

BIEN Range Modeling Meeting Notes

7 June 2012
1:00 PM PST

Attendees

Martha Narro, Nirav Merchant, Edwin Skidmore, John Donoghue
Jim Regetz and Mark Schildhauer via Phone

Notes From the Meeting Organized by Discussion Topic

The following notes summarize the discussion points for the meeting. The handouts are included in subsequent pages. Items in Red on the handouts indicate changes or additions that were made following our discussion.

1. Current range modeling status

- The group discussed the need to document models and model development in a BIEN wiki, likely in the NCEAS' Redmine site unless it's determined that iPlant would be more suitable location.
 - The documentation would focus on how the model's work and how to set up the models to run on TAC's resources so that others wanting to run the models could.
 - Jim posted a link to the NCEAS Redmine site at:
<https://projects.nceas.ucsb.edu/nceas/projects/bien>
- It was also suggested that we set up a repository for the R modeling scripts
 - iPlant uses GitHub (for projects approved to be open source)
 - <https://github.com/iPlantCollaborativeOpenSource>
 - They also use an internal Git server for development projects
 - Jim mentioned that NCEAS uses an internal Git server for E & O scripts that is accessible via the group's Redmine page
 - Jim will set up a repository location where John can store scripts
 - Edwin (after the meeting) asked whether the scripts will be open source at the end of the project. If so, there is some question about what type of licensing they will have: NCEAS or iPlant.
 - Martha will look into this.

2. Proposed range mapping deliverables

- The list needs to be updated to reflect the single-point and two-point area calculation methods that were devised more recently.
 - John will develop a final list of model deliverables.

3. Range modeling decision tree

- John briefly discussed the decision tree and focused on the fact that there will be two new scripts in the final model run. The two new scripts will perform area calculations for species with only single-point and two-point occurrences.
 - John will get new extracts of species occurrences and divide them into jobs
 - 1-Point
 - 2-Point
 - 3-4Point
 - 5-Point
 - John will finalize the scripts to get them ready for running on the cluster
 - He will be meeting with Edwin to discuss the scripts when they are ready

- The Maxent modeling scripts will need to parameterize the Maxent temp directory so we can store the Maxent model temp files for other analyses
- John will work with Edwin to write a scripts to mine the temp directories for the MaxentResults.csv files that Jeff Ott has requested.

4. Script changes and need for retesting new scripts in the cluster

- John discussed some R package updates and additions that are needed to the R environment on the cluster.
 - He will get Edwin a list of R package updates that need to be made on the cluster
 - R version
 - Package versions
- Nirav suggested that iPlant store the R binaries as part of the run package so that the R version and R package versions are always stable

5. Raster to vector conversion issues

- John discussed the problem converting the binary raster range map to a polygon shapefile.
- Jim asked John to send him an example of a raster that can be converted and an example of a problem raster.
 - Jim will look at using gdal directly or using some other methods to perform the conversion

6. Scheduling a model run in the cluster

- Nirav said that he wanted Edwin to work with John (with Jim and Mark following) to teach him how to set up a model run in the cluster so that the BIEN team could run the models as needed and Jim and Mark could learn how to leverage iPlant and TACC resources for other projects.
 - Thus, the next model run will be handled this way.

7. Need to archive Maxent results that are stored in temp path

- John discussed the recent need to save the Maxent model temporary files by species so that the files can be mined for additional information on variable importance and other subsequent analyses.
 - John will estimate the size of the Maxent temp files so we can determine how much space is needed.
 - John will also describe what needs to be saved in the Maxent temp files
 - MaxentResults.csv

8. Scheduling subsequent analysis runs in the cluster

- John mentioned that the BIEN team has already identified some subsequent analyses projects that would require TACC resources after the initial modeling work is done:
 - Range specific climatic attributes (mean temp, precip, variance, etc)
 - Maxent variable importance (this is why we need 5)

9. Range mapping API

- John opened up a short discussion on the range mapping API to get input from the attendees on what was being considered.
- Martha said the API was in response to a long term vision for how BIEN would serve range maps up to the community and how users would access BIEN resources

- She said it would be a web service API centered around how to make data readily accessible to users
- We agreed to table this for a little while, but John tossed out several ideas:
 - Submit a species name – obtain a range map
 - Submit a species name – obtain environmental variables for the species geographic range
 - Submit a location – obtain information on the species that could be there from range maps
 - Perhaps with the logistic probability of suitable habitat (from Maxent raster layers)
- Edwin discussed the eventual idea of performing range modeling on-the-fly
 - He asked whether we planned to expose our R scripts in such an API for modeling on-the-fly
 - John said wasn't sure but could see the need to eventually re-write the modeling scripts (in C or some other language) so they could work in an API environment
 - Mark and Jim suggested looking into Open Modeler framework because it was designed to be a framework in which different modeling algorithms could be written for it.
 - It also worked in an API and server-based environment
 - Jim provided the following link to OpenModeler:
http://openmodeller.sourceforge.net/index.php?option=com_content&task=view&id=2&Itemid=3
 - We briefly discussed the idea of eventually re-write the modeling algorithm in Open Modeler so that it could be made available to the community at large in this kind of widely used framework
 - Martha suggested that we might want to touch base with Rob Guaralick on the status of Open Modeler to see whether it might be a good framework for eventual sharing of our modeling algorithm.
- Martha said that with regards to the API, there's some overlap with the E & O needs, so we should coordinate with them as well.

Proposed Species Range Modeling Products

As of 28 October 2011

This page outlines the species-specific products that we propose to generate and retain from the range modeling effort. This list does not include additional aggregate products, such as diversity maps which will also be generated and retained. The list below only pertains to products specifically generated during the massive species-by-species range modeling effort. Comments from BIEN members are welcome and encouraged so we can ensure that we retain and generate the species range products needed by all members for their science questions.

Goal: Achieve balance between developing products that are needed by group members while minimizing storage space and access/downloading availability.

We propose to generate and retain the following products as part of our range modeling effort:

I'm not sure we really need to store Shapefiles below marked in Grey. They'd be pretty easy to generate with the occurrence data, which we also plan to make available.

- **Convex Hull range** - polygon (SHAPEFILE)
- **Bounding box range** - polygon (SHAPEFILE) - can we just do this as a CSV file (simple file with 4 values)
- **Latitude extent range** - polygon (SHAPEFILE)
Is this in degrees or band across the continent in area (I prefer the first) - in which case this would just be two values (vector)
- **Area of occupancy** - polygon (SHAPEFILE)
Not sure this is useful as a shapefile? Raster, list of grid cells, or ?? Probably redundant with occurrence CSV file
- **Maxent raw probability surface** - environmental suitability grid - raster (GEOTIFF)
Raster layer generated from the Maxent model runs of each species, using a combination of 19 bioclim layers and 19 spatial filters and various thresholds.
- **Maxent binary range map** - from 1 percent training presence threshold - polygon (GEOTIFF)
Raster layer generated from the best modeling method as defined in our range area study.
- **Maxent binary range map** - from 1 percent training presence threshold - polygon (SHAPEFILE)
Polygon map generated from the Maxent raster defined above, perhaps smoothed for better cartographic representation.
- **Species occurrences** - points (CSV FILE)
- **Range size estimates** - text (CSV FILE), contains the following measures of range size:
 - Number of occurrence points
 - Convex hull area

- Clipped convex hull area
- Bounding box area
- Clipped bounding box area
- Latitudinal extent (linear dimension without area)
- Latitude extent area
- Clipped Latitude extent area
- Area of occupancy (area of occupied cells)
- Areas from the three Maxent model types (bioclim, spatial filters, bioclim + spatial filters) and various thresholds defined below:
 - Maxent minimum training presence threshold area
 - Maxent 1 percent training presence threshold area
 - Maxent 5 percent training presence threshold area
 - Maxent max Kappa threshold area
 - Maxent balanced sensitivity and specificity area
 - Maxent maximum sensitivity and specificity area
- **Maxent all thresholds** - text (**CSV FILE**), contains a list of all thresholds generated by Maxent for each species
- **MaxentResults.csv** - text file (**CSV FILE**)
File named maxentResults.csv that is generated in the temp directory by Dismo for each Maxent model run of each species

Range Modeling Decision Tree

The following decision tree will be used to determine how each species is modeled so that we are able to calculate a range area for each species – including those for which we are unable to model using Maxent because they have too few occurrence records in distinct geographic locations (lat/long)

- A. ≥ 5 unique ¹ points..... Model range area using Maxent
- A. < 5 unique points, or failed Maxent..... B

- B. 3-4 unique points or failed Maxent Convex hull
- B. < 3 unique points..... C

- C. 2 unique points..... Two point algorithm
- C. 1 unique point..... Single point algorithm

Two point algorithm:

Create one box around both points or a separate box around each point?

Calculate a box around both points and compare area of box with Gentry area.

- X or Y coordinates of both points are equal..... Fit as 2 single boxes ³
- Box area $>$ Gentry area ($75,000 \text{ km}^2$)..... Fit two boxes ²
- Box area \leq Gentry area ($75,000 \text{ km}^2$)..... Fit single box ³

One point algorithm:

Calculate one box around each point 2

* Fit one Gentry box

Notes:

¹ Unique defined as having distinctly separate geographic locations (lat/long)

² Two point boxes will be constructed using the bounding latitudes and longitudes of both points.

³ Single point boxes will be constructed by setting the point as the center of the box and extending 136.9 km in each direction to create a box of area approx. $75,000 \text{ km}^2$

Gentry cutoff for small range species:

"We defined small-ranged species as those that have never been collected outside the Peruvian department of Madre de Dios, and large-ranged species as those that have. The rationale is that the size of the department of Madre de Dios, $78,415 \text{ km}^2$, fits the most widely used criterion of tropical plant species endemism, $50\text{--}75,000 \text{ km}^2$ (Gentry 1986, 1992)." - From Pitman 1999

Pitman, N. C. a., Terborgh, J., Silman, M. R., & Nuñez V., P. (1999). Tree Species Distributions in an Upper Amazonian Forest. *Ecology*, 80(8), 2651-2661. doi:10.1890/0012-9658(1999)080[2651:TSDIAU]2.0.CO;2

Gentry. 1986. Endemism in tropical vs. temperate plant communities. Pages 153–181 in M. E. Soule', editor. *Conservation biology: the science of scarcity and diversity*. Sinauer Associates, Sunderland, Massachusetts, USA.

Gentry. 1988. Changes in plant community diversity and floristic composition on environmental and geographical gradients. *Annals of the Missouri Botanical Garden* 75(1): 1–34.