A peer-reviewed version of this preprint was published in PeerJ on 13 May 2020.

View the peer-reviewed version (peerj.com/articles/8872), which is the preferred citable publication unless you specifically need to cite this preprint.

Khalighifar A, Jiménez L, Nuñez-Penichet C, Freeman B, Ingenloff K, Jiménez-García D, Peterson T. 2020. Inventory statistics meet big data: complications for estimating numbers of species. PeerJ 8:e8872
https://doi.org/10.7717/peerj.8872

PeerJ Preprints NOT PEER-REVIEWED

Inventory statistics meet big data: Complications for estimating numbers of species

Ali Khalighifar [Corresp., 1, 2], **Laura Jiménez** [1, 2], **Claudia Nuñez-Penichet** [1, 2], **Benedictus Freeman** [1, 2], **Kate Ingenloff** [1, 2], **Daniel Jiménez-García** [1, 3], **A. Townsend Peterson** [1, 2]

[1] Biodiversity Institute, University of Kansas, Lawrence, Kansas, United States

[2] Department of Ecology and Evolutionary Biology, University of Kansas, Lawrence, KS, United States

[3] Centro de Agroecología y Ambiente, Benemerita Universidad Autónoma de Puebla, Puebla, Puebla, Mexico

Corresponding Author: Ali Khalighifar
Email address: a.khalighifar@ku.edu

Abstract

We point out complications inherent in biodiversity inventory metrics when applied to large-scale datasets. The number of samples in which a species is detected saturates, such that crucial numbers of detections of rare species approach zero. Any rare errors can then come to dominate species richness estimates, creating upward biases in estimates of species numbers. We document the problem via simulations of sampling from virtual biotas, illustrate its potential using a large empirical dataset (bird records from Cape May, New Jersey, USA), and outline the circumstances under which these problems may be expected to emerge.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27965v1 | CC BY 4.0 Open Access | rec: 17 Sep 2019, publ: 17 Sep 2019

PeerJ Preprints NOT PEER-REVIEWED

Inventory statistics meet big data: Complications for estimating numbers of species

Ali Khalighifar[1], Laura Jiménez[1], Claudia Nuñez-Penichet[1], Benedictus Freeman[1], Kate Ingenloff[1], Daniel Jiménez-García[1, 2], and A. Townsend Peterson[1]

[1] Biodiversity Institute and Department of Ecology and Evolutionary Biology, University of Kansas, Lawrence, Kansas, USA
[2] Centro de Agroecología y Ambiente, Instituto de Ciencias, Benemérita Universidad Autónoma de Puebla, Puebla, Puebla, México

Corresponding Author:
Ali Khalighifar[1]
1345 Jayhawk Blvd, Lawrence, KS 66045, USA
Email address: a.khalighifar@ku.edu

Abstract

We point out complications inherent in biodiversity inventory metrics when applied to large-scale datasets. The number of samples in which a species is detected saturates, such that crucial numbers of detections of rare species approach zero. Any rare errors can then come to dominate species richness estimates, creating upward biases in estimates of species numbers. We document the problem via simulations of sampling from virtual biotas, illustrate its potential using a large empirical dataset (bird records from Cape May, New Jersey, USA), and outline the circumstances under which these problems may be expected to emerge.

Introduction

Biodiversity measurements have important implications for conservation efforts (Sousa-Baena, Garcia & Peterson, 2014). Biodiversity metrics provide information about community composition, numbers of species, and similarity or dissimilarity of species composition among sites (Colwell & Coddington, 1994), and can allow researchers to separate well-inventoried sites from partially-inventoried sites for macroecological analyses (Lobo et al., 2018). Biodiversity inventories have been implemented at scales ranging from local to global (Moreno & Halffter, 2000; Ballesteros-Mejia et al., 2013), to evaluate and understand biotic responses to changing environmental conditions.

Tracking species richness in biodiversity inventories was originally achieved via visual assessment of asymptotic behavior of species accumulation curves (Karr, 1980), and then with the quantitative assist of non-linear regressions (Clench, 1979; Soberón & Llorente, 1993).

However, for the past 20+ years, non-parametric estimators of numbers of species have been used to estimate species richness, particularly a set of estimators based on sampling theory (Chao, 1987). Diverse data origins and variable data quality pose significant challenges for such analyses, particularly when data are drawn from publicly accessible databases, rather than collected individually by the researcher (Soberón et al., 1996; Lobo, 2008).

However, those same publicly accessible databases offer exciting opportunities for novel analyses (e.g., Cameron et al., 2018; Peterson et al., 2015). Primary biodiversity data connect a particular species with a place and a point in time (Sullivan et al., 2014), and availability of such data records has grown massively, now exceeding 10^9 records (e.g., Global Biodiversity Information Facility, http://www.gbif.org, serving 1,017,227,764 records as of 22 Aug 2018). Although these data are heavily biased in terms of their spatial and temporal distributions, being concentrated massively in Europe and North America and a few other, scattered regions (Yesson et al., 2007; Peterson & Soberón, 2018), the promise of genuine, macroscale, synthetic insights remains, and is growing.

In this contribution, we report on a complication in application of the customary statistics for measuring species richness (Colwell & Coddington, 1994) to very large-scale (e.g., 10^6 records or larger) biodiversity incidence datasets (i.e., records only of presence, and not of abundance). Biodiversity datasets have long been of modest dimensions only, and the field has been built on metrics and methods equipped for those dimensions. In the course of studies of avifaunal change over recent decades in North America that are pending publication, we noted that species richness estimates are affected significantly by what would seem to be negligible numbers of errors among the real data records (see Fig. 1, for an example from a site that is sampled massively by birdwatchers). We present a brief conceptual summary and a demonstration of the problem via a simple simulation; we conclude with an exploration of how such problems can be avoided or mitigated.

Conceptual background

The problem of estimating species richness from samples has been approached via methods that can be separated into three groups according to the statistical approach used to derive a species richness estimator: (1) extrapolating species accumulation curves to their asymptotes (Clench, 1979), (2) fitting parametric distributions of relative abundances (Efron & Thisted, 1976), or (3) using nonparametric techniques based on distribution of individuals among species (or the distribution of species among samples) (Colwell & Coddington, 1994; Colwell, 2013; Chao & Chiu, 2016). We focus on asymptotic versions of these methods *sensu* Chao and Chiu (2016), as we are interested in full inventories of species present at sites; see discussion in Peterson and Slade (1998). Two kinds of data are used in these richness studies: incidence data, in which only presences and absences are recorded for each species and each sample, and abundance data, in which numbers of individuals of each species are recorded within each sample (Gotelli & Colwell, 2011). Abundance data can always be converted to incidence data, whereas the reverse is not generally possible.

The nonparametric approach has been preferred greatly, since it does not make assumptions about underlying distributions of abundances or detection rates of species (Chao & Shen, 2004; Chao & Chiu, 2016). We focus on three nonparametric species richness estimators based on replicated incidence data that estimate numbers of species actually present at a site but not observed in the reference sample. All of the estimators correct observed richness (which is by default a lower bound for a species richness estimator) by adding a term estimating the number of species present but not detected based on numbers of species represented in one sample (uniques), two samples (duplicates), or a few samples only (Gotelli & Colwell, 2011; Colwell et al., 2012).

The reference sample for replicated incidence data consists of a species-by-sample matrix in which each element (m_{ij}) corresponds to either the presence or absence of species i in sample j. The number of columns in this matrix, T, is the number of sampling units in the sample; the number of rows is the observed number of species, S_{obs}. Q_k is the number of species present in exactly k sites of the sample, so the number of species present in the assemblage but not included in the sample (undetected species) is Q_0, the number of species unique to a single sample is Q_1, the number of duplicates is Q_2, and so on.

Chao (1984) originally derived an estimator of species richness S_{obs} for abundance based data that is now called *Chao1*, which she later recast for incidence data (Chao, 1987). This latter estimator, now called *Chao2,* is

$$\hat{S}_{Chao2} = \begin{cases} S_{obs} + \dfrac{\left[\frac{T-1}{T}\right]Q_1^2}{2Q_2}, \; if\ Q_2 > 0 \\ S_{obs} + \left[\dfrac{T-1}{T}\right]\dfrac{Q_1(Q_1-1)}{2(Q_2+1)}, \; if\ Q_2 = 0 \end{cases}, \qquad (1)$$

where T is the sample size available for the overall calculation. The first expression of equation (1) reflects the classic *Chao2* estimator; however, this estimator is undefined when $Q_2 = 0$. The second expression in equation (1) is a corrected form that is always obtainable and defined.

A second estimator of interest, the incidence coverage-based estimator (ICE), is based on the concept of sample coverage: the proportion of the total number of incidences in a set of sampling units that belong to the species represented in the sample. Sample coverage is a measure of the information available regarding occurrence of relatively rare species in the sample (Chao & Chiu, 2016): its estimator depends on the complement of the proportion of singletons, in relation to the total number of incidences of the infrequent species (Colwell, 1994). A third type of species richness estimator is based on the statistical method of jackknifing, a bias reduction technique involving removing subsets of the data and recalculating the estimator with the reduced sample (Chao & Chiu, 2016). Finally, we explored the method developed by Chiu and Chao (2016) for microbial molecular diversity data to account for inflation of numbers of singletons by sequencing errors (akin to identification errors); this method estimates the true value of Q_1 based on Q_2, Q_3, and Q_4, and uses the adjusted value in asymptotic diversity

estimates. It is important to notice that this method defaults to the classic Chao2 estimator when both Q_3, and Q_4 are equal to zero, otherwise the estimator of Q_1 (the true number of uniques) would be undefined. Therefore, its application is only for a certain window of conditions.

Note that, for each of the estimators described above, the estimator does not take advantage of the full frequency distribution of detections for species in an inventory effort—indeed, this partial use of the frequency distribution is the focus of this contribution. Three of these estimators, as well as their corresponding variances and confidence intervals, can be computed using EstimateS (Colwell & Elsensohn, 2014) and a new version implemented in R (Chao & Chiu, 2016); the final estimator can be computed using the R version only. We used EstimateS (version 9.1.0; Colwell & Elsensohn, 2014) for the older three nonparametric estimators, as that platform is that which has seen the greatest use by the biodiversity community, and the R version for the latter estimator.

Materials & Methods

We developed a simple simulation based on large samples from a virtual community of 100 "real" species, by using a log-normal distribution of mean abundances, with parameters μ = 1.5 and σ = 2.0 (the mean and standard deviation of the variable's natural logarithm, respectively). An initial simulation served to illustrate how crucial values (Q_1, Q_2, etc.) approach zero as the frequency distribution of detections of species shifts to higher frequencies of observation, and saturates beyond the few detections on which the inventory estimators focus. Then, to simulate effects of very rare errors in the form of misidentifications or incorrect geographic coordinates on inventory results for sites, in a second phase of simulation, we added 10 "error" species that were designed to mimic occasional, rare errors; this latter set of species had a mean abundance 6 orders of magnitude lower than the 100 real species. To understand sensitivity to distributional assumptions, we also explored log-normal distributions of abundances with parameters μ = 0.3 and σ = 1.2 and μ = 1.0 and σ = 0.5, and gamma distributions with parameters α = 1.8 and β = 1.0, α = 2.5 and β = 2.0, and α = 3.0 and β = 1.2 (where α is the shape parameter, and β is the scale parameter).

We sampled occurrences of the 100 real species in R version 3.2.3 (https://www.r-project.org/) (R Core Team, 2015). To avoid recycling samples and consequent serial dependency among samples, we created independent random samples for each sample size (5, 7, 10, 15, 20, 25, 50, 75, 100, 125, 150, 175, 200, 300, 400, 500, 600, 700, 800, 900, and 1000 samples). We used default settings of EstimateS (Colwell & Elsensohn, 2014) to calculate the Chao2, ICE, jackknife1, and jackknife2 estimators for the 100 replicates x 21 numbers of samples = 2100 simulated populations. Next, we used customized scripts in Python 2.7.11 to separate individual replicate result sets from the combined EstimateS output files, and to select and isolate the final lines from each replicate, to create a final table of results from each simulated population. All code for these analyses is available at http://hdl.handle.net/1808/25686.

Results

The simulation results showed clearly that the estimators converged well on the true value (100 species) in the error-free simulations, and that Q_1 and Q_2 approached zero in increasingly large samples (Fig. 2). The effects of adding the very rare "error" species were also quite clear: early samples lacked error species entirely, as they were just too rare to show up in relatively small samples. Only late in the simulation, after 400-1000 replicates, did these species begin to appear in the analysis datasets (red bars in Fig. 2).

The results of the first phase of the simulation showed that, with ~150 samples, estimates of numbers of species in the community settled at 100 species, which is the correct number of species (Fig. 3, top). However, when rare species were introduced at minuscule abundances compared to the "real" species, even though the results settled initially on the correct answer of 100 species, later—when the rare species begin to appear—a consistent upward bias was noted (Fig. 3, bottom).

The Chiu and Chao (2016) method showed consistent underestimation of true species numbers for modest numbers of days of sampling (Fig. 4), although this bias disappeared with large sample sizes. At modest sampling levels, although analyses of the simulated data with error better approximated the true number of species (100; Fig. 4), the consistent underestimation in error-free analyses suggests that this outcome may represent a balance between downward bias in error-free estimates and upward bias introduced by the errors.

The remaining estimators showed behavior similar to that of Chao2: ICE, Jackknife1 (first-order), and Jackknife2 (second-order) analyses, in the first simulation phase, settled on 100 species at ~100 samples, but in the second phase were biased upwards markedly by 150-250 samples (see Supporting Information). Finally, we explored different abundance distributions for the simulation—indeed, in all log-normal and gamma distributions that we assessed, biases were clear, just as in the results we have presented above.

Discussion

This contribution centers on how inventory statistics need to evolve in the face of larger and larger magnitudes of biodiversity data sets. That is, we have shown that any errors in the data (e.g., misidentifications, misspellings), even at very minor frequencies, can easily end up dominating the estimation process with the common and long-used nonparametric estimators, such as Chao2; the older species accumulation curve approach also would clearly overestimate numbers, given that "error" species would appear as species documented in the inventory. These biodiversity inventory statistics are important, offering crucial additional information to the process of biotic inventories; therefore, updating and amending these approaches to approaches that are less vulnerable to bias, or at least being cognizant of the potential for problems in estimation for big(ger) datasets, is important.

What solutions are available to a researcher with a big data set and the desire to develop detailed analyses of species richness and inventory completeness? Quite simply, a diversity of types of errors is found in pretty much every large-scale biodiversity dataset (Lamb et al., 2009),

and large-scale datasets (see, e.g., Fig. 1) will by nature have more such errors, at least on an absolute scale. A crucial first step is that of reducing spurious and erroneous species names in the dataset (Chapman, 2005). Such names may be misspellings, which can be detected easily by comparison of observed species lists with authority lists (Gueta & Carmel, 2016); this sort of error is well-known to inflate species richness estimates in inventories (Sousa-Baena, Garcia & Peterson, 2013). However, these names may also be real names chosen by accident from controlled pick-lists—such errors may be very hard to detect owing to the fact that they are valid names, but just not represented at the site in question. Similar contrasts in detectability of different error types have recently been documented for ecological niche modeling and species distribution modeling (Simões & Peterson, 2018).

Finally, and particularly for the case of birds and a few other taxa for which species are well documented, a third class of problems regarding species names may arise. Specifically, rare visitors, often termed vagrants, are valid species names, and the species may genuinely be present at the site at some (rare) point in time (see Fig. 1). However, depending on the specific definition of the biota under consideration, these species may not be relevant. That is, detection and documentation of such species depends on continuous, intensive presence of observers or collectors, and also on the presence of the "experts" who will be experienced enough to detect and report such records, and whose records of such species will be believed and accepted. Such dependencies will easily create biases that may make certain sites appear richer in species, when in actuality they are richer only in high-level observers (Dittmann & Lasley, 1992). More generally, this point serves to indicate that biotic inventories need to be defined carefully in terms of a particular point or span of time and space.

The method presented by Chiu and Chao (2016) was developed for application to microbial molecular diversity data to account for inflation of singletons by sequencing errors, which is closely akin to problems created by identification errors in species inventories. This method estimates the true value of Q_1, based on Q_2, Q_3, and Q_4, and uses the adjusted value in asymptotic diversity estimates. This estimator, in our simulation-based assessments, underestimated true species numbers in the absence of error, but estimated the true species number closely when errors were introduced—as such, the Chiu-Chao estimator may offer a useful solution to the problems identified in this contribution for biodiversity inventory estimates.

Conclusions

In summary, in this note, we point out and document a complication with application of the commonly used species inventory statistics, as biodiversity data sets grow to be large. The base observation is that fauna sizes are finite, but sampling effort can grow without limit, which shifts distributions of frequencies of observations of species towards larger and larger numbers—this phenomenon has the effect of rarefying the numbers of relatively rare species that inform inventory statistics. Two processes are involved: (1) estimators depend on the frequencies of detection of the rarer species, which decline to nil in very large datasets; and (2) erroneous

reports come to dominate the estimation process because errors are rare and real species accumulate much larger numbers of observations, such that estimates can come to be based entirely on noise rather than on signal. The first point is a simple consequence of massive-scale sampling of finite biotas; the second, however, derives from the dependence of inventory statistics on information from rare species. Solutions to these problems must involve detailed cleaning and quality control of data, and careful definition of the relevant species pool that is under study. Exploration of new estimators that take into account species with greater numbers of records or that correct for biases in Q_1 (Chiu & Chao, 2016)—may provide solutions to these problems.

Acknowledgements

We thank the University of Kansas Ecological Niche Modeling Group for their support and interest in the course of this project. We thank Jorge Soberón for a helpful review of the manuscript. We also thank Anne Chao for leadership in this field, and for willingness to provide comment and resources necessary for this project.

References

Ballesteros-Mejia L, Kitching IJ, Jetz W, Nagel P, Beck J. 2013. Mapping the biodiversity of tropical insects: Species richness and inventory completeness of African sphingid moths. *Global Ecology and Biogeography* 22:586–595. DOI: 10.1111/geb.12039.

Cameron EK, Martins IS, Lavelle P, Mathieu J, Tedersoo L, Gottschall F, Guerra CA, Hines J, Patoine G, Siebert J, Winter M, Cesarz S, Delgado-Baquerizo M, Ferlian O, Fierer N, Kreft H, Lovejoy TE, Montanarella L, Orgiazzi A, Pereira HM, Phillips HRP, Settele J, Wall DH, Eisenhauer N. 2018. Global gaps in soil biodiversity data. *Nature Ecology & Evolution* 2:1042. DOI: 10.1038/s41559-018-0573-8.

Chao A. 1984. Nonparametric estimation of the number of classes in a population. *Scandinavian Journal of Statistics* 11(4):265–270.

Chao A. 1987. Estimating the population size for capture-recapture data with unequal catchability. *Biometrics* 43:783–791. DOI: 10.2307/2531532.

Chao A, Chiu C-H. 2016. Species richness: estimation and comparison. *Wiley StatsRef: Statistics Reference Online*:1–26.

Chao A, Shen TJ. 2004. Nonparametric prediction in species sampling. *Journal of Agricultural, Biological, and Environmental Statistics* 9:253–269.

Chapman AD. 2005. Principles of data quality (Version 1.0). Copenhagen, Denmark: Report for the Global Biodiversity Information Facility. Retrieved from https://doi.org/10.15468/doc.jrgg-a190

Chiu C-H, Chao A. 2016. Estimating and comparing microbial diversity in the presence of sequencing errors. *PeerJ* 4:e1634. DOI: 10.7717/peerj.1634.

Clench H k. 1979. How to make regional lists of butterflies: some thoughts. *Journal of the Lepidopterists' Society* 33:216–231.

Colwell RK. 2013. EstimateS: Statistical estimation of species richness and shared species from samples. Available at: http://purl.oclc.org/estimates.

Colwell RK, Chao A, Gotelli NJ, Lin S-Y, Mao CX, Chazdon RL, Longino JT. 2012. Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages. *Journal of Plant Ecology* 5:3–21. DOI: 10.1093/jpe/rtr044.

Colwell RK, Coddington JA. 1994. Estimating terrestrial biodiversity through extrapolation. *Philosophical Transactions of the Royal Society of London B 335:101-118.*

Colwell RK, Elsensohn JE. 2014. EstimateS turns 20: statistical estimation of species richness and shared species from samples, with non-parametric extrapolation. *Ecography* 37:609–613. DOI: 10.1111/ecog.00814.

Dittmann DL, Lasley GW. 1992. How to document rare birds. *Birding* 24:145–159.

Efron B, Thisted R. 1976. Estimating the number of unseen species: How many words did Shakespeare know? *Biometrika* 63:435–447. DOI: 10.1093/biomet/63.3.435.

Gotelli NJ, Colwell RK. 2011. Estimating species richness. In: Magurran A, McGill B eds. *Biological diversity: frontiers in measurement and assessment*. Oxford, UK: Oxford University Press, 39–54.

Gueta T, Carmel Y. 2016. Quantifying the value of user-level data cleaning for big data: A case study using mammal distribution models. *Ecological Informatics* 34:139–145. DOI: 10.1016/j.ecoinf.2016.06.001.

Karr JR. 1980. Geographical variation in the avifaunas of tropical forest undergrowth. *The Auk* 97:283–298.

Lamb EG, Bayne E, Holloway G, Schieck J, Boutin S, Herbers J, Haughland DL. 2009. Indices for monitoring biodiversity change: Are some more effective than others? *Ecological Indicators* 9:432–444. DOI: 10.1016/j.ecolind.2008.06.001.

Lobo JM. 2008. Database records as a surrogate for sampling effort provide higher species richness estimations. *Biodiversity and Conservation* 17:873–881. DOI: 10.1007/s10531-008-9333-4.

Lobo JM, Hortal J, Yela JL, Millán A, Sánchez-Fernández D, García-Roselló E, González-Dacosta J, Heine J, González-Vilas L, Guisande C. 2018. KnowBR: An application to map the geographical variation of survey effort and identify well-surveyed areas from biodiversity databases. *Ecological Indicators* 91:241–248. DOI: 10.1016/j.ecolind.2018.03.077.

Moreno CE, Halffter G. 2000. Assessing the completeness of bat biodiversity inventories using species accumulation curves. *Journal of Applied Ecology* 37:149–158. DOI: 10.1046/j.1365-2664.2000.00483.x.

Peterson AT, Navarro-Sigüenza AG, Martínez-Meyer E, Cuervo-Robayo AP, Berlanga H, Soberón J. 2015. Twentieth century turnover of Mexican endemic avifaunas: Landscape

change versus climate drivers. *Science Advances* 1:e1400071. DOI: 10.1126/sciadv.1400071.

Peterson AT, Slade N. 1998. Extrapolating inventory results into biodiversity estimates and the importance of stopping rules. *Diversity and Distributions* 4:95–105. DOI: 10.1046/j.1365-2699.1998.00021.x

Peterson AT, Soberón J. 2018. Essential biodiversity variables are not global. *Biodiversity and Conservation* 27:1277–1288.

R Core Team. 2015. *R:* A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/.

Simões M, Peterson AT. 2018. Utility and limitations of climate-matching approaches in detecting different types of spatial errors in biodiversity data. *Insect Conservation and Diversity*. DOI: 10.1111/icad.12288.

Soberón J, Llorente J. 1993. The use of species accumulation functions for the prediction of species richness. *Conservation Biology* 7:480–488. DOI: 10.1046/j.1523-1739.1993.07030480.x.

Soberón J, Llorente J, Benitez H. 1996. An international view of national biological surveys. *Annals of the Missouri Botanical Garden*, 83(4):562–573. DOI:10.2307/2399997

Sousa-Baena MS, Garcia LC, Peterson AT. 2013. Completeness of digital accessible knowledge of the plants of Brazil and priorities for survey and inventory. *Diversity and Distributions* 20:369–381. DOI: 10.1111/ddi.12136.

Sousa-Baena MS, Garcia LC, Peterson AT. 2014. Knowledge behind conservation status decisions: Data basis for "Data Deficient" Brazilian plant species. *Biological Conservation* 173:80–89. DOI: 10.1016/j.biocon.2013.06.034.

Sullivan BL, Aycrigg JL, Barry JH, Bonney RE, Bruns N, Cooper CB, Damoulas T, Dhondt AA, Dietterich T, Farnsworth A, Fink D, Fitzpatrick JW, Fredericks T, Gerbracht J, Gomes C, Hochachka WM, Iliff MJ, Lagoze C, La Sorte FA, Merrifield M, Morris W, Phillips TB, Reynolds M, Rodewald AD, Rosenberg KV, Trautmann NM, Wiggins A, Winkler DW, Wong W-K, Wood CL, Yu J, Kelling S. 2014. The eBird enterprise: an integrated approach to development and application of citizen science. *Biological Conservation* 169:31–40. DOI: 10.1016/j.biocon.2013.11.003.

Yesson C, Brewer PW, Sutton T, Caithness N, Pahwa JS, Burgess M, Gray WA, White RJ, Jones AC, Bisby FA, Culham A. 2007. How global is the global biodiversity information facility? *PLOS ONE*, 2(11), e1124. DOI:10.1371/journal.pone.0001124

PeerJ Preprints NOT PEER-REVIEWED

Figure 1

Example of an intensively sampled site, Cape May National Wildlife Refuge, NJ, USA.

This example shows how the frequency histogram of number of detections per species reflects large numbers of observations of a finite biota. This histogram summarizes 12,144,561 records for the site, and 436 species detected. We have identified the species having the lowest frequencies of detection, among which can be noted several species that are probably not occurring there naturally, such as Anser anser, Eupsittula canicularis, and Melopsittacus undulatus, all of which are likely there as escapes from captivity.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27965v1 | CC BY 4.0 Open Access | rec: 17 Sep 2019, publ: 17 Sep 2019

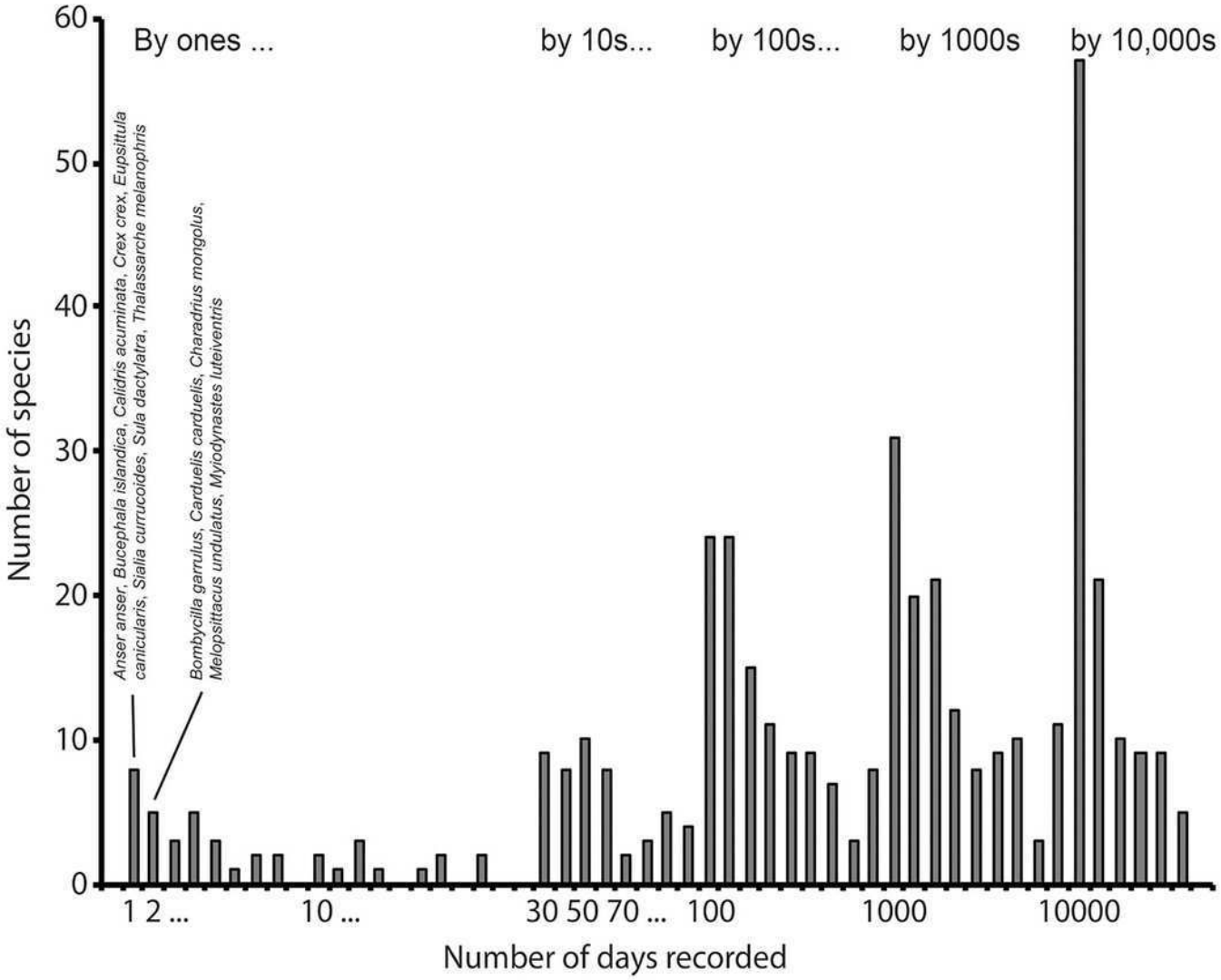
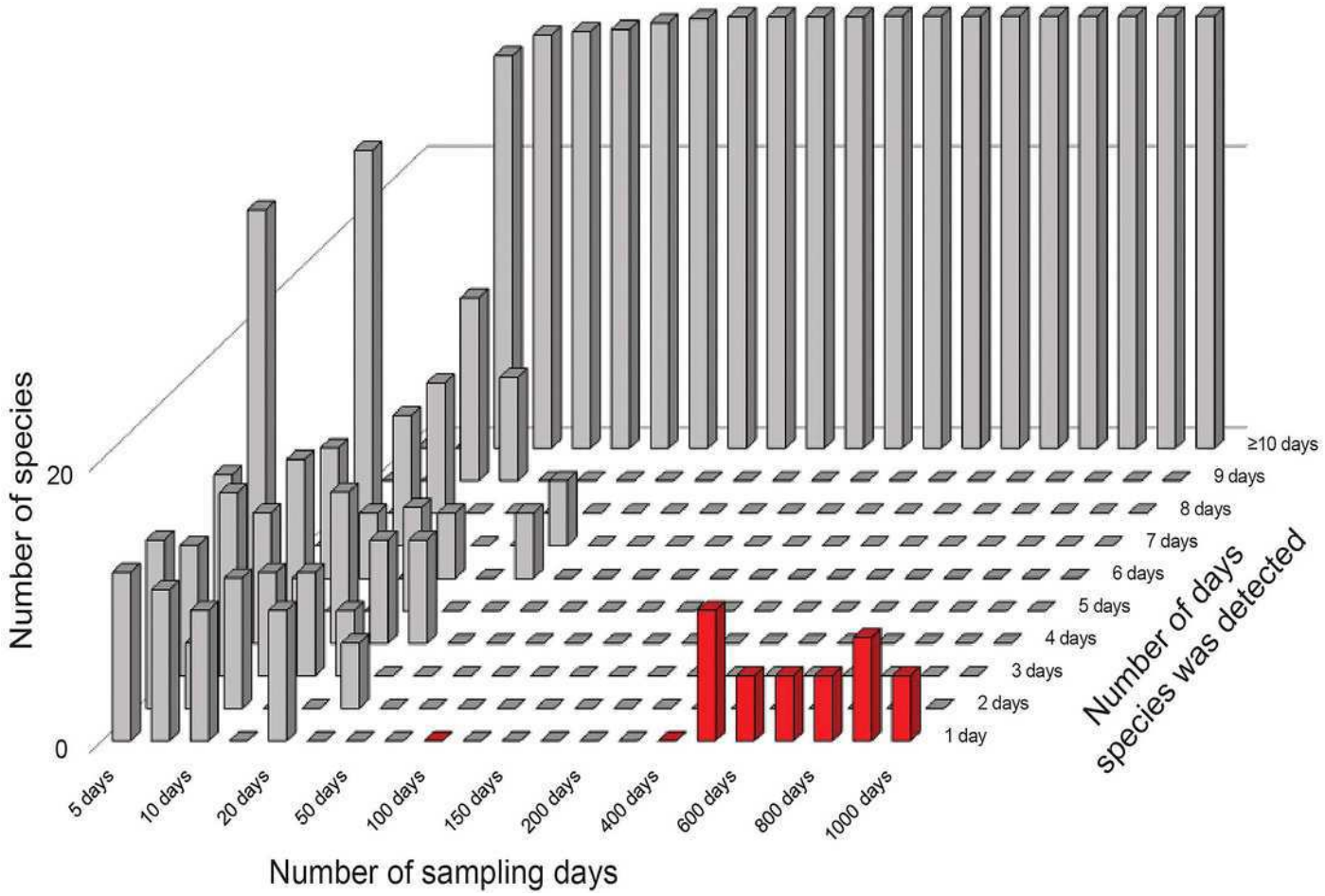

By ones ...
by 10s...
by 100s...
by 1000s
by 10,000s
Anser anser, Bucephala islandica, Calidris acuminata, Crex crex, Eupsittula canicularis, Sialia currucoides, Sula dactylatra, Thalassarche melanophris
Bombycilla garrulus, Carduelis carduelis, Charadrius mongolus, Melopsittacus undulatus, Myiodynastes luteiventris
Number of species
0
10
20
30
40
50
60
1 2 ...
10 ...
30 50 70 ...
100
1000
10000
Number of days recorded

Figure 2

Summary of frequencies of species in inventory samples used in simulation exercises.

The great bulk of these samples had large numbers of detections (the tall bars along the left and back of the figure). Note that by 50-100 days of sampling, no samples are left in the 1-2 detections categories that feed into the Chao2 estimator analyses. Note also the appearance of rare species in the analysis (red bars at front right) when samples became very large.



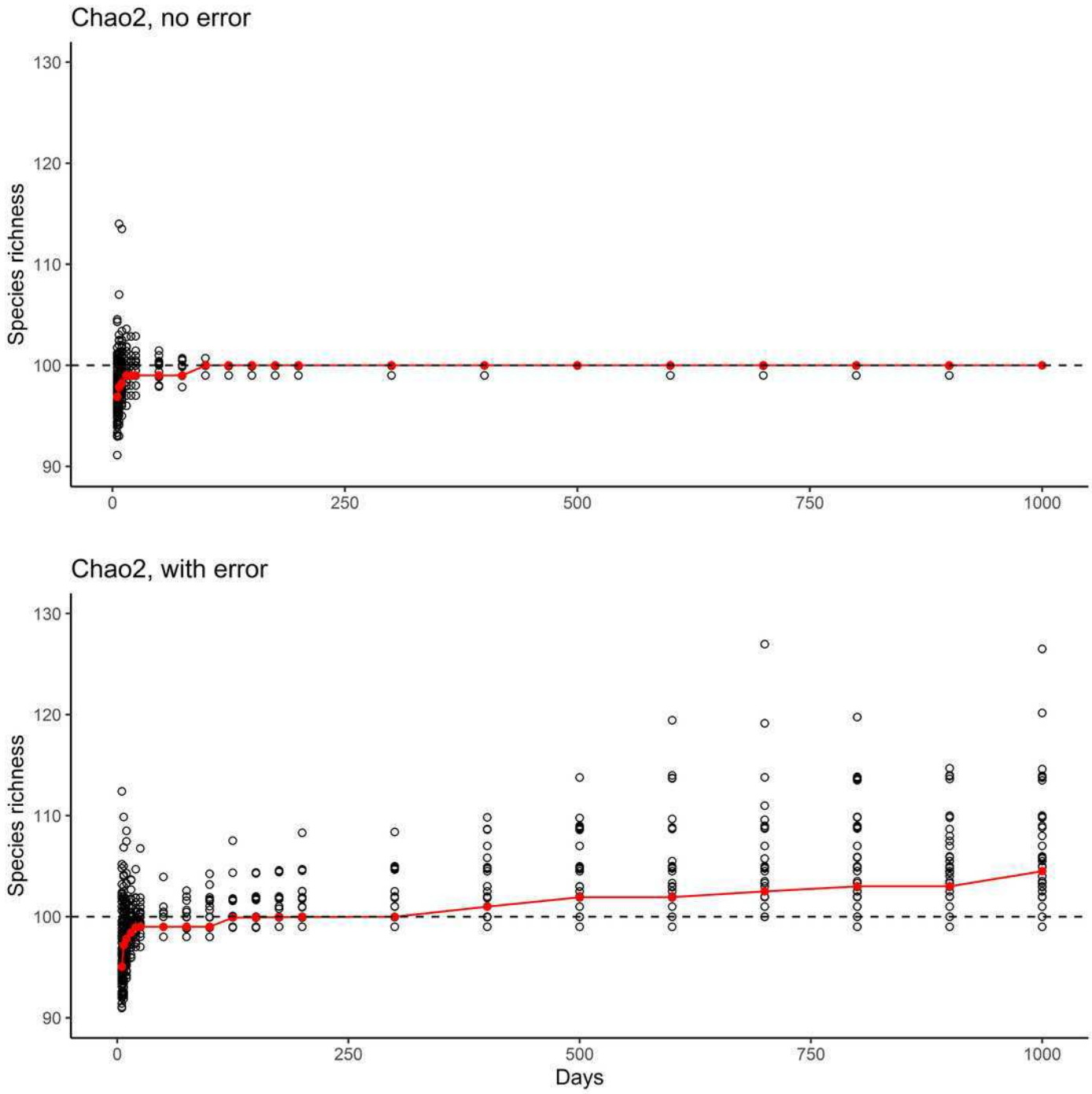
PeerJ Preprints NOT PEER-REVIEWED

Figure 3

Summary of the two phases of simulation results.

Graphics show simulation of accumulation of species in simulated inventories, showing the scatter of individual inventory simulations (black circles) and the median of results (red line). Top: 100 real species, with no error species included. Bottom: 100 real species, with 10 rare species included to simulate errors in identification or geographic references.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27965v1 | CC BY 4.0 Open Access | rec: 17 Sep 2019, publ: 17 Sep 2019


Chao2, no error
Species richness
130
120
110
100
90
0
250
500
750
1000
Chao2, with error
Species richness
130
120
110
100
90
0
250
500
750
1000
Days

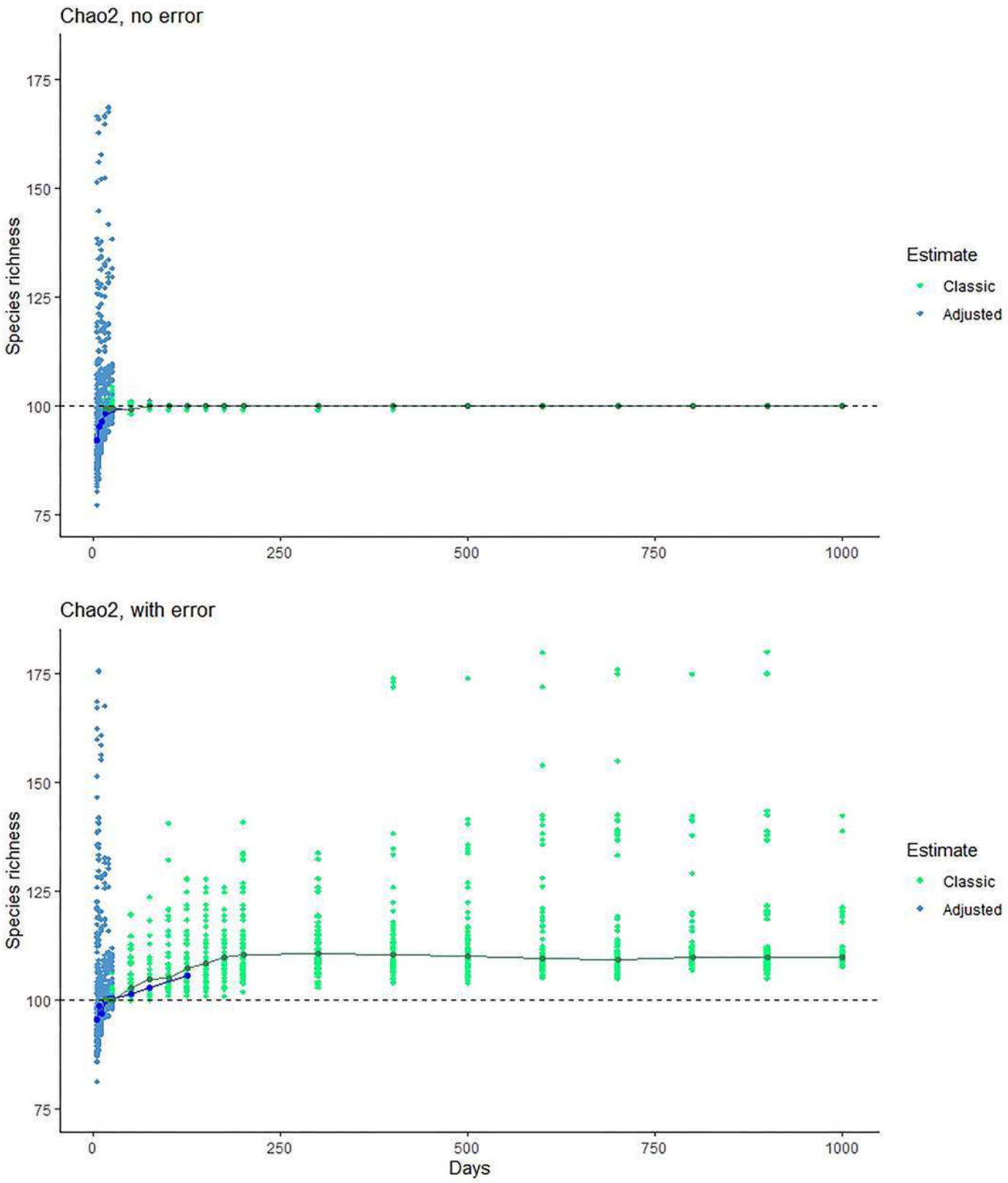
PeerJ Preprints NOT PEER-REVIEWED

Figure 4

Exploration of the estimation method of Chiu and Chao (2016), which takes into account Q1, Q2, Q3, and Q4.

Note that, at larger sample sizes, the Chiu-Chao estimator (blue points) defaults to the Chao2 estimator (green points; Chiu and Chao 2016). We provide (top panel) the results for Chao2 (no error) for purposes of comparison, and then the results from the new estimator in simulations without (middle panel) and with (bottom panel) errors included.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27965v1 | CC BY 4.0 Open Access | rec: 17 Sep 2019, publ: 17 Sep 2019


Chao2, no error
Species richness
175
150
125
100
75
0
250
500
750
1000
Estimate
Classic
Adjusted
Chao2, with error
Species richness
175
150
125
100
75
0
250
500
750
1000
Days
Estimate
Classic
Adjusted