CORRESPONDENCE NOTES:

**PHONE CALL WITH MATT , 4/25/23:**

They do repeat samples 10min apart, and then two visits per year. But having the second visit is what really seems to help. NETN (Aaron Weed) does single visit with removal approach, Jeff Doser.com SECR with point counts—look up what they did.

USGS BBS data 3 min. They do hierarchical analysis with BBS Bayes. Matt thinks standardizing to USGS or using removal model is way to go.

The most useful covariates--% forest cover in 500m radius or 1km for abundance/occupancy side (using NLCD off the internet). On detection side, observer, time-of-day (straight) and julian day, and Decibel meter for background noise. At Site level, also hemlock abundance and hemlock condition b/c there are some species that are hemlock-associated birds, for their abundance.

Matt is a big supporter of just using the raw data b/c he said that almost always the model outputs run parallel to the raw data. He said he is now just focusing heavily on figuring out the most effective covariates. But he did say the repeat visits have been very helpful for him to get better-seeming abundance estimates.

ME TO DO

* COMPARE TO SINGLE-VISIT ESTIMATES USING DISTANCE SAMPLING OR REMOVAL SAMPLING

FOLLOW-UP QUESTIONS ON UNMARKED CHAT:

* Actually is density estimated as (lambda \* phi \* avail)/area? Or possibly (lambda \* phi)/(avail \* area)?
* Pifun question
* best\_mod\_pb <- parboot(best\_mod, fitstats2, nsim = 200, report = 10, ncores = 3) doesn’t work?

RESPONSE FROM KEN KELLER, 4/22/23

> CORRESPONDING ‘yearlySiteCovs’ I ENTERED MEAN COVARIATE VALUES FOR THOSE  
> MISSING SURVEYS B/C gdistremoval() CANNOT HAVE NA’S FOR COVARIATES, IS THAT  
> THE CORRECT WAY TO HANDLE THIS?  
  
That's fine, they are ignored. gdistremoval is a little too aggressive with throwing errors for missing values.  
  
> ·        For  ‘yearlySiteCovs`, my rows are ordered (years stacked) like:  
> Site1Yr2004, Site1Yr2005, Site1Yr2006…. Site1Yr2022, Site2Yr2004,  
> Site2Yr2005… [Q2] THIS SITE-MAJOR, YEAR-MINOR ORDER SEEMS TO BE THE GENERAL  
> RULE FOR `yearlySiteCovs`, DOES IT APPLY TO `unmarkedFrameGDR` ALSO?  
  
yearlySiteCovs are confusingly named. In this case they would be better named periodSiteCovs or something. So it should be site-major, period-minor order, e.g. Site1Sample1, Site1Sample2, Site2Sample1, Site2Sample2, etc. Site in your case being actually a site-year.  
  
> ·        In all models I included year (numeric, scaled) as a site  
> covariate. I am using this estimated year coefficient as a measure of trend  
> (bootstrap w/replacement of sites to get correct 95% CI’s). [Q3] BUT IS  
> THERE A WAY (WITHOUT LEARNING BUGS) TO ALSO ACCOUNT FOR THE UNMODELED YEAR RANDOM EFFECT?  
  
gdistremoval actually does support random effects using the lme4-type syntax. So you could put a year random effect on lambda. This is a very new, and relatively untested feature though so be cautious. [FOLLOW-UP: SO I WOULDN’T NEED TO BOOTSTRAP TO GET CORRECT CI’S IF I ADD THE RANDOM EFFECTS? I WILL CHECK RESULTS WITH AND WITHOUT TO SEE IF ANYTHING SEEMS UNUSUAL WHEN ADDING THE RE] ODDLY, ADDING RE DIDN’T CHANGE RESULTS (SE) ANY…  
  
> ·        I obtained site-survey density estimates with the calculation:  
> (lambda \* phi)/area— [Q4] SHOULD I DO PARAMETRIC BOOTSTRAP (SEPARATE FROM  
> THE NON-PARAMETRIC BOOTSTRAP MENTIONED ABOVE) TO GET A 95% CI ON THAT  
> DENSITY ESTIMATE FOR EACH SITE-VISIT?  
  
Not sure I understand this. If you set output = 'density' in gdistremoval() then the lambda estimates and output from predict() (including a 95% CI) will already be on the density scale. You shouldn't need to adjust with area yourself, and definitely dont need to involve phi. [FOLLOW-UP: I MISPOKE ABOUT DIVIDING BY AREA. IN MY CODE I HAD ACTUALLY JUST MULTIPLIED PREDICTED LAMBDA BY PREDICTED PHI TO GET ESTIMATED DENSITY… PROVIDE LINKS SUPPORTING THIS AND ASK FOR CLARIFICATION. I THINK WE MULTIPLY BY PHI TO GET LOCAL POPULATION DENSITY (I.E. DENSITY IN SAMPLED PLOTS), BUT LAMBDA ITSELF GIVES US.

[Inference about density and temporary emigration in unmarked populations - Chandler - 2011 - Ecology - Wiley Online Library](https://esajournals.onlinelibrary.wiley.com/doi/pdf/10.1890/10-2433.1)

[Problems with GOF and unreasonable predictions in gdistsamp (google.com)](https://groups.google.com/g/unmarked/c/9g1CjKE7BWo/m/l8sOBE2NBwAJ)

SEEMS IF OUTPUT=ABUND, THEN PREDICT() AND RANEF() GIVE ALMOST THE SAME NUMBERS EXCEPT RANEF SEEMS TO TAKE INTO ACCOUNT THE ACTUAL OBSERVED DATA, NOT JUST THE COVARIATE VALUES SO FOR A SINGLE COMMON PREDICTED VALUE, THERE MAY BE MANY DIFFERENT RANEF VALUES BUT ALL JUST CLUSTERED AROUND THE PREDICT. SO I THINK I CAN JUST DIVIDE BY SAMPLE AREA TO GET DENSITY?? EVEN THOUGH THE ESTIMATE IS OF THE SUPERPOPULATION ABUNDANCE.

[gpcount(), estimating difference in abundance between primary periods](file:///C:\Users\echeng\OneDrive%20-%20DOI\Documents\NPS%20PROJECTS_CURRENT\GULN_birds\Resources\Unmarked_chat\gpcount(),%20estimating%20difference%20in%20abundance%20between%20primary%20periods.html)

I THINK RANEF(GDISTSAMP\_MOD) GIVES US EMPIRICAL BAYES ESTIMATE OF SUPERPOPULATION SIZE, SO WE CAN’T CONVERT TO DENSITY…

[Quick guide on how to fit an occupancy model[

STILL SLIGHTLY CONFUSED ABOUT HOW IT DIFFERS FROM PREDICT()—in my gdistamp, is it giving the best estimate of density for that site-year? It’s the superpopulation density??

?gmultmix suggests that if we estimate density as M/area, then we are assuming there is no TE?

> ·        [Q5] IS IT OKAY TO USE HABITAT AS A COVARIATE ON LAMBDA AND ALSO  
> AS A COVARIATE ON DETECTION?  
  
Yes, that's fine.  
  
In general with this model, I recommend building up complexity iteratively. Try fitting the distance and removal parts separately first, with gdistsamp / gmultmix or distsamp/multinomPois. Make sure the combined version makes sense in this context. In simulations I have found it can be very hard to get a good estimate of phi unless you have a lot of data. I am somewhat skeptical with your dataset you'll be able to estimate distance p, removal p, phi, \*and\* a random effect.[FOLLOW-UP: WOULD IT MAKE SENSE TO JUST SET PRIMARY SESSIONS=1 AND NOT TRY TO ESTIMATE PHI FROM THE REPEAT SAMPLES? IN THAT CASE SHOULD I ASSUME THE ESTIMATED DENSITY IS LIKELY AN OVERESTIMATE B/C NOT CORRECTED FOR INDIVIDUALS WITH PARTIALLY OVERLAPPING HOME RANGE TO THE SURVEY CIRCLE?]

Gdistremoval—parboot doesn’t work

Ask about the pcount-glmm question, why is it so wrong? Should we ask our networks to redo their analyses they have already done? (I should do it this way to compare, then ask)

IS PREDICT GIVING DENSITIES INSTEAD OF ABUNDANCES??

QUESTION ON UNMARKED CHAT, 4/21/23

Hi, thank you so much to everyone who puts in (volunteers!) time to develop and maintain the `unmarked` and `ubms` packages and chat pages. I have been reading through the AHME books and google chats for a few weeks now and-- based on what I’ve read-- I think (hope) I’m analyzing our bird point count data correctly. But I would really appreciate it if I can run it by this group for confirmation and also for help with residual questions (NUMBERED QUESTIONS, IN CAPS).

The data are from long-term point count monitoring of birds in national parks. We are primarily interested in estimating long-term trend, and—if possible—also getting an estimate of annual density (with 95% CI). Each species and park will be analyzed separately.

I apologize in advance for so many questions ☹

THE DATA, IN A NUTSHELL:

* We have 20-35 point count locations in a park. Data have been collected since 2004, but in the early years only half the points were surveyed each year, and only once in the year. Since 2019 all points have been surveyed twice each year (2nd survey within 3 weeks of first).
* Point counts are 10 min, and birds are recorded only the first time they are detected (i.e., removal sampling). For each detection, observer records species, distance to detection (bins are 0-25, 25-50, 50-100, 100m+ ), and time-of-detection (1-min intervals for 10min).
* Site covariate data is habitat type (e.g., forest, grass, shrub, developed). Survey event covariate data include julian date, survey start time, and wind condition.

MY APPROACH WITH GDISTREMOVAL():

* I dropped the data collected in the 100m+ bin because I think we can’t analyze distance data with an unknown bin size.
* I stacked years and used the two visits per year to estimate phi (temporary emigration). So for example my `yDistance` data frame is set up with a row for each site-year, and with columns: Visit1Dist1, Visit1Dist2, Visit1Dist3, Visit2Dist1, Visit2Dist2, Visit2Dist3. When a site did not have a second visit that year (like in the years before 2019), I filled those cells with `NA` to indicate missing surveys –[Q1] BUT IN THE CORRESPONDING ‘yearlySiteCovs’ I ENTERED MEAN COVARIATE VALUES FOR THOSE MISSING SURVEYS B/C gdistremoval() CANNOT HAVE NA’S FOR COVARIATES, IS THAT THE CORRECT WAY TO HANDLE THIS?
* For ‘yearlySiteCovs`, my rows are ordered (years stacked) like: Site1Yr2004, Site1Yr2005, Site1Yr2006…. Site1Yr2022, Site2Yr2004, Site2Yr2005… [Q2] THIS SITE-MAJOR, YEAR-MINOR ORDER SEEMS TO BE THE GENERAL RULE FOR `yearlySiteCovs`, DOES IT APPLY TO `unmarkedFrameGDR` ALSO?
* In all models I included year (numeric, scaled) as a site covariate. I am using this estimated year coefficient as a measure of trend (bootstrap w/replacement of sites to get correct 95% CI’s). [Q3] BUT IS THERE A WAY (WITHOUT LEARNING BUGS) TO ALSO ACCOUNT FOR THE UNMODELED YEAR RANDOM EFFECT?
* I obtained site-survey density estimates with the calculation: (lambda \* phi)/area— [Q4] SHOULD I DO PARAMETRIC BOOTSTRAP (SEPARATE FROM THE NON-PARAMETRIC BOOTSTRAP MENTIONED ABOVE) TO GET A 95% CI ON THAT DENSITY ESTIMATE FOR EACH SITE-VISIT?
* [Q5] IS IT OKAY TO USE HABITAT AS A COVARIATE ON LAMBDA AND ALSO AS A COVARIATE ON DETECTION? I have seen `unmarked` examples use the same covariate on different components of the model, but I just want to double-check it’s okay. I did staged model selection by first selecting the best-fit model for the distance function, then using that to compare models with different covariates for lambda. In many cases the model with the lowest AIC included habitat as a covariate for both lambda and distance function, but the 95% CI’s on those estimates were huge.

Thank you for taking the time to read this, and for any thoughts you may be able to provide on the above questions!

EMAIL WITH RICHARD CHANDLER, 4/18/23

Hi Richard! I hope you have been doing well. I'm going to be in Athens in a couple months for the USGS cooperators meeting and hope you'll be around so I can stop by to say hello!

I'm really sorry for bugging you on this—I know you have a million and one things eating up your time—but I'm hoping you can tell me if I'm analyzing some NPS point count data correctly with unmarked. I've gone through the AHME book vol. 2 (unfortunately I don't have vol. 1) and much googling and in the end I'm still not sure if it's possible to use unmarked for data with multiple years and multiple surveys within year. But these are the data and this is what I think I can do with them in unmarked (would you mind checking me on this?) QUESTIONS IN RED.

THE NEED:

* The parks are interested in estimates of long-term trend and annual estimates of density for their most commonly detected bird species.
* They also want to know if they can/should drop the repeat surveys, the distance information, or the time-to-removal information in their data collection. Ideally, they would like to be able to go out only once per year to each site (drop the repeat surveys) and rely on the distance and/or time-to-detection information to get reliable estimates of detection probability. I will be analyzing the data in many ways to see if results are similar when we omit some of the available information that was collected to estimate detection probability.

THE DATA:

* 10-min bird point counts, with (binned) distance and time-to-detection data reported. In a park (they want parks analyzed separately) approximately 20 to 25 point count locations
* Data collected since 2004, but in the early years only half the points surveyed each year, and only once. Starting 2019 they surveyed all points twice each year (2nd survey within 3 weeks of first) ​
* In each point count, distance bins are 0-25, 25-50, 50-100, 100+. Birds are recorded the first time they are seen/heard, and the data indicates which minute that occurs (so ten 1-min intervals).
* Covariate data include habitat type, julian date, survey start time, wind conditions.

QUESTIONS/NOTES ABOUT THE DATA:

* I'm not sure how to deal with data collected in the 100+ bin. (all birds detected beyond 100m, with no upper distance boundary) I'm guessing I just have to drop them when using gdistsamp() or if I'm using gmultmix() then I can include them but report estimated abundances since survey area is unknown? Does that sound about right?
* I know there are issues w/combining vocal and visual detections in an analysis (fitting distance functions or removal sampling functions). They did not distinguish which detections were vocal or visual, but they said the vast majority of detections were vocal.

THIS IS HOW I THINK I WOULD ANALYZE THE DATA WITH GDISTSAMP() IGNORING THE DISTANCE DATA:

* In the response dataframe, the data would be set up with 2 primary sessions and 3 distance bins so the columns would be like: Visit1Dist1, Visit1Dist2, Visit1Dist3, Visit2Dist1, Visit2Dist2, Visit2Dist3
* The data would be year-stacked so each row of data represents a site-year
* For the years pre-2019 when only half the plots were surveyed each year and only once, I would just have NA to indicate missing surveys (but in some examples it seemed like for the SiteCovs and YearlySiteCovs you put the mean covariate values instead of NA for these missing surveys so the model wouldn't "drop the entire site" because of missing surveys? Is it wrong for me to just have NA for missing surveys in the y-data frame and the SiteCovs and YearlySiteCovs?)
* 'Year' as a numeric class variable would be a covariate in SiteCovs and the estimated year effect would be the trend I would want to report
* The estimated density per site-survey would be calculated as (lambda\*phi)/Area because phi corrects for temporary emigration of birds whose home ranges partially overlap the point count survey area
* I should NOT be including site as a factor covariate in SiteCovs b/c then I am estimating every parameter separately for each site and I lose the value of the presumed spatial replication in the model, right?
* I can't incorporate random effect of site and year in the model (and ubms package doesn't have functions for temporary emigration yet, and I don't have the bandwith to learn how to do this all in BUGS right now) but I can "fix" the CI's on the estimates by nonparametric bootstrapping with replacement?
  + This means selecting points with replacement so they match the actual number of point count locations in that park
  + All the data for each selected point will be included in the bootstrap sample (multiple times if the same point is selected more than once). This means that the total number of site-years (rows of data) may differ for the bootstrap vs. original sample b/c not all sites were sampled in all years—is that okay?
  + Rerun the model on the bootstrap sample, grab the estimates (no SE's), then repeat another few hundred times and calculate the 95% CI for each estimate from the distribution of bootstrap estimates
  + That should take care of not having RE of site, but I think nothing can be done about not having RE of year (I'm putting year in SiteCovs but as numeric to get the linear trend)?

REMAINING QUESTIONS:

* I think whatever setup I use for gmultmix() would be similar to what I do for gmultmix() except that instead of time bins I would use distance bins, right?
* I don't think there is a way in unmarked to include both the distance and the time-to-removal information in a single analysis, right? I'll be using the Solymos approach for that, as a comparison to results from unmarked models.

**EMAIL FROM MATTHEW MARSHALL, 4/18/23**

Thanks for reaching out! I am happy to talk bird point count analyses at any time! Distance and time-to-detection were not included in the analyses you are referring to. The “count” of birds for each site-survey was simply the number of birds counted during the five minute sampling period. And, yes, I used the repeat visit information to estimate detection prob and abundance using the pcount function in unmarked.

As you know, I could have included this information but for a variety of reasons chose not to. Which does lead to the questions about why are we collecting this information. For me the answer has more to do with the ability to compare and analyze data across other protocols. For example, some bird monitoring protocols in the NER do not employ repeat visits. So we’ve explored removal models or time of detection models in joint analyses. I could easily be convinced that we should not be assigning birds to distance bands. Especially in forested environments (open grasslands, prairies, etc. it does make sense to use DISTANCE methods). But we have not pulled the trigger in dropping these aspects of data collection. Yet. Very curious to hear more about what you all decide to do.

Hope this helps and happy to chat any time you like!

-Matt

Interior Region 1 North Atlantic - Appalachian

National Park Service

Eastern Rivers and Mountains Network

814-863-0134

c: 814-571-8316

[nps.gov/im/ermn](https://gcc02.safelinks.protection.outlook.com/?url=http%3A%2F%2Fnps.gov%2Fim%2Fermn&data=05%7C01%7Cellen_cheng%40nps.gov%7C8631830c54dd49f89cc708db40144b65%7C0693b5ba4b184d7b9341f32f400a5494%7C0%7C0%7C638174228123488682%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=fQCj%2B%2BdntYmsM%2B6XPnu4aX9YnyfJLibp7HnCxtUsAIw%3D&reserved=0)

**TEAMS CHAT FROM JP, 4/18/23**

[JP]

So far we have mostly done exactly what you say even though we have data (ish) for distance bands and removal counts. I am not sure how much U of D has played with the other methods with our data - but I have a meeting later with them later today and can ask. They have done a lot of work fitting lots of models to lots of species / parks at once and it may be that pcount and occu work better for that - as opposed to focusing on a single species.

I knw they also work with some more western networks where you can take distance data with a rangefinder, and they say the distance models work really well there.

….

[JP]

Just got off the phone with the U of D folks. So they have not really tried to use distance as we only have the 2 bands, but could try it out. They have done some removal model work, but feel that the pcount models fit more species well. There may be some individual species where a removal model would work better, but pcount works with more overall.

[Ellen]

By "work" I think you mean being able to get good fit to the removal function via chi-sq test or something? We have distance bands 0-25, 25-50, 50-100, 100+. I have no idea what to do with the 100+ band data b/c we can't estimate density with an unknown survey radius and the approaches that calculate EDR from infinite distance band data are suspect, IMHO. So I've just truncated the data at 100m but for some species, the 0-25 bin detections are way lower than 25-50 (even accounting for difference in area) and I assume that is b/c some species are avoiding the surveyor. So for those I have to combine the first two bins to 0-50m but then I'm only left with 2 bins and can't do anything with that. Plus our point count surveys are not in any way homogeneous in habitat--the 100m radius circle may have a trail running through it, may have open habitat in some % of it and forested in the other, etc.--which is all going to affect the distances some species are detected at. So I don't think distance models will work for us. Removal models don't seem to fit well either, though, b/c the counts don't "tail off" enough in the last few minutes of the 10-min survey. I'm wondering if they are double-counting some birds in those last few minutes b/c the birds moved from initial location and they lost track or perhaps it's just from new birds moving into the circle during the 10min. With pcount(), though, I've been reading about the gof tests being unreliable (too often saying models are good when they are not) and also issues with overestimating density b/c of the birds whose territories only partially overlap the point count circle. It seems that is why Chandler et el developed the temporary emigration models to "correct" for that, but I honestly don't understand how it's able to figure this out just with the data we provide the model so I'm not convinced it's "better". In the end, my problem is that I can't figure out how to "know" which models are actually more correct. Ugh!

[JP]

I am unaware of the Chandler models - but basically agree with all you wrote. Even just using pcount / occu I am concerned that model misspecification is having a big influence on some outcome. We occasionally get model fits that fluctuate between the species being very rare and the species being very common from year to year. I am concerned that it is just noise, but is being magnified by some fluke in the detectability model. Its all very murky to me.