

Species Discovery and Sampling Effort in Ecological Communities

Abstract

As the community is dominated by some species while others become rare, the effort needed to discover all species increases, especially when some species become more rare. We observed the shift of the most probable minimum effort needed to identify all species as the community became more uneven. This minimum effort depends on all species, not only on the rarest ones. Additionally, the minimum effort doesn't vary with changes in the sampling effort applied at once. We compared three communities on fundamental aspects and differences without knowing the actual distribution, relying solely on samples due to limitations inherent in real systems.

1 Introduction

Our primary aim is to determine the species discovery curve and subsequently identify when the chi-square statistic stopped being significant. To construct the species discovery curve, we employed multiple methods, arriving at various inferences sequentially. Initially, we attempted to identify the species (represented as alphabets) in the given community using Python's `pytesseract` library and `pymupdf`. However, the optical character recognition (OCR) results were suboptimal, yielding unexpected outcomes. Consequently, we utilized an online OCR service to extract the letters and transformed them into a 60x60 matrix, upon which we based our simulations. In this text, we used cumulative area and effort terms simultaneously because it is same in our system but effort is used while finding species discovery curve, on the other hand, cumulative area is used for species area curve. Here "chi square statistic stops being significant" implies when p value becomes less than 0.05 for 95% confidence interval.

2 Methodologies

We utilized the 60x60 matrix, overlaying it with a 3x3 grid placed randomly. Multiple 3x3 grids were applied without overlapping until all species in the community were accounted for. This procedure was replicated with other grid sizes, such as 2x2, 4x4, and 5x5, across all three communities. A single trial provides a point estimate, which may be anecdotal and unreliable. Therefore, we iterated the process multiple times to obtain more reliable and probable data.

3 Analysis(Simulation Based)

Three random experiments were conducted, each corresponding to one community, providing insight into how species discovery increases with sampling effort (cumulative area). Subsequently, we averaged the number of species identified relative to the cumulative area and fitted the power function species-area relationship. To gain statistical insight, we iterated these random trials approximately one million times, which was sufficient for communities one and two. For community three, additional restrictions were applied, which will be explained later. The iterations revealed a right-skewed normal distribution regarding the effort required to detect all 20 species. This skewness arises because there is a minimum effort threshold to discover all species, while the maximum effort is variable, depending

on community distribution and sampling methodology. We also calculated the chi-square statistic for each iteration, averaging it for specific efforts (cumulative area) rather than across all cycles, as individual trial chi-square values may be anecdotal. The advantage of multiple simulations allows for more robust averaged results.

Here, we can infer from this data that the most probable effort is same for all grid size. The most probable cumulative area(effort) in the histogram is the nearest discrete value to the actual most probable effort. We can see the right skewed distribution because there is a lower limit for required number of grids however no restriction exists for maximum number of grids(obviously there is also a higher limit but it depends on the way of grid picking and spatial distribution of the community). Here, we actually averaged the chi square value for the specific effort(cumulative area) in which it reached 20 alphabets.

In case of Community 1, we observed the increase in the value of chi square value as the effort increased and from the box plot , we can see so much outliers because of huge readings for that areas. The chi square value is increasing as the effort is increasing but rate of increase is decreasing that's why it is flattening. Now for higher effort, we don't have much data points , so the averaged values is not actually robust (most probably biased). That's why the chi square is not following any curve for higher values of area, and it is clearly shown in the box plot. Also in the simulation we found that the data point are extremely low so it is not reliable for averaging chi square value. That's why we will ignore those values while fitting the curve (this strategy is used for other two communities as well but curve is obviously different).

In case of community 2, chi square decreased initially and then suddenly started increasing but it is very much flattened and will reach to significance level only when we would have half of population sampled.

While in community 3, we found that the chi square decreased and then flattened which is exactly opposite to what we observed in community 1. Also it exceeds significance level when the sampled population is way less than the most probable effort, while in case of community 1, it exceeds significance level after sampling more than most probable cumulative area. For plotting I have ignored all area values which don't have more than 100 data points.

We even tried to change the chi averaged into p value for individual reading but obviously it is giving different curve fit because of non-linear relation between chi square and corresponding p value. Here, we have 20 categories and conversion of p value 0.05 into chi square is 30.144 (degree of freedom = 19).

In the normal distribution of community 3, we have a large bin at the last because we restricted the search of all individuals and forcefully added that in that bin because it was taking so much time for finding out successful iteration for higher areas.

3.1 Community 1

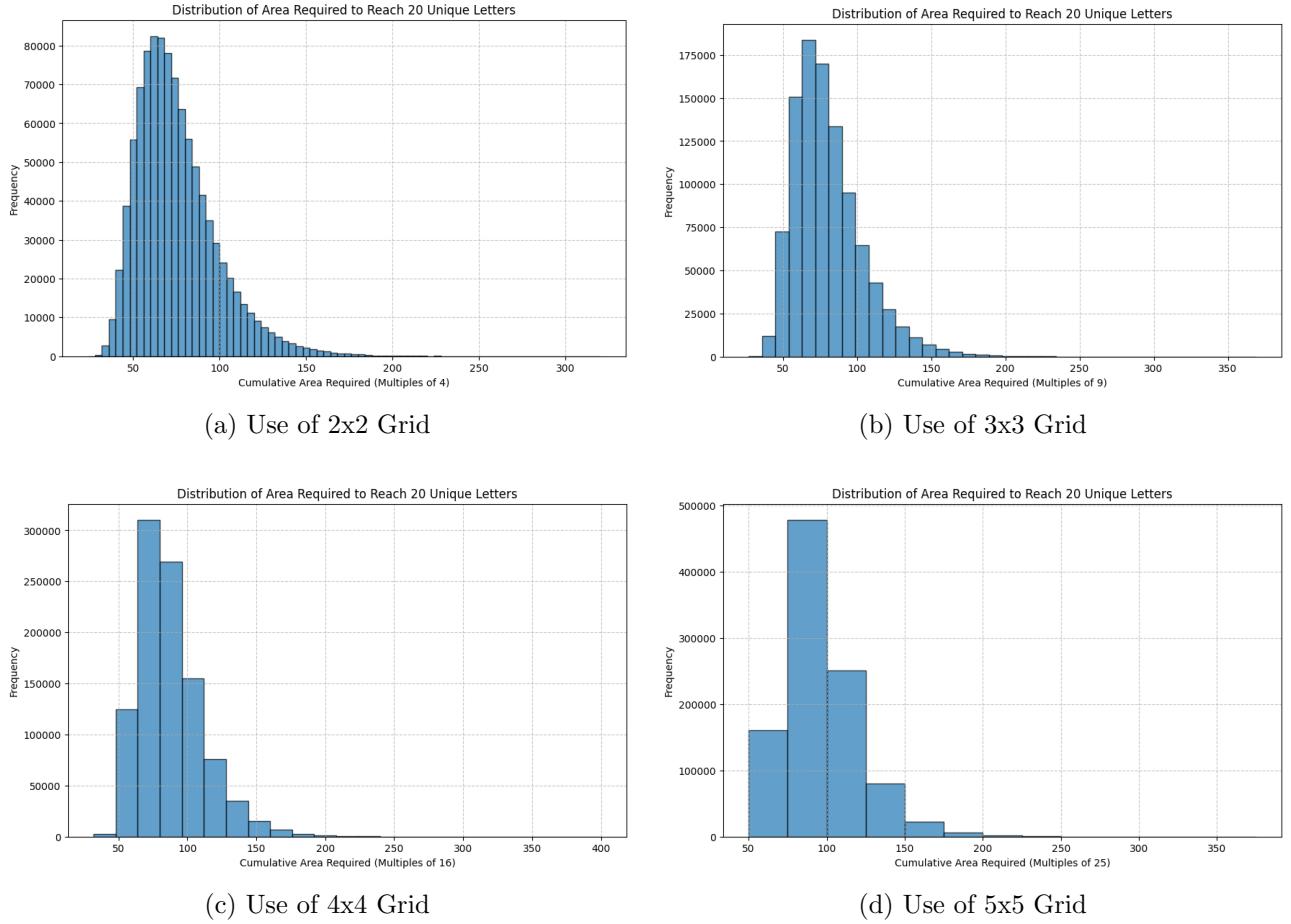
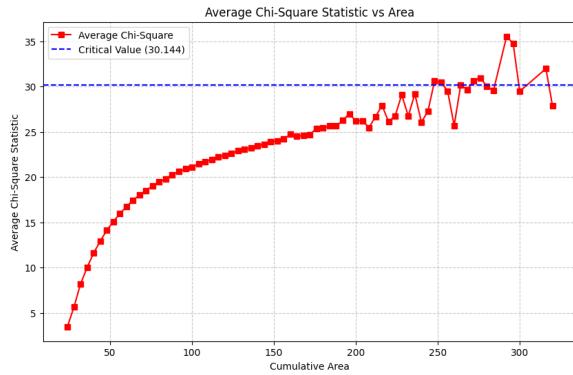


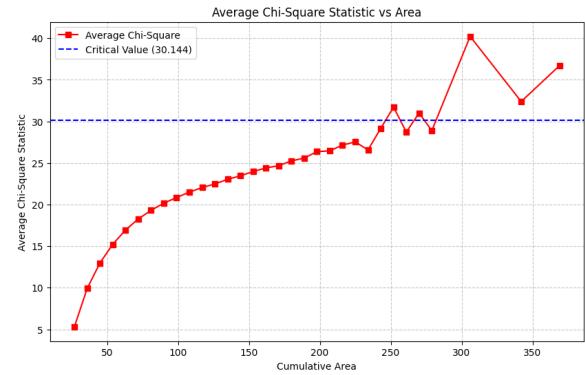
Figure 1: For Community 1

Table 1: Community 1: Five Most Probable Areas with Percentage of Total Trials and Occurrences for Different Grid Types

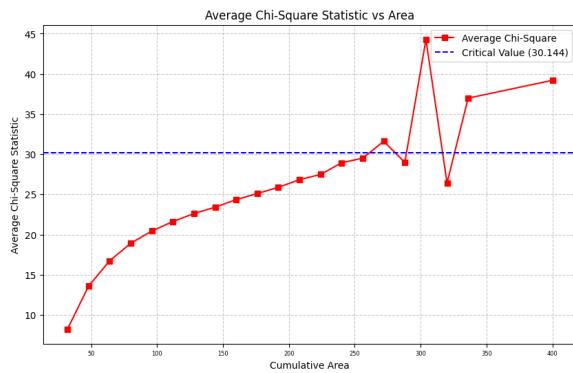
Area	2×2	3×3	4×4	5×5
A1	8.24% (60)	18.38% (63)	31.03% (64)	47.85% (75)
A2	8.21% (64)	17.01% (72)	26.95% (80)	25.05% (100)
A3	7.86% (56)	15.06% (54)	15.47% (96)	16.03% (50)
A4	7.81% (68)	13.36% (81)	12.45% (48)	8.02% (125)
A5	7.17% (72)	9.53% (90)	7.57% (112)	2.23% (150)



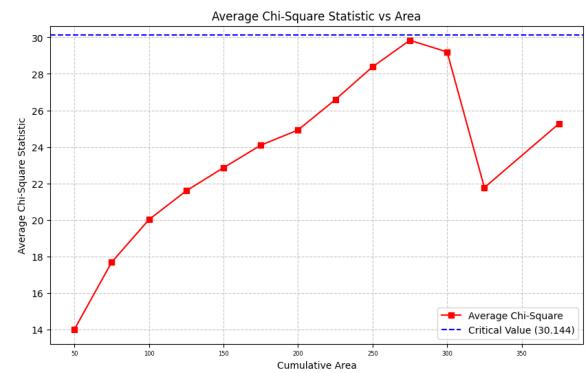
(a) Use of 2x2 Grid



(b) Use of 3x3 Grid

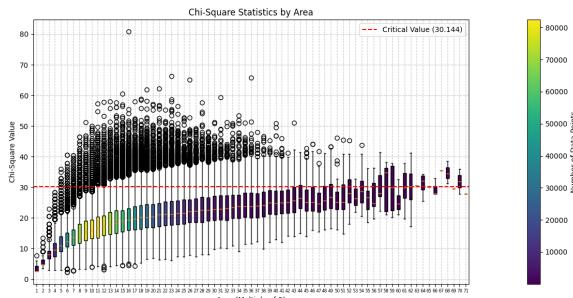


(c) Use of 4x4 Grid

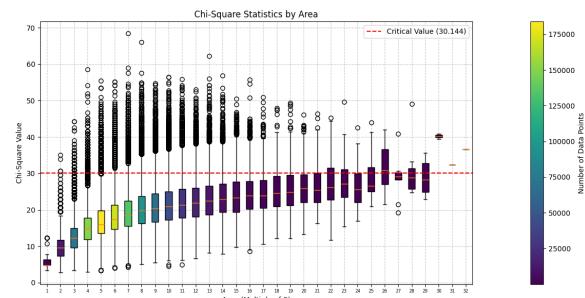


(d) Use of 5x5 Grid

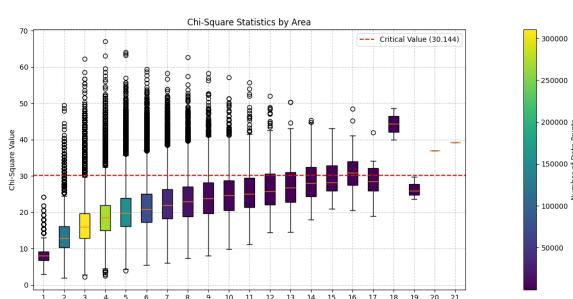
Figure 2: For Community 1



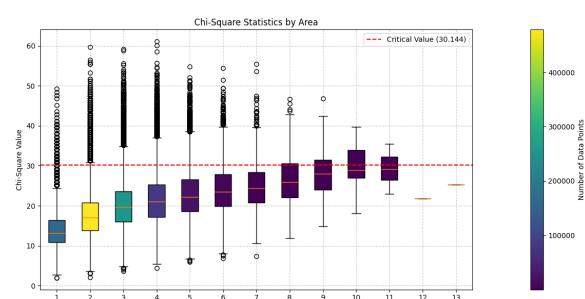
(a) Use of 2x2 Grid



(b) Use of 3x3 Grid

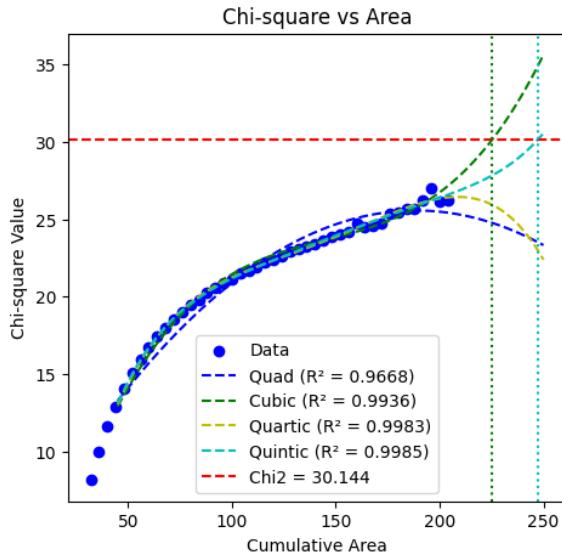


(c) Use of 4x4 Grid

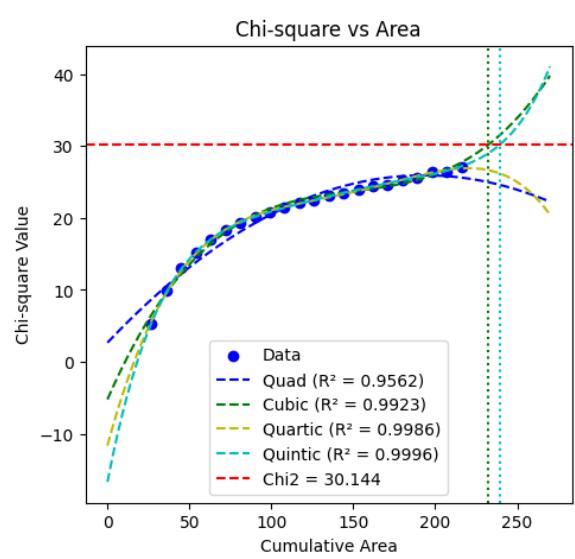


(d) Use of 5x5 Grid

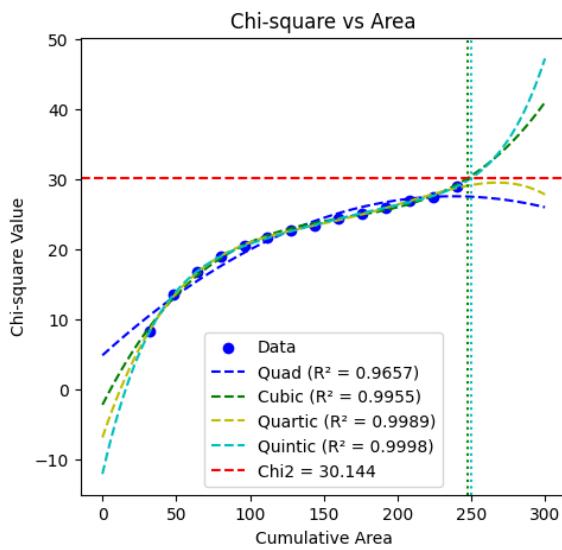
Figure 3: For Community 1



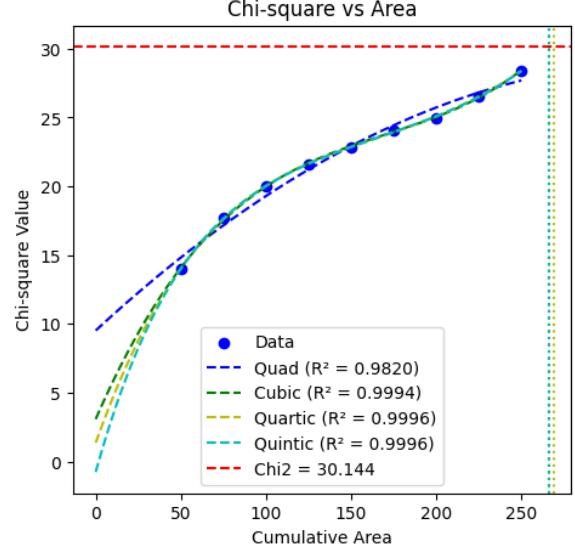
(a) Chi Square fit for 2x2 Grid



(b) Chi Square fit for 3x3 Grid



(c) Chi Square fit for 4x4 Grid



(d) Chi Square fit for 5x5 Grid

Figure 4: For Community 1

Table 2: Community 1 : Polynomial Fits of Areas Where Chi-Square Reaches 30.144 for Different Grid Sizes

Grid Size	Polynomial Fit Degree			
	Quadratic	Cubic	Quartic	Quintic
2×2	None	225.392	None	247.386
3×3	None	232.199	None	239.237
4×4	None	247.747	None	250.169
5×5	None	265.910	269.260	266.478

3.2 Community 2

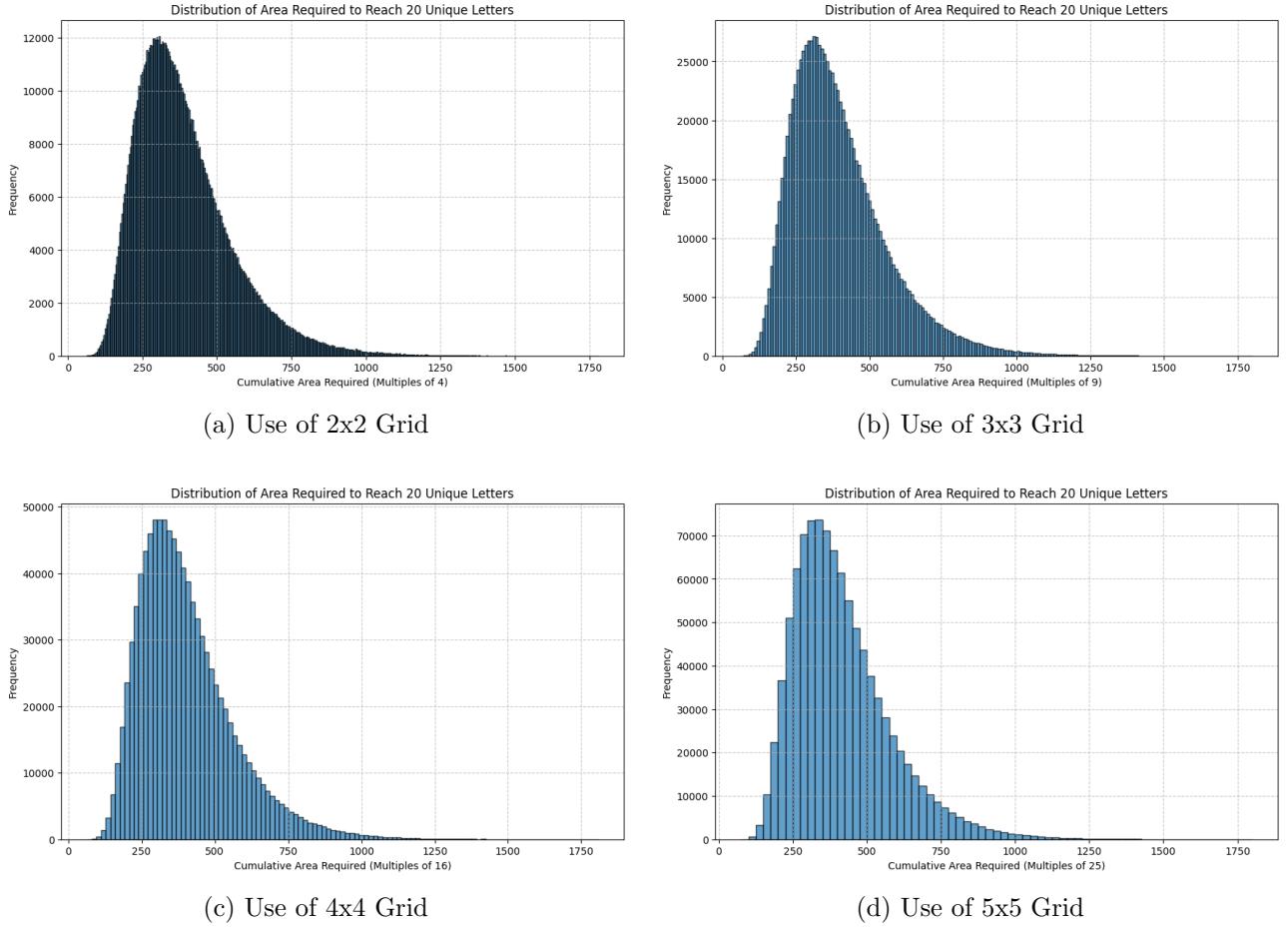
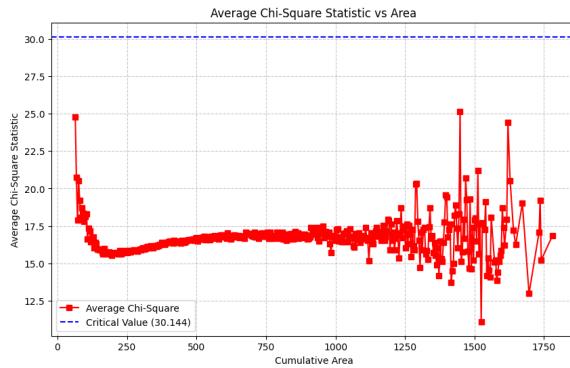


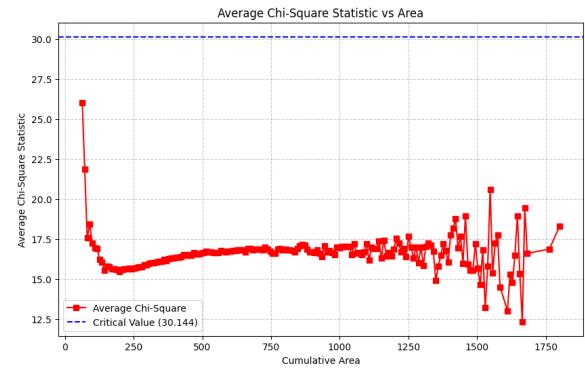
Figure 5: For Community 2

Table 3: Community 2: Five Most Probable Areas with Percentage of Total Trials and Occurrences for Different Grid Types

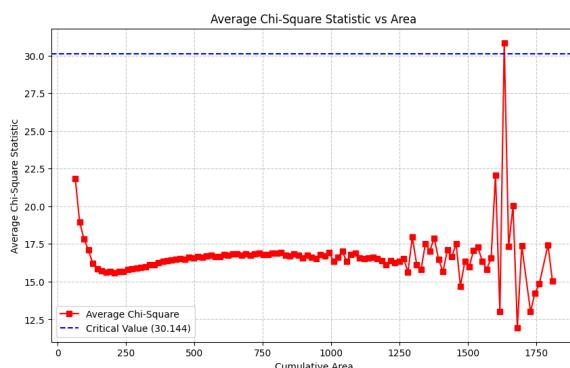
Area	2×2	3×3	4×4	5×5
A1	1.206% (304)	2.714% (306)	4.809% (320)	7.368% (325)
A2	1.204% (296)	2.707% (315)	4.806% (288)	7.351% (300)
A3	1.198% (288)	2.676% (288)	4.802% (304)	7.116% (350)
A4	1.197% (284)	2.673% (297)	4.640% (336)	7.03% (275)
A5	1.191% (300)	2.639% (324)	4.596% (272)	6.662% (375)



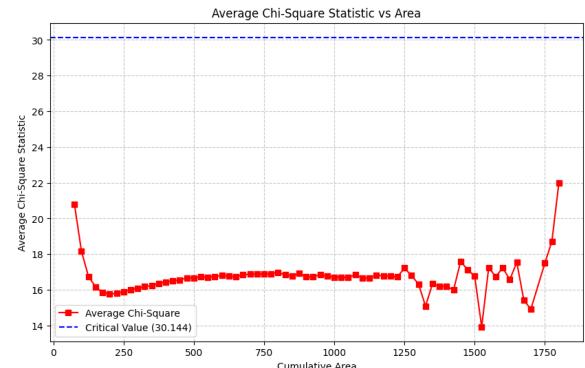
(a) Use of 2x2 Grid



(b) Use of 3x3 Grid

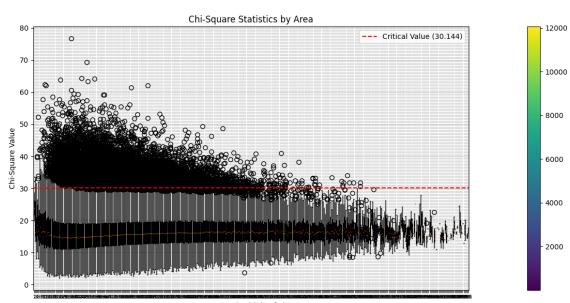


(c) Use of 4x4 Grid

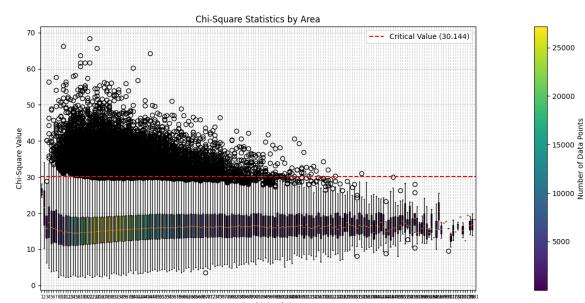


(d) Use of 5x5 Grid

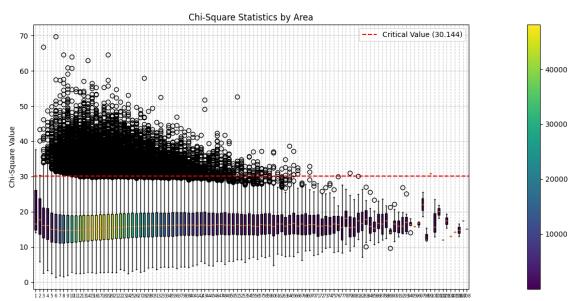
Figure 6: For Community 2



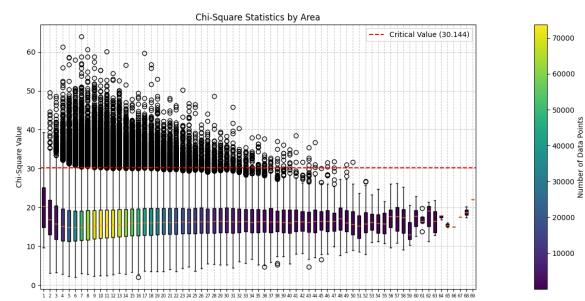
(a) Use of 2x2 Grid



(b) Use of 3x3 Grid



(c) Use of 4x4 Grid



(d) Use of 5x5 Grid

Figure 7: For Community 2

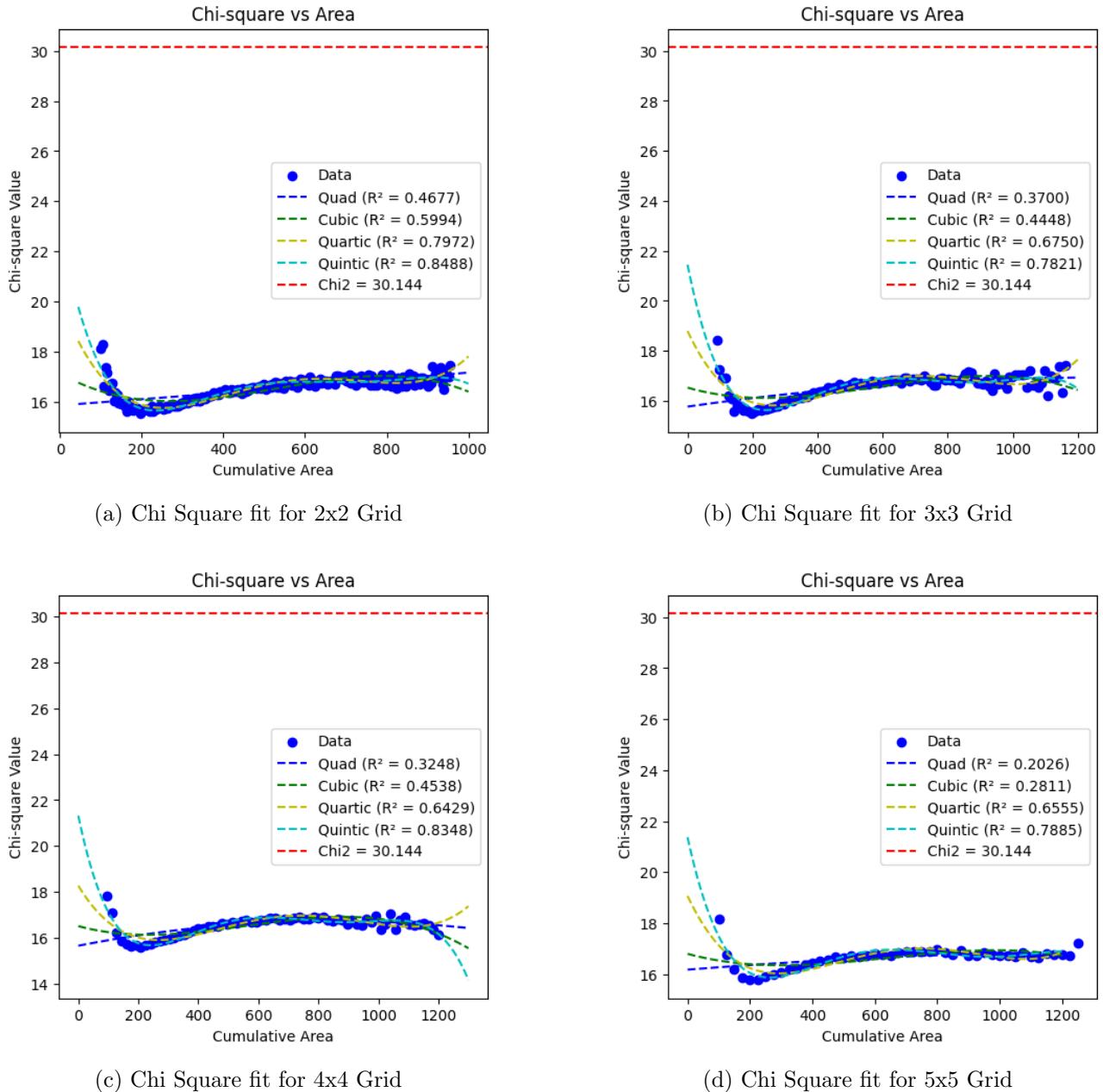


Figure 8: For Community 2

Table 4: Community 2 : Polynomial Fits of Areas Where Chi-Square Reaches 30.144 for Different Grid Sizes

Grid Size	Quadratic	Cubic	Quartic	Quintic
2×2	7165.83	None	1264.16	None
3×3	None	None	1522.83	None
4×4	None	None	1661.02	None
5×5	None	None	1631.63	None

3.3 Community 3

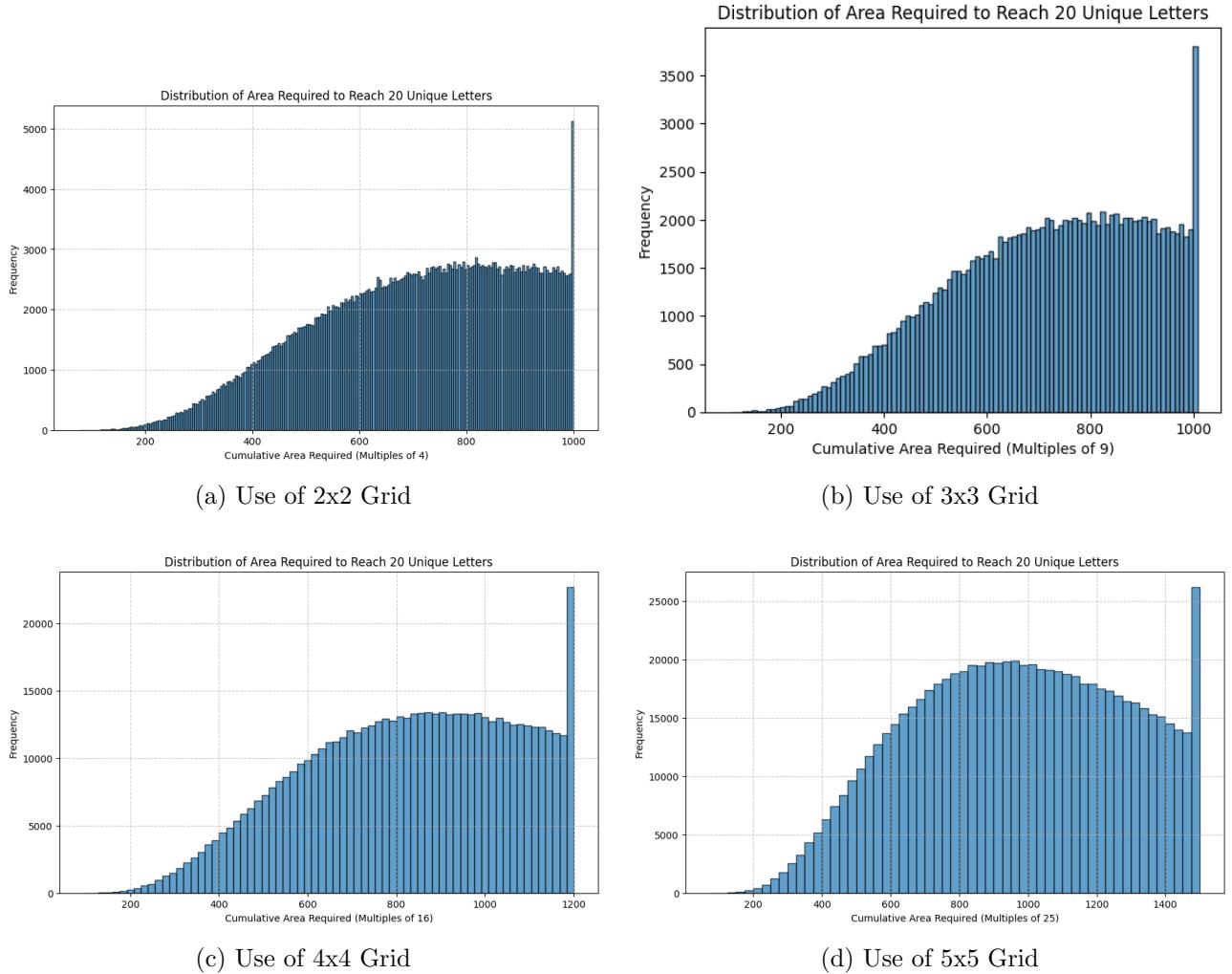
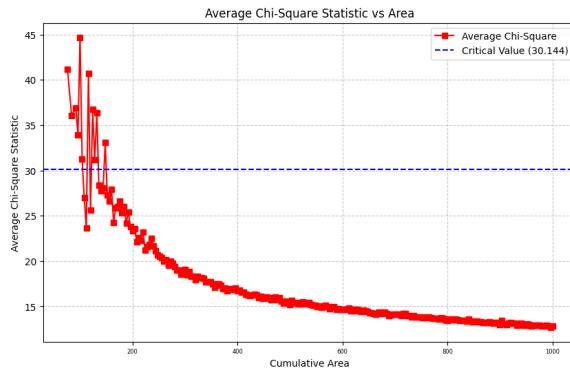


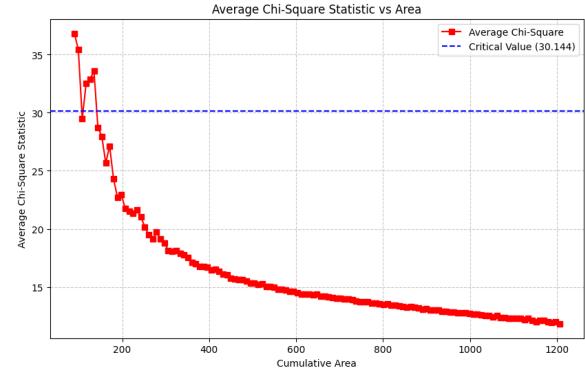
Figure 9: For Community 3

Table 5: Community 3: Five Most Probable Areas with Percentage of Total Trials and Occurrences for Different Grid Types

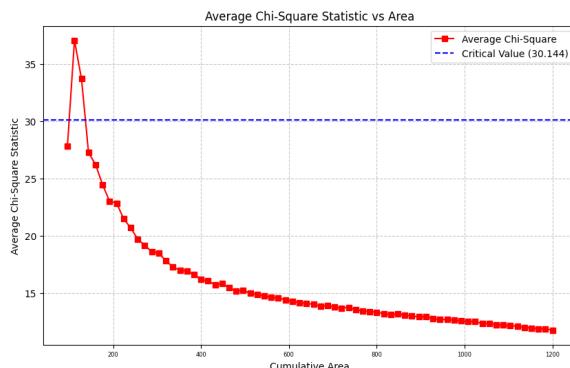
Area	2×2	3×3	4×4	5×5
A1	0.3588% (816)	0.8156% (828)	1.3403% (896)	1.9862% (950)
A2	0.3496% (792)	0.8147% (855)	1.3371% (864)	1.9811% (925)
A3	0.3491% (776)	0.8142% (846)	1.3346% (848)	1.9774% (875)
A4	0.3479% (848)	0.8135% (864)	1.3321% (976)	1.9683% (900)
A5	0.3474% (852)	0.8097% (783)	1.33% (880)	1.96% (1000)
A6	0.3474% (764)	0.8078% (810)	1.3283% (832)	1.95% (825)



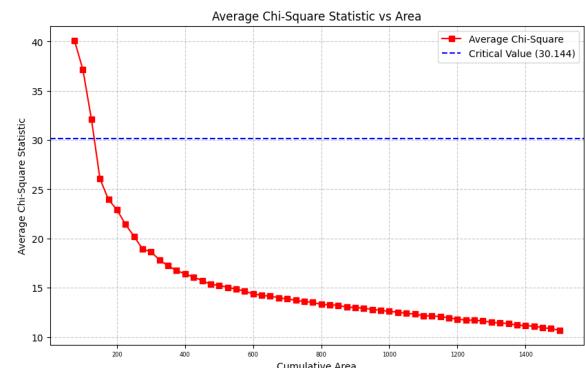
(a) Use of 2x2 Grid



(b) Use of 3x3 Grid

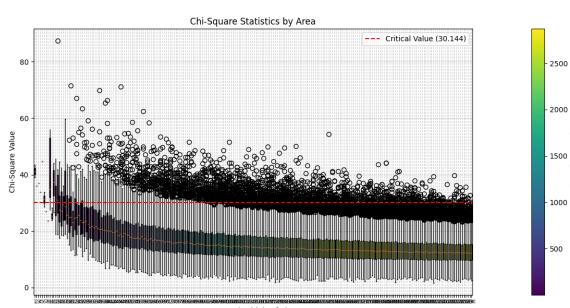


(c) Use of 4x4 Grid

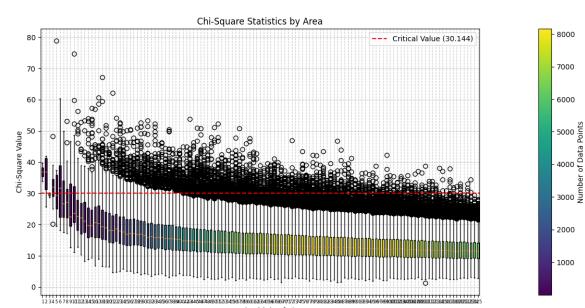


(d) Use of 5x5 Grid

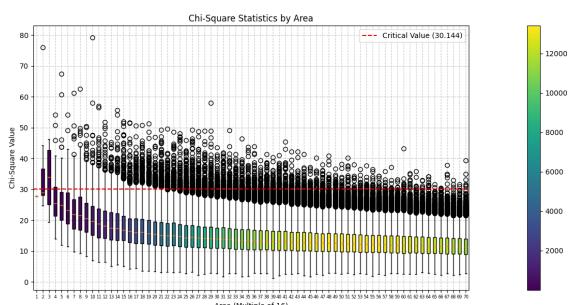
Figure 10: For Community 3



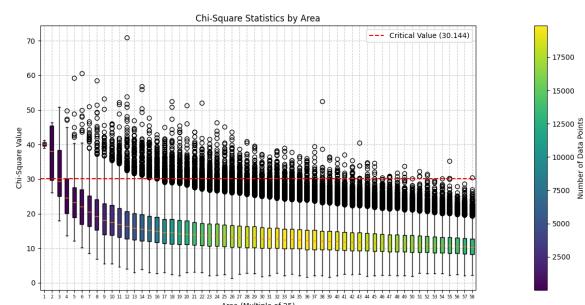
(a) Use of 2x2 Grid



(b) Use of 3x3 Grid



(c) Use of 4x4 Grid



(d) Use of 5x5 Grid

Figure 11: For Community 3

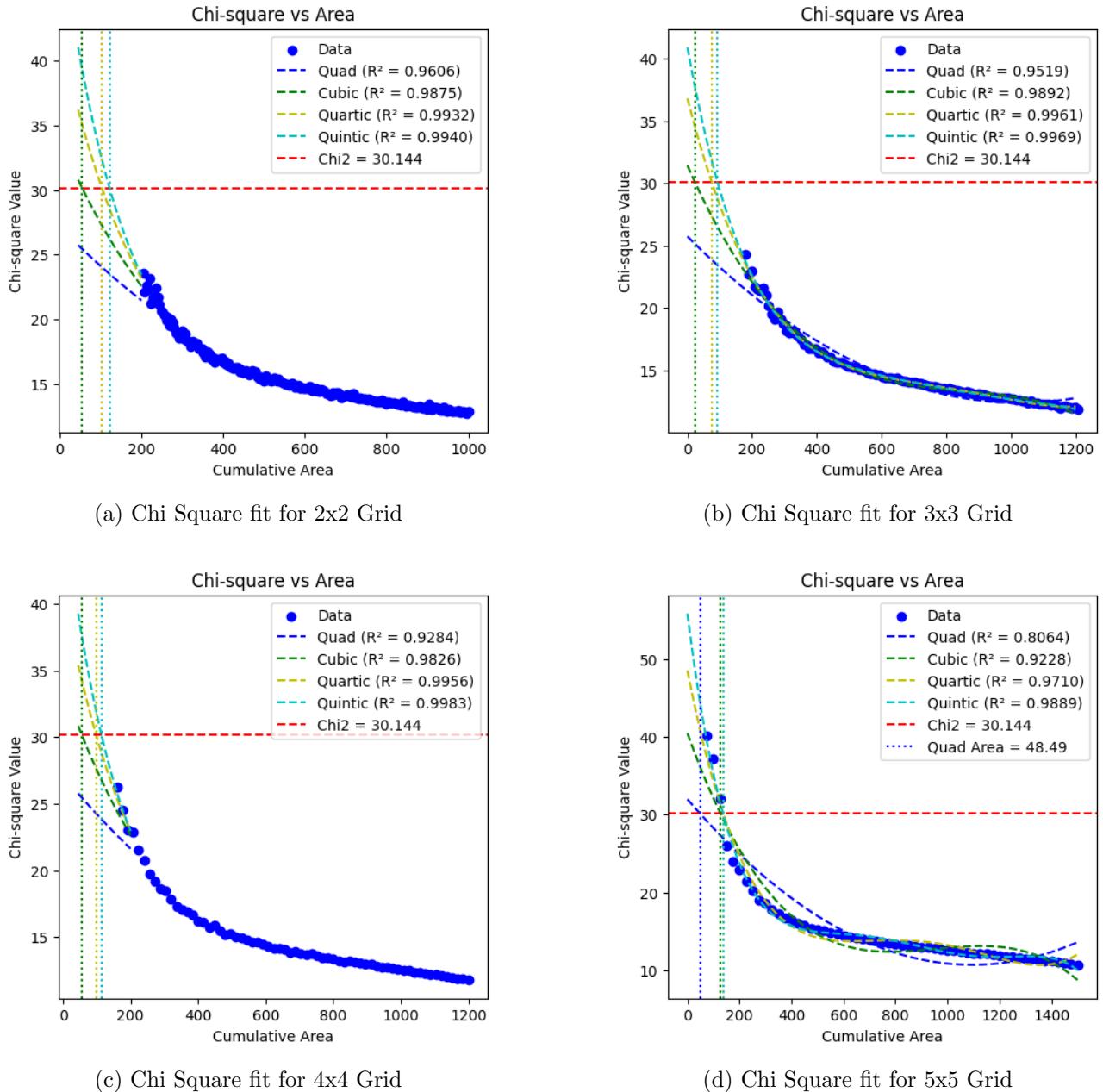


Figure 12: For Community 3

Table 6: Community 3 : Polynomial Fits of Areas Where Chi-Square Reaches 30.144 for Different Grid Sizes

Grid Size	Polynomial Fit Degree			
	Quadratic	Cubic	Quartic	Quintic
2×2	None	55.091	103.401	122.026
3×3	None	22.539	74.697	92.648
4×4	None	55.997	98.441	114.174
5×5	48.489	127.313	138.012	136.668

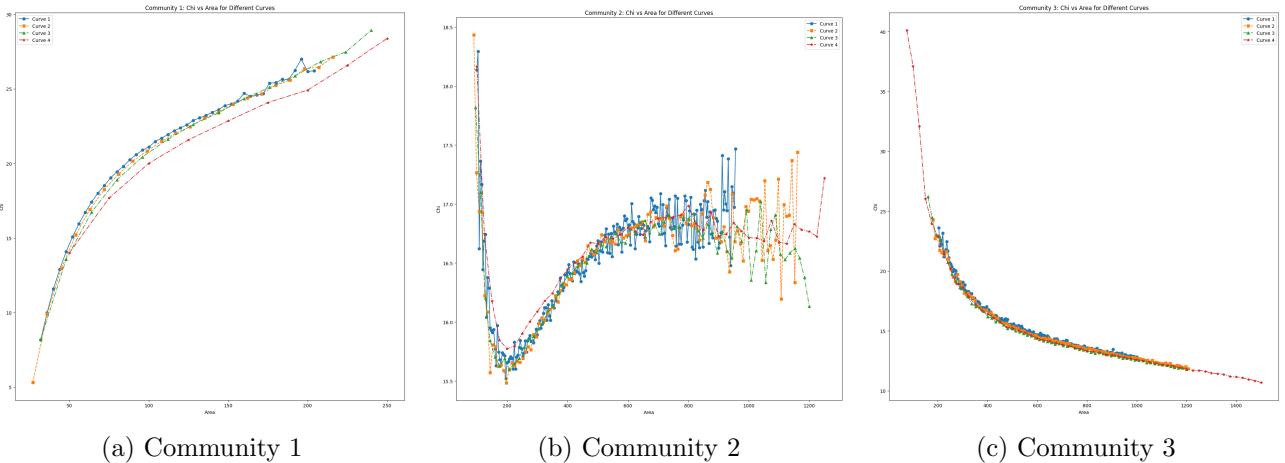


Figure 13: Comparison of all grid types averaged chi square value, showing almost same behaviour even with changing grid sizes

4 Analysis Part 2(a)

We are going to apply some prediction method for getting about the differences between the communities. Chao and Shen (2004) has demonstrated in their paper that how can we predict the total number of species in a population based on a limited sample. They tried to develop a method that not only estimates unseen species but also predicts how many additional species you'd find if you sample more. Their approach is "nonparametric", as it doesn't assume a specific species abundance distribution (like lognormal or logarithmic), making it flexible for real-world data.

Formulas used for non-parametric prediction:

- Sample Coverage for Rare Species:

$$C_{\text{rare}} = 1 - \frac{f_1}{n}$$

- Heterogeneity of Rare Species:

$$CV_{\text{rare}}^2 = \max \left(\frac{S_{\text{rare}}}{C_{\text{rare}}} \cdot \frac{\sum_{k=1}^{\kappa} k(k-1)f_k}{n^2} - 1, 0 \right)$$

- Estimated Unseen Rare Species:

$$f_0 = \frac{S_{\text{rare}}}{C_{\text{rare}}} + \frac{f_1}{C_{\text{rare}}} \cdot CV_{\text{rare}}^2 - S_{\text{rare}}$$

- Predicted Additional Species:

$$\theta_3(t) = f_0 \left(1 - \exp \left(-t \frac{f_1}{f_0} \right) \right)$$

- Total Predicted Species:

$$S_{\text{pred}} = S_{\text{obs}} + \theta_3(t)$$

where S_{obs} is observed species, f_1 is singletons, n is individuals sampled, S_{rare} is rare species (frequency $\leq \kappa$), f_k is species with k individuals, and t is additional sampling effort.

Contributions

They have introduced the method to predict S_{pred} for larger samples using rare species, validated with bird and ant datasets, by selecting species with lower frequency ($\leq \kappa$), which drive accurate estimation of unseen species (f_0). They have used C_{rare} to adjust for sampled proportion, enhancing small-sample predictions. They have also used CV_{rare}^2 to account for abundance variability, improving robustness across communities.

Use for our communities:

We are doing it to find whether it is working for the communities with non-uniform distribution. We assumed the frequency at least 10 for not being counted in rare species. We have done this all with 2 into 2 grid size. From normal distribution, we know the most cumulative areas to get are around 60,300 and 800 for community 1,2 and 3 respectively. So I took sample one third to the most probable effort, so in first community i took minimum 20 cumulative area for prediction but I maintained the proportion for other communities(little disparity in community 3 as I have took more than 1/3 as sample of most probable).So, for samples we took 20,100 and 300 cumulative areas for calculating these values:

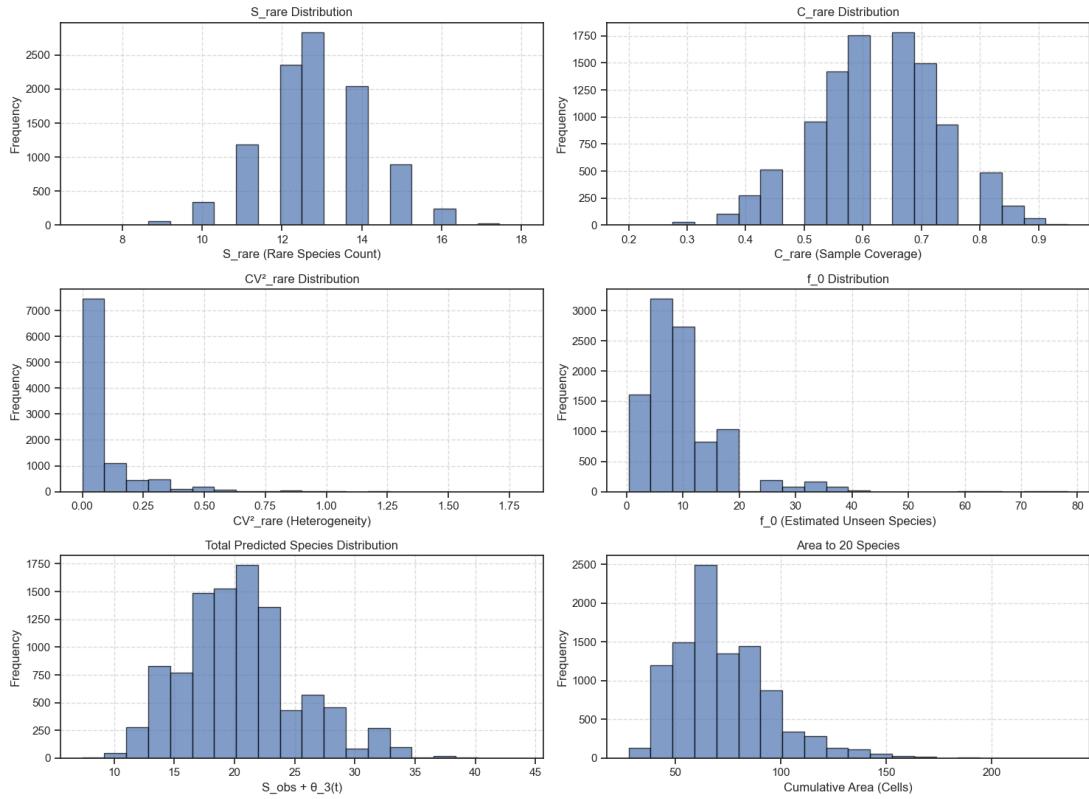
Table 7: Comparison of Mode and Median values Across Communities 1, 2, and 3

Metric	Community 1		Community 2		Community 3	
	Mode	Median	Mode	Median	Mode	Median
S_{rare}	13	13.0	13	13.0	8	8.0
C_{rare}	0.65	0.6	0.95	0.95	0.85	0.85
CV_{rare}^2	0.0	0.0	0.0	0.0	0.0	0.232
f_0	7.0	8.667	0.684	0.632	0.889	2.594
$S_{\text{obs}} + \theta_3(t)$	19.053	20.299	15.684	15.583	8.671	9.455
Area	60	68.0	280	360.0	788	984.0

Conclusion

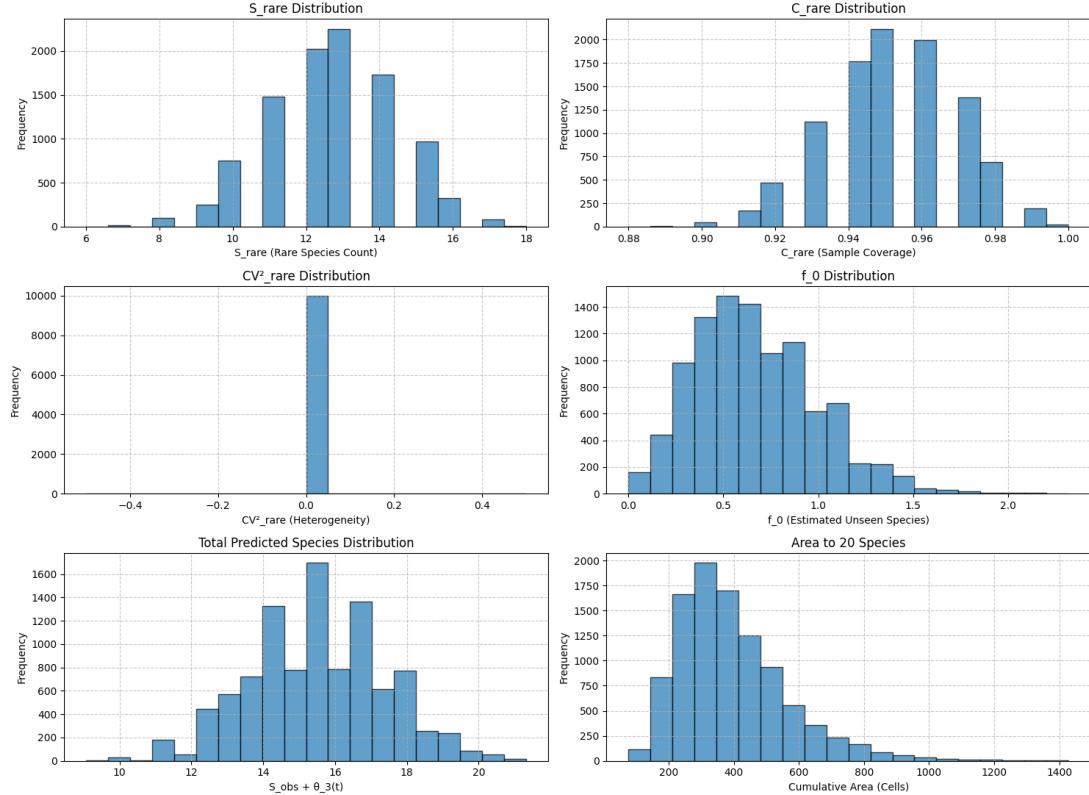
It worked very close for community 1 with equal frequency for all species but it underestimates when community 2 and 3 are there in which some species is so much dominant and some are very rare, and it is clearly showing that in real systems, the actual species figure are underestimated because of the factors like rare species and dominant species.

Distributions from First 5 Subplots (20 Cells) Over 10000 Iterations



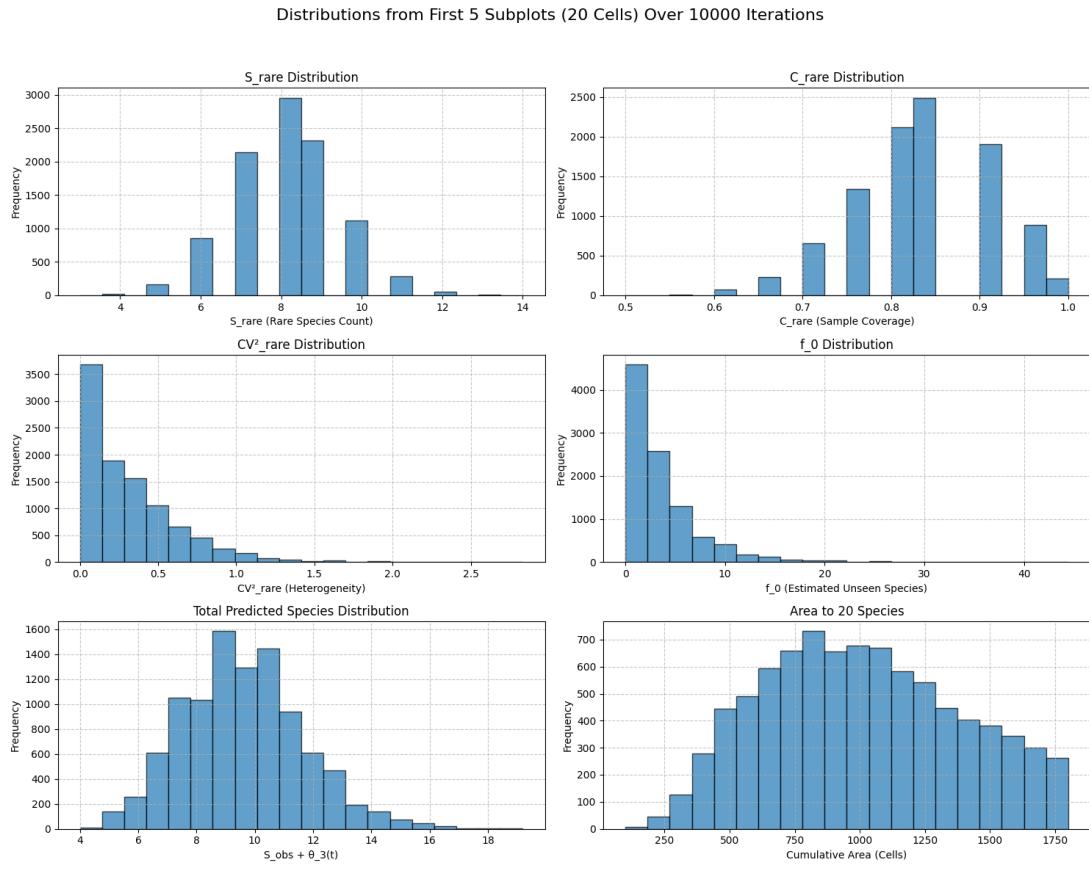
(a) Community 1

Distributions from First 5 Subplots (50 Cells) Over 10000 Iterations



(b) Community 2

Figure 14: Distribution of values over multiple iterations



(a) Community 3

Figure 15: Distribution of values over multiple iterations

5 Analysis Part 2(b)

We used the logarithmic series model from Fisher, Corbet, & Williams (1943) to estimate the diversity index (α) and sampling parameter (x) for a community with $S = 20$ species and $N = 60$ individuals sampled.

Key Formulas

From Fisher et al. (1943):

- Total species:

$$S = -\alpha \ln(1 - x)$$

- Total individuals:

$$N = \frac{\alpha x}{1 - x} \quad \text{or equivalently} \quad x = \frac{N}{N + \alpha}$$

So we solved below two formulas, for comparing α and x :

$$S + \alpha \ln(1 - x) = 0$$

$$x - \frac{N}{N + \alpha} = 0$$

where $S = 20$, $N = 60$, 300 and 800 for community 1,2 and 3. We used `fsolve` to solve it.

Findings in the Paper

Fisher et al. (1943) introduced the logarithmic series to describe species abundance, showing that α reflects diversity and x scales with sampling effort

Conclusion

Table 8: Values of x and α for Communities 1, 2, and 3 Using Fisher et al. (1943)

Community	x	α
Community 1	0.851001	10.505230
Community 2	0.984175	4.823704
Community 3	0.995371	3.720694

We can see the α is decreasing for community 2 and 3 , because the diversity is decreasing either because of dominant species or rarity of some species. That's why the sampling effort is increasing here in response to that.

6 Analysis Part 2(c)

The code applies extrapolation techniques from Colwell & Coddington (1994) , estimating total species richness using nonparametric and parametric methods, plus complementarity between subplots.

Key Formulas

From Colwell & Coddington (1994):

- Chao1 Estimator:

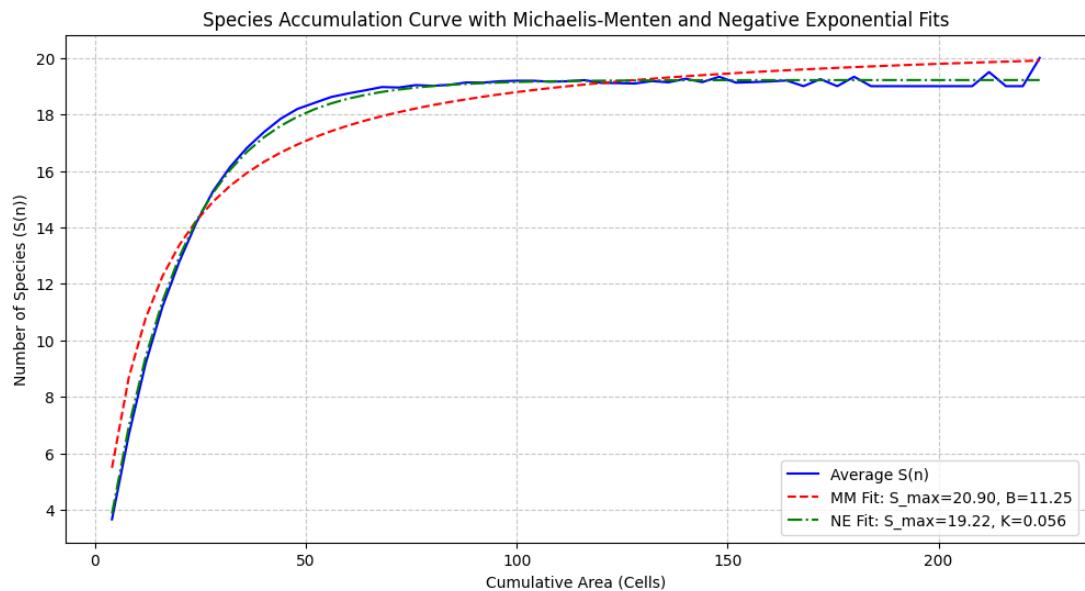
$$S_{\text{Chao1}} = S_{\text{obs}} + \frac{f_1^2}{2f_2}$$

- Complementarity:

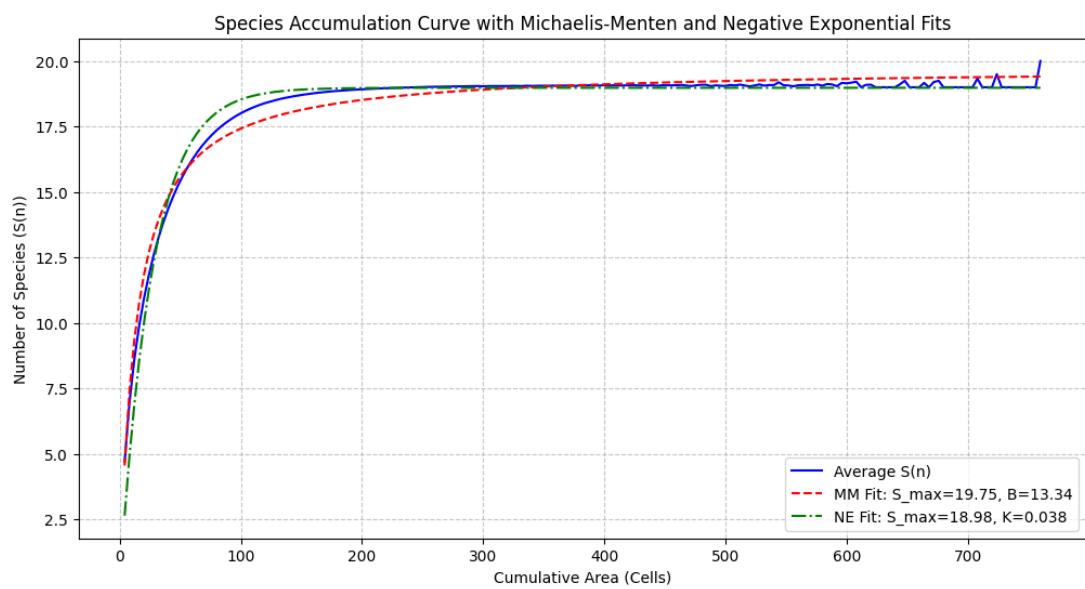
$$C_{jk} = \frac{S_j + S_k - 2V_{jk}}{S_j + S_k - V_{jk}}$$

Additional fits that we tried across iterations:

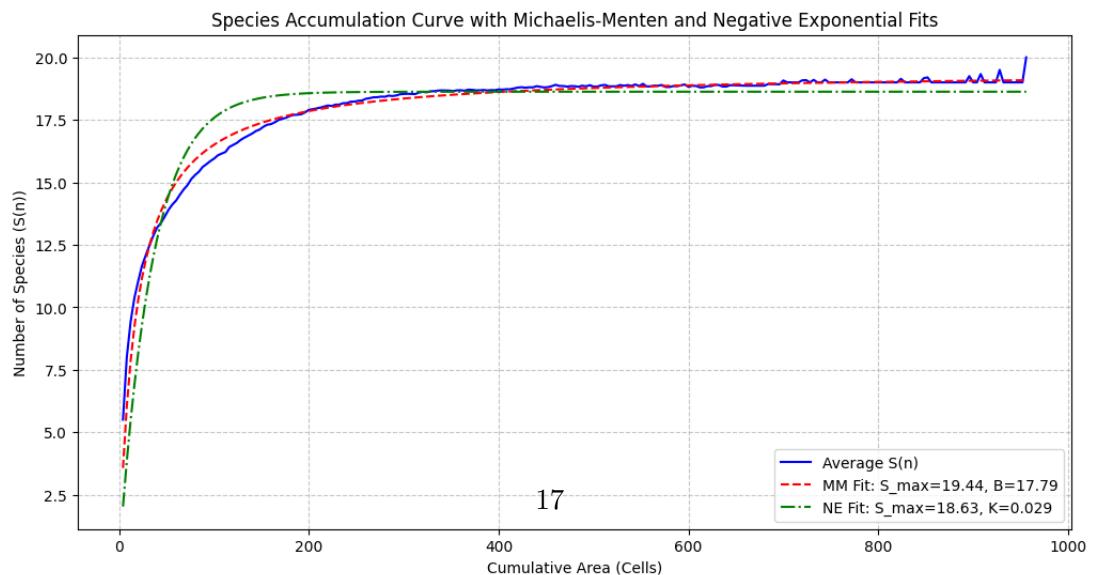
- Michaelis-Menten: $S(n) = \frac{S_{\max} \cdot n}{B + n}$
- Negative Exponential: $S(n) = S_{\max} \cdot (1 - e^{-K \cdot n})$



(a) Community 1



(b) Community 2



(c) Community 3

How we calculated:

We computed S_{Chao1} , for sample we took 20,100 and 300 cumulative areas for calculating these values (same as we used chao 2004 paper). For complementarity , we calculated it across all subplots:

Table 9: Chao1 and Complementarity Metrics for Communities 1, 2, and 3 After 20 Cells

Metric	Community 1		Community 2		Community 3	
	Mean	Median/Mode	Mean	Median/Mode	Mean	Median/Mode
S_{Chao1}	23.92	20.17 / 17.90	20.79	19.90 / 19.50	19.97	19.25 / 20.00
C_{jk}	0.887	—	0.629	—	0.562	—

Here , we can see S_{Chao1} ignores the dominant and rare species effect if total species is same across different communities, because it is only using the effect of singletons and doubletons for calculating that. And for complementarity (going near 0 for identical or near 1 for completely distinct), we can see the value is decreasing because the grids are getting more similar because of some dominant or rare species.

Findings in the Paper

Colwell & Coddington (1994) demonstrated that Chao1 accurately estimates unseen species using rare species frequencies, while parametric fits extrapolate total richness. Complementarity measures sample dissimilarity, aiding biodiversity assessment.

7 Conclusion

This was a theoretical exercise, hence can't give an estimate of optimal sampling methods for real systems.

References

- [1] Chao, A., & Shen, T.-J. (2004). “Nonparametric Prediction in Species Sampling.” *Biometrics*, **60**(2), 524–531.
- [2] 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), 42–58.
- [3] Colwell, R. K., & Coddington, J. A. (1994). “Estimating Terrestrial Biodiversity Through Extrapolation.” *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, **345**(1311), 101–118.

*LLMs were used for generating python code that helped us to simulate. It also helped us to understand the research papers.