Biased-corrected richness estimates for the Amazonian tree flora

Hans ter Steege, Paulo Inácio Prado, Renato a F de Lima, Edwin Pos, Diogenes de Andrade, Lima Filho, Rafael P Salomão, Ieda Amaral, Francisca Dionizia de Almeida Matos, Carolina Castilho, et al.

To cite this version:

Hans ter Steege, Paulo Inácio Prado, Renato a F de Lima, Edwin Pos, Diogenes de Andrade, et al.. Biased-corrected richness estimates for the Amazonian tree flora. Scientific Reports, Nature Publishing Group, 2020, 10, pp.10130. 10.1038/s41598-020-66686-3. hal-02883980

HAL Id: hal-02883980
https://hal.inrae.fr/hal-02883980
Submitted on 29 Jun 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Biased-corrected richness estimates for the Amazonian tree flora

Hans ter Steege1,2, Paulo I. Prado3,4, Renato A. F. de Lima1,2, Edwin Pos5, Luiz de Souza Coelho5, Diogenes de Andrade Lima Filho5, Rafael P. Salomão5,6,7, lêda Leão Amaral5, Francisca Dionízia de Almeida Matos5, Carolina V. Castilho6, Oliver L. Phillips9, Juan Ernesto Guevara10,11, Marcelo de Jesus Veiga Carim12, Dairon Cárdenas López13, William E. Magnusson14, Florian Wittmann15,16, Maria Pires Martins5, Daniel Sabatier17, Mariana Victória Irume4, José Renan da Silva Guimarães12, Jean-François Molino13, Olaf S. Bánki1, Maria Teresa Fernandez Piedad48, Nigel C. A. Pitman19, José Ferreira Ramos5, Abel Monteagudo Mendoza20, Eduardo Martins Venticinque21, Bruno Garcia Luiz22, Percy Núñez Vargas23, Thiago Sanna Freire Silva24, Evlyn Márcia Moraes de Leão Novo25, Neidiane Farias Costa Reis26, John Terborgh27,28, Angelo Gilberto Manzatto29, Katia Regina Casula28, Eurídice N. Honorio Coronado30,31, Juan Carlos Montero30,31, Alvaro Duque32, Flávia R. C. Costa33, Nicolás Castaño Arboleda11, Jochen Schöngart18, Charles Eugene Zartman5, Timothy J. Killeen33, Beatriz S. Marimon34, Ben Hur Marimon-Junior34, Rodolfo Vasquez20, Bonifacio Mostacedo35, Layon O. Demarchi18, Ted R. Feldpausch9,36, Julien Engel17,37, Pascal Petronelli38, Chris Baraloto37, Rafael L. Assis39, Hernán Castellanos40, Marcelo Fragomeni Simon41, Marcelo Brilhante de Medeiros42, Adriano Quaresma18, Susan G. W. Laurance28, Lorena M. Rincón5, Ana Andrade42, Thaiane R. Sousa5, José Luís Camargo42, Juliana Schietti5, William F. Laurance28, Helder Lima de Queiroz43, Henri Eduardo Mendonça Nascimento5, Maria Aparecida Lopes44, Emanuelle de Sousa Farias44,45,46, José Leonardo Lima Magalhães47,48, Roel Brienen5,2,49, Juan David Cardenas Revilla5, Ima Célia Guimarães Vieira7, Bruno Barçante Ladovca Cintra60, Pablo R. Stevenson61, Yuri Oliveira Feitosa53, Joost F. Duivenvoorden53, Hugo F. Mogollón54, Alejandro Araujo-Murakami15, Leandro Valle Ferreira17, José Rafael Lozada15, James A. Comiskey57,58, José Julio de Toledo59, Gabriel Damasco60, Dániel Dávila61, Aline Lopes18,62, Roosevelt García-Villacorta63,64, Freddie Draper37,65, Alberto Vicentini24, Fernando Cornejo Valverde66, Jon Lloyd67, Vítor H. F. Gomes68,69, David Neill70, Alfonso Alonso58, Francisco Dallmeier24, Fernanda Coelho de Souza5,14, Rogerio Gribel7, Luzmiá Arroyo55, Fernanda Antunes Carvalho16,17,22, Daniel Praia Portela de Aguiar18, Dário Dantas do Amaral17, Marcelo Petraíti Pansonato4,15, Kenneth J. Feeley72,73, Erika Berenguer74,75, Paul V. A. Fine60, Marcelino Carneiro Guedes76, Jos Barlow75, Joice Ferreira68, Boris Villa77, Maria Cristina Peñuela Mora78,39, Eliana M. Jimenez79, Juan Carlos Licona31, Carlos Cerón80, Raquel Thomas81, Marco Maas1, Marcus Silveira82, Terry W. Henkel83, Juliana Stropp84, Marcos Rios Paredes85, Kyle G. Dexter86,87, Doug Daly88, Tim R. Baker9, Isac Huamantupa-Chuquimaco9, William Milliken89, Toby Pennington86,87, J. Sebastián Tello90, José Luis Marcelo Pena91, Carlos A. Peres92, Bente Kiltgaard93, Alfredo Fuentes90,94, Miles R. Silman85, Anthony Di Fiore96, Patricio von Hildebrand97, Jerome Chave98, Tinne R. van Andel19,99, Renato Richard Hilário59, Juan Fernando Phillips100, Gonzalo Rivas-Torres101,102, Janaina Costa Noronha103,104, Ana Prieto105, Tharyn Gonzalez105, Rainiellene de Sá Carpanedo103, George Pepe Gallardo Gonzales85, Ricardo Zárate Gómez106, Domingos de Jesus Rodrigues103, Egleé L. Zent107, Ademir R. Ruschel48.
Amazonian forests are extraordinarily diverse, but the estimated species richness is very much debated. Here, we apply an ensemble of parametric estimators and a novel technique that includes conspecific spatial aggregation to an extended database of forest plots with up-to-date taxonomy. We show that the species abundance distribution of Amazonia is best approximated by a logseries with aggregated individuals, where aggregation increases with rarity. By averaging several methods to estimate total richness, we confirm that over 15,000 tree species are expected to occur in Amazonia. We also show that using ten times the number of plots would result in an increase to just ~50% of those 15,000 estimated species. To get a more complete sample of all tree species, rigorous field campaigns may be needed but the number of trees in Amazonia will remain an estimate for years to come.

The lowland rainforest of the Amazon River basin and Guiana Shield, hereafter Amazonia, covers an area of nearly 6 million km² with an estimated total number of 3.9 × 10^4 trees (diameter at 1.30 m - dbh ≥ 10 cm)¹. Sampling such an extensive area has been extremely limited, and accurate estimates of the total number of tree species and their populations have thus been difficult to obtain. Nevertheless, these estimates are important to understand the difference between what we already know and what still needs to be discovered about Amazonian tree diversity. In 2013, we estimated that around 16,000 tree species should occur within Amazonia¹ using the estimated total numbers of plots scattered across the area. A number that has both been criticized²⁻⁴ and accepted as plausible for plot inventories⁵. In the following years there has been considerable progress both in the taxonomy of Amazonian tree species⁶⁻⁸ and the number of forest inventory plots available⁹, which has steadily grown from 1,170 to 1,946. In addition, our previous richness estimate was based on a single estimation method, the linear extension of a hypothetical logseries¹⁰ of the estimated population sizes of the Amazonian tree species present in our plots¹. While the logseries fits the relative abundance distribution of many taxa quite well¹¹, others have recently argued that it tends to overestimate richness and that other models such as the negative binomial would provide...
better estimates\textsuperscript{12}. Other parametric methods, such as the Poisson lognormal\textsuperscript{13,14} have also been suggested to have a better performance than the logseries when used to describe species relative abundances in Amazonian forest\textsuperscript{15}.

Here we present a new species richness estimation of Amazonian tree species based on the revised taxonomy and increased inventory dataset using various parametric estimation methods. Non-parametric methods\textsuperscript{4,16} continue to be used to estimate species richness but as they tend to hugely underestimate richness in sparsely sampled, species-rich areas\textsuperscript{6,17,18}, such as Amazonia, we did not employ them here. All methods assume random sampling of species with random distribution across space, conditions mostly not met in forest inventory data. Rather, limited dispersal and ecological preferences of trees tend to result in aggregated spatial patterns of species distribution\textsuperscript{19–21}. Hence, conspecific aggregation at the sampling scale has also been pointed out as a source of serious bias of parametric estimates of species richness\textsuperscript{12,21}.

The new advances in species taxonomy, improved sampling coverage and richness estimation methods, discussed above, allow us to provide an update on the estimated number of species in the Amazonian tree flora and their estimated population sizes, which are key to the understanding and conservation of the Amazonian tree diversity. Making use of the March 2019 Amazon Tree Diversity Network (ATDN) database, an improved Amazon lowland forest map\textsuperscript{22}, and an updated taxonomy\textsuperscript{4,8}, we provide new estimates of the (i) population sizes for each species and (ii) total species richness of the Amazonian tree flora. We provide species richness estimates using parametric methods based on the logseries (LS), negative binomial (NB) and Poisson lognormal (PLN). For the first time, we use data on species occurrences across plots to assess the impacts of conspecific aggregation in the estimation of tree species richness in Amazonia. Based on the simulation of the sampling from a hypothetical species abundance distribution (SAD) for the Amazonian tree flora, we evaluate the accuracy of each estimation method and provide correction of biases in their estimates. Finally, we evaluate how robust the richness estimates of each method are to increases in the number of forest plots included in the sample and to changes in species taxonomy.

Results

Raw data description. The 2019 version of the dataset (1,946 plots) contained a total of 1,101,368 individuals, 89% of which were identified to a valid species name (Supplementary Table 1). It contained a total of 5,027 tree species (Appendix 1), representing an increase of 63 observed species in comparison to the 2013 version (496 species, if the 2013 dataset would have used the current taxonomy – see also Appendix 2, 3). The total number of trees in Amazonia, based on the tree density modelling, was estimated at 3.06 \times 10^{11}. The 10 most common species were: \textit{Euclea coriacea}, \textit{Euterpe precatoria}, \textit{Pseudomedia laevis}, \textit{Protium allsissimum}, \textit{Iriartea deltoidea}, \textit{Mauritia flexuosa}, \textit{Socratea exorrhiza}, \textit{Astrocaryum murumuru}, and \textit{Pentaclethra macroloba} (species authorities follow ref. \textsuperscript{8}). Six of these species are palms (see Appendix 1 for the population estimates for all species). Most of the hyperdominant species show only small changes in their estimated populations compared to 2013 and only small differences in rank.

Fitted models to empirical SADs. The truncated negative binomial (TNB) was the model with best fit to the empirical SADs for all versions of the data sets (Table 1, Supplementary Table 2) with the logseries providing an equally good fit for the 2013 dataset (Table 1, Supplementary Fig. 1), if we used AIC. Model selection with a Bayesian Information Criterion (BIC) had the same result, except by the best support of LS for the 2013 data set (Supplementary Table 3). Visually, the LS and TNB provided very similar fits, with both models underestimating the abundances of the most abundant species in the sample (Fig. 1). The PLN model had no support from the three data sets, overestimating the abundance of common species and underestimating the abundance of rare species. The LS model had no support in simulated samples drawn from logseries regional SADs with more than 10,000 species with clumping, but had a constant and high support in simulated samples without clumping (Supplementary Fig. 5). For simulated samples from a Negative Binomial regional SAD, the LS model had a poor support for communities with more than 10,000 species, irrespective of conspecific clumping. Thus, the conspecific clumping observed in our data sets causes a strong selection bias against LS, when this is the correct model, but does not cause selection bias for TNB. The best support of TNB provided by the abundances in the sample
could thus be an artefact. In fact, the approximate Bayesian Computation model selection showed that LS models had by far the largest posterior probabilities to approximate the distribution of total population sizes, for all three data sets (Table 2).

For all datasets, simulations of samples with conspecific clumping of a logseries regional SAD had the highest posterior Approximate Bayesian Computation (ABC) probabilities, and simulations of samples from a truncated negative binomial or a lognormal had very low or zero posterior probabilities (Table 2). We thus assumed the simulations of sampling of a logseries regional SAD with conspecific aggregation as the best approximation of the process that generated the ATDN data.

Impacts of sampling effort and species taxonomy. Bias-corrected estimates of species richness, using conspecific aggregation fell between 13,730 (TNB) and 16,741 (LSE) species for the 2019 dataset, a variation of 22% (Fig. 2, Supplementary Table 4). Corresponding figures for the 2013 and updated 2013 data sets were 15,437 – 18,056 (TNB - LSE, 20%) and 11,618 – 15,643 (TNB - LSE, 35%) respectively. The taxonomic update of the 2013 data set led to a decrease in the number and abundance of rarer species (species below the median abundance ranking and with densities below to 2–5 individuals/ha in the sample, Supplementary Fig. 6). The same effect was observed in the distribution of abundances of estimated population sizes (Supplementary Fig. 7). As a consequence, the taxonomic update decreased the estimated number of species by all methods. This reduction ranged from ~7.6% (ABC) to ~24.8% (TNB). The expansion of inventory plots for the 2019 ATDN database increased the number and abundance of rarer species (Supplementary Figs. 6 and 7), which in turn partially reversed the decreasing of the estimated numbers of species (Fig. 2), increasing the estimated richness by 4.9% (ABC) to 18.1% (TNB). The sensitivity of the methods to taxonomic updates and to the expansion of the database followed the same order: ABC < LS < LSE < TNB (Fig. 2). Overall, there was more variation of estimates among methods than among the versions of the database. The variance components of the point estimates were 52% for estimation methods and 38% for the data sets, with a residual component of 10%.

Predicted richness with increasing sample sizes. We predicted species richness for larger plot samples, using a logseries with 15,874 species, which was the average estimation of total species richness for Amazonia from the 2019 data set. Two other logseries with 13,887 and 17,020 species respectively, were used as the lower and upper bounds for the estimated number of species. The simulations predicted that on average 746 additional species would be recorded in the plots if the current sample size was doubled, an increase of 15% (Fig. 4). The expected number species for the same sample size, assuming random dispersion of all species, can be estimated with equation S.3. Assuming a mean density of 528.5 trees ha⁻¹ in Amazon that would amount to 6,110 species, or an increase of 22%. As expected, conspecific clumping, which is stronger for the rare species (Supplementary Fig. 8), decreases the rates of accumulation of species.

If the number of plots was ten times the current sample size, the expected number of species in the plots would be 6,958 (an increase of 38%; lower and upper bounds: 6,517 – 7,884), getting close to half of the species estimated to be present in Amazonia.

Discussion

Our ability to estimate species richness ultimately depends on the capacity to accurately describe patterns of commonness and rarity from samples of local communities and project it to a much larger sample size (see also ref. 18). This is not a trivial challenge and it has been the subject of much empirical and theoretical study23. Here we used nearly 2,000 forest inventory plots to estimate the total number of tree species for the entire Amazonian forest and compare it to earlier estimates1,4,12,22. Our first finding was that despite the updated taxonomy (i.e. reduction of almost 10% of the species) and the addition of nearly 800 plots, the estimates of total species richness by each method were fairly similar. Updated point estimates using LSE and LS had a difference of -5% and 1.3% with our previous estimates1, and the updated estimate using the TNB had a difference of only 0.9% with previous estimates12. Such a small variation was the outcome of the well-known increase of number of species recorded as new plots were added to the sample, compensating the reduction of species by the taxonomic updating. Both processes affected the abundance distributions in the same way, as most of the species removed from the 2013 data set and most of the species added in the 2019 data set are among the less abundant ones.

Therefore, the updates affected the evenness of the abundance distributions in the samples, mainly by changing the number and relative abundances of rare species. These changes directly affect the estimates of species richness by LS and TNB, which rely on the shape of the distribution of abundances in the sample. Moreover, these changes also affect LSE and ABC estimates indirectly, because these methods use the distribution of total population sizes, which in turn depends on the relative abundances of species in the sample and on the occurrence of species across plots.

Also, we also updated the total area of Amazonian forest22 (see methods), causing a reduction of 17% in the estimated total number of trees. This is an additional cause of the decrease in the 2013 estimates of species richness, as all methods we used upscale some abundance distribution to the total size of the community.

The aggregation of species forms another important aspect of species richness estimation24, and as we have shown here, this greatly influences species estimations. All estimators make assumptions about the probability of occupancy, which is then used to estimate the expected number of species recorded if the whole area would be sampled. The occupancy is affected by the distribution of individuals across plots, and is higher under random distribution of species than if species are aggregated (as more aggregation leads to less occupied plots, under the same mean density per plot). Thus, if is there conspecific aggregation the assumption of random distribution underestimates the real number of species, as our simulations have shown. This effect was even greater in our data
Table 1. Delta-AIC for each parametric model fitted to the empirical SAD constructed using the 2013 version of the data (less plots, old taxonomy), the updated 2013 version (less plots, updated taxonomy) and the 2019 version (more plots, updated taxonomy). For each dataset version, the better supported model has 0 Delta-AIC. Models with Delta-AIC < 2 have equal support. DF = number of parameters in the model.

| Model                      | DF | 2013  | 2013 updated | 2019 |
|----------------------------|----|-------|--------------|------|
| Negative Binomial (TNB)    | 2  | 0.00  | 0.00         | 0.00 |
| Logseries (LS)             | 1  | 1.67  | 30.02        | 31.52|
| Poisson-lognormal (PLN)    | 2  | 60.43 | 47.87        | 42.09|

Figure 1. Fit of SAD models to the species abundances from samples from the three ATDN data sets. Left column: observed frequencies of species in each abundance class (octaves, grey bars) and frequencies predicted by analytical Logseries (LS), truncated negative binomial (TNB) and Poisson-lognormal (PLN). Right column: rank-abundance plot in log-log scale of the abundances of species (gray) and the predicted abundances at each rank by the same three models.
Figure 2. Bias-corrected estimates of total species richness for each method and each data set. TNB, LS: upscaling from the fit of Truncated Negative Binomial and the analytical Logseries to abundances in the sample; LSE: linear extension of the distribution of estimated population sizes; ABC: approximated Bayesian Computation for estimated population sizes distribution. ABC and all bias corrections are derived from simulated samples with conspecific clumping from a logseries community. Bars depict bias-corrected 95% confidence intervals or similar (credible interval for ABC).

Figure 3. Extension of assumed logseries of estimated population sizes to predict the number of species in Amazonian forests. Grey dots in both panels are the estimated total population sizes of species recorded in the 2019 ATDN data set. The solid blue line in the main figure is the rank-abundance relationship predicted by a logseries with the average of estimates of number of species (15,874 spp). Dotted lines in the main panel delimit the rank-abundance for minimum and maximum of lower and upper limits 95% of the estimates. The lines in the inset panel are the mean and lower and upper bounds for the values of abundances estimates of the recorded species, also from averaging over all estimation methods.

Table 2. ABC model selection. Posterior probabilities of each combination of regional SAD (PLN, LS, TNB) and conspecific distribution (clumped or random) in simulated samples to approximate observed distribution of estimated population sizes, for the three datasets. ABC posterior probabilities estimates the probability of each simulation model to output an acceptable approximation of empirical data. Sizes of posterior samples: 600 (2013), 508 (2013 updated) and 570 (2019).
set because the rarer the species, the more aggregated it was. This positive relationship between abundance and aggregation was found at local scales in tropical forests. For ATDN data sets this pattern results from larger scale processes, such as environmental differences and dispersal limitation among plots, and is in line with the hyperdominance of a few, abundant species in Amazon, and also with core-satellite hypotheses. The relationship is linear in log-log scale, which means that the degree of aggregation (expressed by the inverse of parameter $k$ in Eq. 8, Fig. S8) increases by a power function with the mean density of species. Thus, the expected occupancy probability of unrecorded species drops quickly as more plots are added to the sample and the more abundant species are recorded, making the collectors curve rise slower than expected by random distribution of species. Thus, completing a list of species for Amazon is a long-term task that depends of increasing effort in data gathering. Such efforts include the expansion of plot-based inventory networks, such as ATDN, the increase of collecting efforts in existing plots, make fertile material available in herbaria for taxonomic experts, support taxonomic work, exchange information between field ecologists and taxonomists. It also depends on other strategies to optimize the chances of adding new records to the list of Amazonian tree species, such as quick species assessments in target areas, and compilation of curated herbarium data.

The main source of variation of richness estimates were the estimation methods. Therefore, which method should we trust? Considering that the current number of tree species recorded for the Amazon region is a little over 10,000 methods that provide uncorrected estimates much below this value are definitely inappropriate for estimating the total richness of the Amazonian tree flora. In fact, PLN had poor support from all data sets. Using non-parametric methods would also result in estimates between 5,000 and 6,000, leading to rejection. Non-parametric methods are based on assumptions that do not meet sampling with tree inventory plots in large areas such as the Amazon, and tend to hugely underestimate richness when sampling is below 40% – sampling here was 0.00035%. We have no good explanation why the PLN had such poor support from the data, however.

TNB provided more accurate estimates of species richness than LS when applied to simulated samples with aggregation taken from a lognormal distribution with 5,000 species, and also when applied to empirical data sets. We added to these results a comparison of LS and TNB with information theoretical criteria, showing that TNB had better support from the distribution of abundances in the samples. Nevertheless, our more comprehensive simulations (simulated samples from LS and TNB, with and without clumping and with species richness ranging from 10,000 to 20,000) showed that the degree of clumping found in our data sets can make TNB fit better even if the sampled distribution is a LS. Although both LS and TNB models assume random sampling (the mean field assumption sensu), TNB has an additional scaling parameter that allows a better fit to different types shapes of SADs, and thus can accommodate better the increased dominance in abundance samples caused by conspecific clumping. Such versatility comes at the cost of a selection bias against LS when the species are not distributed randomly at the scale of sampling units. Indeed, the distribution of total population sizes were best approximated by simulated samples of LS metacommunities than by samples from TNB. Moreover, when fitted to samples from LS, TNB provided estimates with low precision. Combined, our current and previous simulations suggest that benefits of using TNB to estimate species richness and as a model for metacommunity SADs still needs further investigation.

Looking towards the other alternatives (LSE: 16,741; ABC: 14,941; LS: 14,996, Fig. 2), the choice of the best estimate for the entire Amazon still is not an easy one, because differences between them are not trivial in numbers. Differences in thousands of species are around 10% for the Amazon tree species richness but they are bigger than the tree diversity of North America (680) or Europe (250). Recent reviews with simulated data support both LS and LSE as accurate models to upscale SADs from samples to metacommunities. The effectivity of ABC for inferring community parameters has not been evaluated systematically so far (but see for a particular example). Nevertheless, ABC is unique among methods we used because it estimates richness directly from the
simulations that allow different degrees of dispersion among species. As detailed above, these simulations were also used to correct ad hoc the bias of the other methods.

Given all the considerations outlined above, we argue here that using these different methods and taking into account their pros and cons, the most reliable estimate is that of a weighted average providing an upper and lower bound of the estimated species richness.

Because the most up-to-date dataset already contains 5,027 tree species and considering that the 1,946 plots constitute a very small sample of the complete Amazonian forest (0.00035%), estimates close to that number are clearly an underestimate. Any trustable estimate should at least be more than the ~10,000 tree species already collected in Amazonian forests. The LS, LSE and ABC estimations showed a wide range of richness, so which species estimate is most believable? As we cannot make a definite choice for any of the parametric methods (after all, the real number of species is not known), we suggest that the most probable estimate for the number of tree species in Amazonia is 15,874 species, based on a logseries with conspecific aggregation. We have shown that increasing the plot effort two or ten-fold will not contain more than 50% of the total estimated richness. A ten-fold increase of the current number of ~1-ha plots would still represent only 0.00335% of the forest area of Amazonia. Significantly increasing either the plot effort or implementing new, more intensive local sampling schemes (e.g.,) seem to be inconceivable in any near future. Therefore, the number of tree species in Amazonia would remain as much an estimate as it is now and even a rather intensive well planned collecting campaign27,28 will only resolve part of the “dark diversity” of trees in Amazonia.

Methods

To delineate the surface area for estimation, we created a base map of Amazonia, the borders of which were the same as those in our earlier estimate1. Following22, we gridded this landscape into 0.1-degree grid cells (01DGC) and eliminated all 01DGCs that were more than 50% open water31, non-forest vegetation such as open wetlands or savannahs33,34, or >500 m elevation35. We quantified the area of all individual 1-degree grid cells (1DGCs), which varies with latitude due to distance from the equator (~124 km² at the equator, ~106 km² at 14° S, and ~120 km² at 8° N). The final forest map consisted of 46,986 01DGCs, totalling 5.79 million km² of forest area (Supplementary Fig. 10). This is considered the original extent of Amazonian forests. We made no correction for deforestation22, as all plots are in undisturbed forest. Therefore, predictions of the population sizes and species richness estimates relate to the original Amazonian forest cover.

Tree density. Our tree-inventory data are from the ATDN network2,5. March 2019, the ATDN network comprised of 1,946 (1,774 1-ha, 146 < 0.5 ha, 26 > 2 ha) tree inventory plots with information on species composition and abundances, and an additional set of 274 plots for which only tree density is known. These forest plots are scattered throughout Amazonia (Supplementary Fig. 11) and located in all the major forest-soil combinations1. The total plot area of the 1,946 plots with composition data is 2,042 ha. Our composition plot sample thus amounts to 0.00035% of the Amazon forest area.

The methods we used to estimate the density of trees ≥10 cm dbh, species population sizes, and distribution are similar to those of1,22. From the 2,220 (= 1,946 + 274) plots with known tree density we removed outlier plots with less than 200 (59 plots) and plots with over 900 stems (105 plots) (Supplementary Fig. 12). We constructed a loess regression model for tree density (stems ha⁻¹) based on the observed tree density in the remaining 2,056 plots (using latitude, longitude, and their interaction as independent variables). The span was set at 0.5 to yield a relatively smooth average. The model was used to estimate average tree density (D1DGC, stem/ha) in each 1-degree grid cell (1DGC). This average density was then multiplied by the total forested area of each DGC (see above) to obtain the total number of expected trees in the DGC.

Population sizes and species distributions. Analyses of tree species composition were carried out using the 1,946 plots having species composition. Species synonymy was updated following6,9, which resulted in a reduction of almost 10% of the species observed in our sample in comparison to the 2013 version of the ATDN database. Species with a “cf.” identification were accepted as belonging to the named species, while those with “sp.” were tabulated at the genus level and therefore removed from the analysis.

While we assume that identification error is within acceptable limits for common species (see discussion in1), plots vary in the proportion of individuals identified to species. Plots in which this proportion is 75% or greater (1,695 plots) were used for the population estimates of all species. Additionally, for each species we added those plots of the remaining 251 in which the species was identified positively, assuming that where species are known, they usually are locally common and where they are unknown, they are locally rare. In doing so, we assume that this does not add too many false positives. At the same time, we avoid adding too many false negatives, when using the plots with poor resolution in species identifications. Therefore, the number of plots used for the calculation of the population size differed across species.

The number of trees belonging to each species in the 1DGC was estimated following2,22. Abundances of all valid species were converted to relative abundances (fractions) for each plot: RAi = ni / Ni, where ni equals the number of individuals of species i and Ni the total number of trees in the plot (including unidentified trees). For all species with a valid name in the 1,946 plots, we constructed an inverse distance weighting (IDW) model for RAi, with a distance-decay parameter of 4 degrees, a maximum number of plots used for each local estimation of 150, and a maximum distance parameter of 4 degrees. The number of individuals of species i in a given 1DGC was then simply calculated as the total number of trees in the 1DGC (D1DGC) multiplied by the fraction of the species i for that same 1DGC.

Amazonian tree-species richness. We provide estimates for three different versions of the tree inventory data: the 2013 version which contained less plots (1,170) and used an old taxonomy (hereafter the 2013 dataset), the updated 2013 version which also contains less plots (1,162) but uses the updated taxonomy (the updated 2013
dataset), and the 2019 version which includes all 1,946 plots and uses the updated taxonomy (the 2019 dataset). For each version of the data (Table S1), we estimated species richness for the original forest area of Amazonia using two different approaches: (i) extrapolation from the distribution of estimated total population sizes\(^1\); and (ii) parametric methods, using the species abundances recorded in the sampled plots (Fig. 1, right panels). The parametric methods include the fit of the logseries (LS)\(^{12,24,36}\), the Poisson lognormal (PLN)\(^{13,37}\), and the negative binomial (NB)\(^{12}\) to the empirical SAD of the plot data (Fig. 1, right panels) and the upscaling of the model fits to the total area of Amazonia. A summary of each estimation method follows below (more details are available in the supplementary material).

**Logseries extension (LSE).** Using the population sizes of all Amazonian trees as in refs. \(^1,22\) (and as outlined above), this method expands the species abundance distribution of these population sizes down to the species with only 1 individual. Under the assumption of an underlying logseries SAD (see description of logseries below), such expansion is well approximated by a linear extrapolation from the central part of the empirical SAD\(^1,10\). We calculated bootstrap confidence intervals of the total number of species estimated by LSE estimates using the standard deviations of the estimated population sizes, based on 500 bootstraps of the plot data (supplementary info page 5).

**Logseries analytical (LS).** The logseries was among the first attempts to mathematically describe the relationship between the number of species and number of individuals in random samples from ecological communities\(^{36}\) as:

\[
S = \alpha \ln(1 + N/\alpha)
\]

where \(\alpha\) is the single free parameter of the logseries, which can be estimated from the distribution of species abundances in a sample. We fitted the logseries to the empirical SAD and then used the estimated value of \(\alpha\) to estimate number of species predicted by Eq. (1) using \(N\) as the total number of trees estimated for the whole Amazon (supplementary info page 5).

**Zero-truncated poisson lognormal (PLN).** Like the logseries, the PLN was developed based on the sampling theory of SADs\(^{24,38}\). It assumes that the observed SAD can be described as a Poisson sample of a regional SAD that follows a lognormal distribution, which is approximated by the ‘eVil line’ truncation of the lognormal\(^{39}\). The PLN fitted to empirical SADs is truncated at zero, as species with no individuals recorded in the sample are unknown. As any zero-truncated distribution, the PLN fitted to an empirical SAD allows the calculation of the proportion of species that have not been sampled, thus allowing to estimate the total number of species in the sampled community\(^{13,37}\) (supplementary info page 6).

**Zero-Truncated Negative binomial (TNB).** The TNB also results from a Poisson sample, yet from a Gamma distribution. It has two free parameters, \(r\) and \(\xi\), and at the limit \(r \to 0\) the TNB converges to the LS\(^{12,24,36,38}\). After fitting the TNB to the empirical SAD, we used the sampling intensity \(p\) (the proportion of all individuals included in the sample or the proportion of total area covered by the sample) to estimate the number of species in the entire Amazon (\(S\)) using the following equation\(^{12}\):

\[
S = S_p(1 - (1 - \xi)^r)/(1 - (1 - \xi)^r)
\]

where, \(\xi = \xi_p/(p + (1 - p) \xi_p)\).

The fit of the LS, PLN and TNB models to the empirical SAD, was performed using maximum likelihood techniques with functions from the ‘sads’ R Package\(^40\) (supplementary info page 13). The support that each data set provide for each of these three competing models was gauged by Akaike Information Criterion (AIC) and also by the Bayesian Information Criterion (BIC). (supplementary info page 6).

**Adding conspecific aggregation.** Assuming that the trees in our plots constitute a sample from the unknown regional Amazonian SAD, we tested if sampling from this theoretical SAD could provide accurate estimates of the calculated population sizes and total species richness. We performed this assessment by upscaling the LS, PLN and TNB for the entire Amazonia. We then simulated 1,946 random draws of 1-ha plots from regional SADs generated by these models, with and without conspecific aggregation. In both cases, we assumed that the expected abundance of each species in each plot was its mean density (ha\(^{-1}\)) estimated for the whole Amazon (ha\(^{-1}\) to 20,000 samples). For each simulation and each estimation method (LSE, LS, PLN or TNB), we applied the same methods described above to estimate the total species richness. We then estimated the bias of each method, defined here as the mean difference between the known values of species richness in the theoretical SAD and the richness estimated by each method. We used the estimated bias of each method to calculate their bias-corrected species richness.

For simulations with conspecific aggregation, we allowed species to have different degrees of spatial aggregation, that is, different values of the dispersion parameter \(k\) of the negative binomial distribution. We obtained species-specific values of \(k\) based on the relationship between the estimated the values of \(k\) and the mean density of each species observed in the 1,946 plots (supplementary material page 7).
Moreover, we assessed which combination of theoretical SAD (LSE, LS, PLN or TNB) and sampling scheme (random or aggregated) best approximated the distribution of estimated population sizes. We evaluated the performance of each combination by comparing their proportion among of the set of simulations that best approximated the distributions of calculated population sizes, using Approximate Bayesian Computation (ABC)\(^4\), which was also used to build the credible intervals for the species richness estimates (supplementary material page 9).

**Influence of increasing sample size.** We also used simulations to estimate the number of additional species that would be recorded if we increased the sample size (i.e., adding new 1-ha plots). We simulated the sampling procedure from a logseries with conspecific clumping, as described above. This logseries had a total number of trees equal to the estimated number of trees from the ATDN data set from 2019 (estimate presented in the results). We simulated samples with sample size varying between 1,046 and 19,460 1-ha plots. This range of sample sizes correspond to 90% of the number of plots in the 2013 updated data set (1,162 plots) to ten times the number the plots in the 2019 data set (19,460 plots). For each sample size we repeated the simulations 100 times and then calculated the mean number of species recorded in the simulated sample. The 100 simulations were also used to calculate the lower (5%) and upper bounds (95%) for each sample size.

Received: 21 November 2019; Accepted: 19 May 2020; Published online: 23 June 2020

References

1. ter Steege, H. et al. Hyperdominance in the Amazonian tree flora. Science 342, 1243092, https://doi.org/10.1126/science.1243092 (2013).
2. Fung, T., Villain, L. & Chisholm, R. A. Analytical formulae for computing dominance from species-abundance distributions. Journal of Theoretical Biology 386, 147–158, https://doi.org/10.1016/j.jtbi.2015.09.011 (2015).
3. Ricklefs, R. E. How tree species fill geographic and ecological space in eastern North America. Annals of Botany 115, 949–959, https://doi.org/10.1093/aob/mcv029 (2015).
4. Cardoso, D. et al. Amazon plant diversity revealed by a taxonomically verified species list. Proceedings of the National Academy of Sciences 114, 10695–10700, https://doi.org/10.1073/pnas.1706756114 (2017).
5. Harte, J. & Kitszes, J. Inferring regional-scale species diversity from small-plot censuses. PloS One 10, e0117527, https://doi.org/10.1371/journal (2015).
6. ter Steege, H. et al. Estimating species richness in hyper-diverse large tree communities. Ecology 98, 1444–1454, https://doi.org/10.1002/ecy.1813 (2017).
7. ter Steege, H. et al. The discovery of the Amazonian tree flora with an updated checklist of all known tree taxa. Scientific Reports 6, 29549, https://doi.org/10.1038/srep29549 (2016).
8. ter Steege, H. et al. Towards a dynamic list of Amazonian tree species. Scientific Reports 9, 3501, https://doi.org/10.1038/s41598-019-40101-y (2019).
9. ter Steege, H. & al. e. Amazon Tree Diversity Network, http://atdn.myspecies.info/ (2019).
10. Ulrich, W. & Ollik, M. Limits to the estimation of species richness: the use of relative abundance distributions. Diversity and Distributions 11, 265–273 (2005).
11. Baldrige, E., Harris, D. J., Xiao, X. & White, E. P. An extensive comparison of species-abundance distribution models. PeerJ 4, e2823, https://doi.org/10.7717/peerj.2823 (2016).
12. Toivo, A. et al. Upscaling species richness and abundances in tropical forests. Science Advances 3, e1701438, https://doi.org/10.1126/sciadv.1701438 (2017).
13. Bulmer, M. G. On fitting the Poisson lognormal distribution to species abundance data. Biometrics 30, 101–110 (1974).
14. Engen, S., Lande, R., Walla, T. & DeVries, P. J. Analyzing spatial structure of communities using the two-dimensional Poisson lognormal species abundance model. The American Naturalist 160, 60–73, https://doi.org/10.1086/340612 (2002).
15. Duque, A. et al. Insights into regional patterns of Amazonian forest structure, and dominance from three large terra-firme forest dynamics plots. Biodiversity and Conservation 26, 669–686, https://doi.org/10.1007/s10531-016-1265-9 (2017).
16. Chao, A., Colwell, R. K., Lin, C.-W. & Gotelli, N. J. Sufficient sampling for asymptotic minimum species richness estimators. Ecology 90, 1125–1133, https://doi.org/10.1890/07-2147.1 (2009).
17. Branco, M., Figueiras, F. G. & Cermeño, P. Assessing the efficiency of non-parametric estimators of species richness for marine microplankton. Journal of Plankton Research 40, 230–243 (2018).
18. Ulrich, W., Kusumoto, B., Fattorini, S. & Kubota, Y. Factors influencing the precision of species richness estimation in Japanese vascular plants. Diversity and Distributions, https://doi.org/10.1111/dad.13049 (2020).
19. Condit, R. et al. Spatial patterns in the distribution of tropical tree species. Science 288, 1414, https://doi.org/10.1126/science.288.5470.1414 (2000).
20. Plotkin, J. B. & Muller-Landau, H. C. Sampling the species composition of a landscape. Ecology 83, 3344–3356, https://doi.org/10.1890/0012-9658 (2002).
21. Plotkin, J. B. et al. Species-area curves, spatial aggregation, and habitat specialization in tropical forests. Journal of Theoretical Biology 207, 81–99, https://doi.org/10.1006/jtbi.2000.2158 (2000).
22. ter Steege, H. et al. Estimating the global conservation status of over 15,000 Amazonian tree species. Science Advances 1, e1500936, https://doi.org/10.1126/sciadv.1500936 (2015).
23. McGill, B. J. et al. Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework. Ecology Letters 10, 995–1015 (2007).
24. Green, J. L. & Plotkin, J. B. A statistical theory for sampling species abundances. Ecology Letters 10, 1037–1045 (2007).
25. Hanski, I. Dynamics of regional distribution: the core and satellite species hypothesis. Ökos 87, 210–221 (1982).
26. Magurran, A. E. & Henderson, P. A. Explaining the excess of rare species in natural species abundance distributions. Nature 422, 714–716 (2003).
27. Baker, T. R. et al. Maximising Synergy among Tropical Plant Systematists, Ecologists, and Evolutionary Biologists. Trends in Ecology & Evolution 32, 258–267, https://doi.org/10.1016/j.tree.2017.01.007 (2017).
28. Hopkins, M. J. G. Are we close to knowing the plant diversity of the Amazon? Anais da Academia Brasileira de Ciências 91, e20190396, https://doi.org/10.1590/0011-3765201920190396 (2019).
29. Elias, T. S. The Complete Trees of North America. Field Guide and Natural History. (Times Mirror Magazines., 1980).
30. Kunin, W. E. et al. Upscaling biodiversity: estimating the species–area relationship from small samples. Ecological Monographs 88, 170–187, https://doi.org/10.1002/ecm.1284 (2018).
31. Jabot, F. & Chave, J. Analyzing tropical forest tree species abundance distributions using a nonneutral model and through approximate Bayesian inference. The American Naturalist 178, E37–47 (2011).
Acknowledgements
This paper is the result of the work of hundreds of different scientists and research institutions in the Amazon over the past 80 years. Without their hard work this analysis would have been impossible. We thank Charles Zartman for the use of plots from Jutai. HtS, VFG, and RS were supported by grant 407232/2013-3 - PVE - MEC/MCTI/CAPES/CNPq/FAPs; PIP had support for this work from CNPq (productivity grant 310885/2017-5) and FAPESP (research grant #550373/2010-1 and # 550373/2010-1 and # 457515/2012-0, and JLLM were supported by grant CAPES/PDSE # 88881.135761/2016-01 and CAPES/Fapespa #1530801; The Brazilian National Research Council (CNPq) provided a productivity grant to EMV (Grant 308040/2017-1); Floristic identification in plots in the RAINFOR forest monitoring network have been supported by the Natural Environment Research Council (grants NE/B503384/1, NE/D01025X/1, NE/I029882/1, NE/F005806/1, NE/D005590/1 and NE/I028122/1) and the Gordon and Betty Moore Foundation; BMF is funded by FAPESP grant 2016/25086-3. BSM, BHMJ and OLP were supported by grants CNPq/CAPES/FAPs/BC-Newton Fund #441244/2016-5 and FAPEMAT/0589267/2016; TWH was funded by National Science Foundation grant DEB-1556338. The 25–ha Long-Term Ecological Research Project of Amacayacu is a collaborative project of the Instituto Amazónico de Investigaciones Científicas Sinchi and the Universidad Nacional de Colombia sede Medellín, in partnership with the Unidad de Manejo Especial de Parques Nacionales Nacionales and the Center for Tropical Forest Science of the Smithsonian Tropical Research Institute (CTFS). The Amacayacu Forest Dynamics Plot is part of the Center for Tropical Forest Science, a global network of large-scale demographic tree plots. We acknowledge the Director and staff of the Amacayacu National Park for supporting and maintaining the project in this National Park.

Author contributions
H.t.S. initiated the study; P.I.P. and H.t.S. carried out the analyses, H.t.S., P.I.P., R.A.F.L. and E.P. wrote the manuscript. All members of ATDN provided tree inventory data. All authors reviewed and added comments and additions on/to the manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-66686-3.

Correspondence and requests for materials should be addressed to H.t.S. or P.I.P.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2020