Parametric estimates of total tree richness in Amazon

Han Ter Steege, Paulo Inácio Prado, Renato Lima, ATDN

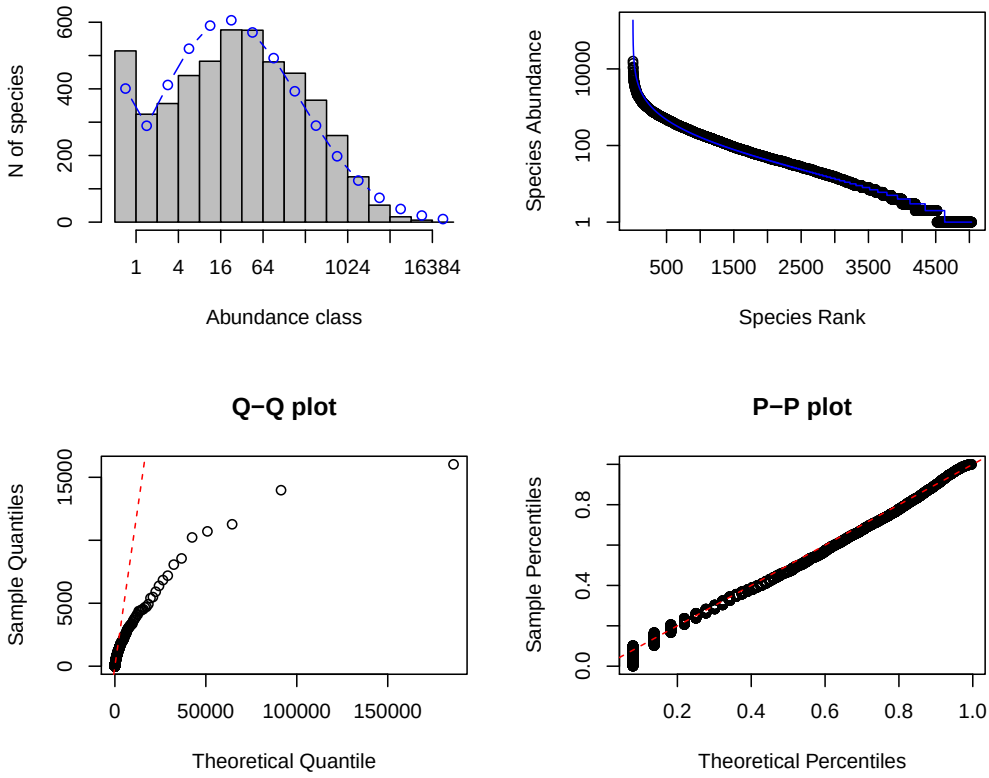
November 16, 2018

```
dados <- read.csv2("data.csv", as.is=TRUE)
y <- dados$N.ind
Sobs <- length(y)
## Total number of trees (average density x area)
Tot.t <- 567*5.5e8
## Proportion of total trees in the sample
p1 <- sum(dados$N.ind)/Tot.t
## Total number of plots
N.plots <- 1945
```

Poisson lognormal

Seems to overestimate the abundance of the most abundant species (see qq-plots)

```
pln <- fitpoilog(y)
par(mfrow=c(2,2))
plot(pln)
par(mfrow=c(1,1))
```



Species richness estimation from the underlying lognormal

Still, let's use the estimated coefficients to estimate total species richness. The estimated parameters of the Poilog are:

```
(pln.cf <- coef(pln))

##       mu      sig
## 2.787362 2.491355
```

In this model the observed SADs is a Poisson sample of a lognormal SAD of the whole community. The parameter σ of the Poilog estimates the same parameter of the underlying lognormal. The parameter μ of the Poilog has

the following relationship with the corresponding parameter of the sampled lognormal (M):

$M = \mu - \ln(p)$

where p is the proportion of the community that has been sampled (Bulmer, 1974; Saether et al., 2013).

As the estimated total number of trees in Amazon is 3.1×10^{11}, and the number of identified trees in the sample is 986577, $p = 3.1636267 \times 10^{-6}$. Therefore, we have $M = 15.4511534$.

The mean of the lognormal distribution is

$E[X] = e^{M+\frac{\sigma^2}{2}}$

And then this model says that the regional SAD is lognormal with an average number of trees/species of

```
pln.cf <- unname(pln.cf)
Mu <- pln.cf[1]-log(p1)
(mean.ln <- exp(Mu + pln.cf[2]^2/2))

## [1] 114327511
```

The estimated richness is then the total number of trees in the region divided by the mean above:

```
## Estimated number of species
(pln.S <- Tot.t/mean.ln)

## [1] 2727.69
```

Rather disappointing. I guess that the problem is the assumption that the pooled samples can be approximated by a Poisson sample of the entire tree community of Amazon (pretty unrealistic indeed).

Estimating species richness directly from PLN

An alternative way [1] to estimate total number of species in the sampled metacommunity is simply to calculate the value of the Poisson-lognormal fitted to the data at the zero value. That is, to estimate the probability

[1] I guess this is the way recommended by developers of the *poilog* package, to be checked

value assigned to species that had zero abundance in the sample. Simple as that:

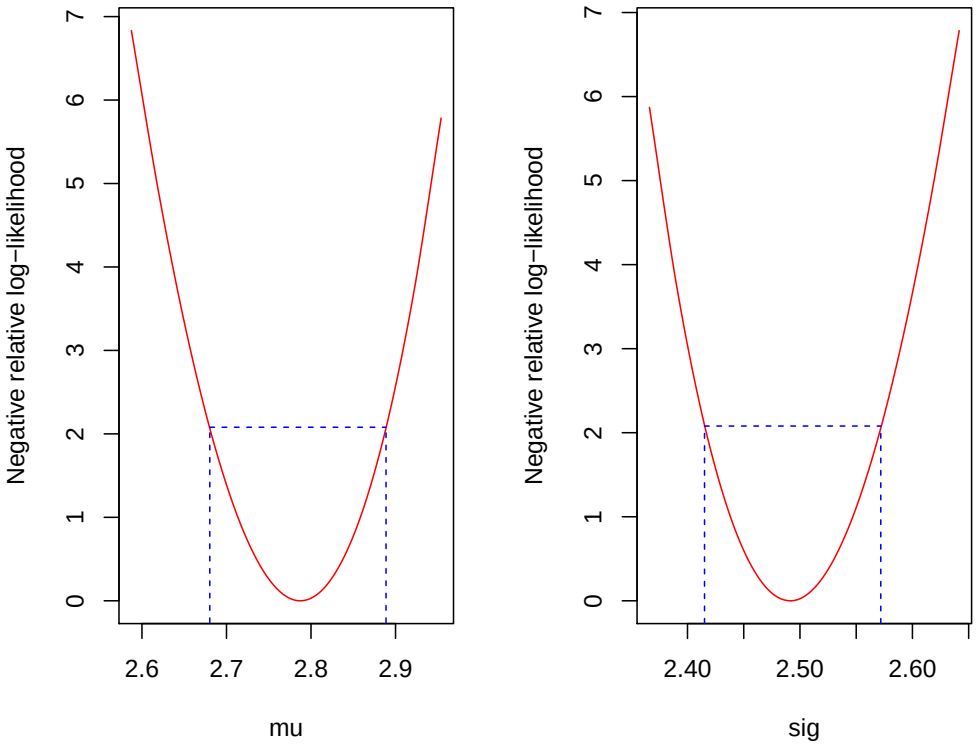
```
(pln.d0 <- dpoilog(0, mu = pln.cf[1], sig=pln.cf[2]))

## [1] 0.1124546
```

That is, an estimate that the recorded species are 89% of the total number of species. This gives an estimate of 5671 species.

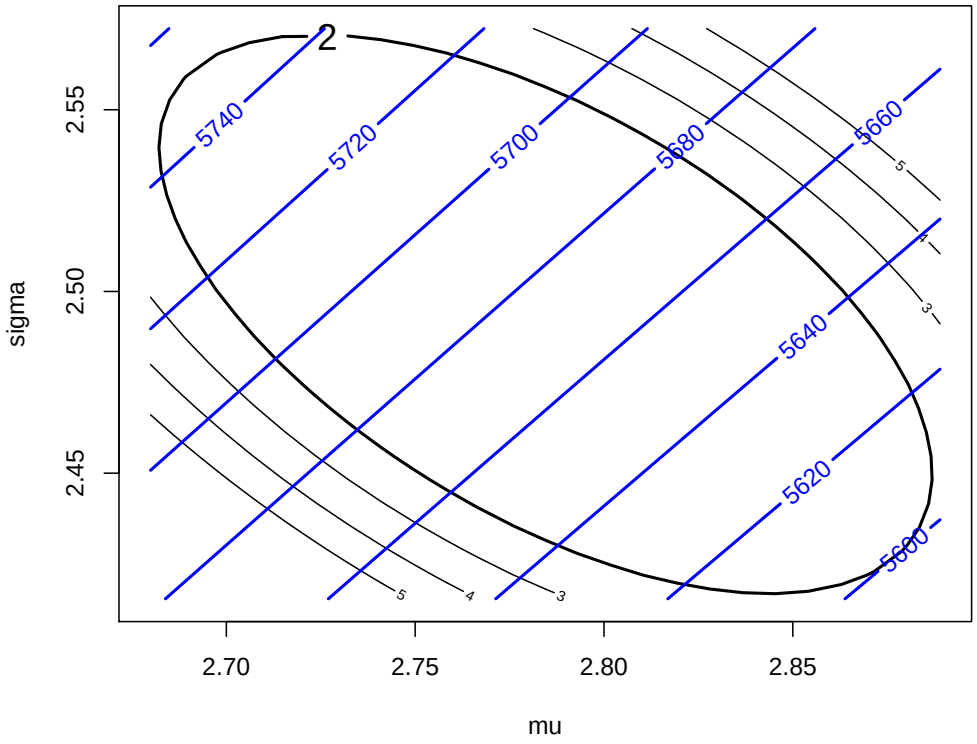
Likelihood and confidence intervals

Confidence profiles for both parameters are quite well-behaved (parabolic) and so match frequentist confidence intervals:

```
pln.prf <- profile(pln)
par(mfrow=c(1,2))
plotprofmle(pln.prf)
par(mfrow=c(1,1))
(pln.ci <- confint(pln.prf))

##        2.5 %   97.5 %
## mu  2.684205 2.885074
## sig 2.418213 2.569101

(pln.li <- likelregions(pln.prf))

## Likelihood regions for ratio = 2.079442
##
##  mu:
##         lower   upper
## [1,] 2.679937 2.88898
##
##  sig:
##         lower   upper
## [1,] 2.415443 2.57232
```



The likelihood interval of total species richness can be estimated visually from the likelihood surface. Using the likelihood ratio of two (bold black isoline) we can see the isolines that are tangent tothe likelihood region.



Because confidence intervals of the estimated parameters are very close to likelihood intervals we have a similar result using one or another interval to estimate a interval for the total species richness. But the likelihood surface help us by showing that we have to use the upper bound of μ CI with the lower bound of σ CI to get the lower bound of species richness (and the opposite for the upper bound):

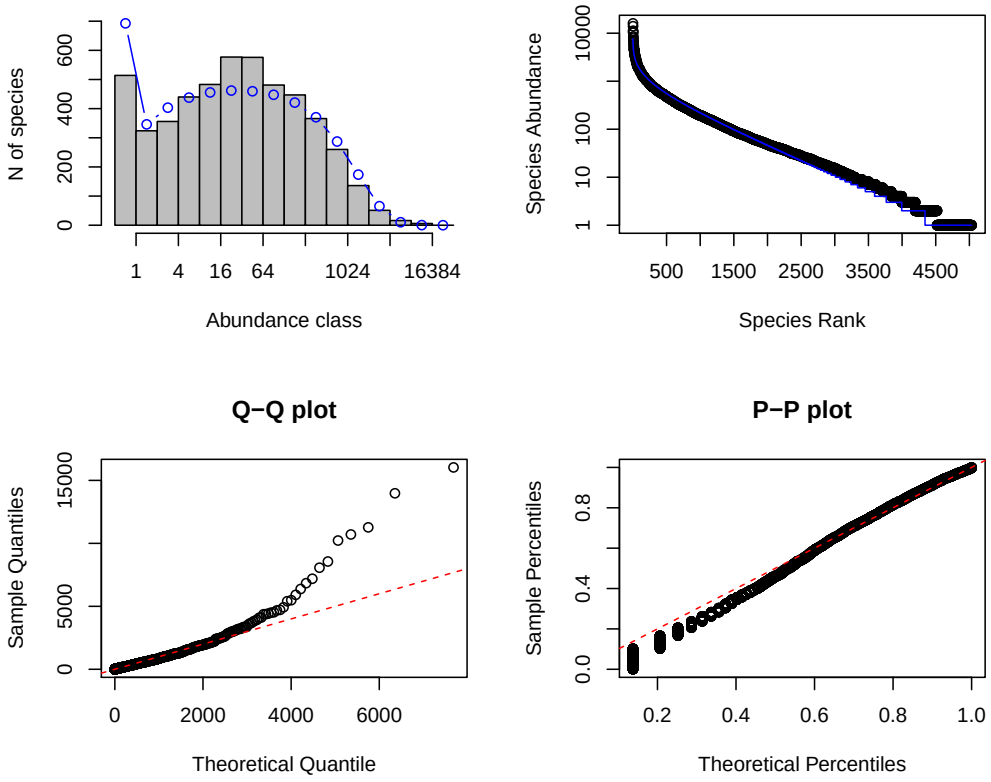
```
## Lower bound of S
Sobs / (1 - dpoilog(0, pln.ci[1,2], pln.ci[2,1]))

## [1] 5592.471

## Upper bound
Sobs / (1 - dpoilog(0, pln.ci[1,1], pln.ci[2,2]))

## [1] 5758.637
```

Still much lower than the estimates provided by log-series or negative binomial.

Log-series

Log-series seems to overestimate the number of singletons and to underestimate the abundance of species with intermediary abundances (between 16 and 64 individuals, see octave plot). Also, qq-plots show that the model underestimates the larger abundances.

```
y.ls <- fitls(y)
par(mfrow=c(2,2))
plot(y.ls)
par(mfrow=c(1,1))
```



Estimate of species richness

Well, you have that stuck in your brain, I suppose. But just for the records:

```
alpha <- coef(y.ls)[[2]]
(S.ls <- alpha*log(1 + Tot.t/alpha))

## [1] 13809.86
```

And here is the confidence interval for Fisher's α and the interval for estimated total richness from these values

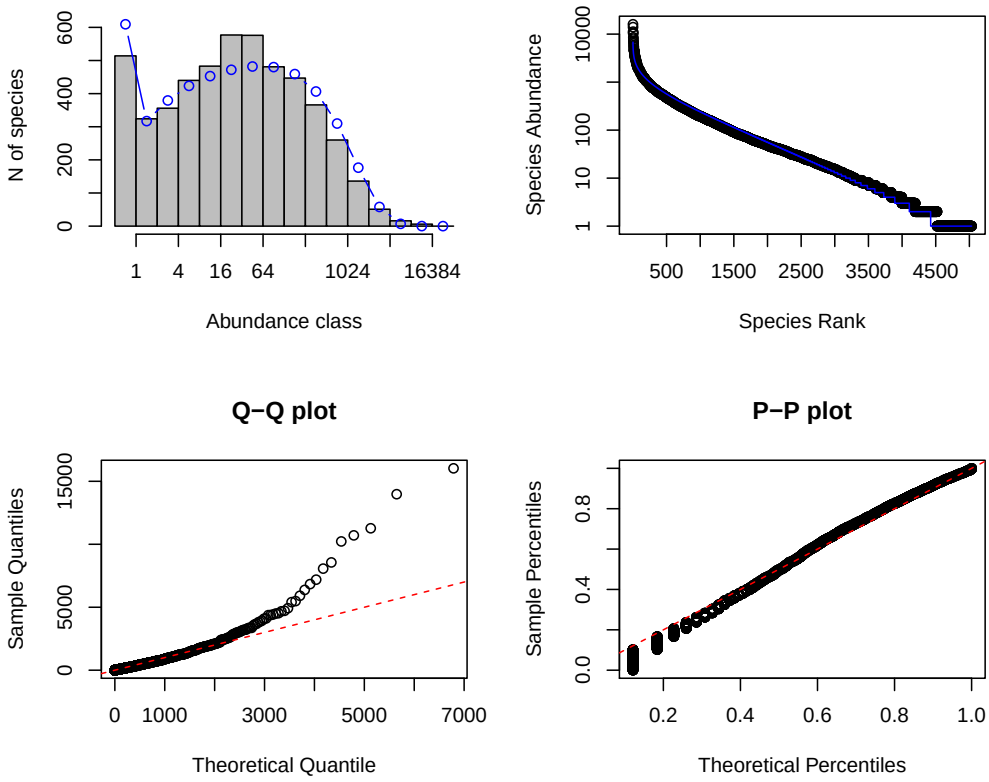
```
(ls.ci <- confint(y.ls))

##    2.5 %   97.5 %
## 638.9166 750.1000

## Estimated species richness for lower bound of alpha's IC
ls.ci[1]*log(1 + Tot.t/ls.ci[1])

##    2.5 %
## 12782.17

ls.ci[2]*log(1 + Tot.t/ls.ci[2])

##   97.5 %
## 14886.17
```

Negative binomial

Here I use the method proposed by Tovo et al. (2017). The first step is to fit a negative binomial to the abundances in the sample. I did that with the *VGAM* package (as the authors did) and also with the *sads* package. The results were similar, and the fit looks identical to those provided by the log-series (but has much lower AIC value, see below).

```r
## With VGAM
y.nb <- vglm(y ~ 1, posnegbinomial)
## With sads
y.nb2 <- fitnbinom(y, start.value=c(size=0.3, mu=mean(y)))
## Comparing:
exp(coef(y.nb))

## (Intercept):1 (Intercept):2
##   48.96347392    0.04048581

coef(y.nb2)

##        size          mu
##  0.04153996 50.08844429
```

```r
par(mfrow=c(2,2))
plot(y.nb2)
par(mfrow=c(1,1))
```



Estimate of species richness

Following the recipe of Tovo et al. (2017):

```
cf.nb <- coef(y.nb2)
csi.p <- unname(cf.nb[2]/(sum(cf.nb)))
csi <- csi.p/(p1+(1-p1)*csi.p)
## Estimated number of species
S.est <- Sobs*(1-(1-csi)^cf.nb[1]) / (1-(1-csi.p)^cf.nb[1])
unname(S.est)

## [1] 11038.67
```

I did a simple function to automate the calculations and to return the confidence intervals, based on the confidence intervals of the coefficients of the NB fit (see file `functions.R`)

```
tovo(fit = y.nb2, p = p1, CI=TRUE)

## Estimated species richness: 11038.67
##  95% CI: 10328.09 - 11977.2
```

Model selection

Among the three models, the Negative Binomial provides the best fit:

```
AICtab(pln, y.nb2, y.ls, base=TRUE)

##       AIC     dAIC    df
## y.nb2 54530.9     0.0 2
## y.ls  54557.5    26.6 1
## pln   54577.6    46.6 2
```

Parametric estimates from occupancies

Here goes a quick exploration of models for the distribution of occupancies (that is, the proportion of plots in which each species has been recorded). Such distribution captures a bit of the spatial aggregation of species [2].

I tried two tentative approaches. The first one was to calculate the empirical occupancies (that is, the proportion of plots in which each species has been recorded) and then I fitted a beta distribution to these values. To take into account undetected species I tried to truncate this continuous distribution at different points.

[2] check if what has already been done with these distributions; maybe we can call them SODs (Species Occupancy Distributions). Moreover, following core-satellite hypothesis, occupancies should correlate with abundances and them we might think on modelling a bivariate SAD x SOD distribution.

The second approach was to fit the distribution of frequencies (the number of plots out of the total of 1945 in which each species has been recorded) to a beta-binomial distribution. Because this is a discrete distribution, truncation is at a single point (species with zero frequency in the sample).

In both cases I assumed that the occupancies of the species follow a beta distribution, and that the observed frequencies were the result of independent 1945 trials for each species, in which the number of occupied plots follow a binomial distribution. Hence the compound beta-binomial distribution that underlies both approaches. With this distribution we can estimate the proportion of undetected species and then the total number of species.

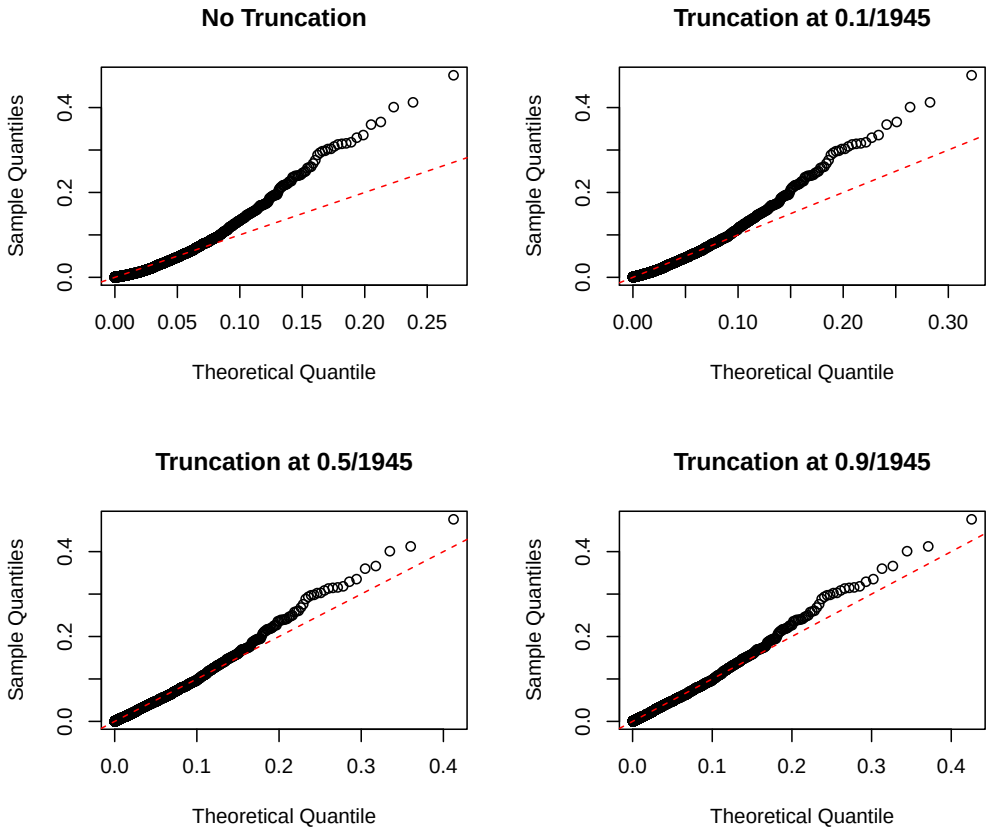
Truncated Beta distribution

The beta distribution is used to describe a distribution of bounded continuous variables like probabilities. We can think observed occupancies as estimates of the probabilities of occurrence of each species in a plot.

Because species with low occupancy values have lower detectability, the distribution should give more weight for the lower values of occupancies, as they are underrepresented in the sample. A first-order approximation is to truncate the beta distribution at 1/(number of plots).

I then fitted the distribution of occupancies to beta distributions without truncation and truncated at 0.1/1945, 0.5/1945 and 0.91945. The code for fitting with maximum likelihood is in the source file of this document. Below I show the QQplots of predicted x observed values, which shows that truncation above 0.5/1945 clearly improves the fit:

```
##        s1        s2
##  0.473385 23.882701
##         s1         s2
##  0.3083174 17.9980203
##          s1          s2
##  0.07090589 11.67846020
##           s1           s2
## 5.217535e-05 1.086736e+01
```



And the model selection shows that the truncation at 0.9/1945 provides a much better fit:

```
AICtab(f.beta, f.betat.1, f.betat.5, f.betat.9)

##           dAIC   df
## f.betat.9    0.0 2
## f.betat.5 1112.7 2
## f.betat.1 2528.3 2
## f.beta    3270.3 2
```

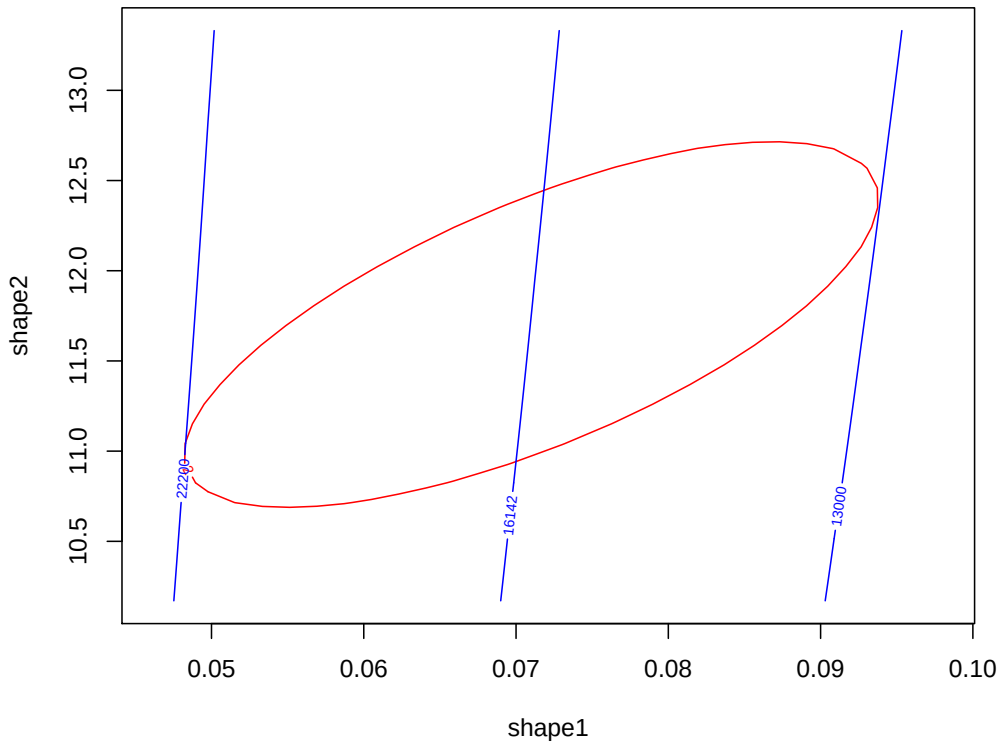
Estimate of species richness

If we assume that species occur in the plots independently with different values of occupancies (described by the beta distribution fitted above), the dis-

tribution of frequencies (number of occupied plots) follows a **beta-binomial distribution**. Because species that have not occurred in any plot are undetected, the beta-binomial distribution of observed frequencies is truncated at zero. The full distribution nevertheless returns the probability of zero frequency, which estimates the proportion of species not recorded. With this and the number of observed species we can estimate the total number of species. I did a function to do this (see scripts at `functions.R`) and got the following:

Truncation	Estimated richness
None	5642
0.1/N plots	6458
0.5/N plots	16142
0.9/N plots	18099397

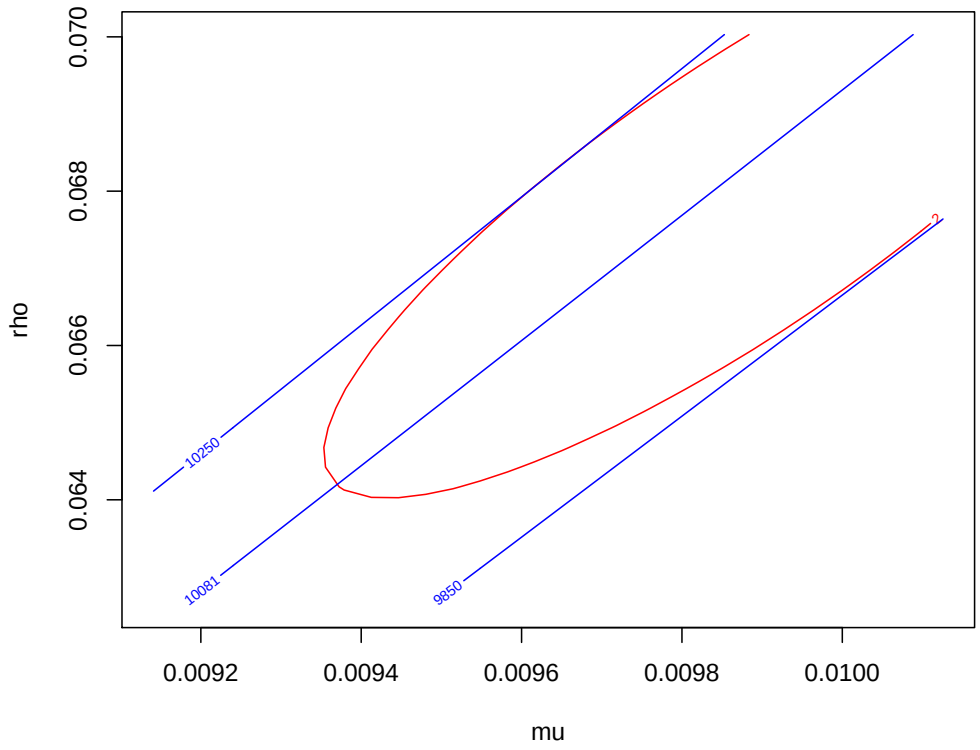
The distribution truncated at 0.9/1945 gives unrealistic estimates, so I'll follow with the model with truncation at 0.5/1945. With the likelihood surface we get an idea of likelihood interval:



Truncated beta-binomial

Another easy way to get species richness estimates from the beta-binomial is to fit this distribution directly to the observed frequencies. This distribution is zero-truncated to account for the unobserved species. In this case only the binomial distribution that composes the beta-binomial is truncated. To truncate also the beta compounding distribution we would need Bayesian models.

With this model I got an estimated total richness of 1.0081×10^4, with an likelihood interval below:



Non-parametric estimates

Results of the non-parametric indices available in the package *SPECIES* (Wang, 2011). All around 5000 species.

```
Ab <- as.data.frame(table(dados$N.ind))
names(Ab) <- c("j", "n_j")
Ab$j <- as.integer(as.character(Ab$j))
##jackknife method
(Sj <- jackknife(Ab,k=5))

##
## Your specified order is larger than that determined by the test,
```

```
## Therefore the order from the test is used.
## $JackknifeOrder
## [1] 2
##
## $Nhat
## [1] 5737
##
## $SE
## [1] 55.53377
##
## $CI
##        lb   ub
## [1,] 5628 5846

##ACE coverage method
(SChao92 <- ChaoLee1992(Ab,t=10, method="all"))

## $Nhat
## [1] 5399 5485
##
## $SE
## [1] 33.21531 44.06319
##
## $CI
##         lb   ub
## ACE   5334 5464
## ACE-1 5400 5573

##Chao1984 lower bound estimator
(SChao84 <- chao1984(Ab))

## $Nhat
## [1] 5441
##
## $SE
## [1] 47.05694
##
## $CI
```

```
##        lb   ub
## [1,] 5358 5544

##Chao and Bunge coverage-duplication method
(SChaoB <- ChaoBunge(Ab,t=10))

## $Nhat
## [1] 5542
##
## $SE
## [1] 55.42057
##
## $CI
##        lb   ub
## [1,] 5435 5652

##Penalized NPMLE method
(SNPMLE <- pnpmle(Ab,t=15))

## Method: Penalized NPMLE method by  Wang and Lindsay 2005.
##
##         MLE=                                    5551
##         Estimated zero-truncated Poisson mixture components:
##         p=                                      0.9153796 2.048484 4.198693 9.77843
##         pi=                                     0.3390338 0.0835869 0.2598503 0.317
## $Nhat
## [1] 5551

##Unconditonal NPMLE method
(SUNPMLE <-unpmle(Ab,t=10))

## Method: Unconditional NPMLE method by Norris and Pollock 1996, 1998,
##         using algorithm by Wang and Lindsay 2005:
##
##         MLE=                                    5608
##         Estimated Poisson mixture components:
##         p=                                      0.370657 1.411825 6.183764
##         pi=                                     0.1811724 0.4804983 0.3383293
## $Nhat
## [1] 5608
```

```
##Poisson-compound Gamma method
(SPG <- pcg(Ab,t=20))

## Method: Poisson-Compound Gamma method by Wang 2010.
## Alpha grid used: 1 2 3 4 5 6 7 8 9 10 .
##
##         MLE=                                    5576
##         Selected alpha model:                   Inf
##         Estimated Gamma components:
##         p=                                      0.734205 2.246373 6.79769 14.67896
##         pi=                                     0.2218206 0.2631206 0.286423 0.2286
## $Nhat
## [1] 5576
##
## $AlphaModel
## [1] Inf
```

Wrap up

Findings

- There is a gap between the estimates, rather than a more uniform spread over the range of values. We have a cluster of estimates around 5000-6000 and other around or above the known number of species.

- Using a strict approach of selecting the best SAD model with AIC, negative binomial wins. As noted by Tovo et al. (2017), NB provides a lower total richness estimate compared to Log-series. But now it is still a realistic value (1.1038668×10^4). The diagnostic plots show that is so because LS overestimate singletons more than NB. If we go for NB we would check if/how to make the hypothetical SAD for the whole Amazonia from a Negative Binomial, as the codes by Steege et al. do.

- The realistic estimates (that is, above the known number of species) came from methods that approximate some effect of a large beta-diversity. Such beta-diversity effect is approximated in different ways. Log-series assumes an open metacommunity with an infinite number of

species, negative binomial describes SADs with a variance higher than those expected by an independent distribution of species, beta-binomial models capture spatial aggregation of species.

Open questions

- Which models to include?
  - I would not use the 1st method to estimate S from Poisson-Lognormal, nor the truncated beta method. Agree?
  - On the other hand, not so sure if we should include truncated beta-binomial. **Pros**: raises an simple, alternative way to estimate (occupancy-based); realistic estimates based on some assumption of beta-diversity. **Cons**: shift the focus from SADs (best keep for another paper?); need to search in the literature to check if/how this method has already been used.
- Keep the estimation based the reconstruction of the regional SAD? If so, try to do that for negative binomial? If we stick to the current reconstruction based on LS, how to deal with the better performance of NB from model selection?

More (and some redundant) thoughts

Oldies, kept for the records

I can see a paper that reviews comprehensively the species richness estimates we can get from this data. Of course one important information is the known number of species recorded in the region. There is a gap between the estimates, rather than a more uniform spread over the range of values. We have a cluster of estimates around 5000-6000 and other around the known number of species. Another idea is to do the same for Mata Atlantica. A nice main figure would be a point-and-error-bars of all estimates for both ecorregions.

The estimates provided by non-parametric and also the parametric methods with strong assumptions of random and independent dispersion of species across space (which implies in describing the plots as Poisson samples) were consistently below the number of species currently recorded for the Amazon.

On the other hand, parametric estimates based on models that approximates some effect of a large beta-diversity provided consistently larger values, which encompassed the known number of species. Such beta-diversity effect is approximated in different ways. Log-series assumes an open metacommunity with an infinite number of species, negative binomial describes SADs with a variance higher than those expected by an independent distribution of species, and beta and beta-binomial models capture spatial aggregation of species. I think that the key message here is that beta-diversity matters a lot to get a better estimate. And also that those models that performed better are still rough approximations.

Also, I think that the comparison of the saturation curves of the estimators you did for the Amazon is nice. Many of them seem to stabilize well below the known number of species, while others do show an increasing trend. I think this may be valuable to discuss the quality of the estimates available for the Amazon. I would not venture in general evaluations of the estimators because there is already a ton of papers about that, so I think we do not need the simulated data.

Next steps

Not sure if for the same paper, but I would proceed with the idea of a model that improves assumptions about a large beta-diversity. My first guess is a minimum model that assumes different degrees of aggregation for each species. My first choice is a compound Negative binomial-Lognormal distribution. This is hard to fit by traditional methods, but it is doable with simulation and ABC (Approximate Bayesian Computation). To parameterize simulations we can use the empirical matrix of species x sites. I started to work in this idea with Renato and I have some preliminary R codes of a bootsrap estimator. The *SADISA* package uses a similar approach (using the information of replicate samples from the same metacommunity).

A related topic is the idea of SOD's (Species Occupancies Distributions), which depends of checking if we are not reinventing the weel. But I think that combining SODs and SADs in a bivariate model has not been fully explored yet.

References

Bulmer M. G. On fitting the Poisson lognormal distribution to species abundance data // Biometrics. 1974. 30. 651–660.

Saether Bernt-Erik, Engen Steinar, Grøtan Vidar. Species diversity and community similarity in fluctuating environments: parametric approaches using species abundance distributions. // The Journal of Animal Ecology. apr 2013. 721–738.

Tovo Anna, Suweis Samir, Formentin Marco, Favretti Marco, Volkov Igor, Banavar Jayanth R, Azaele Sandro, Maritan Amos. Upscaling species richness and abundances in tropical forests // Science Advances. 2017. 3, 10. e1701438.

Wang Ji-Ping. SPECIES: An R Package for Species Richness Estimation // Journal of Statistical Software. 2011. 40, 9. 1–15.

Wang Ji-Ping Z, Lindsay Bruce G. A penalized nonparametric maximum likelihood approach to species richness estimation // Journal of the American Statistical Association. 2005. 100, 471. 942–959.