws.gov
- Carrie Kiel - ckiel@rsabg.org
- João de Deus Vidal Jr. - jd.vidal@unesp.br
- Ratidzayi (Rati)) Takawira-Nyenya - rnyenya@yahoo.com
- Roy Vera-Vélez - roy.vera@usask.ca
- Watchara Arthan - krich.scpl@gmail.com
- Megan Brown - mbrow241@kent.edu
- Christine Rose-Smyth - mcrscay@gmail.com
- Levent Can - levent.can@daad-alumni.de / lev.can@gmail.com
- Dina Clark - dina.clark@colorado.edu
-

General Questions

- I am currently using data from GBIF. I am trying to understand the main differences between GBIF and iDigBio. Is this something that we can address? - Cara Hastings
- I think this google doc file is also very informative. It would be great if this file is also available after the workshop. Thanks! (Miao)
- Ditto the above! - Vanessa Handley
    - iDigBio focuses on specimen records, while GBIF includes additional basis (observation, ect.). They serve slightly different purposes.
    - There are many data aggregators, here is a table I made to help differentiate: https://biodiversity-specimen-data.github.io/specimen-data-use-case/AggregatorsTable (Shelly)
- Is it possible to add my student Kira Lindelof to the workshop? She was late for registration and did not get in?
    - I can ask Johanne, will update you shortly. (Shelly)
    - Waiting for an update, but hoping your student gains access any min.
    - Thanks, and I suggested that Kira contact Johanne directly and say that we said it's fine for her to join.
    - Heather said Kira should have complete access now! (Shelly)
    - I'm in! Thank you all! (Kira)

- @Shelly, why are there so many data points in the ocean? Are they the aquatic species in salt water? Do the data include only plants?
    - iDigBio doesn't just include plants! Sometimes locality points may be incorrect, we will talk more about this later! (Shelly)
    - Most of those points are either fish/aquatic species, or poor locality data that place species without Lat/Lon data at 0 degrees N 0 Degrees W (Andre)
- How to get or know the taxonomic backbone used (e.g., accepted family names, order names, etc.? "Tracheophyta" recognized? Angiospermae/Magnoliophyta? Or is there not a specific backbone, and we will need to do multiple searches? Peter Fritsch
    - Currently there is not. I hear there may be plans to link the Catalog of Life in the future. (Shelly)
        - You can search multiple synonyms at once! (See example in R here) (also Shelly)
        - Thanks! PF
    - iDigBio has followed multiple 'community' taxonomic backbones, most recently EOL's. If those terms are part of a record, the specimens will be found, but who puts Plantae on an herbarium record? We have added Plantae to all plant specimen records so it is possible to download that way, but other taxonomic

levels have not been applied. As Shelly indicated, we will likely adopt COL+ for consistency with GBIF in the near future. (Pam)

- Does a program like IrisBG send the information to iDigBio for other's to use or is it a separate upload/resource? (Dasha Horton)
    - Asking iDigBio people now, will update (Shelly)
    - Thank you (DMH), I asked my herbarium curator and she said yes this collection management program does upload to GBIF, iDigBio, and some others. If you already asked, still happy to hear their answer too for verification.
    - iDigBio people let me know that you could export from IrisBG and send to iDigBio, but we are only harvesting data directly from collections rather than from other aggregators (like IrisBG, Symbiota portals, ect.) (Thanks to Erica Krimmel)
    - Thank you, good to know.
- Are locations for rare specimens/populations included in iDigBio or are they not included to protect the location of the population? (Sheila Lyons-Sobaski)
    - Usually not. I, for instance, work on several critically endangered species and their locality data is restricted due to security reasons, although images of the digitized specimens might be included in iDigBio. What I've ended up having to do is reaching out to herbaria that have vouchers for these species and state my need for those locality data. Most herbaria managers will be happy to help! (Andre)
    - I also made custom functions for my CURE class to identify these specimen and create csv for each herbarium (see more here: https://github.com/mgaynor1/CURE-FL-Plants) (Shelly)
        - So i've heard of this issue before where the locality data is protected but then the images are there, what's the point of protecting this then if others can still see the images): (Megan King)
            - That is a really good question, I think this is specific to a herbariums practice. UF blacks out a label if they share a rare species image, but not all herbariums do this. (Shelly)
    - Thank you! I was looking for sites of rare populations I study and I didn't see them (Sheila).

- How to deal with different accepted names in search (Dina Clark -U of Colorado)
    - iDigBio does not use a synonym database, you will need to search for each of the names associated with your species of interest. You can search multiple synonyms at once! (See example in R here) (Shelly)
        - Great resource Shelly, thank you (Megan King)
    - We formerly had the EOL synonymy functional, but it broke (without us being notified until recently). We were going to fix it, but because a new taxonomy is coming soon, it has been put on hold to accommodate the new taxonomy. (Pam)
- **General Question:** I don't know if this was mentioned already but will this recording be shared somewhere to be viewed again in the future. I think I would like to review this before beginning to work in R in the future (Megan King)

    - We hope so! We will do our best to get a copy to post on the iDigBio website. Or you may be able to access it as a registered Botany participant - we are not yet sure, but will find out! (Pam)

      Thanks Pam!
- **GENERAL R ISSUE:** I know many of you said you had low comfort with R, but I'll point out that Shelly had freshmen use R to do all of this during the spring semester, and they had no experience with R before they started. Shelly's materials at her github site will walk you through doing all of this in R. Once you're comfortable with R, you'll find that this is a lot easier - especially because all of the needed scripts can be used from the examples. Also, in the Demos folder, there are both R and non-R versions of how to do everything. (Pam)
- Can I join this workshop? I registered this workshop.
    - Can you not join the meeting? If not, please contact Johanne. (Shelly)
    - Yes, I didn't join the meeting. How to get in this workshop?
    - There should be a button that shows this, if not you need to contact Johanne at johanne@botany.org

      Ok, I'll contact Johanne.
      You should have access now (Shelly)
- I have registered for this workshop but did not receive the email for demo files. Please help me out (Jahnabi)
    - The link is above! **Workshop Material can be found: https://github.com/soltislab/BotanyENMWorkshop2020**
    - **Thank you-Jahnabi**
    -

Data Download

Please indicate your name at the end of your question.

- Can we get multiple families at the same time through rq-input in R?- Roy-UofSaskatchewan
    - Yes! If you look at the functions in this github, you will see how you could easily do that (https://github.com/mgaynor1/long-winded-scripts/tree/master/OccurrenceData)(Shelly)
    - Awesome! - Thanks - Roy

- How to maintain the DOIs of records from both iDigBio and GBIF during the process, and later cited in the publication? Thanks! (Miao Sun)
    - Download dois? If so, there is a download DOIs from iDigBio, but it is not maintained. For GBIF you can do an easy loop after to cite the databases properly (https://gist.github.com/mgaynor1/a1df02c28cdeb6e7cc6b74c952d13934) (Shelly)
    - Great! Thanks! - Miao
- Can you show how to save the data downloaded again-just see the slide one more time?-Jenny

Saving downloaded data

- save as a csv
- name including date

```
write.csv(spocc_LC_df,
"data/download/spocc_LC_df_050219.csv",
row.names = FALSE)
write.csv(iDigBio_LC,
"data/download/iDigBio_LC_050219.csv",
row.names = FALSE)
write.csv(iDigBio_LC_family,
"data/download/iDigBio_LC_family_050219.csv",
row.names = FALSE)
```

- I have registered for this workshop but did not receive the email for demo files. Please help me out (Jahnabi)
    - The link is above! **Workshop Material can be found https://github.com/soltislab/BotanyENMWorkshop2020**
    - **Thank you-Jahnabi**
    -
- What does the False mean-Jenny
    - We don't want a column with row numbers in the csv we write, so we make row.number = FALSE (Shelly)
    - Thanks Shelly!

Data Cleaning

Please indicate your name at the end of your question.

- Symbiota related - No set status (processing status) yielding no records, but says there are records… any information on this (not 2.0, specifically Mid-atlantic Herbaria - I know this might be an Ed Gibert question)? (Megan King)
    - I'm not entirely sure about Symbiota (2.0?) - do you know what databases it query? - I asked the experts and they agree it is an Ed question. (Shelly)
        - Thanks shelly!
- All packages work with any R version? - Roy
    - Hi Roy, I am not sure in combination but below are the individual packages with their version requirements, hopefully that helps? (LW) Agreed with Lauren, some of these packages need more recent R versions to run properly, rcurl especially (AN)
    - dplyr Depends: R (≥ 3.2.0)
    - tidyr R (≥ 3.1)
    - rjson R (≥ 3.1.0)
    - rcurl (**≥ 3.4.0),** methods
    - raster sp (≥ 1.4.1), R (≥ 3.5.0)
    - sp R (≥ 3.0.0), methods
    - This helps a lot! Thanks-Roy - No problem!!

- Are there coordinates for every botanical garden? Or do we have to find them? - Lev
    - Coordinate cleaner, ropensci - Shelly
    - Wikipedia has a great resource keeping track of herbaria worldwide https://osm4wiki.toolforge.org/cgi-bin/wiki/wiki-osm.pl?project=en&article=Category%3ALists_of_botanical_gardens_by_country&l=0
    - Cool, Thanks! - Lev
- Are there data points from Human observations in IdigBio? In GIBF, there are those records. They have a choice on website to filter data automatically. Do these functions use the Packages you were talking about?(Jenny)
    - iDigBio doesn't have human observation normally (it is designed that way!). But, for GBIF and iDigBio records, you can look at the basis column from the raw download (see https://www.tdwg.org/standards/dwc/ to learn more about the Darwin Core Format) . Using dplyr, you can filter the basis row (ex filter(basis != "Human Observation")) (Shelly)
    - Thanks Shelly! (Jenny)
- Does the location cleaner distinguish between cultivated gardens and natural areas in botanical gardens? (Nadia Cavallin)

    - Im not sure how specific this is, but the CoordinateCleaner package flags the points that fall within the gardens and you have to choose which to remove. You could always manually look up these points and determine if the point should be included or not (I often do this). (Shelly)
- Is the code for the ggplot (of the Florida Asclepias) you made in the R script? I am not seeing it and it was really useful! - Andrea
    - Yes! You can see this in the HTML file - CrashCourse_2020.html (Shelly)]
        - Ah I see it now, thank you!!! I was reading the R script -Andrea
- How do you clean duplicated data if there are any? (Wuu Kuang Soh)
    - Yes, dplyr has many ways to alter/clean the data and remove duplicate points/values - LW dplyr and tidyr cheat sheets
- Are there resources for what you mentioned about correcting for spatial sampling bias? (Wendy Van Drunen)
    - Anthony has a good collection of scripts (https://github.com/meltonae/ENM) related to spatial sampling bias. There are many papers on this in relation to ENM's and there is mix perspectives. I am not sure what the best practice is - I always reduce to 1 point per pixel, while Anthony always does that + thins the points (https://github.com/meltonae/ENM/blob/master/Functions/thin_max.R) (Shelly)
        - Thanks! - Wendy
- How did you make the HTML document with the R-code?
    - It is an R Markdown, it took a long time to knit. I suggest using R Notebook! Or Bookdown! (Shelly)

Georeferencing

Please indicate your name at the end of your question.

- Georeferencing is time consuming, is there a way to make sure the work we invest in georeferencing can be incorporated into the source databases? (Josh Der)
    - GBIF just added a new way to do this, but iDigBio does not have an automatic method at this point. You should contact the collection manager to have this added to the specimen! (Shelly)
    - Ideally, any annotations like this will go back to the provider, and the provider will supply it to iDigBio in regular updates. iDigBio does not alter the records but shares what has been provided from the institution (but we flag fields that appear to have errors). (Pam. Hi Josh!)(Hi Pam! ~Josh)
- Is it possible to obtain elevation data from coordinate points? (Caitlin Bumby)
    - Yes, we will show more examples of how to extract variables at each point. If you have a raster of elevation, this will be easy (Shelly) - Thank you!(Caitlin)
- There was a comment about discarding occurrence points with high uncertainty (larger than your climate data resolution). Is the uncertainty output when we download our data from iDigBio or GBIF? (Josh Der)

    - You do get a column with uncertainty (the darwin core website has specifics on this name (I also show which column this in a recent workshop https://github.com/mgaynor1/BCEENET-DataCleaning)) - this column is in the Occurrence_raw.csv (webportal download) file or the ~180 dataframe from the R package (Shelly)
    - But we don't get that column using ridigbio or spocc, right? How can we filter these out in an R workflow. Maybe this is not the right time to ask. (Josh)
        - You can get that column with certain downloads. Here is how with the idigbio R package - you cannot get this field with the spocc package, thus my long winded script example (https://github.com/mgaynor1/long-winded-scripts/blob/master/OccurrenceData) (Shelly)[Thanks!]

```
spocc_idigbio <- ridigbio::idig_search_records(rq = list(scientificname=synonyms_list),
                                               fields = idigbio_fields)
```

- Is there a batch-manner/programing based (e.g., R) geolocated workflow, other than doing it in a browser? (Miao)
    - There isn't any great R methods,Rob Guralnick from UF has some papers on this (Shelly)
    - Thanks!
- Can I link Geolocate to other map, e.g. old map? (Wuu Kuang Soh)
    - Hi! To my knowledge GeoLocate doesnt currently have any way of importing maps built on other platforms (ArcGIS, QGIS, etc), since it uses a Google base map. But if there was a way to extract the coordinates of the points found on your other maps then GeoLocate should be helpful. By old maps, do you mean physical paper maps, or files? (Andre)
    - Thanks, yes, old paper map or image of old map (Wuu Kuang Soh)
        - To georeference a past distribution map- you can use the Freehand Raster Georeferencer plugin in QGIS, add raster and then there is a 2 point georeferencing option and you drag the map to where it should be (Shelly). This tutorial may help too: https://www.qgistutorials.com/en/docs/georeferencing_basics.html

- Any thoughts on the difference between GoogleMaps and GoogleEarth? (Lauren)
    - Google maps requires you to put your credit card information in when trying to use via the API, so I avoid google maps/earth often. Andre, do you know? (Shelly)
    - Interesting! I heard it only does that over a certain number of records? (L)
        - Any new accounts require it! It is not very accessible (Shelly) Old accounts still work without, I believe!
        - Oh yuck! But old accounts still okay? (L)
        - Thanks! (L)
    - If you have to use Google, is Maps or Earth preferable / does one give you more data? (Sterling)
        - Hi Sterling, I prefer Google Earth because of the ability to place locality points in groups/generally more complete user interface.

        - Okay, thanks!

- I don't really know if the question has to do with georeferencing, but I have a problem with the coordinates. When uploading the coordinates to ArcGis, the points come out on one side and the region map on a different side do not overlap. What I can do? (Enmily)
    - You might be selecting the wrong columns for lat/long, maybe try flipping them? Or the wrong datum? (Shelly)
    - Yes, sometimes some programs want points in a long/lat column arrangement as opposed to a Lat/Lon. Best to try both ways. I know in QGIS there is a way for you to specify which column refers to the x (longitude) and y (latitude) axis. (Andre)
    - Okay. I will do so, thank you very much! (Enmily). perhaps the data format may be wrong?
        - Maybe, ArcGIS is picky! I use QGIS > ArcGIS - here is a great tutorial on QGIS: https://github.com/reptilerhett/GIS_Tutorial (Shelly)
        - Yes it is! I'll try QGis thank you Shelly (Enmily)

Climate Layer Processing

Please indicate your name at the end of your question.

- Do you know of any good data bases for global soil data?(Emily)
    - Yes SoilGrid - here is a python script to download the new 2.0 version (but just for North America, you would need to change the extent indicated) https://github.com/mgaynor1/long-winded-scripts/blob/master/SoilGridDownload/SoilGrids-PythonDownload.ipynb (Shelly)
    - Thank You!
    - Harmonized World Soil Database v 1.2
- Do you know whether there is data apart from climatic data available for other points in time (e.g. last glacial maximum)? I would be especially interested in vegetation data. (Mareike)
    - I am not sure if there is, if you find any we would love to know! (Shelly)
    - Hey there, yes there is! **EarthEnv is a great resource for land cover**… Was supposed to be in my slides but I guess it didn't save properly!
    - **http://www.earthenv.org**
- Is it possible to extrapolate data from soil databases (e.g., Soil Grid) to a particular georeferenced point(s) as we do with WorldClim? - Roy
    - If you are asking if we can point sample soil, the answer is Yes. We will show this method in Climatic Niche, instead of the bioclim raster you would use the soil raster (Shelly). - Yes - Thanks!
    - There are also other soil variables like soil geochemistry that USGS has that can be interpolated between the points, and they are at varying scales; I've worked with some USGS people to access the data and perform the interpolations. I'm not sure how much of this is currently public. (Pam)

- Need a bit of clarification, after I download the climate data of a certain grid resolution, I can't use it as it is? But I need to do something like apply convex hull to it? What is the purpose of Convex Hull?(Wuu Kuang Soh)
    - **Y**ou can change the resolution using the R raster package to downscsale
        - https://datacarpentry.org/r-raster-vector-geospatial/03-raster-reproject-in-r/
        - Here is how you change the resolution:

```
#aggregate from 40x40 resolution to 120x120 (factor = 3)
meuse.raster.aggregate <- aggregate(meuse.raster, fact=3)
res(meuse.raster.aggregate)
#[1] 120 120

#disaggregate from 40x40 resolution to 10x10 (factor = 4)
meuse.raster.disaggregate <- disaggregate(meuse.raster, fact=4)
res(meuse.raster.disaggregate)
#[1] 10 10
```

    - Hope this helps (Shelly) Thanks (Wuu Kuang)
    - The convex hull is essentially the model training region, off which Maxent will build the "ecological niche" of your species of interest with the climate/soil variables + the georeferenced points. From there, maxent can project that ecological niche on to any geographic area you want. (Andre)
- We have to re-run the correlation after removing variables? When you have two that are correlated, how do you choose which variable to remove? (Josh)
        - No, we just showed the excel sheet after the variables were removed. (Shelly)
        - You may want to choose variables relevant to your species, we also show how to randomly pick as well. (Shelly) This is when background information about your group of interest comes into play. Probably also best to include an even mix of temperature based and precipitation based variables because of some issues with overfitting Maxent if you have too much of one variable type. (Andre)
- How big should the buffer around occurrence points be? (Josh)
    - I think it depends on how unsure you are about your data sources (Maria). Are you talking specifically about Georeferencing? No, the convex hull for climate data.
    - I normally make my buffer width = 0.2 which a simulation study found optimal, but you should think about dispersal ability of your species and how well sampled your species may be (Shelly)

What is the source of the altitude data? USGS? I am using STRM data but am interested if you know another source that is good for the Neotropics, thanks- (Cara)

    - I believe so, but I need to confirm. I know GTOPO30 is a USGS database, but not sure if what we have in our scripts derive from this source (Maria). I used GTOPO30 for the Neotropics and it worked well.
    - http://www.earthenv.org/topography is another great resource for topographic data layers across a variety of resolutions.(Andre)
- Does SoilGrid and/or WorldClim have API connections for R or do they have to be downloaded? (Chris Tyrrell)
    - SoilGrids is not available via R at this time - But you can use python to get https://github.com/mgaynor1/long-winded-scripts/blob/master/SoilGridDownload/SoilGrids-PythonDownload.ipynb
    - The raster package R can be used to get bioclimatic variables (but maybe not the newest) - (Shelly)

Climatic Niche

Please indicate your name at the end of your question.

- MESS and Mahalanois distances are only for climate or for species data too? - Roy
  - Thanks, I'm not sure either, it's kind of difficult to decide which distance works better- Roy
    - These are looking only at climatic area, this is due to the whole training region being used to build the model via MaxEnt - so if you train in one area, is it okay to project into a different area (ie. will the model have enough information to be informative in the new area). (Shelly)
    - I know what a MESS analysis refers to in a completed ecological niche model… essentially shows you areas on your projection where the Maxent model extrapolated (not a good thing) so that you can subsequently interpret areas of suitable habitat that might have been incorrectly inferred as suitable.

- I am planning to project my model to past climatic conditions. I understand that the MESS analysis tells me how different climatic conditions are in two areas. Does this tell me whether projections are viable? So whether the projection is viable depends on how well the model performs in current conditions? (Mareike)
    - Mess will show you where the climates are really different and where you'd have to extrapolate for the projection. I dont know if theres a cutoff, but it would give an idea of where the predictions should be taken with a grain of salt.
    - I do not see this used for temporal projections, only for geographic projections. Instead temporal projections are evaluated based on model statistics which Maria will present about today (Shelly)
    - Thanks a lot for clarifying! (Mareike)

- How to find out the variations that count for the PC1 or PC2 of each species? - Weixuan
    - Loading values can be seen in the $rotation column, but you can not look at single groups PC within a multiple group PCA (Shelly)
- I think that my ENM is overestimating habitat, is it appropriate to use PCA to start to divide my specimens into climatic regions to better understand the conditions important to my focal species. I just want to make sure I am thinking about this correctly. We suspect that we have a cryptic species complex…), thanks! (Cara) Your answer makes sense, I'm just thinking about model refinement while waiting for molecular data! The morphology has been investigated. We think SEM of Pollen might help, but there are not any other informative characters we have been able to ID- maybe with the molecular data we can go back and look at morphology again.
    - My answer was that you should use more than environmental data to determine which specimen belong to which species. Morphology is a great thing to investigate before you have DNA (see the APPs special issue for specimen based AP methods to maybe help differentiate which specimen belong to which species) (Shelly)
    - I think trying to understand the biological factors that influence how you differentiate the taxa within your cryptic species complex is very important, besides distribution. It may help you think about all the environmental factor and how they contribute to the distribution (Maria)
    - I am at a place where I think I have my Bioclim data figured out, but when I look at my model, the habitat far exceeds occurrence, I think a lot of this is due to access (Amazon basin) but I started coding the points for different regions (very broad) and then I start to see real differences in the areas…
    - Cara, what source of molecular data are you using?
        - I just got sequences for 21 samples from the Angiosperm353 baits, I hope to get another 60 or so specimens in the mix including types and types of suspected synonyms.
        - I think you might have a better idea of what is going on once you get the molecular data back, given you said morphology is not useful in this context. What about pollination?
            - Bees on continental areas and flies on islands, I have not found great data for this. I am open to any suggestions! I did find some really great data for the bat genus that is the most common for distribution but its range exceeds that of my focal species so it did not add refinement…
            - Cara, we can talk more about this if you want. Your system looks very interesting and I have worked with cryptic species for my masters, but under a population and phylogeography approach (Maria)-
                - That would be great! Maybe I can drop in the sessions you guys mentioned at the beginning sometime in the near future. (Cara)

- Layers from different sources may have different resolutions - Thanks. Weixuan
    - You can change the resolution using the R raster package to downscsale
        - https://datacarpentry.org/r-raster-vector-geospatial/03-raster-reproject-in-r/
        - Here is how you change the resolution:

```
#aggregate from 40x40 resolution to 120x120 (factor = 3)
meuse.raster.aggregate <- aggregate(meuse.raster, fact=3)
res(meuse.raster.aggregate)
#[1] 120 120

#disaggregate from 40x40 resolution to 10x10 (factor = 4)
meuse.raster.disaggregate <- disaggregate(meuse.raster, fact=4)
res(meuse.raster.disaggregate)
#[1] 10 10
```

    - Hope this helps (Shelly)
- As you mentioned earlier that we can use paleoclimatic data, is it possible that we would predict ecological niche with paleoclimatic layers by using present occurrence data? (Watchara)
    - You will always train your models in the current conditions (current points + bioclim variables) to define the species niche (or infer fundamental niche) and then you can project to the future (aka how well does the current niche work in the future). I believe Pam will talk about this more after Lunch! (Shelly)
    - Yes, I was just going to reply. There are a couple of ways to deal with this. One is that you use the current conditions, as Shelly just mentioned, and you can project onto the future or past (or other geographic region, for example, if you want to consider whether a species might be an effective invader). The other thing you can do is reconstruct the ancestral niche and obtain distributions of variables at ancestral nodes, and you can project those nodal variables onto the past conditions. Andre did a version of this in his dissertation, which I will mention super-briefly, but he may want to comment further (but he is 'giving' his talk now and won't be back for awhile). (Pam)
    - Thank you very much for the suggestions! Could I have another question here even if we are not in modeling yet? (Watchara)'
    - Sure!
    - My interest paleoclimatic data is in Pleistocene which I think it's too old for the current climatic data there. So, are there any ways to do this? (Watchara)
        - There are layers available via Paleoclim! Ryan Folk and I are almost done with a project using layers from 0 - 3.3 mya, so it's possible and can be very informative - our preprint should be posted soon with interesting stuff (Shelly)
        - Yes, and the example that I will show has the Interglacial and Last Glacial Maximum. WorldClim also has data for the Holocene (but as Andre said,

PaleoClim has better data, and as Shelly said, new layers will be coming!).
        - Ah! I see. Thank you very much Shelly! I am new to this field and sometimes just do not know how to get some data from.
            - Totally understand! The PaleoClim website is fairly new: http://www.paleoclim.org/
            - Thanks a lot! I'll check it out and will check your preprint too!

- The QGIS in the tutorial requires version 2.18.23, why is that? And I understand that this version uses Python 2, will that affect my current Python version, which is version 3, if I install QGIS version 2.18.23? (Wuu Kuang Soh)
    - The tutorial we show is for 2.18.23, we haven't updated it to any newer version, but we probably should if you are having these issues! Thank you for sharing; we will make sure to look into this and try to update this at some point. We do link to the installation for this version in our word document. But, since Python2 is no longer maintained we totally understand and should update it! (Shelly) Thanks (Wuu Kuang Soh)
    - I haven't install this version of QGIS, I just read the installation instruction and realised that it uses Python 2. I currently use Python 3 in my computer. (Wuu Kuang Soh)
- A general question about downloading/using idigbio data, is the data in idigbio complete, say for example for plants at a particular county in US, should I also download data from herbaria for that region, say via symbiota portal?
    - I normally download from iDigBio, GBIF, and BISON (USGS) - I have found that these databases differ in records. The tweet on this page and table may be helpful. https://biodiversity-specimen-data.github.io/specimen-data-use-case/AggregatorsTable (Shelly)
        - I also do not know the specific databases Symbiota now has, but it may be the similar or completely overlapping with GBIF and/or iDigBio (also Shelly) Thanks (Wuu Kuang Soh)
        - Symbiota is a data management system that can be aggregated among institutions. Symbiota systems generally contribute their data to both iDigBio and GBIF; their databases are their own local management systems rather than something to be searched (although this is possible through some Symbiota portals), but the general workflow is for Symbiota systems to contribute to larger aggregators. (Pam) Thanks! (Wuu Kuang Soh)

Ecological Niche Modeling

Please indicate your name at the end of your question.

- Which presentation slide is this, in the presentation folder, Maxent? (wuu Kuang Soh)
- All presentations are in the folder (Shelly) We are currently on #8. Ah, Thanks (Wuu Kuang Soh)
- What is the difference between logistic and cloglog output formats in Maxent (Lydia Soifer)
    - Logistic is 0-1 and is the exponential function of the environmental variables, while cloglog is 1 minus the exponential of the negative of the cumulative threshold times the raw percentage. More information here: https://biodiversityinformatics.amnh.org/open_source/maxent/Maxent_tutorial2017.pdf (Shelly)
- Could you elaborate a bit more on what the clamp grid does? (Mareike)
    Maxent extrapolates points and clamping allows variables to be limited in the training region (summarizing Andre)
- Instead of choosing the test samples randomly, can you ask for the software to target the specific samples to test for the model? (Weixuan)
    - Probably? But you should compare to a random sample - see more about maxent in this great tutorial https://biodiversityinformatics.amnh.org/open_source/maxent/Maxent_tutorial2017.pdf (Shelly)
    - For example, if 20% of the data is field collected data - very precise, and 80% are specimens. I wish to use the 20% to build the model and 80% to test it. (Weixuan)
        - You can not do this with MaxEnt GUI, but you could probably assign these to be background points in R - Anthony Melton has scripts you could modify to do this. I dont think this is a good thing to do, since the random points are testing your models fit. (Shelly)
        -
- Are you familiar with the Dismo package? I know it is fundamentally different from MaxEnt, just curious about your thoughts
    - Dismo requires rJava, and is included in our R demo - but we will not walk through this because of the rJava requirement (Shelly)-thanks!
    - Are you thinking about MaxNet? Which is different from MaxEnt, but Dismo can do both! (Shelly)

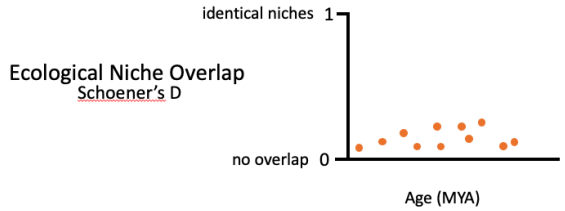
Interpreting ENM Results + Post ENM

Please indicate your name at the end of your question.

- What does it mean variables have to be in the same dimension? It means perhaps standardized - transformed? - Roy

    - It means that the projection layers all have to be the same size and resolution (cropped to the same area). This is why we run Maxent species by species, so that we can use one specific set of layers - Got it! Thanks - Roy
- So, plant species can be modelled using Niche models, what about plant community in general, can we model changes in plant composition? - Roy
    - If you overlay models for each species and if you predict models across time you can have an idea of how a community changed within a given time frame. Does that make sense? Yes- that makes sense, I can try that - Thanks. Roy. Sure! Biotaphy should be very helpful for that as well!
- Is it possible to use ENM to study niche on a hill, say on one limestone karst, looking at microhabitat? Let say if I measure the microclimate environment and species distribution on a hill. Do you know of any papers that study microhabitat in a very restricted area like one hill etc. using ENM? Thanks everyone (Wuu Kuang Soh)
    - Might be helpful to look into papers that use Ecospat R package and maybe Humboldt R package
        - Off the top of my head I don't know any papers that use ENM at a microhabitat level, but you can investigate Climatic Niche (rather than model ecological niche). (Shelly)
- Are there any "rules of thumb" for the suitability cutoff of a binary map? For example, if using this for a species distribution map in a monograph, or is it more common to present the probabilistic/gradient map? (Chris Tyrrell)
    - Predicted Niche Occupancy might be a better way to look at this than binary - it weights the suitability and shows the spread of the data. Niche space might be more informative to a species to look at the spread/weight for each bioclimatic variable. PNOs can be calculated via the phyloclim R package, as well as through other methods (like Ryan Folks github repository https://github.com/ryanafolk/pno_calc) (Shelly)
- Sorry could you please one more time explain/interpret the results of your last three slides for the phylogenetic correlation (the summary of the glm.aoc object) thank you!
    - I put some slides explaining what this is looking out. So at each node, you are looking at the overlap - or Ecological Niche Overlap - these are high are the nodes. Overall the slope is showing sympatric speciation (see the slide below). You can also look at multiple types of overlap (range vs points) to infer speciation modes more specifically. I use this method to look at Diapensiaceae in a recent publication (Gaynor et al. 2020 JSE) + my Botany talk if you want to read/watch this being applied (Shelly) Thank you!
    - So the other example before that (object range.aoc) shows allopatric?
    - This is only 4 nodes with 10 reps, it is not the best example. Look at the slope/intercept plots - the points are highly dispersed on the plot! So, take the conclusions with a grain of salt in this example. We include both based on the inferences you could make compare the two metrics (see the Cardillo and Warren paper) (Shelly)

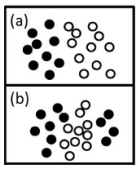
Age-overlap correlation test



Mode of Speciation	Intercept	Slope
Sympatric	> 0.5	
Allopatric	< 0.5	positive
Parapatric	< 0.5	near or below 0

Fitzpatrick & Turelli. 2006. Evolution.

Age-overlap correlation test



Spatial overlap measure			Example of possible interpretations of overlap pattern
Range Overlap	Point Overlap	Local co-occurrence	
low	low	low	Allopatric speciation with broad geographic barrier as isolating mechanism
high	low	low	Allopatric speciation with finer-scale landscape features as isolating mechanism

Cardillo and Warren. 2016. Global Ecology and Biogeography.

-

Phylogenetic Diversity

Please indicate your name at the end of your question.

- Say you have a phylogeny of all species in large monophyletic genus where species span a wide variety of habitat types - some species are weedy while others may occur in particular ecoregions. Is it possible to examine PD for that particular group or do you need full community data as you described in your lecture? Carrie

    - Phylogenetic diversity studies generally focus on a specific region, rather than on a clade. That said, there are sometimes studies of PD for members of a clade in a specific region (for example, the early work of Jeannine Cavender-Bares on oaks of Florida). In this sort of study, you might use PD to infer properties of community assembly - and the roles of filtering and competition. These are really different uses than we've described but may also be done with the same data types and methods. Does that make sense? (Pam)
    - It does - thank you - this has been a fantastic overview. Looking forward to applying these methods (Carrie)

BiotaPhy

Please indicate your name at the end of your question.

- I have some morphology data that makes me suspect ecological differentiation may be happening between related species in a region. I'm wondering if PD and/or BiotaPhy could be integrated with these data to understand the evolutionary picture of these species (Laura)
    - It can be very informative to combine trees, occurrences, models, and traits (see the recent paper by R. Folk et al. on Saxifragales diversification (PNAS 2019); in that paper we showed that rates of evolutionary change differed for species diversification, niche divergence, and trait divergence. Interestingly, species diversification occurred prior to shifts in niche and morphology. I think you may have the types of data to do the same type of comparison! (Pam)
    - Thanks!

End Comments

- I just want to thank you for such an excellent workshop! WOW! (Dina)
- Great workshop ever!! Thanks! (Miao)
- Thank you so much! You gave me a lot to consider moving forward in my research. (Megan)
- Thank you all very much! Really appreciated the workshop, even a virtual one. (Sheila L-S)
- Thanks a million everyone! From Dublin, Ireland (Wuu Kuang Soh)
- This was great! Thank you so much, so many ideas and future thoughts! Greetings from Germany Levent
- I will second that. Superb workshop! (Mareike)
- Thank you so much for this great workshop!!!! It was such a great introduction. (Andrea B)
- Thank you so much! I think that Shelly mentioned some weekly working group meetings- can you provide a little more info on that? Thanks!!!

    - Biweekly working group for DigBio data API: https://www.idigbio.org/content/open-office-hours-hosted-api-user-group-r-based
- Thank you for such an amazing session!! I learned a lot in just a few hours! Great work!! (Roy)
- Thank you very much! I learned a lot today (Taylor)
- Thank you very much! This is super useful. I got a lot of ideas from this workshop for my PhD project. (Watchara)
- Thank you for giving us a great opportunity to participate in a great workshop. I really enjoyed it!!! (Chung Hyun)
- Thank you for a great workshop. I would love to see this work extended to the Caribbean. Christine.
- I can't thank you enough for organising this workshop. Shout out to the whole team!It was excellent. Thank you so so much! (Jahnabi)
- Fantastic workshop, thank you so much! (Chris Tyrrell)