Questions for - Using Digitized Collections-Based Data in Research:
Applications for Ecology, Phylogenetics, and Biogeography
Botany 2021

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.

Github: https://github.com/soltislab/BotanyENMWorkshops
HTML version of our R markdown can also be viewed here:
https://soltislab.github.io/BotanyENMWorkshops/Demos/Rbased/CrashCourse_2021.html

Workshop Schedule

9:00	Welcome and Overview of the Workshop – Pam
9:10	iDigBio Portal – Pam
9:20	Data Downloads – Lauren
9:35	Data Cleaning – Shelly
10:15	Outlier Detection – Tal
10:20	Georeferencing – Andre
10:50	Break
11:05	Climate Processing – Shelly
11:55	Climatic Niche – Shelly
12:15	Lunch
12:40	*Question Session*
1:00	Applications of ENMs – Doug
1:10	Ecological Niche Models – Shelly
1:40	Interpreting ENM Results – Shelly
2:00	Post-ENM analysis – Shelly
2:30	*Break/Question Session*
2:45	Intro to BiotaPhy – Doug
2:55	Overview of Phylogenetic Diversity (PD) – Hannah
3:10	Hands-on PD Demo – Maria
3:30	Overview of Alpine Biodiversity Project – Hector
3:40	Overview of OCBIL (Old, Climatically Buffered, Infertile Landscapes) Project – Maria
3:50	Correlating Chromosome and Climate Evolution – Jon
4:00	Computing with Heterogeneous Species Occurrence Data for Global Analyses – CJ
4:40	*Question Session*
5:00	End

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
- Tal Kinser: tkinser@ufl.edu
- Makenzie Mabry: mmabry44@gmail.com
- Hector Figueroa: hecfox@umich.edu

- Hannah Marx: marxh@umich.edu
- Jon Spoelhof: spoelhof.jon@ufl.edu
- Ryan Folk: rfolk@biology.msstate.edu
- CJ Grady: cjgrady@ku.edu

Participant General Information

Add your **name** and **email** here so we can follow up:
- Tanja M. Schuster: tanja.schuster@nhm-wien.ac.at
- Kurt Neubig: kneubig@siu.edu
- Nourin Aman: nourinaman@gmail.com
- Ayress Grinage: adg223@cornell.edu
- Andrii Tarieiev andrii.tarieiev@forst.uni-goettingen.de
- Hannah Stanford hannahstanford@knights.ucf.edu
- Ludwig Baldaszti baldasztil@gmail.com
- Bashir Bolaji Tiamiyu: tiamiyu.bb@unilorin.edu.ng
- Elizabeth White elizabethwhite1@ufl.edu
- Prasanna NS: prasanna16@iiserb.ac.in
- Todd Osmundson: tosmundson@uwlax.edu
- Kanthraj A.S.: kan7th@gmail.com
- Anne Kasperski: annekasperski@hotmail.com
- Carrie Kiel ckiel@calbg.org
- Millicent Oulo: millicentoulo@gmail.com
- Vinita Gowda: gowdav@iiserb.ac.in
- Fernanda Schmidt Silveira:okologie_natur@hotmail.com
- Kailin Sun: sun.kailin@campus.lmu.de
- Claudia Ciotir: claudiaciotir2@trentu.ca
- Lisa Boucher: lisadboucher@austin.utexas.edu
- Gauri Shankar Bhandari bhandarigs@snu.ac.kr
- Karthikeyan karthikeyan.sbsmku@gmail.com
- Laura Calvillo Canadell laura.calvillo@ib.unam.mx
- Jim Cohen, jamescohen@weber.edu

iDigBio

Data Download

- I don't see how to run CrashCourse, though. It popped open a window but doesn't seem to have a console into which to type commands or any buttons on a GUI.
    - Do you have R studio installed? Did you open the .Rproj to open the project? I put a picture below to help. I opened the .Rproj file and then opened the first two scripts. -Shelly
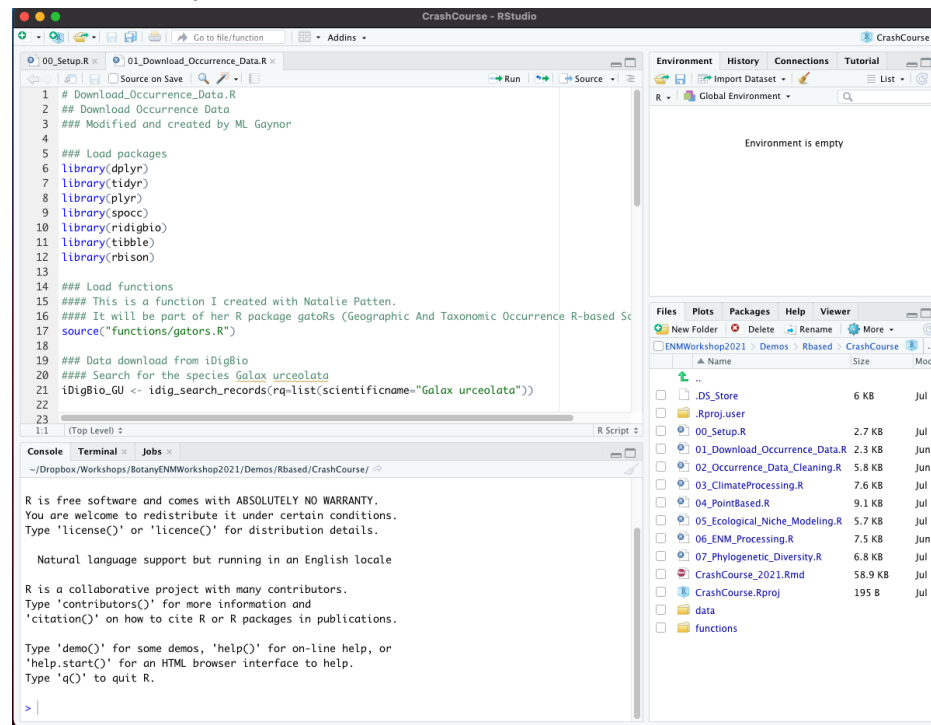
    -
    - Question: I opened the first r scripts first instead of opening the Rproject can I still open the R project even though I already started an unnamed session?
    - No, I would close and reopen using the .Rproj. Luckily we have not loaded any needed objects just yet. And packages installed will stay. (hope this helps and makes sense) - Shelly

Data Cleaning

- Intro to what rasters are if you are like me and never heard of them before!: https://datacarpentry.org/organization-geospatial/01-intro-raster-data/
    - We will go over them more today too! -Shelly
- DOI for manuscript explaining more on spThin: https://doi.org/10.1111/ecog.01132

- - I also have all the papers cited in the powerpoints listed on our github: https://github.com/soltislab/BotanyENMWorkshops -Shelly

- **Why do you remove outliers? Won't that make your data incorrect at the end?**
  - In some cases, outliers are removed/detected because they may not contribute to the "true" range, it could be a far-placed exotic, or in a botanical garden, or ornamental, or an error! Also, this could over-fit (correct?) the model if you include points from a larger area of occurrence than what is actually true to your species...if that makes sense!! -LW
  - To account for sampling bias: See Daru et al. 2017. New Phytologist. for more background. -Shelly

Georeferencing

- Did I hear correctly that we can still do georeferencing without coordinates? I am thinking about old herbarium specimens that only have descriptions.
  - Yes! Andre will go over this and our instructions are in "Demos/Manual/Georeferencing". This is great for older specimens (with only descriptions of locality)! -Shelly
-
- How would georeferencing work with newer maps that do not show older specimen locality information?
  - He will show that too! There are a lot of tools you can use to figure out the updated terms for the locality! - SG
- If you georeferenced your coordinates for a specimen (s), how can you show that these coordinates were obtained through georeference methods in index herbariorum?
  - Great question. So you would need to contact the herbarium managers to get this added to the databases. GBIF does have a feature where you can add the data to the specimen. However, this is a major question/concern in the collection management/digitized natural history specimen world! There isn't a great way to do this currently, but the best way currently is contacting herbarium managers - Shelly (Pam may have more insight!)
  - Good question and good answer! How to apply annotations to specimen records and who should do it is definitely an issue in flux! Most collection managers do not want the shared record changed by anyone other than themselves. In iDigBio, annotations, such as georeferences with uncertainties or suggested changes to an identification, are to be sent to the manager of the collection that houses the specimen, and there is contact info for doing that. If the manager chooses to accept the annotation (which will typically happen for georeferences but may not always happen for suggested name changes), the manager will update the record and then re-share the records with iDigBio, GBIF, etc. As Shelly noted, GBIF provides some annotation tools, as do some other

aggregators, but the collection community in the US generally does not want specimen records in iDigBio to be annotated directly by the community (and this has driven iDigBio's response).

- How useful will it be for the downstream analysis to georeference with a pretty wide uncertainty? Yes, thank you!
    - That's a good question! If you already have a decent set of points, then another record with a wide uncertainty may not be helpful and could even be detrimental. But if you don't have very many specimen records, it might be worth it to have some, even if there is wide uncertainty. The important thing is to have a record of the uncertainty associated with the coordinates. Does that help? - Pam
    - The rule of thumb is that you do not want your uncertainty to be more than the precision needed. So not more than 1000 meters when rounding to 2 decimals and using 30 arc-sec layers. But, like Pam said, sometimes the point may still be helpful. - SG
    - Yes, good point, Shelly! - PS

Climate Layer Processing

- A fun video that helps explain (at least for me) why we are converting the dataframe to a spatial dataframe : https://www.youtube.com/watch?v=kIID5FDi2JQ
- I'm having errors visualizing the plots after defining the hull:
> plot(hull[[1]], col=transparentColor('gray50', 0.5), border = NA) Error in plot.new() : figure margins too large
> points(x = alldf$long, y = alldf$lat, cex = 0.5, pch = 3) Error in plot.xy(xy.coords(x, y), type = type, ...) : plot.new has not been called yet
    - From Carrie Kiel: Expand your window for the plot if using RStudio

- What is the difference between your projections and species layers?
    - The layers are traditionally projected at WGS (degrees). When we make our species coordinates, we match the layers. BUT, when we want to know distances in meters we convert both species and climatic layers to cea (meters). - SG
- "Do you need to choose projection based on the location of the samples? like if we are doing data from Arctica?"
    - It is recommended, but not necessary. As I mentioned earlier, certain projections are focused/centered on particular regions. Most datums commonly used in ENM development are equidistant projections that are not focused on one particular region. I think the only issues pertaining to geographic-centric projections will come from georeferencing (as a source of uncertainty) and when visualizing niche model results. EDIT: It seems there's some issue with datums elsewhere: "Deriving topographical variables from geographical coordinate systems is sometimes difficult to track if the methodology is not described sufficiently. Articles frequently do not specify the coordinate reference system, with ellipsoid

and datum, of their spatial data, and one can suspect that if authors are using the altitude map included in the WorldClim dataset, topographical variables were also obtained directly from the geographical unprojected map." from Sillero & Barbosa 2021 -Andre

●
● Did anyone else get a strange looking equivalency plot? My values were different

Equivalency
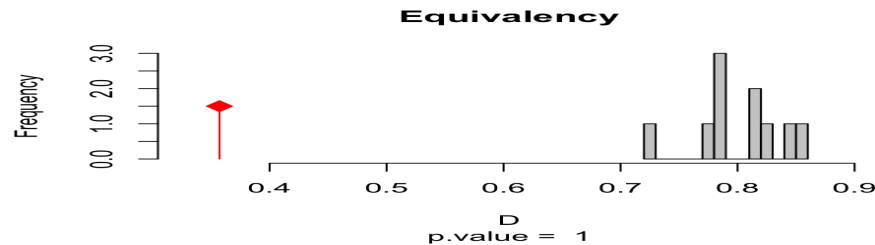
than demo
  ○ What species did you compare? That is weird! Can you show me the code you ran?-SG
    ■ ## Niche Overlap
    ■ ### Schoener's D ranges from 0 to 1
    ■ ### 0 represents no similarity between niche space
    ■ ### 1 represents completely identical niche space
    ■ overlapD <- matrix(ncol = 2, nrow = 7)
    ■ n <- 1
    ■ for(i in 1:3){
    ■   for(j in 2:4){
    ■   if(i != j){
    ■   overlapD[n, 1]<- paste0("z", i, "-", "z", j)
    ■   overlapD[n, 2]<- ecospat.niche.overlap(zlist[[i]], zlist[[j]], cor = TRUE)$D
    ■   n <- n + 1
    ■   }
    ■   }
    ■ }
    ■
    ■ overlapDdf <- data.frame(overlapD)
    ■ overlapDdf
    ■
    ■ ## Niche Overlap Visualization # 1 and 4 have the highest percent overlap
    ■ par(mfrow=c(2,1))
    ■ ecospat.plot.niche.dyn(z1, z4, quant=0.25, interest = 1

- - - ■ , title= "Niche Overlap - Z1 top", name.axis1="PC1", name.axis2="PC2")
      - ■ ecospat.plot.niche.dyn(z1, z4, quant=0.25, interest = 2
      - ■ , title= "Niche Overlap - Z4 top", name.axis1="PC1", name.axis2="PC2")
      - ■
      - ■ ## Niche Equivalency Test
      - ■ ### Based on Warren et al. 2008 - Are the two niche identical?
      - ■ #### Hypothesis test for D, null based on randomization
      - ■ #### H1: the niche overlap is higher than expected by chance (or when randomized)
      - ■ eq.test <- ecospat.niche.equivalency.test(z1, z4, rep = 10, alternative = "greater")
      - ■ ecospat.plot.overlap.test(eq.test, "D", "Equivalency")
    - ○
    - ○ I wonder if the lack of reps did this. Your D value shouldnt be negative though, hmmmmmm. - SG
    - ○ I increase the reps to 50 and got roughly the same answer. The similarity test was the same as your demo value however.
- Okay - I think I know the answer. I think it has to do with the random background points being different between you and me. Since this creates a null based on background points (I believe). Try to generate background points again and see what happens. Your plot will probably differ from your last if the background points differ. Both our plots didn't have significant p values, so though they looked different, the interpretation is the same. - SG

- For subsetting training data, is it the package that does that randomly or you need to enter a specific range.
  - ○ So here we were subsetting the data frame for each species, but we aren't partitioning the dataset (if this makes sense). BUT, we will show how to data partition after the break (like Maria said). If I am totally misunderstanding your question, please let me know! - SG

Climatic Niche

- Within the ecospat package, is there also a function that lets you calculate climatic niche breadth?
  - ○ Maybe…I'm not sure. A quick google says maybe, but I cannot find which function you would use. - SG

- - ○ If I understand correctly, I think you can use ecospat. For example page 15 and 16 in the Vignette below (figure 2.5.6.2)
      - ■ https://cran.r-project.org/web/packages/ecospat/vignettes/vignette_ecospat_package.pdf
  - Can we record this segment of the workshop?
    - ○ We are recording it. We forgot to pause for lunch! - SG

Ecological Niche Modeling/Interpreting ENM Results/Post-ENM Analysis

- Any advantage over the GUI vs R for the program?
  - ○ Using ENMeval allows data partition + variable features & regularization multipliers. This allows model comparisons. For models in dismo or the GUI, they are equivalent and it is a user preference. I personally used to run models on the command line, it was easier to set up (no checking of boxes, just text!). - SG
  - ○ ENMeval also allows for model testing using several different performance statistics
-
-

```
>
> library(hypervolume)
Loading required package: rgl
Registered S3 methods overwritten by 'rgl':
  method            from
  knit_print.rglId
  knit_print.rglOpen3d
Error in dyn.load(dynlib <- getDynlib(dir)) :
  unable to load shared object '/Library/Frameworks/R.framework/Versions/4.1-arm64/Resources/library/rgl/libs/rgl.so':
  dlopen(/Library/Frameworks/R.framework/Versions/4.1-arm64/Resources/library/rgl/libs/rgl.so, 6): Symbol not found: _hb_ft_font_create
  Referenced from: /Library/Frameworks/R.framework/Versions/4.1-arm64/Resources/library/rgl/libs/rgl.so
  Expected in: flat namespace
 in /Library/Frameworks/R.framework/Versions/4.1-arm64/Resources/library/rgl/libs/rgl.so
Error: package or namespace load failed for 'rgl':
 .onLoad failed in loadNamespace() for 'rgl', details:
  call: rgl.init(initValue, onlyNULL)
  error: OpenGL is not available in this build
Error: package 'rgl' could not be loaded
In addition: Warning messages:
1:     Loading rgl's DLL failed.
       This build of rgl depends on XQuartz, which failed to load.
 See the discussion in https://stackoverflow.com/a/66127391/2554330
2: Trying without OpenGL...
> |
```

-
-
    - ○ xQuartz: https://www.xquartz.org/

- - The error message links the solution!: https://stackoverflow.com/questions/66011929/package-rgl-in-r-not-loading-in-mac-os/66127391#66127391
      - Basically uninstall xQuartz, remove rgl, remove two files listed on the stack overflow, reboot your system, reinstall XQuartz 2.7.11, and then reinstall rgl.
- There is an error with phangorn as well
  - `The work around may be remotes::install_github("KlausVigo/phangorn")`
  - I cannot find many others with this error, and since my computer didnt, I can only guess what the workaround will be. This is my suggestion, but it may not be correct. - SG

Biotaphy

Phylogenetic Diversity

- Phytools wasn't necessary to have installed in this script (Comment in chat from Andrew Simpson)

Alpine Biodiversity Project

-
-

OCBIL Project

-
-

Correlating Chromosome and Climate Evolution

- Gymnosperms are (with some exceptions) a declining group. Assumptions of the lack of changes in evolutionary rate, especially speciation and extinction models, cause severe problems with inferring past evolutionary rates. Is this a concern for rate shift correlations with polyploidy in gymnosperms? - Andy Simpson

- Good point. None of the models I've used incorporate speciation and extinction explicitly, so those variables are less of an issue, but it's still important to take that concern into account and make conservative conclusions with clades like this. - Jon

Computing with Heterogeneous Species Occurrence Data for Global Analyses

-
-