A review of Non-Parametric and Parametric Models for Species Richness Estimation.

Evans Otieno Ochiaga[1] & Frederic Ntirenganya[2]

[1] Phastar Limited Company, Nairobi, Kenya

[2] University of Lay Adventists of Kigali, Rwanda

Corresponding Author: evansochiaga@aims.ac.za, fredo@aims.ac.za

ABSTRACT

Species richness estimation is one of the key concepts in conservation biology. Many models have been developed to estimate species richness: ranging from commonly used non-parametric to parametric models. However, not all the models give excellent prediction of number of species in the community. Therefore, in this paper we present and compare the performances of 5 commonly used non-parametric and 9 parametric models. In this research we use Barro Colorado Island (BCI) dataset as the assemblage with 10%, 5%, 2%, and 1% of individuals being drawn for the estimations. The overall performances of the models were done using Akaike Information Criterion variances at 100 simulations. Five non-parametric models underestimate the species richness and nine parametric models overestimate the species richness. Among all the models, abundance coverage estimate model performed the best.

Key words: Non-parametric and parametric models, species richness

INTRODUCTION

Information about the total number of species in a community, ecosystem or geographical area plays a key role in ecology and biogeography (Chiarucci 2012). In fact, knowing how many species live in a region is not only useful for biodiversity conservation but it also contributes to the species extinction risk assessment. These risks can either be triggered by internal biotic interactions and external human disturbances (Xu, et al. 2012).

Number of species can be obtained by conducting census in each community or heterogeneous habitat. This enumeration is often designed to estimate species component of the community (Thompson, et al. 2003). Although, full enumeration of species in the community might give a good approximation of species richness, it is only possible if the community is small enough

for census to be done with reasonable effort. In reality, most ecological communities are very large such that a full censuring process is not possible; therefore, this calls for efficient species richness estimation techniques (Jobe 2008).

Despite our need for accurately assessing species richness, it is a difficult variable to measure (Gotelli and Colwell 2011). Nevertheless, quite a number of estimators are classified in three categories; extrapolation or rarefaction, non-parametric and parametric estimations have been proposed by different ecologist in predicting species richness. Although, these models give good estimates of total number of species in the community, their evaluation and identifying which model performs the best is still of research interest.

Non-parametric models provide a good platform for estimating species richness, however, they underestimate total number of species in the assemblages since rare species might not be detected in the sample (Colwell, Chao, et al. 2012). Thus, in recent years, ecologists estimate species richness by considering the relationship between the number of detected species and sampling effort (i.e. time, number of individual sampled, area, accumulation of samples). This relationship is referred as species accumulation curve with y axis representing increasing number of species and x axis increasing sampling effort (Jobe 2008).

The paper is organized such that in method section we presented some nonparametric and parametric species richness models. The discussed models use individual based (abundance) data (Colwell, Chao, et al. 2012) with number of individuals sampled as the sampling effort as well as incidence-based data. Then we discussed the evaluation of the models using BCI data as the habitat where samples were picked from. The detailed performance of the models is presented in result and discussion sections.

METHOD

Species richness estimation

For long, species richness has been estimated using different models proposed by different ecologists, Dengler and Colwell in (Dengler 2009) and (Colwell, Chao, et al. 2012) respectively provide complete lists of both parametric and non-parametric models respectively. However, for the sake of this paper, the major models presented in table 1 were considered.

Models	Formulae	References
Jackknife1	$S = S_{obs} + \left(\frac{n-1}{n}\right) f_1$	(Heltshe and Forrester 1983)
Jackknife2	$S = S_{obs} + \left\lceil \frac{f_1(2n-3)}{n} - \frac{f_2(n-1)^2}{n(n-1)} \right\rceil$	Colwell, Chao, et al. 2012)
Chao1	$S = S_{obs} + \left(\frac{n-1}{n}\right) \frac{{f_1}^2}{2f_2}$	Colwell, Chao, et al. 2012)
Chao2	$S = S_{obs} + \left(\frac{n-1}{n}\right) \frac{f_1(f_1-1)}{2(f_2+1)}$	Colwell, Chao, et al. 2012)
Boot trap	$S = S_{obs} + \sum_{k=1}^{S_{obs}} (1+P_k)^n$	Colwell, Chao, et al. 2012)
Abundance coverage estimate	$S = S_{rare} + S_{abundance}$	Colwell, Chao, et al. 2012)
Clench and Eadie	$S = (az)/(1+(bz))$	(Clench 1979)
Linear dependence	$S = a/b(1-exp(-bz))$	(LlorenteB 1993)
Negative exponential	$S = a(1-\exp(-bz))$	(Miller and Wiegert 1989)
Exponential	$S = a + (blog_{10}(z))$	(Gleason 1992)
Power law	$S = az^b$	(Preston 1962)
Logarithmic B	$S = \log(1+(abz))/b$	(Longino and Colwell 1997)
Asymptote	$S = a - (b(c^z))$	(Thompson, et al. 2003)
Chapman-Richards	$S = a((1-\exp(-bz))^c$	(Thompson, et al. 2003)
Rational	$S = (a+(bz))/(1+(cz))$	(Thompson, et al. 2003)

Table 1: Table showing major species richness estimate models (both non parametric and parametric), f_1 and f_2 are number of singletone and doubletone

species respectively and n is the sample size collected from the assemblage. S_{obs} is observed species in the sample, a, b and c are model parameters. P_k is the proportion of samples that contains species k and z is the number of individual sampled.

RESULTS

Curve fitting and evaluation of models

For the analyses, the models in table 1 with exception of boot trap model, were selected for estimating species richness. Due to lack of enough data, 9 parametric models were fitted using non-linear least square method to individual based species accumulation curves, which was simulated from the Barro colorado island (BCI) dataset. The BCI dataset used in this simulation was for 1983 census. Non-parametric models were done numerically based on observed and rare species.

Despite the differences in the method used in estimating species richness in parametric and non-parametric models, their joint performance evaluation was done using variation in expected number of species estimated by each model after 100 simulations. However, for more detailed evaluation of parametric models; Akaike information criterion (AIC) value was considered. The sampling of individuals from BCI dataset were done such that 1%, 2%, 5% and 10% of the total individuals were picked for the study. The BCI dataset has a total of 263896 individuals with 306 species. All the analyses were done in R (version 3.1.3) and the result summarised in table 2 and table 3.

Models	$AIC_{1\%}$	$AIC_{2\%}$	$AIC_{5\%}$	$AIC_{10\%}$
Jackknife1	-	-	-	-
Jackknife2	-	-	-	-
Chao1	-	-	-	-
Chao2	-	-	-	-
Abundance coverage estimate	-	-	-	-
Rational	175.7127	348.408	874.1975	1750.255
Exponential	233.9497	439.8705	1005.651	1864.907
LogB	167.5907	327.9936	828.0001	1686.479
Power	187.5839	376.1392	960.8377	1934.887
Asympote	286.5732	556.6169	1362.41	2675.757
Clench and Eadie	175.7213	354.6209	909.1418	1853.634
Linear dependence	196.9011	409.834	1074.684	2199.895
Negative exponential	196.9011	409.834	1074.684	2199.895
Chapman-Richards	263.314	502.218	1184.473	2269.498

Table 2: Table showing average Akaike information criterion (AIC) values for parametric models for 1%, 2%, 5% and 10% of individuals sampled from Barro colorado dataset. The simulation was done 100 times.

Models	Var1%	Var2%	Var5%	Var10%
Jackknife1	84.82365	59.25159	39.09453	24.1955
Jackknife2	67.213	45.14618	31.8966	17.41533
Chao1	86.67947	63.18633	44.88974	28.39127
Chao2	89.26711	65.45029	46.58909	30.24011
Abundance coverage estimate	52.39483	31.21559	22.25079	12.30863
Rational	129.1592	102.6612	72.00083	52.9845
Exponential	139.2684	106.9952	68.73661	45.59281
LogB	124.2826	95.83008	62.9926	42.80687
Power	119.2382	90.5652	58.36299	39.12588
Asympote	169.5602	140.0214	103.5548	79.56218
Clench and Eadie	129.8653	103.8484	73.96719	55.43799
Linear dependence	135.2307	111.0755	82.63107	64.25028
Negative exponential	135.2307	111.0757	82.63132	64.25022
Chapman-Richards	154.3367	128.2711	96.00202	74.4657

Table 3: Table showing variance in predicted species richness of both nonparametric and parametric models. The variance is given by; $\sqrt{\frac{\Sigma_{i=y}^{n}(y-\widehat{y_i})^2}{n}}$, where y is the expected species richness (in this case is 306), $\widehat{y_i}$ is the predicted species richness estimated using models in each simulation and n is the number of simulations (n = 100). The simulation is done 100 times using 1%, 2%, 5% and 10% of individuals sampled form Barro colorado island dataset.

Tables 2 and 3 summarize the evaluation of all models used in estimating species richness. Using AIC values only in table 2 for evaluation of parametric models, it is observed that log B model with smallest AIC value across all number of individuals sampled performs the best and asymptote model which has the largest AIC value performs poorly. On the same context, rational model performs best as compared to power model with AIC value of 1934.887 which is greater than 1750.225 for rational model when 10% of the individuals are sampled for the study.

A model that was proposed by Clench and Eadie also did well; it's the third after rational model with AIC of 1853.634 and performs better than linear dependence and negative exponential models which are the fifth with AIC value of 2199.895 each for 10% of individuals sampled. Exponential model was the fourth with AIC value of 1864.907 for 10% of individual sampled and performs best as compared to Chapman-Richards model which was the seventh with AIC value of 2269.498. Despite the fact that only 10% of individuals sampled are used in these explanations, the performance of the

models based on their AIC values were consistent across all the individuals sampled; 1%, 2%, 5% and 10%.

The parametric models considered in the analyses in this work, predicts species richness of the community through extrapolation method; for example, individuals are sampled till the predicted number of species becomes asymptotic and the asymptotic value is the expected species richness in the assemblage. For clear understanding how parametric models fit species accumulation curve, the following plots for models were obtained.

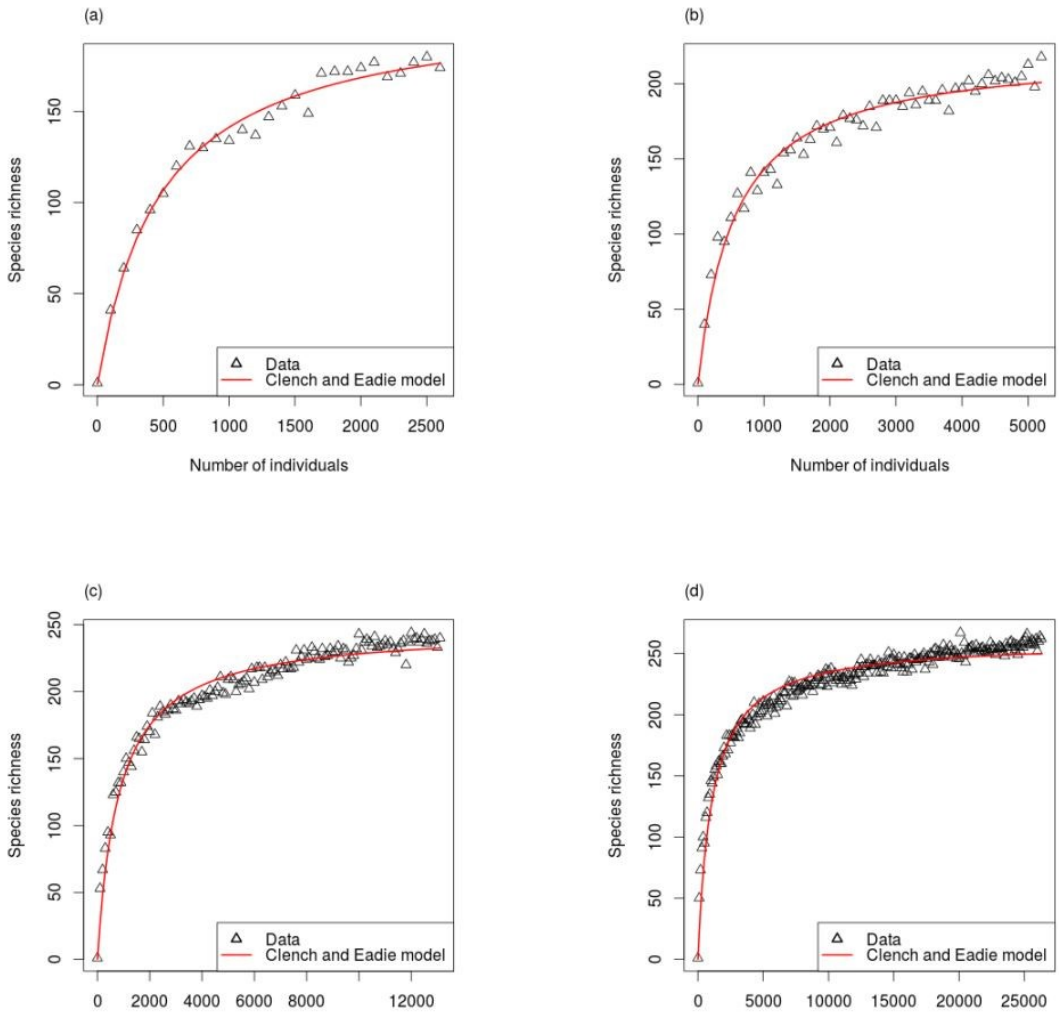

Figure 1: Figure showing fit of Clench and Eadie model to species accumulation curve with different number of individuals sampled from the BCI dataset; (a): is for 1% of individuals sampled; (b): is for 2% of individuals sampled; (c): is for 5% of individuals sampled; and lastly (d): is for 10% of individuals sampled. As the total number of individuals sampled increases the species richness estimates approaches asymptotic value.

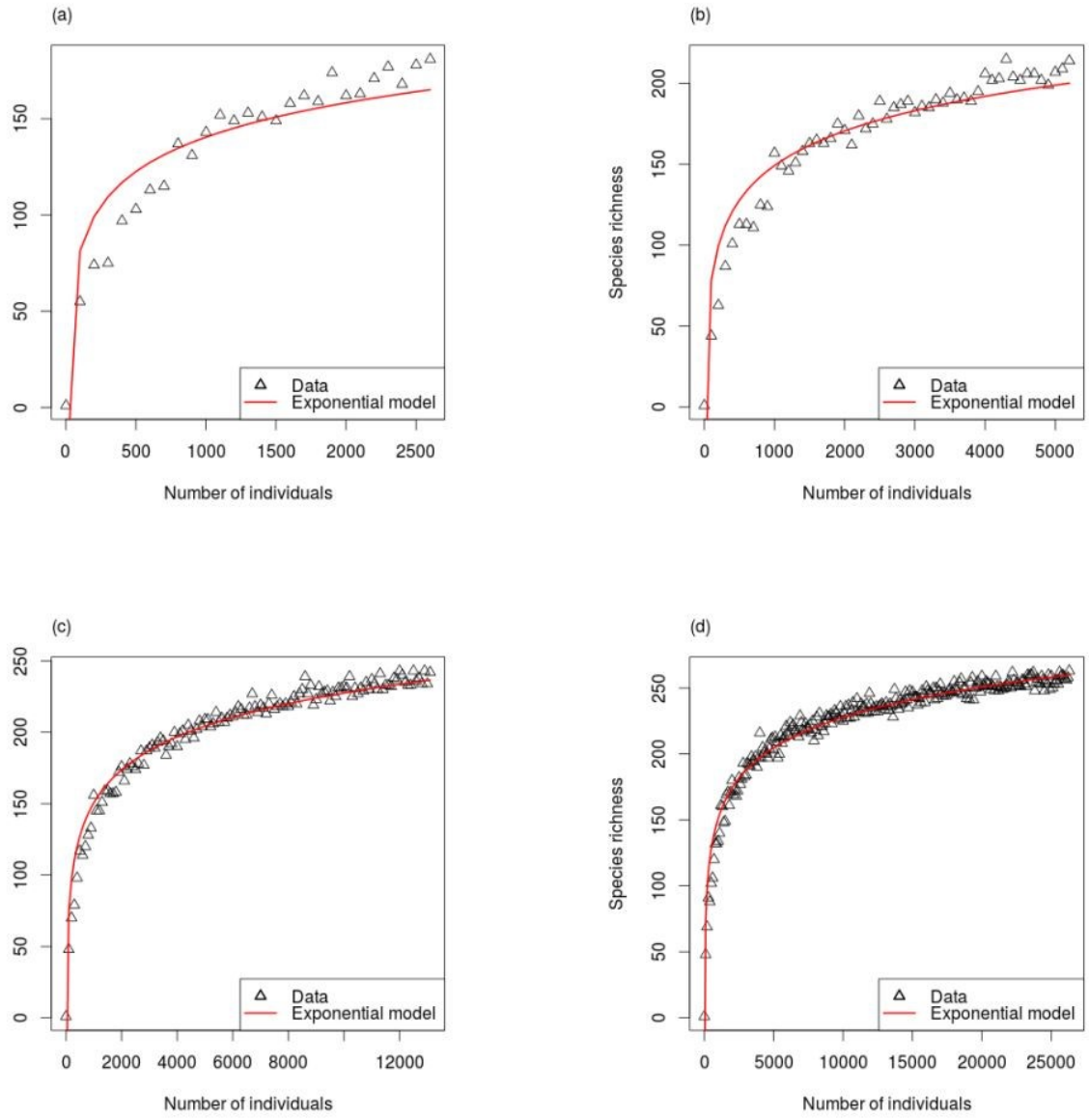

Figure 2: Figure showing fit of exponential model to species accumulation curve with different number of individuals sampled from the BCI dataset; (a): is for 1% of individuals sampled; (b): is for 2% of individuals sampled; (c): is for 5% of individuals sampled; and lastly (d): is for 10% of individuals sampled. As the number of individuals sampled increases the species richness estimates approaches asymptotic value.

Although, non-parametric models lack AIC values, since they predict species richness numerically (not fitted to species accumulation curve), their accuracy alongside parametric models in predicting expected species richness is evaluated using the variation in the predicted species richness as shown in table 3. Among all the models, it is observed that abundance coverage estimate model with small variance across all the individuals sampled performs the best. The comparison of the

results also showed that all the non-parametric models underestimate species richness while parametric models overestimated it. However, in general non-parametric models performed well as compared to parametric models due to their small variance value as opposed to the one for parametric models.

Despite the fact that, non-parametric models' predictions were solved numerically, it was possible to plot the predicted species richness for every individual sampled as shown in the figures that follows;

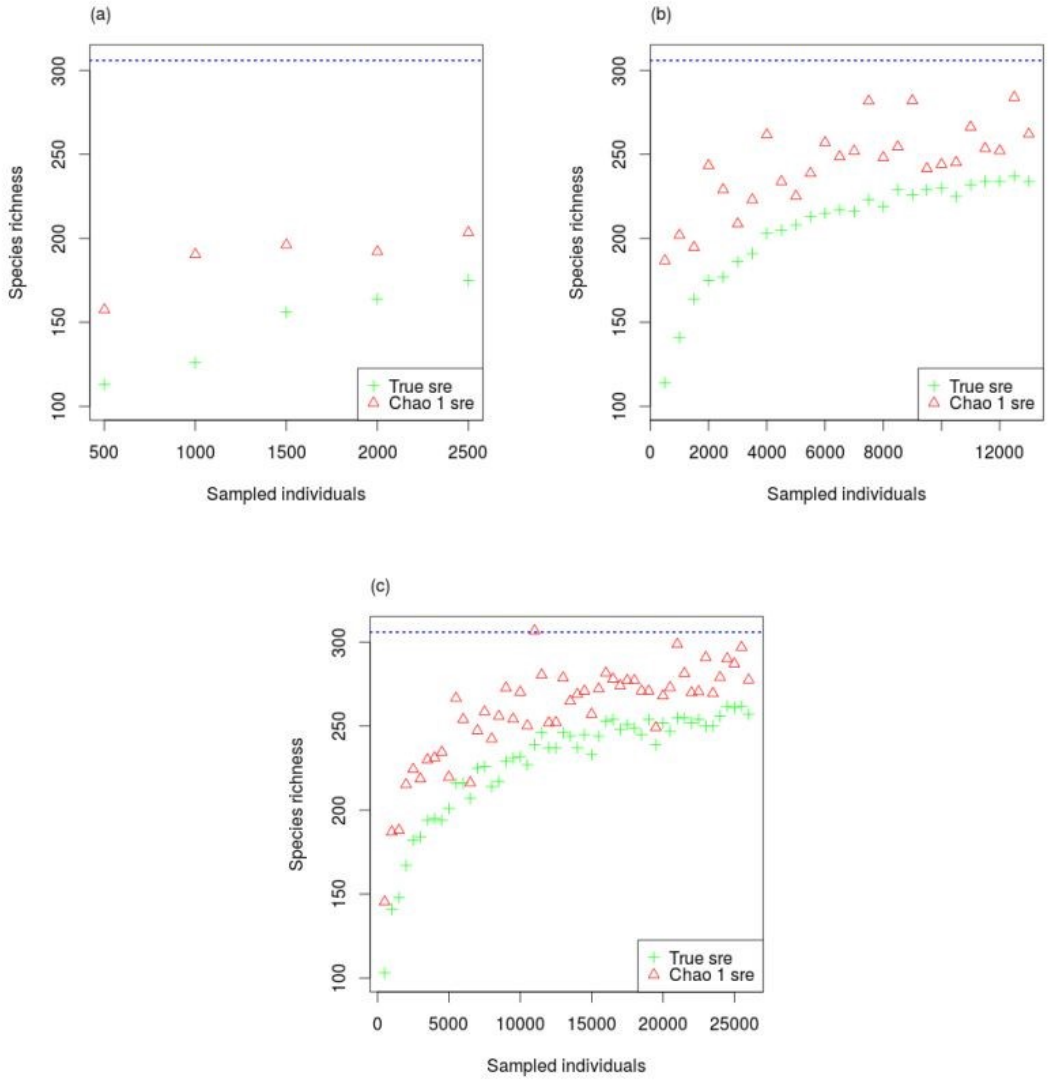

Figure 3: Figures showing plots for species richness as estimated by Chao 1 model; (a): is for 1% of individuals sampled (b): is for 5% of individuals sampled and (c): is for 10% of individuals sampled. Blue dashed line is the expected species richness which is 306 and True sre is the observed species richness in the sample of individuals

collected. The sampling was done from Barro colorado island dataset. As the percentage of individuals sampled increases, predicted species richness move close to expected species richness of the community

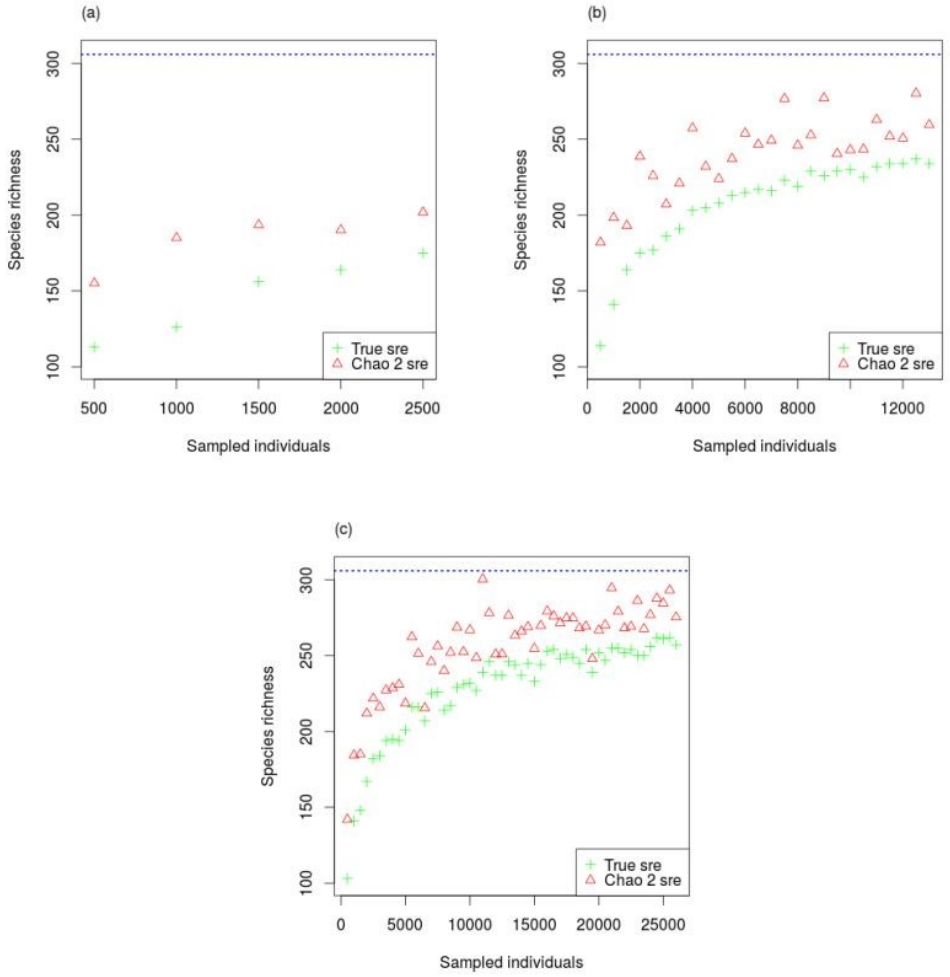

Figure 4: Figures showing plots for species richness as estimated by Chao 2 model; (a): is for 1% of individuals sampled (b): is for 5% of individuals sampled and (c): is for 10% of individuals sampled. Blue dashed line is the expected species richness which is 306 and True sre is the observed species richness in the sample of individuals collected. The sampling was done from Barro colorado island dataset. As the percentage of individuals sampled increases, predicted species richness move close to expected species richness of the community.

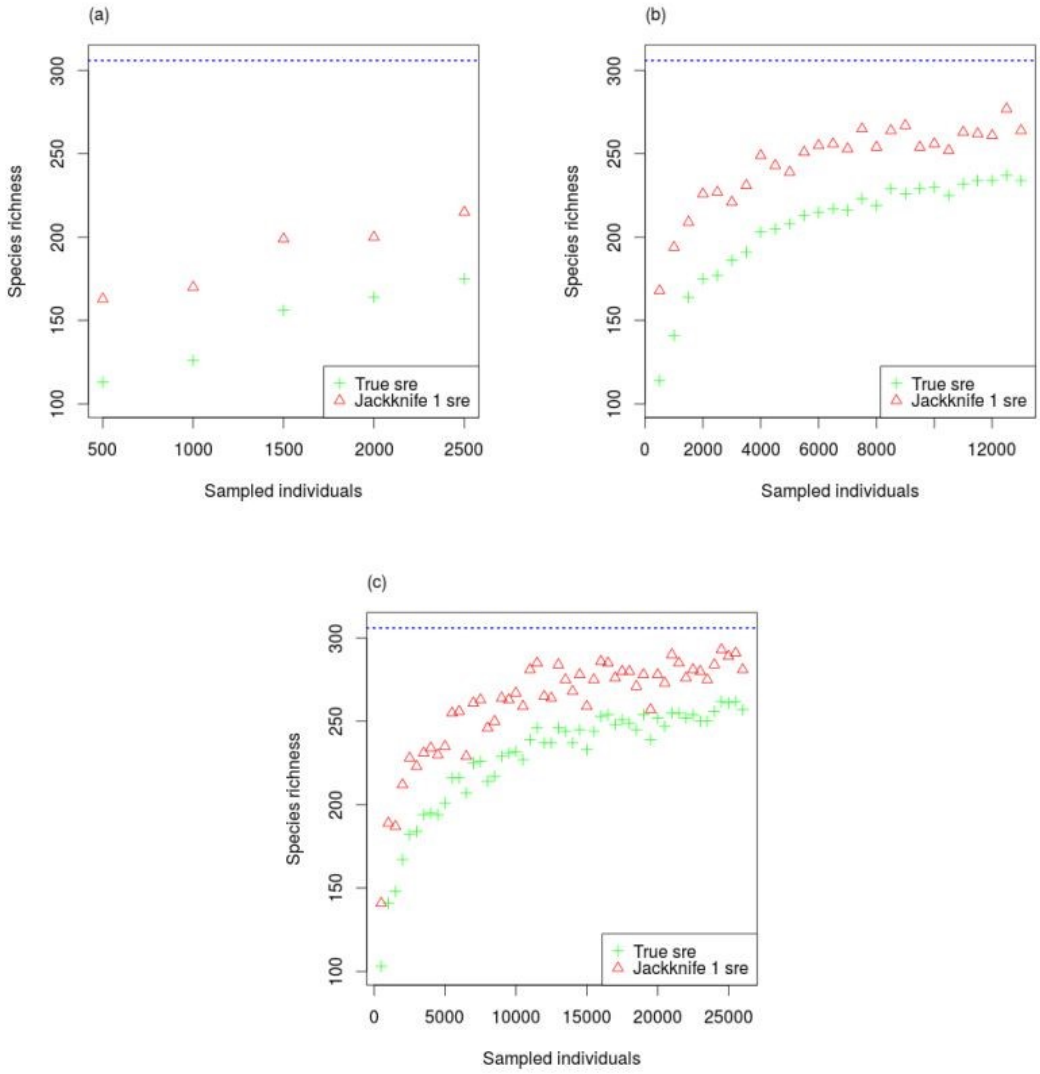

Figure 5: Figures showing plots for species richness as estimated by Jackknife 1 model; (a): is for 1% of individuals sampled (b): is for 5% of individuals sampled and (c): is for 10% of individuals sampled. Blue dashed line is the expected species richness which is 306 and True sre is the observed species richness in the sample of individuals collected. The sampling was done from Barro colorado island dataset. As the percentage of individuals sampled increases, predicted species richness move close to expected species richness of the community.

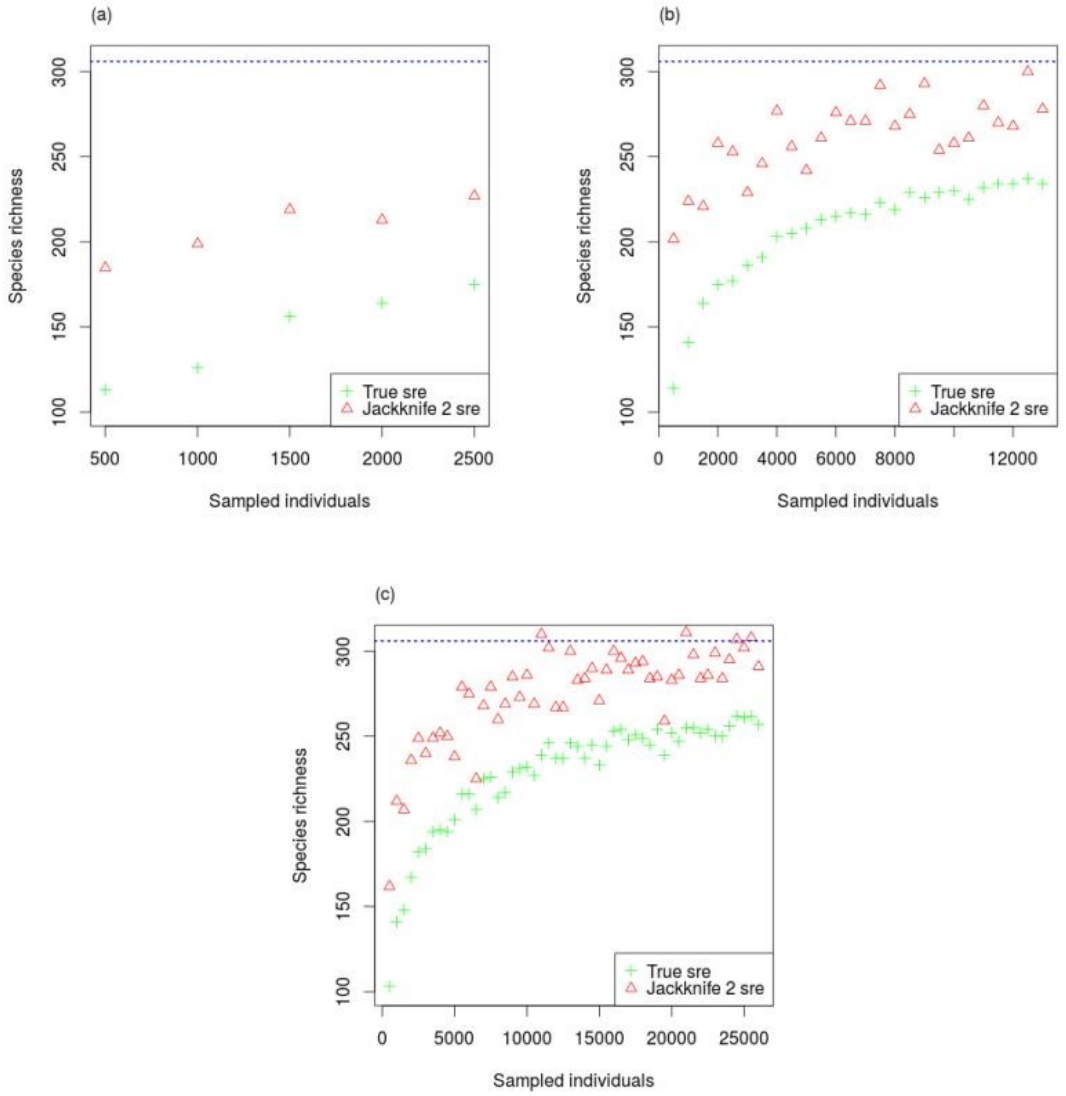

Figure 6: Figures showing plots for species richness as estimated by Jackknife 2 model; (a): is for 1% of individuals sampled (b): is for 5% of individuals sampled and (c): is for 10% of individuals sampled. Blue dashed line is the expected species richness which is 306 and True sre is the observed species richness in the sample of individuals collected. The sampling was done from Barro colorado island dataset. As the percentage of individuals sampled increases, predicted species richness move close to expected species richness of the community.

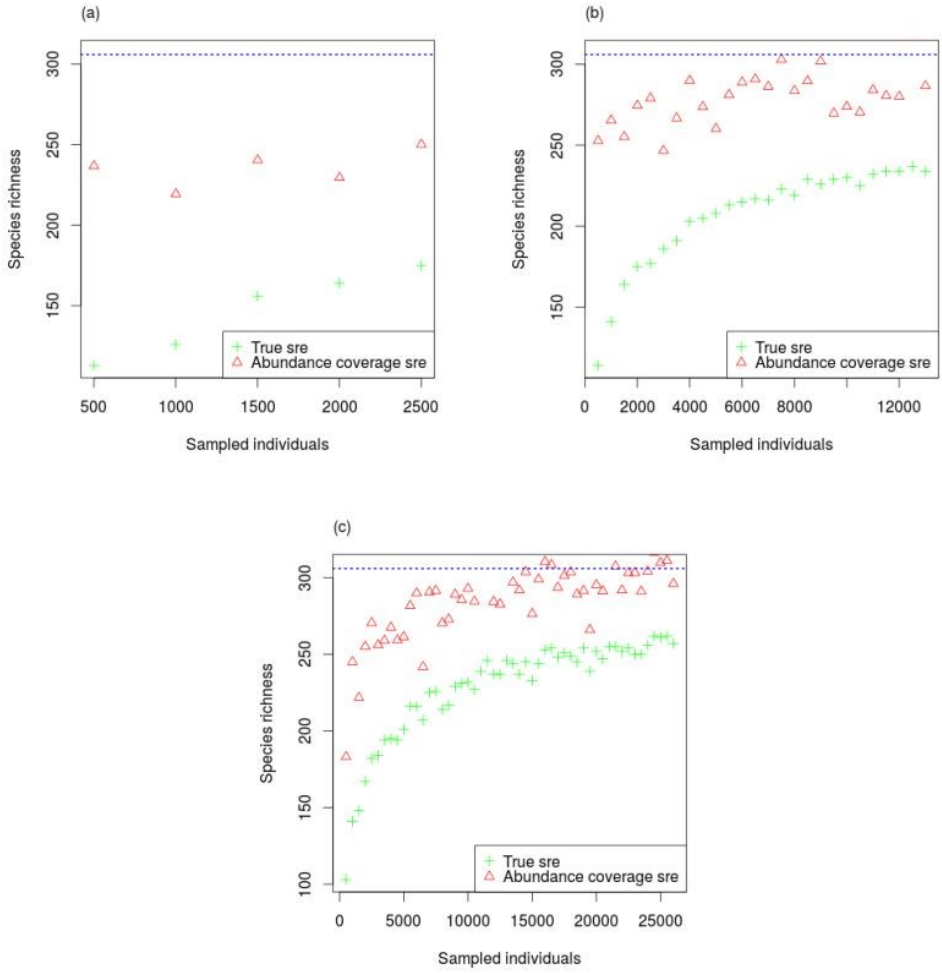

Figure 7: Figures showing plots for species richness as estimated by abundance coverage model; (a): is for 1% of individuals sampled (b): is for 5% of individuals sampled and (c): is for 10% of individuals sampled. Blue dashed line is the expected species richness which is 306 and True sre is the observed species richness in the sample of individuals collected. The sampling was done from Barro colorado island dataset. As the percentage of individuals sampled increases, predicted species richness move close to expected species richness of the community.

DISCUSSION

Species richness is a key issue in biodiversity (Hortal, Borges and Gaspar 2006). That is, estimating number of species in a community helps in assessing impact of human disturbance and climatic change on biodiversity (Xu, et al. 2012). Since total number of species in the community is unknown, estimators are used to help in describing species richness.

Although non-parametric and parametric models have been proposed by ecologists, not all of them give satisfactory extrapolated species richness (Palmer 1990). In fact none of them gives an exact

prediction, for clarity, non-parametric models which utilize rare and observed species approximately underestimate number of species in the community while parametric models which are proposed to analyse increase in number of species with increase in sampling effort overestimate species richness estimation (Gotelli and Colwell 2011) .

Despite limitations of these models, they are still used provided they give a good picture of species richness estimation (Hortal, Borges and Gaspar 2006). However, using AIC values in table 2, it is important to note the good performance of logarithmic B and power models which are also observed in (Dengler 2009). On the same context, it is also reasonable to note the weak performance of asymptote and Chapman-Richards models proposed in (Thompson, et al. 2003), due to their large AIC values. Therefore, the use of these models in predicting species richness shouldn't be given first priority until their performance in future proves to be convincing. All the same, with the exception of these two models, the rest performs fairly well. Exponential model has better performance; however, it lacks upward asymptote. Despite all the limitations associated with both non-parametric and parametric models, model that was developed by Clench and Eadie arguments performs fairly good and similar pattern is observed in (Hortal, Borges and Gaspar 2006).

In estimating species richness using species accumulation curve, it's believed in ecological literature that asymptotic value of the model is the expected number of species in the community under study (Dengler 2009). This theory might not be the case in all the models; for example, we observe that some models lack upper asymptote but they give a good prediction. For clarity, rational model which delays to reach its' asymptotic value is the second best among parametric models considered in this work. On the other hand, Chapman Richards model with asymptotic value performs poorly. This pattern represents behaviour of other species richness models which are not part of this work.

The parametric models were evaluated using AIC value, however, in table 3 it is observed that the variance in predicted species richness by the models varies based on the number of individuals sampled. This implies that a model can perform better as far as species richness estimation is concerned, though there might be inconsistency in the variation in expected species richness, therefore sample size must be good enough to evaluate their performance.

For a couple of years, non-parametric models has been questioned in estimating species richness since they underestimate the number of species in the community (Hortal, Borges and Gaspar 2006). Nevertheless, they play a central role in species richness estimation in the absence of species complete count and they can't be ignored completely. In fact, in general they perform better as compared to parametric models as shown in table 3. Among non-parametric models, abundance coverage model which was derived based on abundance of individuals in the sample proves to be good as compared to the rest of the model.

All the species richness models; non-parametric and parametric models, none of them give exact estimation of species richness in the community i.e. some overestimate while others underestimate. However, despite the current development in theoretical framework of the species richness models, finding a model that gives exact species richness is still difficult. Although, Darwin describes how species originated, we are still unable to figure out how many species goes extinction process due to human activities so that we get the total number of species in the community (Chiarucci 2012). Hence in conclusion, species richness is still a long way to go and more research should be done in this area.

Conclusion

Conservation biology is a field that plays a role in proper management and planning of our community. In this paper species richness which plays a key role in conservation biology has been discussed. Knowing how many species are there in the community helps in proper management of biodiversity, i.e. it would be easy to identify which species goes extinction process and whether to introduce new species in the community or not. In ecological literature, species richness has been estimated using both non-parametric or parametric process. Non-parametric methods use rare and observed species during sampling and parametric procedures use species accumulation curve with sampling effort being; area, individual sampled and duration of time taken during sampling, in this paper individual sampled was considered. In conclusion, for proper species richness estimation, there is still much more to be done. All the models in ecology doesn't give exact estimation of species, therefore estimating species richness is still a long way off.

Acknowledgment

First and foremost we would like to thank the almighty God for this far He has brought us, indeed He is Ebenezer. Secondly we would wish to thank Mr Frederic Ntirenganya for his collaboration

to ensure this paper is a success. Last but not least, we give thanks to our families for their support during build up of this paper.

References

Colwell, R K, A Chao, N J Gotelli, S Y Lin, C X Mao, R L Chazdon, and J T Longino. 2012. "Models and estimators linking individual-based and samplebased rarefaction, extrapolation and comparison of assemblages." *Journal of* 5 (1): 3-21.

Chiarucci, A. 2012. "Estimating species richness: still a long way off." *Journal of Vegetation Science* 23 (6): 1003-1005.

Clench, H K. 1979. *Journal of the Lepidopterists' Society.*

Colwell, R K, and J E Elsensohn. 2014. "EstimateS turns 20: statistical estimation of species richness and shared species from samples, with non-parametric extrapolation." *Ecography* 37 (6): 609-613.

Dengler, J. 2009. *Journal of Biogeography* 36 (4): 728-744.

Dengler, J. 2009. *Journal of Biogeography* 36 (4): 728-744.

Gleason, H A. 1992. "On the relation between species and area." *Ecology* 3 (2): 158-162.

Gotelli, N J, and R K Colwell. 2011. "Estimating species richness." *Biological diversity: frontiers in measurement and assessment* 12: 39-54.

Heltshe, J F, and N E Forrester. 1983. "Estimating species richness using the jackknife procedure." *Biometrics* 1-11.

Hortal, J, P A Borges, and C Gaspar. 2006. "Evaluating the performance of species richness estimators: sensitivity to sample grain size." *Journal of Animal Ecology* 75 (1): 274-287.

Hui, C, R Veldtman, and M A McGeoch. 2010. "Measures, perceptions and scaling patterns of aggregated species distributions." *Ecography* 33 (1): 95-102.

Jobe, R T. 2008. "Estimating landscape-scale species richness: reconciling frequencyand turnover-based approaches." *Ecology* 89 (1): 174-182.

LlorenteB, J. 1993. "The use of species accumulation functions for the prediction of species richness." *Conservation biology* 7 (3): 480-488.

Longino, J T, and R K Colwell. 1997. "Biodiversity assessment using structured inventory: capturing the ant fauna of a tropical rain forest." *Ecological applications* 7 (4): 1263-1277.

Miller, R I, and R G Wiegert. 1989. *Ecology* 16-22.

Palmer, M W. 1990. "The estimation of species richness by extrapolation." *Ecology* 1195-1198.

Preston, F W. 1962. "The canonical distribution of commonness and rarity." *Ecology* 43 (2): 185-215.

Thompson, G G, P C Withers, E R Pianka, and S A Thompson. 2003. "Assessing biodiversity with species accumulation curves; inventories of small reptiles by pit-trapping in western australia." *Austral Ecology* 28 (4): 361-383.

Xu, H, S Liu, Y Li, R Zang, and F He. 2012. "Assessing non-parametric and areabased methods for estimating regional species richness." *Journal of Vegetation* 23 (6): 1006-1012.