Finding value in SAD moments: a novel approach to upscaling species abundance distributions

Hercules Araclides

A thesis submitted in partial fulfilment of the requirements for

the degree of Master of Science of Imperial College London

and Diploma of Imperial College London

September 2012

Abstract

Predicting species abundance distributions at scales larger than the available data is a key tool in estimating biodiversity. However, there are few techniques available that can do this successfully, with most approaches being far too general to be of practical use, and many techniques requiring *a priori* estimation of parameters. One of the key insights generated by research in this area is that the shape of a species abundance distribution is a function of sample scale. To this end, the Tchebichef method is a new approach that uses Tchebichef polynomials to approximate a probability density function, using only the scaling behaviour displayed by moments of sample areas. To date, this has been tested once, against tree and shrub species from Barro Colorado Island, with great success. Tests here against neutral model populations show encouraging results despite limitations of the available data. Attempts were also made to use this method in the context of bacterial communities, for which an appropriate scaling parameter was not determined. Results suggest that further testing will determine the Tchebichef method, with its minimal requirements, to be a valuable tool in upscaling species abundance distributions.

Contents

Introduction

The importance of biodiversity is beyond doubt, and the value of a better understanding is not just an academic pursuit. Ecosystems display greater stability at higher diversities (McCann, 2000), and human-mediated processes have already produced marked changes on species diversity (Chapin *et al.* 2000). As social and political interest in biodiversity continues to grow, there is an increasing need to develop better tools and measures for study, with an eye towards conservation efforts.

One of the key tools in describing biodiversity is relative species abundance (Hubbell 2001). This species abundance distribution (SAD) is presented as the number of species in a community with a given abundance (this abundance often represented in $\log_2$). Overwhelmingly, SADs indicate a prevalence of rare species, with comparatively fewer highly abundant species, a pattern often referred to as the 'hollow curve' (McGill *et al.* 2007). Historically, the development of the understanding of SADs began with Fisher *et al.* (1943), whose famous logseries distribution for Lepidoptera abundance followed from probabilistic considerations of the negative binomial distribution. Preston (1948) went on to claim that SADs follow a lognormal distribution, and that the rarest species are likely to either be underrepresented or unrepresented by a partial sample of a community, resulting in what appears to be a logseries distribution. Preston introduced this as the concept of the 'veil line', with an apparent logseries distribution becoming progressively lognormal as the sample size increases and the "veil-line" is shifted further left (figure 1).

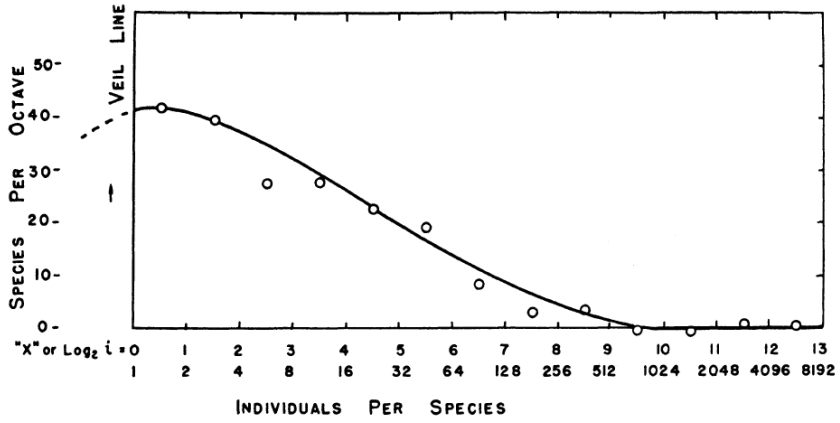

Figure 1. Preston's 'veil line', demonstrating the convergence to lognormality as sample size increases, and the 'veil line' shifts leftwards.

However, the existence of the "veil line" has been disproven theoretically (Dewdney 1998, Williamson and Gaston 2005), and also has empirical evidence against it (Forster and Warton 2007).

All told, since initial work by Fisher, dozens of distributions and models have been suggested as the best candidate to explain this hollow curve (the review by McGill *et al.* 2007 outlines 27 such models). Of the research done, some consensus has been reached. Some of the most important conclusions have been that a) rather than a lognormal distribution, SADs regularly show a definite left-skew (Nee *et al*. 1991, Hubbell 2001), and b) the shape of a SAD is in fact a function of sampling scale (McGill 2003, Borda-de-Água *et al.* 2012), particularly when we assume spatial aggregation of species, or equivalently, non-random sampling regimes (Green and Plotkin 2007). Further insights have been provided by Hubbell's unified neutral theory of biodiversity and biogeography (UNTB), in particular, Hubbell's zero-sum multinomial (ZSM) distributions, which better fit empirical data (Forster and Warton 2007), as well as allowing for the excess of singletons empirically observed (Borda-de-Água *et al.* 2012). In addition, work by Etienne *et al.* (2007) further refined the ZSM to a dispersal-limited multinomial (DLM), showing the importance of dispersal in determining the shape of SADs.

Of course, when obtaining SADs for different communities, a complete census is almost always out of the question for practical reasons including time and resources, and we must make the best of what information is obtainable. To this end, there has been interest in predicting SADs for areas at larger scales than data is available for. It has been shown that the shapes of SADs at different scales may be a result of a spatial analogy of the central limit theorem, with local SADs converging to regional SADs as scale increases (Sizling *et al.* 2009a) and that predictions based on upscaling are a valid avenue of interest (Kunin 1998, Zhang 2006, Sizling *et al.* 2009b). The upscaling of SADs is still in its infancy, however. Harte *et al.* (2009) used a maximum-entropy (akin to maximum likelihood) technique, and though promising, this approach required *a priori* estimation of parameters including metabolic rate, where present concern is predicting SADs using only existing SADs of known scale. Zillio and He (2010) used Bayesian techniques, and though there were also some *a priori* assumptions made, these were using probabilistic arguments relating to number of individuals and sampling technique, which are soluble in the present context. Results were highly variable, but again, they show some promise.

One particularly striking technique (and the inspiration for the present work) is the Tchebichef method of Borda-de-Água *et al.* (2012), which uses the scaling properties of the moments of the distribution in order to upscale SADs. The method has been shown to successfully predict SADs of tree species on Barro Colorado Island (BCI), Panama, and one essential improvement over other methods is its ability to retain the rare singletons across all scales. The present work will attempt to test this method against datasets generated via a spatially explicit neutral model, across different speciation rates, dispersal distances and area sizes, provided by Dr. James Rosindell, and bacterial community datasets collected by Dr. Thomas Bell. The latter is data taken from distinct tree holes (communities) rather than subsets of a regional pool, and so in addition to testing the predictive power of this method on scales vastly dissimilar to eukaryotes, it also aims to test the method in a setting more similar in context to islands of different scale.

Methods

Moments and SADs

Predicting SADs for scales larger than the available data requires taking existing moments, scaling over available areas, extrapolating the predicted moments, converting these to Tchebichef moments, and then using these to approximate the abundance distribution for the desired area.

I will provide the appropriate background for the method and mathematics behind it, based on the works of Mukundan *et al*. (2001) and Borda-de-Água (2012), which give further details.

A moment is a descriptive measure of a distribution (Crawley 2005). For example, the first four moments characterise the mean, variance, skew and kurtosis of a distribution, respectively. In present context, the moment of order n, M_n of a SAD is given by Eq. (1),

$$M_n = \frac{1}{S}\sum_{j=1}^{S} x_j^n$$

where the total number of species is given by S, and the jth species contains x individuals in $\log_2$.

With moments calculated for each dataset, the $\log_{10}$ moments are plotted against $\log_{10}$ area, as scaling laws are observed for the moments (Borda-de-Água 2002), and so available areas can be used to calculate predicted moments of a new area for which data is unavailable. To this end, regression is used and residuals are analysed in order to decide the appropriate scaling region, and the highest order moment to use when converting the Tchebichef moments to a probability density function (pdf).

As a given SAD can be seen as a pdf, this pdf can be linked to the moments of the distribution via a characteristic moment function (Morrison 1995), whereby the function is in a polynomial form such as to allow the reconstruction of the distribution. However, conventional moment functions have their drawbacks, such as computational demands (Borda-de-Água *et al*. 2012), limitations of domain (e.g. Legendre polynomials require the domain to be within the bounds [-1,1], which require further transformation of the distribution), and larger ranges of error at higher order moments (Mukundun *et al.* 2001).
As such the technique used within is based on a method of discrete orthogonal Tchebichef moments which do not have the above mentioned drawbacks, introduced by Mukundan et al. (2001) and first used within this context by Borda-de-Água et al (2012). Equations for calculating the Tchebichef moments, and relating these moments to a pdf and thus a SAD, can be found in the Appendix.

When presenting the SAD as a histogram, choice of bins (abundance classes) was informed by Williamson and Gaston (2005), who suggest centering the bins on the logarithms of the powers of 2, with boundaries at $\log(2^{n\pm0.5})$. This serves the dual purpose of ensuring that the boundaries exactly double, and as these are multiples of the root of 2 (and thus irrational), no species can contain a number of individuals which falls exactly between adjacent bins.

In addition, as the Tchebichef moments generate a probability density function, the total number of species over which to apply this function need also be extrapolated, as does the number of bins, to accommodate larger abundance classes. Extrapolation is done with regressions of $\log_2$ abundance and $\log_{10}$ number of species against $\log_2$ area.

Data

The method of scaling moments was used on two datasets, Dr. James Rosindell's spatially explicit neutral model populations, and bacterial datasets provided by Dr. Thomas Bell.

The neutral model populations allow application of the method on population sizes not obtainable in the field and also serve as a spatial test of neutral models. Area lengths used were 2, 4, 8, et seq. up to 2048, with area size and number of individuals being 4, 16, 64, et seq. Speciation rates varied from $1x10^{-7}$ to $3x10^{-1}$, with dispersal distances ranging from 1 to 64. In total, the Tchebichef method was applied to 2210 populations. It should be noted that each dataset contained the species-abundance in the form of pooled data from between 50-500 readings, and while the moments from pooled data are accurate representations of the true moments, this led to issues and considerations in extrapolating the number of species for the new desired area, which will be covered in the results.

The bacterial datasets were collected from tree-hole communities, which allow a test of the method on prokaryotes, whose spatial distribution and dispersal are dissimilar to those of conventionally studied species, and also allow examination of the behaviour of the moments on a scale unlikely to behave triphasically (Horner-Devine *et al*. 2004). In total, 199 datasets obtained from 6 sites around the U.K. were used. Samples were collected after homogenizing (stirring) the tree-hole water. Genomic DNA was extracted, and PCR techniques were used in order to amplify regions of the 16S rRNA locus. Fragments of approximately 400 base pairs from the variable V6 and V7 regions were sequenced on a Roche 454 sequencing platform and sequences were processed using the Clovr pipeline, down to a genus level, of which there were found to be approximately 750 genera.

Of the 6 sites, one contained only 3 samples and so was discarded. Of the remaining sites, sequences were generated from 10ml samples, and one of the sites also contained sequences generated from 25 ml samples. An exploratory analysis of species abundance and moments was performed first by comparing abundance against parameters of tree-hole volume (l), depth (cm), width (cm) and length (cm) for all sites of the same sample (ml) size, using the highest resolution of data available, that of genus. These were then divided by site in order to examine what, if any, scaling behaviour was present.

Results

Neutral model populations

For the neutral datasets, areas from area length 2 up to those of length 512 or 1024 were used in order to extrapolate the SAD for the area of length 1024 or 2048, respectively, dependent on the largest available area. Within each parameter combination of speciation and dispersal, the maximum number of moments that could be used was equal to the number of bins in the largest area.

Though Borda-de-Água *et al.* (2012) found a power-law behaviour for the logarithm of the moments against the logarithm of the area, visual inspection makes it clear that there is a definite curvature to the $\log_{10}$ moments generated by a neutral model, and that at larger areas, the moments are lower than those predicted by a linear power-law. This is a result of the moments being calculated from pooled data, which overestimate species number at larger areas, more about which in the discussion section.

As a result of this, moments were primarily extrapolated using polynomial regression, which gave R^2 values of 0.99 and above, with $p<0.05$ for all available parameter combinations, though p-values were used only as a first approximation due to the fact that moments are not independent. See Fig 1. for comparison of linear (1a) and polynomial (1b) fits.

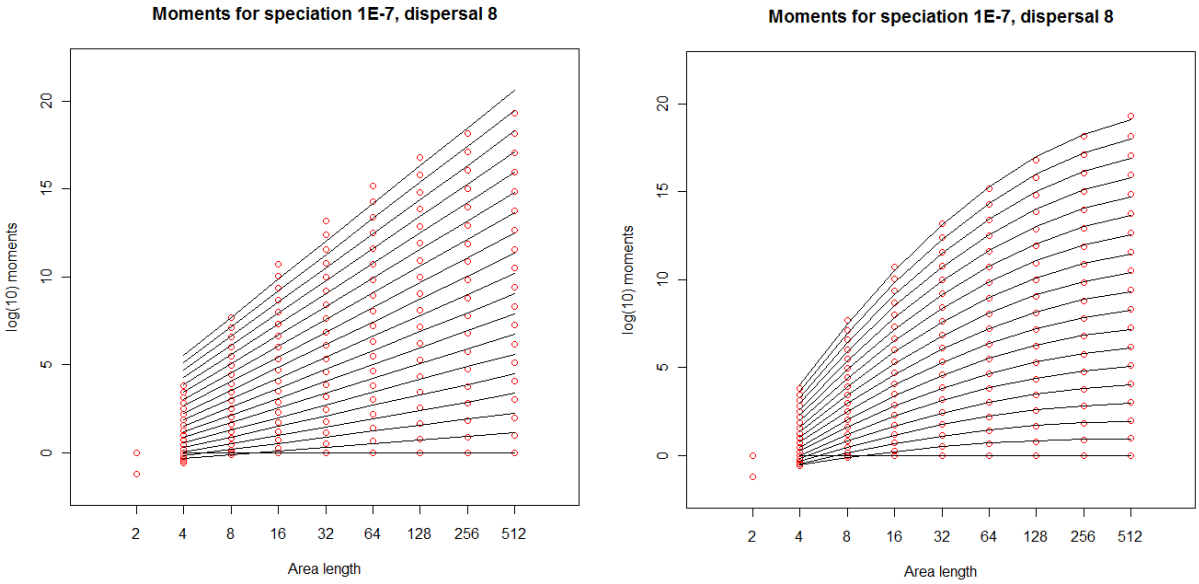

Figure 1. (a) R^2 is 0.92 for linear, $p>0.05$ (b) R^2 is 0.99 for polynomial, $p<0.05$

The scaling region was initially informed by evaluating which of the lower area datasets only contained one or two abundance classes. While moments of order M_0 are always 1, moments of order M_1 and above produce 0 when only one abundance class is present, which evaluated to negative infinity when taking the $\log_{10}$ moments. In cases where there were only two abundance classes, these were of course 1 and 2, which gave $\log_2$ abundance classes of 0 and 1, and as a result, moments of order M_1 and above all evaluated to the same value, which led to regression lines being heavily inflected downwards and reducing fit significantly. As a result, a minimum of 3 abundance classes were required as a low bound for scaling region (thus area length 2 not being used for the line-of-fit in Figure 1 above). With this area bound (A_l) determined, polynomial regressions were performed on all possible scaling regions, and the moments predicted from intercept and slope coefficients were compared to the original moments, using sum of squared residuals (SSR). This showed that when using polynomial regression, there was no deviance in the behaviour of the moments at larger areas, and as a result, the best fit was in fact the scaling region from A_l to the largest available area.

In addition, assuming that the lower moments at larger areas were an artefact of calculating moments based on pooled data from multiple readings, linear regressions were also performed on the assumption that the lower areas behaved in accordance with a power-law (figure 2.), and ignoring the larger areas which show a pronounced curvature (aided by visual inspection and examination of residuals of the moments, see figure 3.), with an eye to seeing how linear and polynomial predictions fit to the actual data for the largest area.

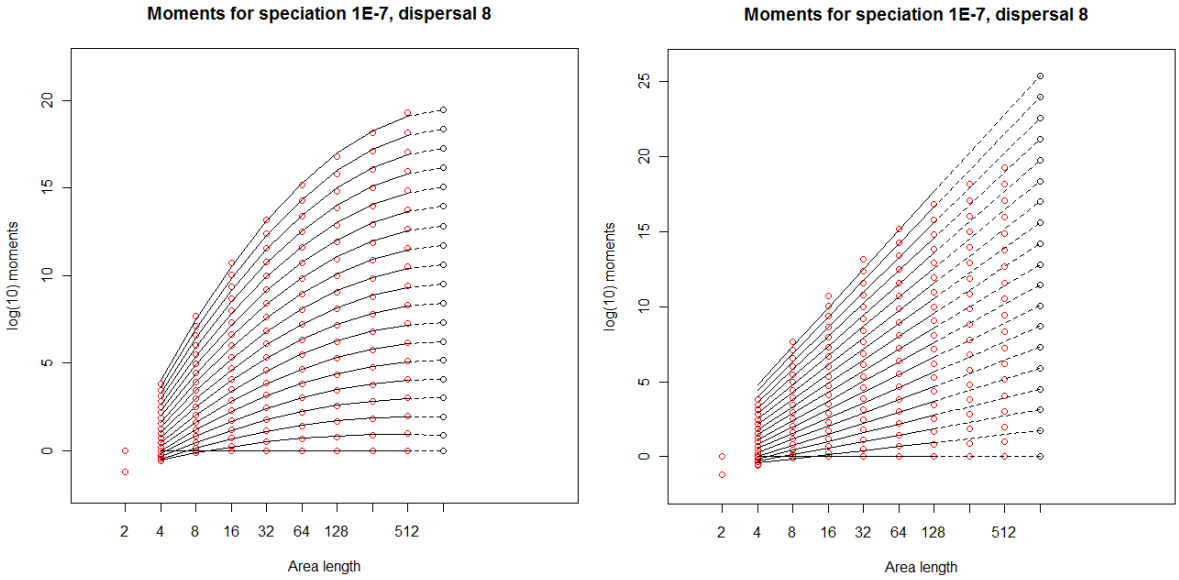

Figure 2. (a) Polynomial regression fit for A_{1024} R^2 range = 0.99-1.00, $p<0.05$ (b) Linear regression fit for A_{1024} R^2 range = 0.38-0.88, $p>0.05$

In both instances, the solid line is the scaling region used, and the dashed line and black points are the extrapolated predictions.

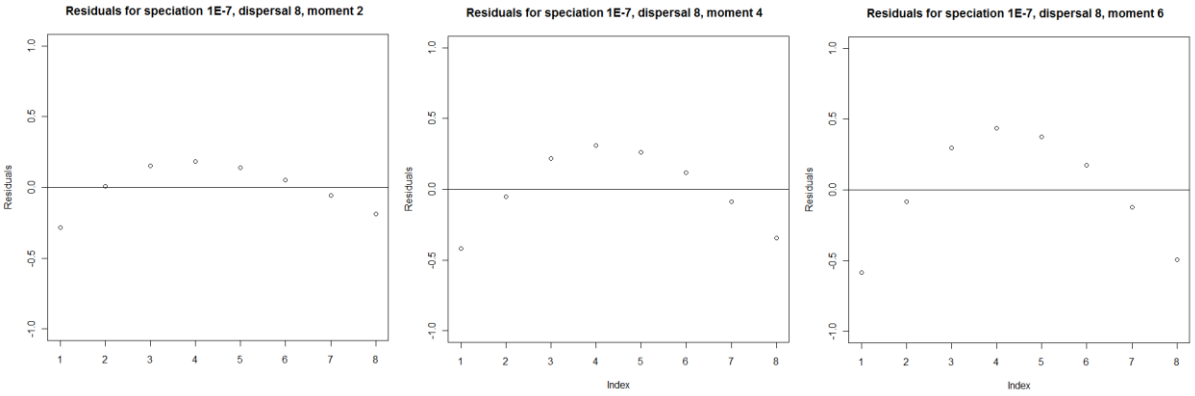

Figure 3. Residuals of the moments M_3, M_5 and M_7, showing downward curvature at larger areas. In this case, this led to the two largest area lengths 256 and 512 not being used for linear regression extrapolation of moments.

Choice of the highest order moment to use was informed by further investigation of the residuals of the moments. From moment order M_6 and above, the residuals tended to increase by one order of magnitude, though compared to the actual values of the moments, they remained reliably small. This led to the choice of using moments of order M_0 to M_5, which was given mild support by p-values (despite the non-independence of moments) often becoming non-significant from M_6 and above. This is in accordance both with a preliminary study performed on an artificially generated and

repeatedly subsampled dataset, which gave clearer indication of moment behaviour than the 10 areas in the neutral dataset, as well as findings by Borda-de-Água *et al.* (2012), and recommendation by Boyd (2001).

In addition to the predicted moments obtained by choice of scaling region and number of moments, the maximum abundance class and the number of species for the new predicted area (length 1024) need to be extrapolated. The former is required for binning of abundances, and the latter is required as the Tchebichef method results in a probability density function, and so this must be scaled over the appropriate number of species.

As abundance classes and area lengths were both best represented in $\log_2$, a simple linear regression was adequate in order to extrapolate the number of bins needed for the predicted area. This extrapolation was reliable and consistent for all parameter combinations (a typical example is demonstrated in Figure 4a). For the extrapolation of the number of species, polynomial regression was unsuitable for reasons pertaining to estimation based on pooled data, as further explained in the discussion. With this in mind, larger areas were ignored after inspection of residuals, and linear regression was performed over the region which displayed power-law behaviour in the associated moments (Figure 4b).

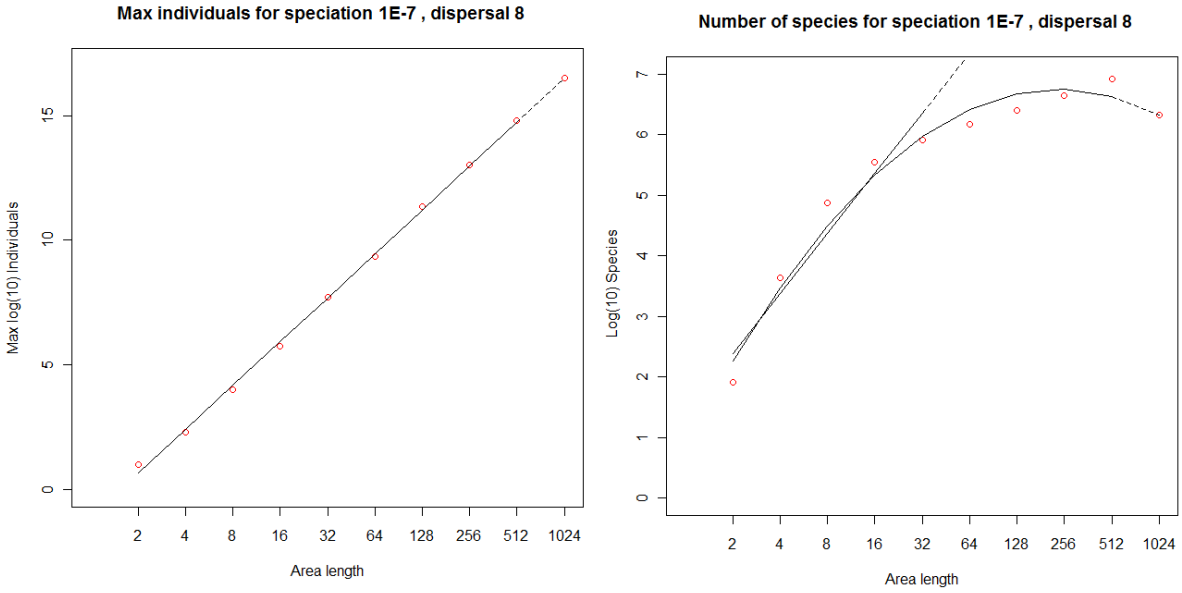

Figure 4. (a) Extrapolation of abundance classes was consistent and reliable. (b) Both linear and polynomial regression fail to give a fair estimate of species number.

Unfortunately, both linear regression and polynomial regression greatly overestimated or underestimated the actual number of species present when compared with actual species data for the predicted area (sometimes on the order of 2 or 3 magnitudes), even when appropriately rescaled for the number of readings. When converting the extrapolated moments to Tchebichef moments and applying this pdf over this number of species, the predicted SAD displayed erratic behaviour, either not fitting within reasonable confines of the plot (for linear regression, see Figure 5a), or being distributed over so low a species number that the distribution was hugely misrepresentative (in the case of polynomial regression, Figure 5b).

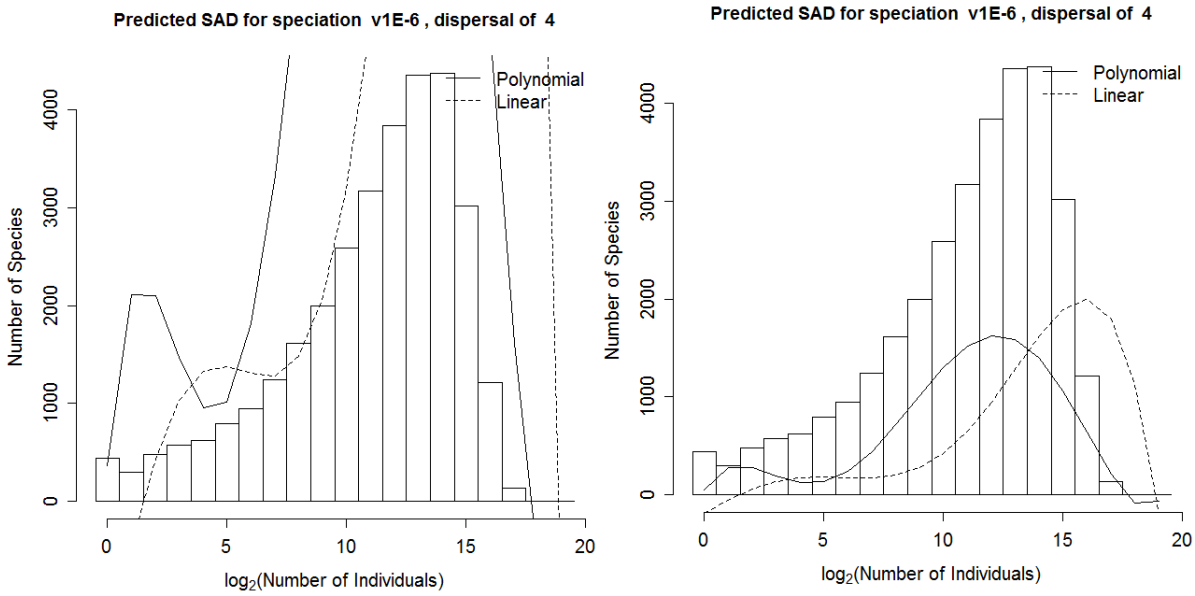

Figure 5. (a) Linear regression of species number (b) Polynomial regression of species number
In both cases, the SAD is the actual SAD of the predicted area, while the polynomial and linear lines refer to the method of regression applied to the moments.

Despite limitations imposed by pooled data, in order to test the efficacy of the Tchebichef method in principle, the predicted Tchebichef moments and associated pdf were applied using the total number of species (appropriately rescaled for number of readings), obtained from the actual data for the largest area. Figure 6a, 6b and 6c respectively display the predicted SAD based on linear and polynomial regression, for low ($1x10^{-6}$), high ($1x10^{-4}$) and realistically accepted (Rosindell and Cornell 2007, Pigolotti and Cencini 2009) speciation rates ($1x10^{-5}$), at 3 dispersal lengths (4, 16 and 64).

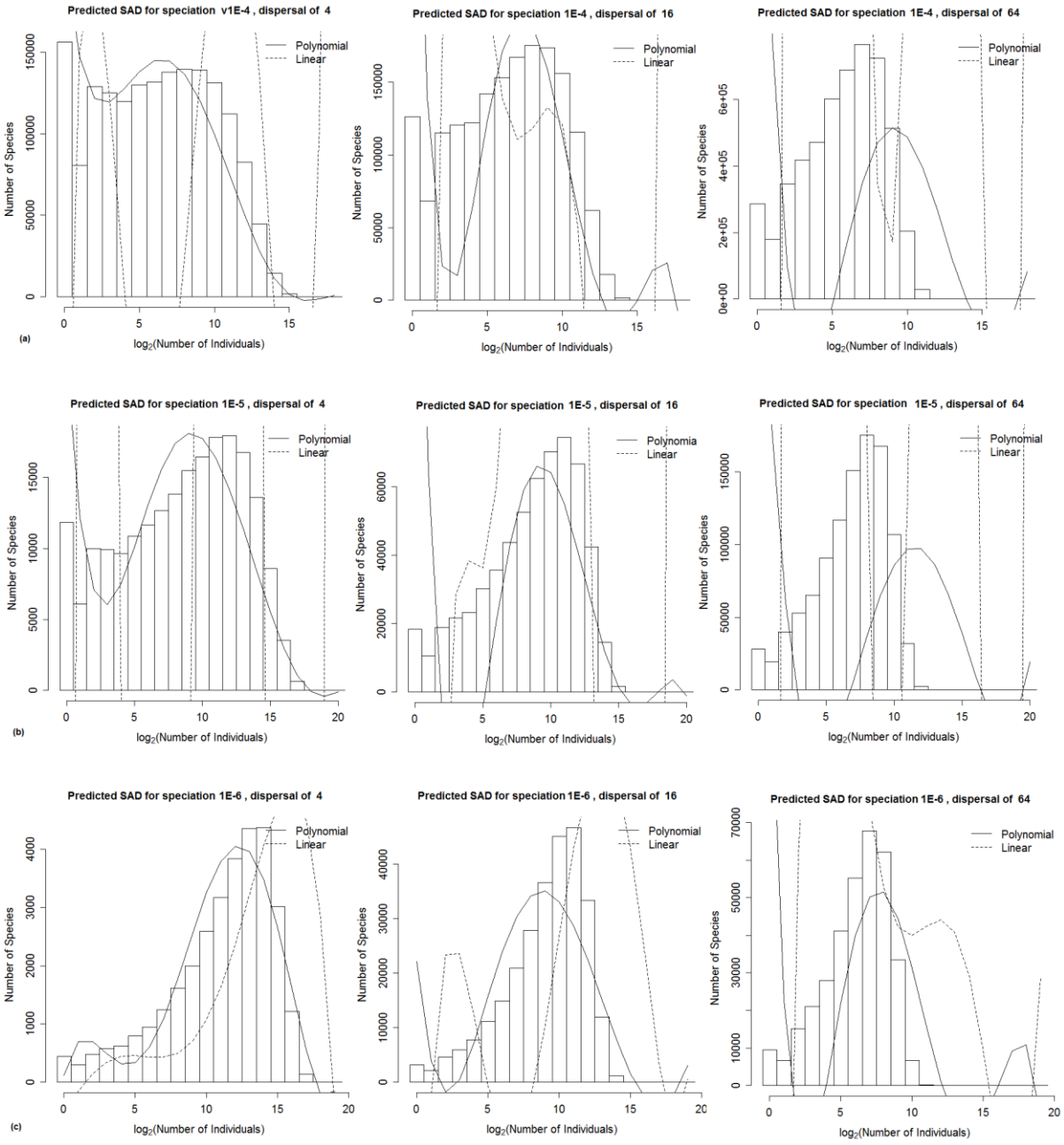

Figure 6. In all cases, the SAD is the actual SAD for the largest area for that parameter combination, with solid and dashed lines presenting the predicted SAD, respectively.

As can be seen, the best fitting predicted SAD is that of the lowest speciation, and lowest dispersal distance (6c, dispersal 4). Obviously, as dispersal and speciation increase, so too does total number of species present, and as previously described, an increase in the number of species contained within pooled data has serious effects on both the value of the moments and their power-law behaviour, which appears clearly in the predicted SADs other than the lowest speciation and dispersal distance.

With this in mind, the lowest speciation rate ($1x10^{-7}$) and dispersal distance (1) was investigated and a SAD fitted, Figure 7. This is in accordance with the best fit in Figure 6, suggesting that the Tchebichef method has potential when certain conditions are met.

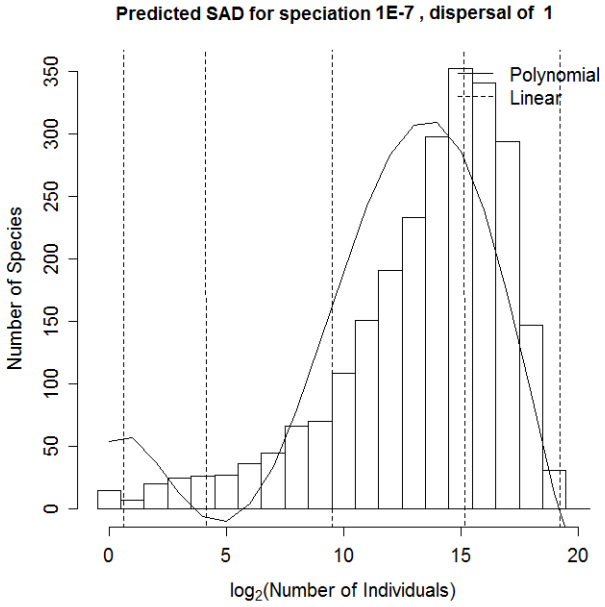

Figure 7. Predicted SAD for the lowest speciation rate and dispersal distance.

Bacterial communities

An examination of the moments of species abundance against the parameters of tree-hole volume (l), width (cm), depth (cm) and length (cm) was first performed against all available sites of each sample (ml) size. Figure 8 shows the behaviour of the moments using the volume parameter.

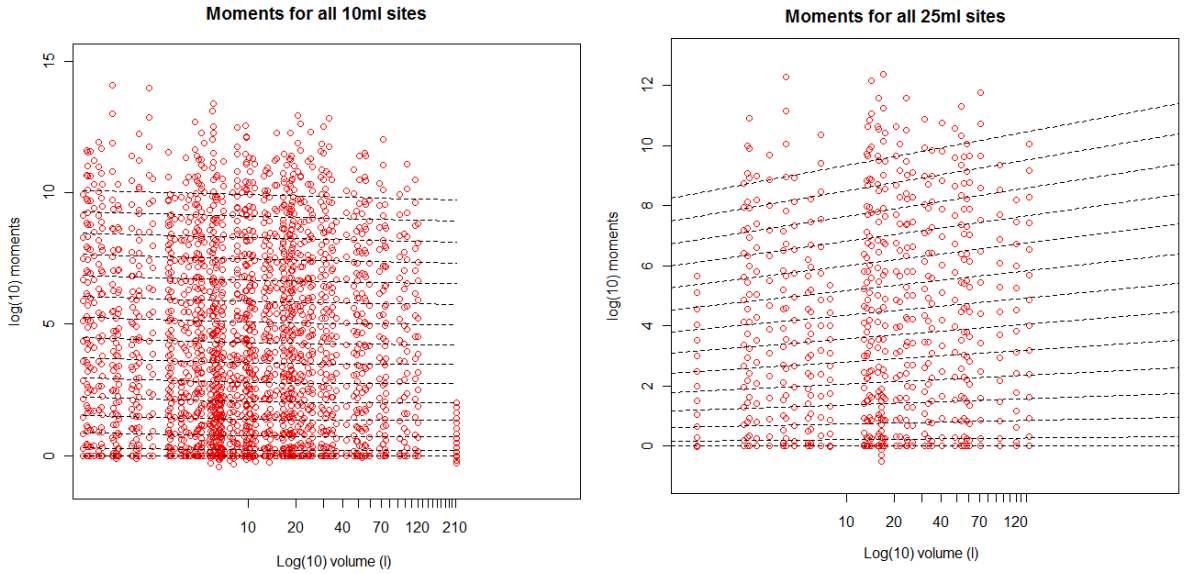

Figure 8. Moments for all sites of a given volume, with linear regression lines.

As is clear, no scaling behaviour exists for total sites, with 10ml sites having an R^2 between 0.001 and 0.02, and 25ml sites having an R^2 of between 0.04 and 0.09.

Similar lack of fit occurred for the parameters of tree-hole width, length and depth. Subsequently, samples were split by site and site-specific scaling behaviour was investigated. However, no behaviour presented itself. Table 2 shows the minimum and maximum R^2 values by site, using the volume parameter, and Figure 9 shows moments for 2 of the sites. Again, similar lack-of-fit was found when examining other parameters for scaling behaviour.

Site	Sample (ml)	Min R^2	Max R^2
Ashridge	10	0.001	0.04
Burnham Beeches	10	0.05	0.05
Warburg	10	0	0.002
Wychwood	10	0.01	0.02
Wytham	10	0.002	0.01
Wytham	25	0.04	0.09

Table 2. Site-specific fit of moments.

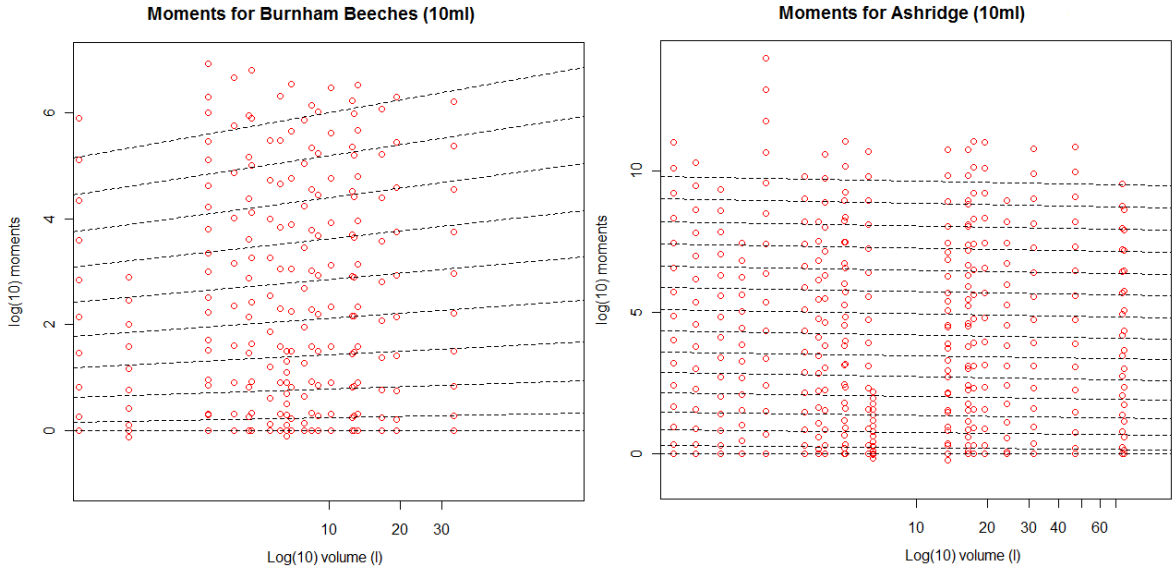

Figure 9. Moment-parameter behaviour for two of the 10ml sites.

With no such scaling behaviour found, no moment extrapolation could be performed, and so no predicted SADs could be fitted.

Discussion

As mentioned in the results, estimation based on pooled data led to a misrepresentation of the true moments and of the true species number, with this effect being significantly pronounced at larger areas. The reason for this becomes clear when considering the manner in which the data was generated, and the form of the moments equation, Eq. (1). The original work (Borda-de-Água *et al.* 2012) calculated the moments of individual repeated subsamples, and then averaged them, whereas in the present work the individual readings which made up the pooled data could not be teased apart. Pooled data has no problem producing the true average moment, as shown in Table 1a.

A		B		Pooled	
Abundance	S	Abundance	S	Abundance	S
1	4	1	5	1	9
2	3	2	2	2	5
3	2	3	2	3	4
4	1	4	1	4	2
Total I	20	Total I	20	Total I	39
Total S	10	Total S	10	Total S	20
M_1	2	M_1	1.9	M_1	1.95

Table 1a. The first-order moment of two separate communities A and B, and of the pooled data

Using moment M_1 for simplicity (no need to raise to higher powers, though this works with any order of moment), the moment is simply the total number of individuals over the total number of species. It is clear here that the moment of the pooled data, 1.95, is of course the average of each moment of community A (2) and community B (1.9). However, there is one important caveat: this works ideally when each community has the same number of individuals and species. With the pooled data from the neutral datasets, the number of species present among a given number of individuals (i.e. in a given area size) will not be constant.

At smaller areas, the possible combination of abundances is limited, for example in an area containing 4 individuals, the only possible abundances are obviously between 1 and 4. At larger areas, there are a greater amount of abundance classes, and more freedom within the partitioning of species needed to make up the necessary amount of

individuals, with the partitioning of any area size or number of individuals into abundance classes increasing exponentially (Folsom *et al*. 2012).

Table 1b demonstrates effect of species partitioning when there are a larger number of abundance classes:

A		B		Pooled	
Abundance	S	Abundance	S	Abundance	S
1	4	1	1	1	5
2	0	2	0	2	0
3	2	3	0	3	2
4	0	4	2	4	2
5	1	5	0	5	1
6	0	6	1	6	1
Total I	15	Total I	15	Total I	30
Total S	7	Total S	4	Total S	11
M_1	2.14	M_1	3.75	M_1	2.73

Table 1b. The first-order moment of communities A and B, and of the pooled data when species number differ.

Here, communities A and B have the same number of individuals, yet the number of species differs. The average of the moments for A and B is 2.95. However, the moment for the pooled data is 2.73, underestimating the true average. As the species abundances of the neutral datasets were totalled from between approximately 50-100 readings, it is evident that at larger areas this led to the moments being lower than would be predicted by a power law.

This was further confirmed by comparing moments for the same area, but a different number of readings. For example, looking at data for speciation rate $1x10^{-}7$, the largest number of readings for area length 512 was 356, while the lowest was 142. These were for the dispersal distances of 8 and 2, respectively. The moments predicted by linear regression were expressed as a fraction of the actual moments for that area, in order to have a proportional measure which also takes into account the fact that different dispersal distances have different moments. For the lower number of readings, predicted values for the first 6 moments ranged from a minimum of 90.0 to 92.7% accuracy, while for the great number of readings, accuracy was between -35.6 to 44.9%, further suggesting that pooled data creates issues at larger areas.

As a result, the use of pooled data is recognized as a poor alternative to use of separate readings for which the number of species can be obtained and reliably averaged. It also casts serious doubt on the accuracy of the ordinary moments in circumstances with a large number of species.

Despite this, the extreme case of using the true species value with the predicted probability density function for low speciation rates (Figures 6 and 7) hints that is potentially a powerful and promising method under the right conditions, in addition to the already suggestive results with empirical data, found by Borda-de-Água *et al*. (2012). Species abundance downscaling is well-covered ground (He and Reed 2006), but species abundance upscaling is still a new (and long overdue) technique, with the only methods suggested so far being the maximum-entropy approach of Harte *et al*. (2009), as well as a Bayesian method developed by Zillio and He (2010). The approach by Harte *et al* required *a priori* parameter estimation, and was more concerned with species richness, so it is difficult to compare their results to the current work. Zillio and He (2010), however, tested their method on the same BCI data as Borda-de-Água *et al*. (2012), and so can be directly compared to a use of the Tchebichef method which meets the appropriate conditions (namely, averaging from individually available readings rather than pooled data), in Figure 10. It can be seen that with appropriate data structure, the Tchebichef method not only approximates the desired SAD to a much closer extent, but also has the benefit of predicting the presence of rare singletons, which neither the maximum-entropy nor Bayesian methods do. As such, the Tchebichef method promises to be a useful technique for both future research, and in practical use of predicting species abundances in a biodiversity context. Also in its favour is the ability to use it 'straight out of the box', as requiring only a set of SADs across different scales makes it a very inviting technique, particularly in the latter application of conservation efforts.

In addition, despite not finding the appropriate moment behaviour for bacterial communities, I remain confident that further work and analysis of the parameter space may yet provide a workable scaling parameter, allowing the Tchebichef method to provide insight into the spatial scaling of bacterial species abundance distributions.

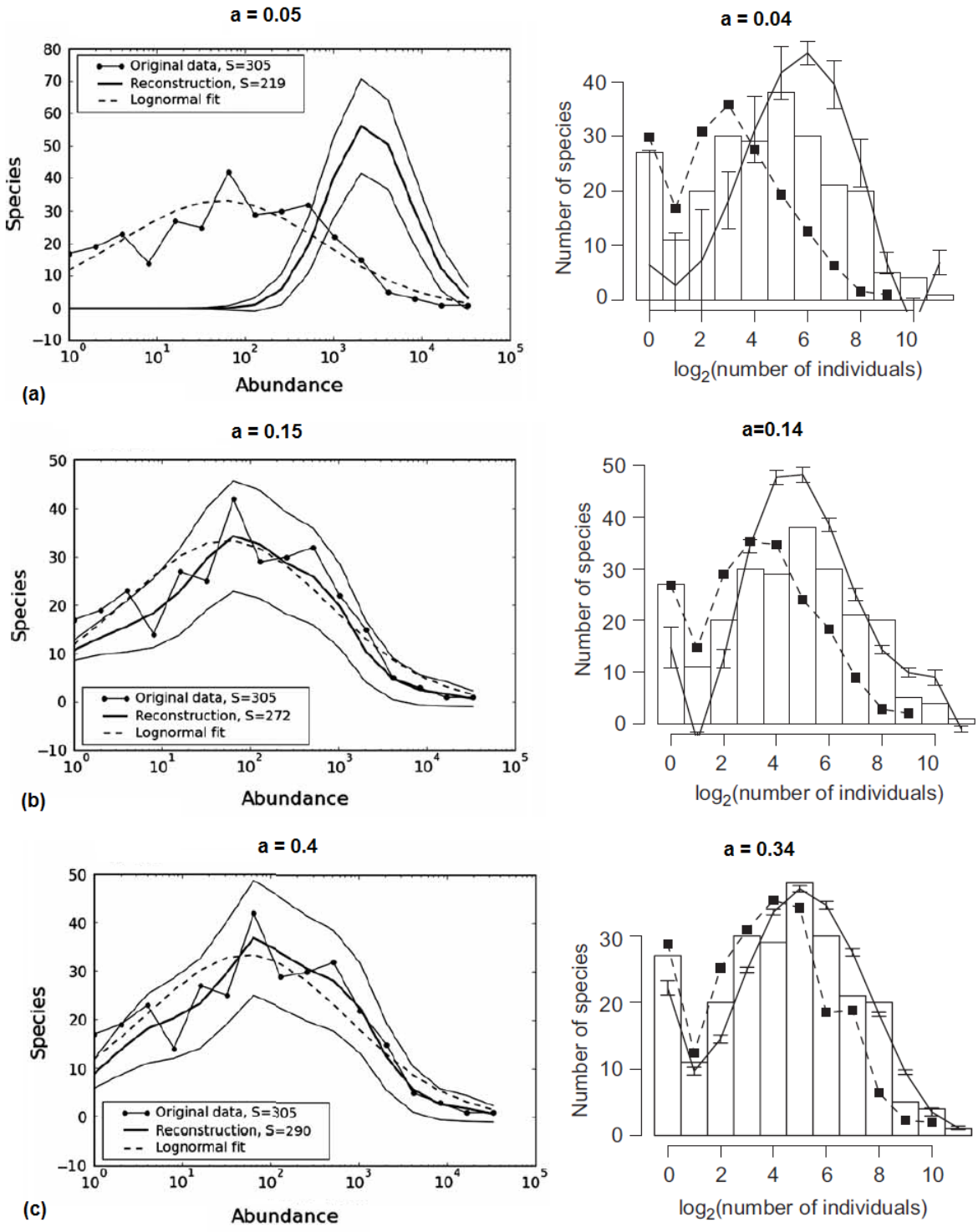

Figure 10. The SADs on the left are from Zillio and He (2010), and those on the right are from Borda-de-Água *et al*. (2012). In both cases, the measure *a* expresses the fraction of the desired area used in order to predict for that area.

Acknowledgements

Thanks to Dr. Thomas Bell for agreeing to supervise my thesis and kindly providing me with the bacterial community datasets. I would also like to thank Dr. James Rosindell for his support and encouragement throughout, as well as his neutral model population data. Special thanks to Dr. Luís Borda-de-Água, whose work and suggestions inspired this project, and for providing the code for implementing the Tchebichef method. Thanks also go to Dr. Dan Reuman, whose Quantitative Biology course equipped me with the necessary skill set needed to take on this project.

References

Borda-de-Água, L., Hubbell, S., & McAllister, M. (2002). Species-area curves, diversity indices, and species abundance distributions: A multifractal analysis. *American Naturalist, 159*(2), 138-155.

Borda-de-Água, L., Borges, P. A. V., Hubbell, S. P., & Pereira, H. M. (2012). Spatial scaling of species abundance distributions. *Ecography, 35*(6), 549-556.

Boyd, J. P. (2001). *Chebyshev and fourier spectral methods* (2nd ed.). New York: Dover.

Chapin, F. S.,III, Zavaleta, E. S., Eviner, V. T., Naylor, R. L., Vitousek, P. M., Reynolds, H. L., et al. (2000). Consequences of changing biodiversity. *Nature (London), 405*(6783), 234-242.

Crawley, M. J. (2005). *Statistics : An introduction using R*. Chichester, West Sussex, England: J. Wiley.

Dewdney, A. (1998). A general theory of the sampling process with applications to the "veil line". *Theoretical Population Biology, 54*(3), 294-302.

Etienne, R. S., Alonso, D., & McKane, A. J. (2007). The zero-sum assumption in neutral biodiversity theory. *Journal of Theoretical Biology, 248*(3), 522-536.

Fisher, R. A., Corbet, A. S., & Williams, C. B. (1943). The relation between the number of species and the number of individuals in a random sample of an animal population. *Journal of Animal Ecology, 12*(1), pp. 42-58.

Folsom, A., Kent, Z. A., & Ono, K. (2012). l-adic properties of the partition function. *Advances in Mathematics, 229*(3), 1586-1609.

Forster, M. A., & Warton, D. I. (2007). A metacommunity-scale comparison of species-abundance distribution models for plant communities of eastern australia. *Ecography, 30*(4), 449-458.

Green, J. L., & Plotkin, J. B. (2007). A statistical theory for sampling species abundances. *Ecology Letters, 10*(11), 1037-1045.

Harte, J., Smith, A. B., & Storch, D. (2009). Biodiversity scales from plots to biomes with a universal species-area curve. *Ecology Letters, 12*(8), 789-797.

He, F., & Reed, W. (2006). In Wu, J Jones, KB Li, H Loucks,OL (Ed.), *Downscaling abundance from the distribution of species: Occupancy theory and applications*. Dordrecht, Netherlands: Springer.

Horner-Devine, M., Lage, M., Hughes, J., & Bohannan, B. (2004). A taxa-area relationship for bacteria. *Nature, 432*(7018), 750-753.

Hubbell, S. P. (2001). Monographs in population biology. the unified neutral theory of biodiversity and biogeography. *Monographs in Population Biology.the Unified Neutral Theory of Biodiversity and Biogeography, 32*, i-375.

Kunin, W. (1998). Extrapolating species abundance across spatial scales. *Science, 281*(5382), 1513-1515.

McCann, K. (2000). The diversity-stability debate. *Nature, 405*(6783), 228-233.

McGill, B. (2003). Does mother nature really prefer rare species or are log-left-skewed SADs a sampling artefact? *Ecology Letters, 6*(8), 766-773.

McGill, B. J., Etienne, R. S., Gray, J. S., Alonso, D., Anderson, M. J., Benecha, H. K., et al. (2007). Species abundance distributions: Moving beyond single prediction theories to integration within an ecological framework. *Ecology Letters, 10*(10), 995-1015.

Morrison, K. E. (1995). Cosine products, fourier-transforms, and random sums. *American Mathematical Monthly, 102*(8), 716-724.

Mukundan, R., Ong, S., & Lee, P. (2001). Image analysis by tchebichef moments. *IEEE Transactions on Image Processing, 10*(9), 1357-1364.

Nee, S., Harvey, P., & May, R. (1991). Lifting the veil on abundance patterns. *Proceedings of the Royal Society of London Series B-Biological Sciences, 243*(1307), 161-163.

Pigolotti, S., & Cencini, M. (2009). Speciation-rate dependence in species-area relationships. *Journal of Theoretical Biology, 260*(1), 83-89.

Preston, F. W. (1948). The commonness, and rarity, of species. *Ecology, 29*(3), pp. 254-283.

Rosindell, J., & Cornell, S. J. (2007). Species-area relationships from a spatially explicit neutral model in an infinite landscape. *Ecology Letters, 10*(7), 586-595.

Sizling, A. L., Storch, D., Reif, J., & Gaston, K. J. (2009b). Invariance in species-abundance distributions. *Theoretical Ecology, 2*(2), 89-103.

Sizling, A. L., Storch, D., Sizlingova, E., Reif, J., & Gaston, K. J. (2009a). Species abundance distribution results from a spatial analogy of central limit theorem. *Proceedings of the National Academy of Sciences of the United States of America, 106*(16), 6691-6695.

Williamson , M., & J. Gaston, K. (2005). The lognormal distribution is not an appropriate null hypothesis for the species?abundance distribution. *Journal of Animal Ecology, 74*(3), 409-422.

Zhang, Y., Ma, K., Anand, M., & Fu, B. (2006). Do generalized scaling laws exist for species abundance distribution in mountains? *Oikos, 115*(1), 81-88.

Zillio, T., & He, F. (2010). Inferring species abundance distribution across spatial scales. *Oikos, 119*(1), 71-80.

Appendix

Equations for calculating Tchebichef moments and polynomials

Here the relevant equations and formulas for the Tchebichef method are reproduced. Full mathematical details of the Tchebichef method can be found in Mukundan *et al* (2001), and also specific to the current context, Borda-de-Água *et al.* (2012).

For N bins, the Tchebichef moments of order *n*, *Tn*, with $0 \leq n < N$, are given by Eq. (1)

$$T_n = \frac{1}{N^n \tilde{\rho}(n,N)} \sum_{k=0}^{n} C_k(n,N) \sum_{i=0}^{k} s_k^{(i)} M_i \ , \qquad (1)$$

Here, *Mi* is the moment of order *i*, estimated using Eq. 1 (from main text).
$S_k^{(i)}$ are the Stirling numbers of the first kind.
$C_k(n,N)$ is given by Eq. (2)

$$C_k(n,N) = (-1)^{n-k} \frac{n!}{k!} \binom{N-1-k}{n-k} \binom{n+k}{n}, \qquad (2)$$

and $\tilde{\rho}(n,N)$ is given by Eq. (3)

$$\tilde{\rho}(n,N) = \frac{N\left(1-\frac{1}{N^2}\right)\left(1-\frac{2^2}{N^2}\right)\left(1-\frac{n^2}{N^2}\right)}{2n+1}; \qquad n=0,1,\ldots,N\text{-}1. \qquad (3)$$

The scaled Tchebichef polynomials $\tilde{t}_n(x)$ are calculated with the following recurrence formulas

$$\tilde{t}_0(x) = 1$$

$$\tilde{t}_1(x) = (2x + 1 - N)/N$$

$$\tilde{t}_n(x) = \frac{(2x-1)\tilde{t}_1(x)\tilde{t}_{n-1}(x) - (n-1)\left(1 - \frac{(n-1)^2}{N^2}\right)\tilde{t}_{n-2}(x)}{n} \qquad n=2,3,\ldots,N\text{-}1.$$

The Tchebichef moments and polynomials can then be related to the probability density function $f(x)$ using Eq. (4)

$$f(x) = \sum_{n=0}^{N-1} T_n \tilde{t}_n(x) \qquad \textbf{(4)}$$

(The functions in R to implement the Tchebichef calculations were provided by Dr. Luís Borda-de-Água, and are available upon request).