TRY 2 SUCCESSIVE 5-MIN POINT COUNTS WITH TIME-TO-DETECTION? BUT JUST FOCUS ON VOCALIZATIONS NOT SIGHTINGS (BUT CAN RECORD SIGHTINGS SEPARATELY BUT THEN RECORD THE FIRST TIME IT VOCALIZES)—OTHERWISE SECOND 5-MIN ROUND MAY BE BIASED B/C YOU ALREADY KNOW WHERE THE BIRDS ARE. OR IF THEY ARE SUPPOSEDLY ALREADY KEEPING TRACK OF WHICH BIRDS THEY ALREADY COUNTED, THEN IN THE SECOND 5 MIN THEY SHOULD ALSO RECORD IF IT’S A NEW BIRD OR NOT. RESTRICT TO 100M.

NMixture points/questions:

* “free lunch” of estimates from unmarked data—the trade-off is high sensitivity to violations of model assumptions. For example, unmodeled heterogeneity in p (detection prob.) is known to bias abundance estimates a lot
* For point counts, simply using pcount() on the repeat surveys overestimates density estimates b/c we are actually getting an estimate of the superpopulation of the site, which includes birds whose home ranges just partially overlap the point count circle. Therefore, the survey area (denominator) really is not the correct area to divide by. If we DON’T have repeat surveys and we do something like distance sampling or removal sampling (multinomial data) we get estimates of “instantaneous” density, which may be lower than the TE-estimated density and have larger SE (NOTE ALSO likely overestimate in any case due to animals moving into the survey area during the 10-min survey, or accidental double counting of animals)
* If we use the wrong N-mix models (i.e., model assumptions are not met), we likely will get very wrong abundance estimates but estimates of trend should be somewhat robust. So we can just think of it as relative abundance instead of absolute abundance.
* Stacked years gives CI’s that are too small b/c doesn’t account for repeat sampling of sites—need to do bootstrapping to get appropriate CI’s for density and all estimated parameters. NOTE: I guess with unmarked we should not used fixed effect of site b/c we are supposed to use sites as spatial replicates. But with ubms we can use random effect of sites, so does that address the issue of incorrect CI’s with data stacking?????
* In the book, when some covariate data are missing, they set the values to the means (to 0 for scaled covariates) saying that otherwise the entire site will be deleted. Is that the case when covariate data are missing b/c there was not a survey?? E.g., site was surveyed only once in the year instead of twice? I thought in that case it was fine to put NAs for those missing surveys
* Is there an umarked option for DISTANCE + REMOVAL, with year stacking and no repeat surveys within year?
* Is it appropriate to just take the estimates from pcount and then put them in GLMM with year trend and RE, as others have done??
* In general, because we KNOW we have unmodeled heterogeneity in detection (habitat covariates missing, etc.) we may have to just view Nmixture model (and DISTANCE, etc.) results as relative abundance that may or may not provide more insights than GLMM. Certainly we are lacking covariates to appropriately account for habitat impacts on probability of detection (via distance function)
* Kery et al. believe that even when N-mixture models have unreliable estimates of abundance and are therefore used to model relative abundance (e.g., b/c we have unmodeled heterogeneity in detection), they are still superior to GLMMs b/c N-mixture allows covariates separately for detection and abundance. Basically, when N-mixture model assumptions are not met, the abundance estimates may be very wrong but trend estimates should be fairly robust so we can just think of it as relative abundance instead of absolute abundance.

Models to try:

* Gmultmix for REMOVAL + TE OVER REPEAT SURVEYS, and with stacked years (bootstrap for CI’s)
  + Year (scaled, continuous) as a site variable to estimate trend in lambda, where lambda is the site-year superpopulation (with the understanding the for surveys within each site-year the predicted density is a subset of lambda). The estimated trend is on the superpopulation, which has an unknown area denominator.
  + Each row of y-df would be visit1dist1, visit1dist2, visit1dist3, visit2dist1… (can’t deal with unknown bin sizes)
  + To estimate the predicted density per site-survey, it’s estimated lambda multiplied by estimated phi
  + To get corrected CI’s on the stacked data, do nonparametric bootstrapping by selecting site with replacement and then including all data for the selected sites (so the number of site-surveys will differ among bootstrap samples, but the number of sites will not)—this is okay to do even when not all sites surveyed all years, and not all sites surveyed twice in a given year, right?
  + Okay to leave in the missing surveys as NA, or should I limit analyses to starting in 2019 when all sites were surveyed twice per year?
  + Would pretty much do it the same way for gdistsamp, right? I know there is now an open population version, but I’m not interested in estimating recruitment and survival, just getting trend estimates
  + To then compare to estimates when I just use one survey per year, I should have each row represent a site and the years would be the primary survey, then use year covariate on phi and that estimated year coefficient would then represent the estimated year trend?
  + How about this approach…gpcount on stacked data to get an estimated mean density per site-survey, as (lambda \* phi)/Area, then plugging those estimated densities (ignoring CI’s) into GLMM with year as fixed effect (for trend) and year and site as random effects, bootstrapping to carry the uncertainties in pcount() etsima. As long as those two changes made, would it be a valid way to be able to incorporate year as RE and fixed effect into the model?

**NMIXTURE MODELS**

* **NMIXTURE multinomPois() for REMOVAL. NOTE!! Start with this to see if I can get a decent fit to removal data. I’m guessing not UNLESS I CAN WRITE A PIFUN FOR 2-CALL-RATES <<<<<< If I can, then do the actual analyses with UBMS stan\_multinomPois()** and include site RE, year RE, and year FE. My guess is that anyway the removal models won’t fit when when assuming a single homogeneous call rate.
  + 100m truncation and full 10min, only first survey
  + 100m truncation and only 5min, both surveys
* **UBMS stan\_distsamp()** for DISTANCE with stacked years and include site RE, year RE, and year FE.. If we go with this, it will be with the idea of dropping repeat surveys so not using full 10min and both surveys. This may not work well with species that exhibit “avoidance” of researcher.
  + 100m truncation and full 10min, only first survey (3 separate distance bins)
  + 100m truncation and only 5min, both surveys (3 separate distance bins)
* **UBMS stan\_pcount()** b/c that is SIMILAR TO what others do but perhaps better (?) b/c I will include site RE, year RE, and year FE. I’m not stacking years so I don’t need to bootstrap to “correct” CI’s and I don’t need separate GLMM b/c I’m already putting the RE’s and year trend in the N-mixture model, right?
  + 100m truncation and only 5min, both surveys
  + All data
* **NMIXTURE gdistremoval()**. Fit the model of Amundson et al. (2014) to point count datasets containing both distance and time of observation data. The Amundson et al. (2014) model is extended to account for temporary emigration by estimating an additional availability probability if multiple counts at a site are available. Abundance can be modeled as a Poisson, negative binomial, or Zero-inflated Poisson. Multiple distance sampling key functions are also available. [R: Fit the combined distance and removal model of Amundson et... (r-project.org)](https://search.r-project.org/CRAN/refmans/unmarked/html/gdistremoval.html)
  + 100m truncation and only 5min, both surveys
  + 100m truncation and 10m, both surveys (this may be the assumed most “reliable” estimate for density)
* **NMIXTURE gpcount()** to account for TE—THIS ONE CONFUSING. It says if availability is due to vocalization rate AND TE, then density can’t be estimated and can only estimate superpopulation size. Perhaps use this to see what is the estimated phi?? Then GLMM.
  + 100m truncation and only 5min, both surveys
  + 100m truncation and 10m, both surveys (this may be the assumed most “reliable” estimate for density)
* **NMIXTURE pcount()** for reference b/c that is what others do. Then GLMM as they do. ERMN just reports abundance, not density, b/c they have unlimited distance bins too I think. This is mostly to see if results differ from stan\_pcount()
  + 100m truncation and only 5min, both surveys
  + 100m truncation and 10m, both surveys (this may be the assumed most “reliable” estimate for density)
  + All data

**WORKFLOW FOR NMIX:**

* Models w/Poisson on abundance
* AIC to identify best model
* Larger K’s on best model
* GOF on best model. If fit statistics suggest it’s a poor model, try with NB on abundance
* Nonparametric bootstrap for CI’s in stacked models

**OTHER MODELS – QUICK RUNS**

* **DISTANCE then GLMM**
  + 100m truncation and full 10min, only first survey (3 separate distance bins)
  + 100m truncation and only 5min, both surveys (3 separate distance bins)
  + Everything else I already did… (trying artificial distance boundaries, combining first bin or not; including ALL DATA)
* **SOLYMOS then GLMM**
  + 100m truncation and full 10min, only first survey (3 separate distance bins)
  + 100m truncation and only 5min, both surveys (3 separate distance bins)
  + Everything else I already did… (trying unlimited bins for EDR, combining first bin or not; including ALL DATA)
* **GLMM**
  + 100m truncation and only 5min, both surveys
  + 100m truncation and 10 min, both surveys
  + All data

**DATA SUMMARY**

* **If we restrict to 100m, what is detection probability? (is it sufficiently high that maybe we don’t need to model the distance function). Does the best model usually include a covariate on the distance function (if yes, we would want to be able to correct detection for that)**
* **If we survey for 5 vs 7 vs 10 min, what is availability? (is it sufficiently close to 1 that maybe we don’t need to model the removal function). Does the best model usually include a covariate on the removal function (if yes, we would want to be able to correct for that)**
* **Regardless, it may be a good idea to continue collecting distance and time-to-detection information b/c: 1. Future analyses may be developed that can work with these, or a better QE may be able to work with it, 2. With more data, we may be able to work with these.**
* **BUT I’m thinking we may want to truncate sampling at 100m. AND perhaps instead of replicate samples, can do back-to-back replicate samples limited only to vocalizations to avoid bias from “knowing” where a bird was initially located—can they do this without being confused about whether or not they reported the bird in a second 5- or 7-min round yet (perhaps take a 2-min break btwn the surveys)**
  + **How do Solymos density estimates differ when analyzing 5- vs 7- vs 10-min surveys, when only using availability offset to correct?**