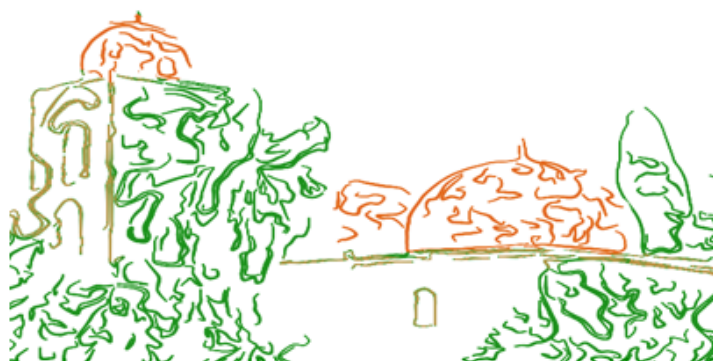


PROCEEDINGS

Edited By: Giada Adelfio and Antonino Abbruzzo

---

# GRASPA 2023



## GRASPA-SIS BIENNIAL CONFERENCE

The Researcher Group for Environmental Statistics of The Italian Statistical Society

## TIES EUROPEAN REGIONAL MEETING

The International Environmetrics Society

**Palermo, 10-11 July, 2023**

Dipartimento di Scienze Economiche Aziendali e Statistiche, Università degli Studi di Palermo

---



**Università  
degli Studi  
di Palermo**

Proceedings of the GRASPA 2023 Conference  
Palermo, 10-11 July 2023  
Edited by: Giada Adelfio and Antonino Abbruzzo

–  
Palermo: Università degli Studi di Palermo.

**ISBN:** 979-12-210-3389-2

Questo volume è rilasciato sotto licenza Creative Commons  
**Attribuzione - Non commerciale - Non opere derivate 4.0**



© 2023 The Authors

Sponsored by:



Stazione  
Zoologica  
Anton Dohrn  
Napoli



ISTITUTO NAZIONALE  
DI GEOFISICA E VULCANOLOGIA



---

---

# Contents

<b>1</b>	<b>Keynote Sessions</b> .....	<b>9</b>
1.1	Modeling Extremal Streamflow using Deep Learning Approximations and a Flexible Spatial Process. <i>R. Majumder, B. J. Reich and B. A. Shaby</i>	9
1.2	Estimating Covid-19 transmission time using Hawkes point processes. <i>F. P. Schoenberg</i>	11
<b>2</b>	<b>Complex data and models</b> .....	<b>13</b>
2.1	O2S2 for uncertainty quantification in natural background level concentrations. <i>A. Menafoglio, L. Guadagnini, A. Guadagnini and P. Secchi</i>	13
2.2	Penalized multivariate hidden semi-Markov models for time series environmental data. <i>M. Mingione, P. Alaimo Di Loro, F. Lagona and A. Maruotti</i>	16
2.3	Statistical analysis of complex networks. <i>N. Pronello, A. Cucco, R. Ignaccolo and L. Ippoliti</i>	22
<b>3</b>	<b>Statistical Methods and Environmental Studies</b> .....	<b>27</b>
3.1	ARPA Sicily: collection of environmental data, for an integrated vision of the regional territory and to support the "One Health" approach. <i>G. Cuffari and A. Conti</i>	27
3.2	Exploring the effects of temperature on demersal fish communities in the Central Mediterranean Sea using INLA-SPDE modeling approach. <i>C. Rubino, G. Adelfio, A. Abbruzzo, M. Bosch-Belmar, F. Colloca, M. Di Lorenzo, F. Fiorentino, V. Gancitano and G. Milisenda</i>	33

3.3	Analytical Methods of the Respiratory Epidemiological Surveys carried out by the National Research Council. <i>G. Viegi, S. Maio, G. Sarno, S. La Grutta and S. Baldacci</i>	39
<b>4</b>	<b>Measuring and modelling of environmental</b>	<b>45</b>
4.1	A Bayesian Time Series Model for Reconstructing Hydroclimate from Multiple Proxies. <i>N. Cahill, J. Croke, M. Campbell, K. Hughes, J. Vitkovsky, J. E. Kilgallen and A. Parnell</i>	45
4.2	Some empirical results on nearest neighbour pseudo-populations for re-sampling from spatial populations. <i>R.M. Di Biase, A. Marcelli, S. Franceschi and L. Fattorini</i>	52
4.3	Kriging Riemannian data for environmental applications. <i>A. Menafoglio, D. Pigoli and P. Secchi</i>	58
<b>5</b>	<b>Spatio-temporal modelling for urban</b>	<b>61</b>
5.1	Stochastic reconstruction of a spatio-temporal Hawkes process with isotropic excitation: an application to road accidents. <i>P. Alaimo Di Loro and M. Mingione</i>	61
5.2	ARPALData: retrieving and analyzing air quality and weather data of Lombardy (Italy). <i>P. Maranzano and A. Algieri</i>	68
5.3	Mechanistic spatio-temporal modeling of infectious diseases and crime data on urban environments. <i>J. Mateu and A. Brizi</i>	74
<b>6</b>	<b>Advances in statistical ecology</b>	<b>81</b>
6.1	Spatial occurrence models using Integrated Nested Laplace Approximation. <i>S. Martino, J. Belmont Osuna, J. Illian and H. Rue</i>	81
6.2	Data analysis of photogrammetry-based mapping: the seacucumbers in the Giglio Island as a case study. <i>G. Mastrantonio, D. Ventura, E. Casoli, A. Rakaj, G. Jona Lasinio and A. Pollice</i>	83
6.3	A species distribution modelling framework for combining citizen science data from different monitoring schemes. <i>J. Belmont Osuna, S. Martino, G. Panunzi, and J. Illian</i>	84
<b>7</b>	<b>Controlling for unmeasured confounding</b>	<b>85</b>
7.1	On the effect of spatial confounding: an approach based on the theory of quadratic forms. <i>F. Greco and M. Narcisi</i>	85
7.2	A Flexible Bayesian Time-varying Coefficient Regression Model in Health Applications. <i>C. Zaccardi, P. Valentini and L. Ippoliti</i>	92
7.3	Confounder-Dependent Bayesian Mixture Model: Characterizing Heterogeneity of Causal Effects in Air Pollution Epidemiology. <i>D. Zorzetto, F. J. Bargagli-Stoffi, A. Canale and F. Dominici</i>	98
<b>8</b>	<b>Posters</b>	<b>99</b>
8.1	Spatio-temporal analysis of the distribution of larval round sardinella in Mediterranean coastal area. <i>A. Abbruzzo, A. Granata, B. Patti, A. Cuttitta and M. Torri</i>	100

- 8.2 A Novel Spatio-Temporal Estimation Method for Occurrences Over Planar and Curved Regions . *B. Begu , S. Panzeri , E. Arnone and L. M. Sangalli* 107
- 8.3 Addressing Sequestration Bias in Estimating Malaria Parasite Age Distribution: A Bayesian Statistical Model. *E. Bortolato, P. E. Jacob, and J. A. Watson* 113
- 8.4 Accounting for Misreporting in Spatial Zero-Inflated Poisson Models. *C. Calculli, S. Arima and A. Pollice* 118
- 8.5 Penalized quantile regression for spatial distributed data. *C. Castiglione, E. Arnone, M. Bernardi, A. Farcomeni, L. M. Sangalli* 124
- 8.6 Sign-Flip tests for the nonparametric component in Spatial Regression with PDE regularization. *M. Cavazzutti, E. Arnone, F. Ferraccioli, L. Finos, C. Galimberti, L.M. Sangalli* 130
- 8.7 Environmental exposure indicators and respiratory health in asthmatic children: a case study. *G. Cilluffo, G. Ferrante, S. Fasola , V. Malizia, L. Montalbano, A. Ranzi , C. Badaloni, G. Viegi and S. La Grutta* 135
- 8.8 Non-parametric density estimation over linear networks. *A. Clemente, E. Arnone, J. Mateu and L. M. Sangalli* 141
- 8.9 Selecting the Kth nearest-neighbour for clutter removal in spatial point processes through segmented regression models . *N. D'Angelo and G. Adelfio* 146
- 8.10 Modeling linkage errors in species diversity estimates: an ABC approach. *D. Di Cecco, and A. Tancredi* 152
- 8.11 The effect of deforestation on infant health: a multilevel mediation analysis. *C. Di Maria* 158
- 8.12 Geostatistical methods comparison: an application on the concentration of PM2.5 in Lombardy. *A. Fassò, J. Rodeschini and A. Fusta Moro* 164
- 8.13 Application of the hierarchical variance decomposition approach to an ecological case study with mixed effects. *L. Ferrari, M. Ventrucci and A. Laini* 170
- 8.14 Variables selection in a P-spline regression model for flood frequency analysis. *A. Gardini* 176
- 8.15 Assessment of death risk for asbestos cancers using functional regression models among dockworkers exposed to asbestos in northeastern Italy. *P. Girardi, V. Comiati, V. Casotto, C. Gaetan, M.N. Ballarin, E. Merler and U. Fedeli* 182
- 8.16 Confidence predictive distributions: an application to temperature forecasting in Veneto. *F. Giummolè and V. Mamei* 188
- 8.17 Particulate matter PM10 assessment by meteorological conditions to monitor extractive activities' impact: a case of study. *E. Iakimova, F. Condino and F. Domma* 193
- 8.18 Health and environmental risk perception measurement and classification: a case study in Gela (Sicily). *G. Ilardo, A. Pandolfo, V. Malizia, G. Cilluffo, S. Fasola, A. Bonomolo, G. Piva, L. Cori, G. Viegi and S. La Grutta* 199
- 8.19 Severe storms events' reproduction in the United States of America: evaluation from the marked self-exciting point processes point of view. *F. Lagona and M. Mingione* 205

- 8.20 A parametric hidden semi-Markov model for toroidal time-series. *G. Lo Galbo, G. Adelfio and M. Chiodi* 211
- 8.21 Design-based mapping of land use/land cover classes with bootstrap estimation of precision by nearest-neighbour interpolation. *A. Marcelli, R.M. Di Biase, P. Corona, S.V. Stehman and L. Fattorini* 217
- 8.22 BayesSizeAndShape: a Julia package for Bayesian estimation of Size and Shape data regression models. *G. Mastrantonio and G. Jona Lasinio* 223
- 8.23 Space-time clustering of seismic events in Chile. *O. Nicolis, L. Delgado, B. Peralta and M. Díaz* 228
- 8.24 Forecasting Air Pollutants through Artificial Neural Networks. *E. Nissi and L. Saliaj* 234
- 8.25 Functional principal component analysis for space-time data. *A. Palummo, E. Arnone, L. Formaggia and L. M. Sangalli* 239
- 8.26 Association between vegetation index and asthma in children and adolescents in North-Central Europe. *A. Pandolfo, G. Ilardo, V. Malizia, G. Squillacioti, S. Fasola, F. Ghelli, G. Ferrante, S. La Grutta, G. Viegi and R. Bono* 245
- 8.27 A taxonomy for Random Forest in the spatial regression framework. *L. Patelli, M. Cameletti, N. Golini and R. Ignaccolo* 251
- 8.28 Fish characteristics and microplastic ingestion: a mediation analysis of fish length and trophic level in Western Mediterranean pelagic demersal fish. *M. Sciandra, A. Albano, A. Plaia, C. Di Maria, C. Andolina, S. Vizzini and M.C. Fossi* 257
- 8.29 UAV plant image Classification Using Combined Machine Learning And Deep Learning Models. *A. Simonetto, G. Tariku and G. Gilioli* 262
- 8.30 A Bayesian study of temporal changes in seismicity. *E. Varini and R. Rotondi* 268
- 8.31 On time lag detection between time series sampled by eddy covariance systems. *D. Vitale and D. Papale* 273

# Modeling linkage errors in species diversity estimates: an ABC approach

D. Di Cecco<sup>1,\*</sup> and A. Tancredi<sup>1</sup>

<sup>1</sup> Department Memotef, University of Rome La Sapienza; [davide.dicecco@uniroma1.it](mailto:davide.dicecco@uniroma1.it), [andrea.tancredi@uniroma1.it](mailto:andrea.tancredi@uniroma1.it)

\*Corresponding author

---

**Abstract.** *The estimation of species diversity of ecological communities relies on surveying species abundances, that is, counting the number of units by species in a sample. Diversity estimators are particularly sensitive to rare species, that is, to low abundance cases. In microbial studies, rare species, in particular singletons, often represent the vast majority of the specimens in a sample. Many studies hypothesize the spurious nature of these cases, and various methodological contributions focus on estimating and eliminating the spurious singletons to avoid a gross overestimation of the total diversity of a community. We present a different approach that treats the spurious singletons as the result of false negative errors in the clustering step of the RNA sequencing. We demonstrate that the estimation of the total number of species under our scenario is equivalent to that one can obtain by discarding spurious cases. On the converse, diversity as measured by Shannon's index for example, can differ considerably. The computation of such index requires to estimate all true abundances counts, which appears to be computationally challenging. We then propose a likelihood-free Bayesian approach to the problem.*

**Keywords.** *Microbial Diversity; Sequencing Errors; Linkage Errors; Approximate Bayesian Computation.*

---

## 1 Introduction

The development in recent years of next generation high-throughput sequencing technology has provided the capability of analyzing an unprecedented number of DNA and RNA sequences. This technology has greatly benefited the study of microbial communities. In microbial diversity studies, environmental samples are processed in order to detect, amplify and sequence RNA genomes. The sequences are clustered into distinct species (or Operational Taxonomic Units) according to their similarities in specific genomic regions. The species are then counted by abundance, (i.e., by the number of specimens representing it). Those counts are utilized to estimate the microbial community diversity. The main interest is in estimating the “total diversity”, that is, the total number of distinct species, captured or not, and in evaluating the heterogeneity of the species structure of the community. For the latter purpose, Hill's numbers are a common choice, representing a family of indexes with different degrees of sensitivity to the species abundances. Shannon's and Simpson's indexes represent particular cases. This kind of analysis can provide an indication of an ecosystem health; for example, in analysing the human gut microbiota, one can measure and compare the richness of the same community in different occasions (e.g., before and after an antibiotic treatment).

The species problem in the microbial context appears to have some peculiarities. While in macroecological studies rare species represent a small portion of the total abundance, microbial datasets are characterized by a large number of low abundance cases. In particular, singletons, that is, species represented by one specimen in the sample, often constitute the vast majority (up to 80%) of the number of observed species.

Several authors seem to agree on the fictitious nature of many rare species in microbial diversity studies (see, e.g., [8]). Even if there is no clear consensus on the frequency of such errors, or on the way their distribution is associated to the sample composition, their presence is confirmed in several ways. Experiments on samples from population with known diversity, both simulations with mock databases, and communities cultivated *in vivo*, confirmed an important overcount of the rare species, with consequent overestimation of the observed and total diversity.

An error in any phase of the bioinformatics pipeline can produce spurious singletons [7]. Errors in RNA sequencing produce artefactual sequences (known as chimeric sequences). Those sequences will constitute, in general, novel species which cannot be matched with any other sequence. A chimera removal step is included in many of the widely used pipelines [4]. Nonetheless, apparently, their presence is confirmed even in various curated databases of S16 sequences [5].

Since the complete removal of fictitious sequences in a pre-analysis step does not appear to be feasible at the moment, various methodological contributions aim at estimating and removing the spurious cases count (in particular singletons). See, for example, [2], [10], [11], [3]. We will call this approach where we remove units (possibly) affected by error, “discounting”. This approach is consistent for a model where spurious singletons are added to the baseline counting distribution of the true abundances. This model then results in a mixture of a counting distribution (truncated in zero as the number of uncaptured species has to be estimated), and a Dirac’s measure modeling the spurious singletons. Various authors propose to completely ignore the observed singletons and base the estimation solely on the other (supposedly error-free) counts.

We believe that the nature of the spurious cases is best described by linkage errors. That is, we assume that random errors occurring in sequencing result in the impossibility of a correct classification of the specimen, which cannot be associated to the correct existing species. Therefore, we can describe these cases as false negative linkage errors (or missing links), which are added to the true singletons. This approach implies a re-estimation of the “real” frequency counts for all the abundances, not just the singletons. We found that treating the excess of singletons in this way leads to significant differences in the diversity estimates with respect to the discounting approach.

In this work we focus on a secondary perspective to the linkage problem, that is, we assume the (not uncommon) condition of not having access to the actual record linkage process. Instead, we admit the possibility that our observed counts data are affected by linkage errors, and include them in our estimation process. Modeling record linkage errors in this secondary setting, appears as a computationally formidable task. We fix some simplifying assumptions on the type of error in order to tackle the issue. In particular, we assume that

- we just have missing links and no false positive record linkage errors;
- missing links create additional singletons only;
- each captured specimen has the same probability of being mistakenly classified as a singleton independently of the other.



Despite these assumptions, we resorted to a Bayesian likelihood-free approach as the most convenient method in order to estimate the true observed abundances.

## 2 The missing links model

Say we observed  $n$  species in our sample with abundances  $y_1, \dots, y_n$ . Let  $n_j$  be the number of species with  $j$  observed captures, such that  $\sum_{j \geq 1} n_j = n$ . We denote as  $D$  the observed data  $(n_1, n_2, \dots)$ . We assume that these counts can be affected by linkage errors, in particular, a portion of the singletons are due to missing links errors, and the number of true distinct captured species  $n^*$  is a portion of  $n$ . Let  $n_j^*$  be the number of true species with  $j$  captures,  $j \geq 0$ , and let  $N^*$  be the total number of distinct species (captured or not),  $N^* = \sum_{j \geq 0} n_j^*$ . Let  $X_i^*$  be the latent true number of times species  $i$  has been captured in the sample,  $X_i^* \geq 0$ ,  $i = 1, \dots, N^*$ . The generating process of our missing links model is the following: each specimen has the same probability  $\mu$  of being missclassified as a singleton. Then, for each species  $i$  captured  $X_i^*$  times we have  $M_i$  missing links, such that the registered abundance for species  $i$  is reduced from  $X_i^*$  to  $X_i = X_i^* - M_i$ . Then,  $M_i$  has the following Binomial conditional distribution:

$$P(M_i = m_i | X_i^* = x_i^*) = P(X_i = x_i^* - m_i | X_i^* = x_i^*) = \binom{x_i^*}{m_i} \mu^{m_i} (1 - \mu)^{x_i^* - m_i}. \quad (1)$$

Let us denote as  $f^*(\theta)$  the baseline counting distribution of the true captures  $X_i^*$ , defined up to a parameter  $\theta$ :  $\{f_j^*(\theta)\}_{j=0,1,\dots} = \{P(X^* = j)\}_{j=0,1,\dots}$ . Clearly, a missing links model is completely defined by  $f^*(\theta)$  and the probability  $\mu$ .

In order to estimate all counts  $n_0^*, n_1^*, n_2^*, \dots$ , we adopt a Bayesian likelihood-free approach exploiting the generating process of our model described above. We first adopt an ABC rejection sampling and then a sequential ABC to accelerate the estimation procedure as described in [6].

Mixtures of Poisson are a popular parametric choice for the baseline  $f^*(\theta)$ . For the sake of simplicity, in the present work we just consider Poisson and Geometric families for the baseline distribution  $f^*(\theta)$ . The prior setting is the following: We set  $\pi(\theta)$  as the appropriate conjugate prior depending on  $f^*$ :  $\lambda \sim \text{Gamma}(\alpha_\lambda, \beta_\lambda)$  for a Poisson baseline of parameter  $\lambda$ , and  $p \sim \text{Beta}(\alpha_p, \beta_p)$  for a Geometric of parameter  $p$ . For the linkage error probability, we set  $\mu \sim \text{Beta}(\alpha_\mu, \beta_\mu)$ . Finally, we set an improper prior over  $N^*$  in the family  $\pi(N^*) \propto 1/(N^*)^k$  (see, e.g., [1] for a sensible choice of  $k$ ). All priors are independent of one another.

### 2.1 Rejection ABC algorithm

Note that the total number of captured specimens in our sample, call it  $s$ , remains unaltered under a missing links model. That is, we have

$$\sum_{i=1}^n y_i = \sum_{i=1}^{n^*} x_i^* = \sum_{j \geq 1} j n_j = \sum_{j \geq 1} j n_j^* = s.$$

As a consequence, while the ABC rejection algorithm, in its simplest form, utilizes draws of the parameters from the (independent) priors, in order to exploit all the available information, we want to generate values of the parameters conditionally on the observed value  $s$ . That is, we adopt the following scheme:

1. generate values for  $(\theta, N^*)$  given  $s$  and the priors  $\pi(\theta)$  and  $\pi(N^*)$
2. generate values  $(n_0^*, n_1^*, n_2^*, \dots)$  conditional on  $N^*$ ,  $\theta$  and  $s$
3. generate a value for  $\mu$  from the prior  $\pi(\mu)$
4. generate values  $\tilde{D} = (n_0^*, \tilde{n}_1, \tilde{n}_2, \dots)$  by simulating missing links over  $(n_0^*, n_1^*, n_2^*, \dots)$  given  $\mu$
5. retain the current generated values if a measure of distance  $\rho$  between the generated data  $\tilde{D}$  and the observed data  $D$  is below a certain threshold  $\varepsilon$ :

$$\rho(\tilde{D}, D) < \varepsilon.$$

In the first step, under a Poisson or a Geometric baseline distribution, we can derive the analytical form of  $P(N^* | s, \pi(\theta), \pi(N^*))$  and generate values of  $N^*$  accordingly.  $(N^*, s)$  constitute a sufficient statistic for the Poisson and the Geometric. Hence, the posterior distribution  $\pi(\theta | N^*, s)$  is easily derived and values of  $\theta$  can be easily sampled.

In step 2., we note that the distribution of  $(n_0^*, n_1^*, n_2^*, \dots)$  conditional on  $N^*$  and  $s$  is independent of  $\theta$ . In fact, the joint distribution of  $N^*$  independent Poisson having fixed sum  $s$  is Multinomial with constant probabilities:

$$P((x_1^*, \dots, x_{N^*}^*) | N^*, s) = \text{Mult}(s, (1/N^*, \dots, 1/N^*)),$$

and, consequently,

$$P((n_0^*, n_1^*, \dots, n_s^*) | N^*, s) = \binom{N^*}{n_0^* \dots n_s^*} \binom{s}{y_1^* \dots y_{N^*}^*} \frac{1}{(N^*)^s}.$$

Then we can easily generate values for  $(n_0^*, n_1^*, n_2^*, \dots)$ . Similarly, under a Geometric assumption, all vectors  $(x_1^*, \dots, x_{N^*}^*)$  having fixed sum  $s$  have the same probability regardless of  $p$ , which is then equal to the reciprocal of the number of possible nonnegative integer  $N^*$ -vectors summing to  $s$  (or weak  $N^*$ -compositions of  $s$ ):  $\binom{N^* + s - 1}{s}^{-1}$ , that is,

$$P((n_0^*, n_1^*, \dots, n_s^*) | N^*, s) = \binom{N^*}{n_0^* \dots n_s^*} \binom{N^* + s - 1}{s}^{-1}.$$

As a consequence, we can generate  $(x_1^*, \dots, x_{N^*}^*)$  with fixed sum  $s$  with an algorithm to generate a random compositions, (see, e.g., [9]).

In step 3. we simply generate from the prior as  $\mu$  is independent of  $N^*$  and  $s$ . In step 4., we simply generate missing links at random according to the binomial described in (1), modify accordingly the observed counts, and increment the number of singletons. (Note that the missing links mechanism has no effect on  $n_0^*$ ). In step 5. we use the simple euclidean distance.

### 3 The effect of missing links on diversity

To illustrate the effect of (ignoring) a missing links mechanism on the estimation of diversity, we show the results of a simulation. As a measure of diversity we consider Shannon's diversity  $H$  (see, e.g., [3]) calculated as:

$$H = \exp \left( - \sum_{j \geq 1} n_j \frac{j}{s} \ln \frac{j}{s} \right). \quad (2)$$

We fix a scenario with  $N^* = 10000$  true species and a baseline Poisson distribution  $f^*$  of parameter  $\lambda = 3$ . We simulated a million datasets  $(n_1^*, n_2^*, \dots)$  from the baseline distribution, then, for each such sample, we generated 30 datasets  $(n_1, n_2, \dots)$  by simulating the effect of missing links errors for 30 different values of  $\mu$  ranging in  $[0, 0.25]$ . We calculated (2) for the original data and for the data affected by missing links. In addition, we considered the discounting procedure presented in [3] where the spurious singletons are eliminated from the observed data. The true number of singletons is estimated on the basis of this formula:

$$\tilde{n}_1 = \frac{2n_2^2}{3n_3} + 2n_2 \left( \frac{2n_2}{3n_3} - \frac{n_3}{4n_4} \right), \quad (3)$$

and Shannon’s diversity is calculated over the “adjusted” data  $(\tilde{n}_1, n_2, n_3, \dots)$ .

Figure 1 summarizes the simulation results. The horizontal line indicates the average value of  $H$  (8332) over the baseline datasets  $(n_1^*, n_2^*, \dots)$ . The continuous curve going upward represents the average values of  $H$  based on the simulated  $n_1, n_2, \dots$ , that is, what we would obtain by ignoring any one–inflation mechanism. The continuous curve going downward represents the average values of  $H$  based on the adjusted data obtained by substituting the value  $n_1$  with (3). That is, what we would obtain by assuming a discounting approach. The gray areas limited by red dotted lines represent the 95% confidence intervals.

Our Bayesian approach implies the generation of all counts  $(n_1^*, n_2^*, \dots)$  at each iteration of the algorithm. This allows us to estimate with ease the expectation of  $H$  under the missing links model. However, the computational complexity of the ABC approach does not allow the same number of simulations presented above. Then, we replicated the simulation 20 times with  $\mu = 0.1$  and 20 times with  $\mu = 0.2$  with 5000 generations for each ABC. To obtain 5000 accepted generations, each replication of the ABC required about 50 millions tries. The results were quite encouraging, producing an average of  $H$  equal to 8290 for  $\mu = 0.1$  and 8420 for  $\mu = 0.2$ .

In conclusion, we believe that our error model represents a sensible hypothesis for the generation of additional singletons. Ignoring an existing one–inflating mechanism of this kind, implies a severe overestimation of the diversity. A discounting approach reduce sensibly the error one commits, but still leads to different results than what can be achieved with an ABC simulating the actual generating process.

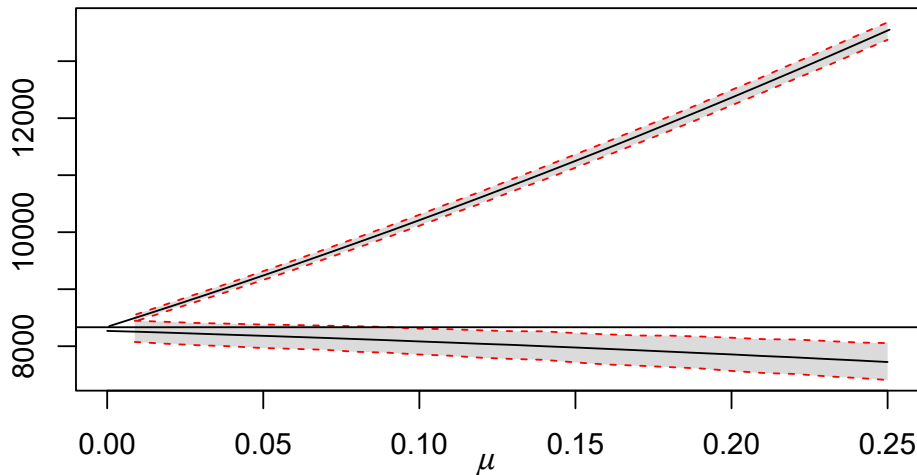


Figure 1: Shannon diversity over simulated data under different values of missing link probability  $\mu$  ranging in  $[0, 0.25]$ .

## References

- [1] Barger, K., Bunge, J. (2010). Objective Bayesian estimation for the number of species. *Bayesian Analysis*, **5(4)**, 765–785.
- [2] Bunge, J., Böhning, D., Allen, H., Foster, J. A. (2012). Estimating population diversity with unreliable low frequency counts. In *Biocomputing 2012: Proceedings of the Pacific Symposium* (pp. 203–212). World Sci. Publ. Hackensack, NJ.
- [3] Chiu, C. H., Chao, A. (2016). Estimating and comparing microbial diversity in the presence of sequencing errors. *PeerJ*, **4**, e1634.
- [4] Edgar, R. (2016). Unoise2: improved error-correction for illumina 16s and its amplicon sequencing. *BioRxiv*, page 081257.
- [5] Haas, B.J., Gevers, D., Earl, A.M., Feldgarden, M., Ward, D.V., Giannoukos, G., Ciulla, D., Tabbaa, D., Highlander, S., Sodergren, S., et al. (2011) Chimeric 16s rRNA sequence formation and detection in Sanger and 454-pyrosequenced pcr amplicons. *Genome research*, **21(3)**, 494–504.
- [6] Marin, J. M., Pudlo, P., Robert, C. P., Ryder, R. J. (2012). Approximate Bayesian computational methods. *Statistics and computing*, **22(6)**, 1167–1180.
- [7] Porter, T. M., Hajibabaei, M. (2018). Scaling up: A guide to high-throughput genomic approaches for biodiversity analysis. *Molecular ecology*, **27(2)**, 313–338.
- [8] Quince, C., Lanzen, A., Davenport, R. J., Turnbaugh, P. J. (2011). Removing noise from pyrosequenced amplicons. *BMC bioinformatics*, **12(1)**, 1–18.
- [9] Stojmenović, I. (1992). On random and adaptive parallel generation of combinatorial objects. *International journal of computer mathematics*, **42(3–4)**, 125–135.
- [10] Willis, A., Bunge, J. (2015). Estimating diversity via frequency ratios. *Biometrics*, **71(4)**, 1042–1049.
- [11] Willis, A. (2016). Species richness estimation with high diversity but spurious singletons. *arXiv preprint arXiv:1604.02598*.