Species richness estimation revisited – An effective computation method

Song Qian[1], Mark R. DuFour[2], Sabrina Jaffe[1], Corbin Hilling[2]

[1]The University of Toledo and [2]USGS

SFS 2024 Philadelphia PA
June 6, 2024

Materials covered in this presentation will be uploaded to `https://github.com/songsqian/missingSP`

In This Talk

- A Bayesian hierarchical model for Estimating biodiversity/species richness
    - An efficient computation strategy
    - Capable of pooling data from multiple sources
- Model evaluation (comparing to other existing methods)
- Potential of evaluating biodiversity changes over time and space
- Discussions

Species Richness – difficult to observe and estimate

- Catching/observing rare species is always hard

Species Richness – difficult to observe and estimate

- Catching/observing rare species is always hard
    - Physically

Species Richness – difficult to observe and estimate

- Catching/observing rare species is always hard
  - Physically
  - Analytically

Species Richness – difficult to observe and estimate

- Catching/observing rare species is always hard
    - Physically
    - Analytically
- Species richness is always an approximation

Species Richness – difficult to observe and estimate

- Catching/observing rare species is always hard
  - Physically
  - Analytically
- Species richness is always an approximation
  - We can never observe the true richness

A Mixture Model – Data format

- Fisher et al (1943)
- Data format –
  - y_i – number of individuals from species/taxon i
  - Often recorded as n_j number of species (n) with j individual(s) in the sample.

A Mixture Model – Parametric model

- A Poisson-gamma mixture model

$$\begin{aligned} y_i \mid \lambda_i &\sim \mathit{Pois}(\lambda_i) \\ \lambda_i &\sim \mathit{gamma}(\alpha, \beta) \end{aligned}$$

A Mixture Model – Parametric model

- A Poisson-gamma mixture model

$$\begin{aligned} y_i \mid \lambda_i &\sim Pois(\lambda_i) \\ \lambda_i &\sim gamma(\alpha, \beta) \end{aligned}$$

- Predictive distribution of y_i: a negative binomial (NB) model

$$\pi(y_i) = \int_{\lambda_i} \pi(y_i \mid \lambda_i)\pi(\lambda_i) d\lambda_i$$

$$\Pr(y_i = k) = \frac{\Gamma(\alpha + k)}{\Gamma(\alpha) k!} \left(\frac{\beta}{\beta + 1}\right)^{\alpha} \left(\frac{1}{\beta + 1}\right)^{k}$$

A Mixture Model – Parametric model

- A Poisson-gamma mixture model

$$\begin{aligned} y_i \mid \lambda_i &\sim Pois(\lambda_i) \\ \lambda_i &\sim gamma(\alpha, \beta) \end{aligned}$$

- Predictive distribution of y_i: a negative binomial (NB) model

$$\pi(y_i) = \int_{\lambda_i} \pi(y_i \mid \lambda_i)\pi(\lambda_i) d\lambda_i$$

$$\Pr(y_i = k) = \frac{\Gamma(\alpha + k)}{\Gamma(\alpha) k!} \left(\frac{\beta}{\beta + 1}\right)^{\alpha} \left(\frac{1}{\beta + 1}\right)^{k}$$

- Reparameterizing: $\mu = \alpha/\beta$ and $r = \alpha$:

$$\Pr(y_i = k) = \frac{\Gamma(r + k)}{\Gamma(r) k!} \left(\frac{r}{\mu + r}\right)^{r} \left(\frac{\mu}{\mu + r}\right)^{k}$$

Meanings of μ and r

- μ – mean of y, average number of individuals among all species
- r – the dispersion parameter, measuring the differences/variance in number of individuals among species
    - A larger r – more evenly distributed and less clustering
    - A smaller r – more clustering

A Mixture Model – Species richness

- Under the NB model

$$\Pr(y = 0) = p_0 = \left(\frac{r}{\mu + r}\right)^r$$

A Mixture Model – Species richness

- Under the NB model

$$\Pr(y = 0) = p_0 = \left(\frac{r}{\mu + r}\right)^r$$

- The total number of species S:

$$S = S_{obs} + S_0 = S_{obs} + Sp_0$$

or

$$S = S_{obs}/(1 - p_0)$$

A Mixture Model – Hierarchical modeling

- Combining data from multiple sources (l)

$$\begin{aligned} y_{il} &\sim NB(\mu_l, r_l) \\ \log(\mu_l) &\sim N(\theta, \tau^2) \end{aligned}$$

Data for Model Evaluation

- Butterflies (Fisher et al 1943)

Data for Model Evaluation

- Butterflies (Fisher et al 1943)
  - Rothamsted Experimental Station (R): 15,609 individuals of 240 species from the Lepidoptera family (1933-1936)

Data for Model Evaluation

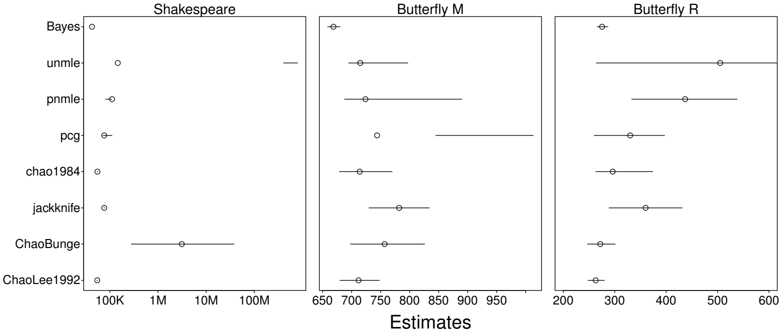
- Butterflies (Fisher et al 1943)
  - Rothamsted Experimental Station (R): 15,609 individuals of 240 species from the Lepidoptera family (1933-1936)
  - Malaysia (M): Part of the effort of documenting all butterfly species in Malaysia, only species with up to 24 individuals were included in the data. 119 species with $>$ 24 specimens were recorded as having 25 specimens.

Data for Model Evaluation

- Butterflies (Fisher et al 1943)
  - Rothamsted Experimental Station (R): 15,609 individuals of 240 species from the Lepidoptera family (1933-1936)
  - Malaysia (M): Part of the effort of documenting all butterfly species in Malaysia, only species with up to 24 individuals were included in the data. 119 species with > 24 specimens were recorded as having 25 specimens.
- Shakespeare's Vocabulary – A total of 884,647 words in Shakespeare's known works tabulated in n_j for $j = 1, \cdots, 100$. For words with > 100 appearances: one recorded as appeared 1,305 times and 845 appeared 815 times.

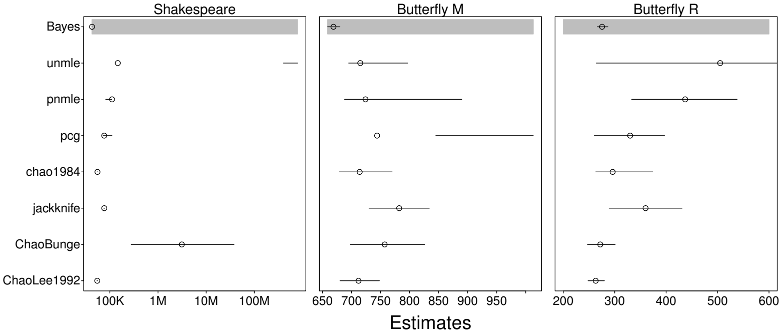
Model Evaluation – Comparisons

Comparing to 7 methods included in R package `SPECIES`



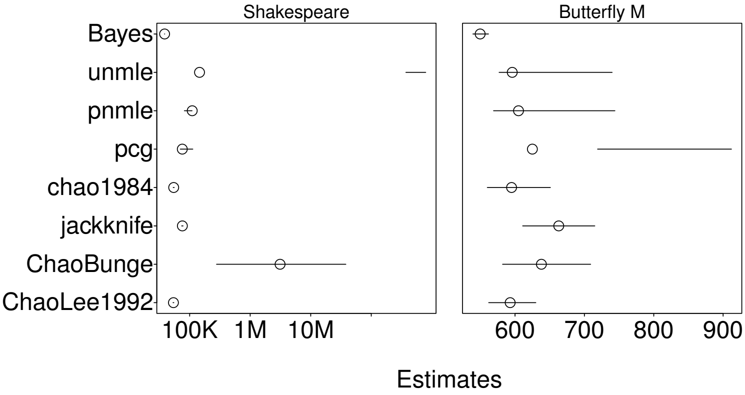
Model Evaluation – Comparisons

Comparing to 7 methods included in R package `SPECIES`



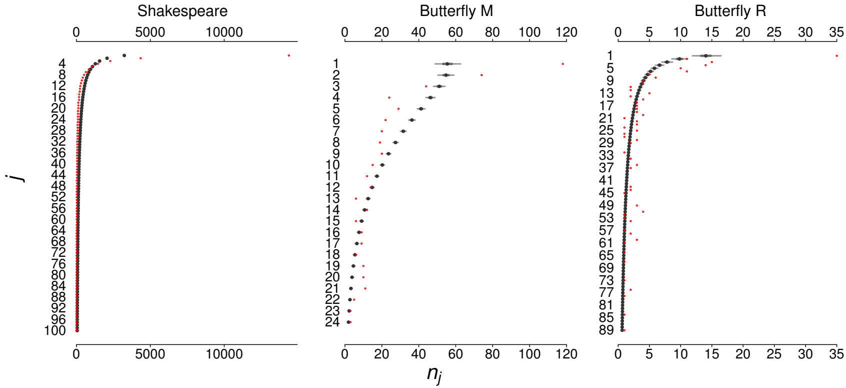
Model Evaluation – Comparisons

Using only accurately recorded data from SK and Butterfly M



Model Evaluation – Goodness-of-fit

Posterior simulation: model predicted number of species n_j with $j, j = 1, 2, 3, ..., J$ specimens



Data for Illustrating BHM

- USGS trawl fishing data for evaluating fish recruitment

Data for Illustrating BHM

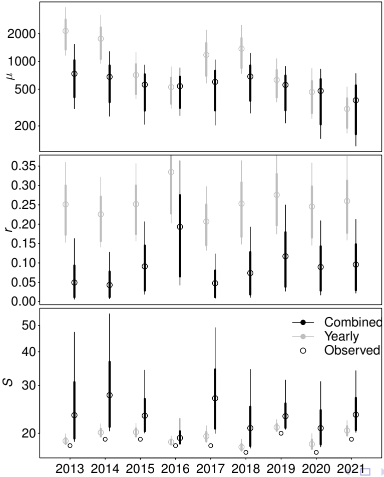
- USGS trawl fishing data for evaluating fish recruitment
    - Data used 2013-2021, age-0 catch data

Data for Illustrating BHM

- USGS trawl fishing data for evaluating fish recruitment
  - Data used 2013-2021, age-0 catch data
  - Numbers of species represented in each year's data are small ($\sim 10 - 20$)

Combining Data – BHM

Comparing results from BHM and individual year fits



Conclusions

- Estimated S is comparable to competing MLE-based methods when data were recorded accurately
- Much faster computation and smaller estimation uncertainty
- Combining data from similar sources is effective in evaluating changes in community structure

Parametric versus Nonparametric

- Parametric models are more sensitive to data error than nonparametric ones
- Issues with the Butterfly M and Shakespeare data sets are no longer important – record all samples accurately

The Dispersion Parameter r

- Likely overestimated (underestimate the dispersion) with small sample size
- BHM is a sensible way to share information among multiple sources of data

“True” Richness Impossible

- Cannot compare the model estimate to the observed – true richness unobservable
- Model parameters μ and *r* describe species composition structure
  - Changes of μ and *r* over time or space can be more informative

Acknowledgment

Funding of the project provided by GLRI through CESU and Great Lakes Science Center, USGS

Kate Curtis, Lou Trois, Tatsaneewan Phoesri, and Ly Trung Nguyen for comments and discussion

Thank You

Questions/Comments?

Computational details are at `github.com/songsqian/missingSp`