

# Using Species Richness to Estimate Benthic Sampling Effort

Nicole P. Maher & Robert M. Cerrato  
Marine Sciences Research Center  
Stony Brook University, Stony Brook, NY 11794-5000

A Report Prepared for the Hudson River Estuary Program of the New York State Department of  
Environmental Conservation

January 20, 2006

## Introduction

The pilot studies conducted in the Hudson River (Strayer et al 2005; Maher and Cerrato 2004) and ongoing analysis of the data from these and other areas have demonstrated significant relationships between geophysical data (e.g., side-scan, multibeam, subbottom data) and benthic faunal distributions. These relationships suggest that that it should be possible to develop an efficient, parsimonious design for field studies that uses the geophysical data collected during a first stage of a study to direct subsequent environmental and faunal sampling. Independent of how the first stage is carried out, there is a need to determine the number of faunal samples that should be collected at the second stage to guarantee that an adequate amount of data are available to identify and describe faunal community structure. We suggest that an estimate of species richness can serve as the basis for the determining sample size. An appropriate criterion might be, for example, to set sample size within an environmentally and biologically homogenous area large enough such that at least 70% of the species that are present are collected. The goal of present study is to identify an appropriate species richness estimator, use it to estimate the fraction of species collected in environmentally and biologically homogenous areas, and examine the results to recommend a practical sample size for faunal studies.

## Method

### Species Accumulation Curves and the Chao 2 Species Richness Estimator

Species accumulation curves are plots of the cumulative number of species vs. the number of pooled samples. They are generated from samples of a local species assemblage collected from within a biotope, an area with homogeneous environmental and biotic characteristics. The curve of the cumulative number of species vs. the number of samples pooled increases initially as the number of samples increases, but it eventually levels off as it approaches the species richness for the biotope. Since the order that samples are added affects the shape of the curve, the analysis requires generating an ensemble of curves by randomly permuting sample order 100-200 times and calculating the mean of the ensemble.

Terrestrial ecologists have used a number of approaches to estimate species richness, including parametric fitting of exponential and hyperbolic models to species accumulation curves in order to find the asymptote, as well as nonparametric techniques based on the empirical distributions. Most of these approaches have not been applied to estuarine or marine systems. A comparison of species richness estimators by Colwell and Coddington (1994) suggested that an estimator designated as Chao 2 and a second order jackknife estimator worked extremely well to predict species richness. These two nonparametric estimators were also particularly well suited for small sample sizes (< 25).

We evaluated two parametric models (exponential and hyperbolic) and the two nonparametric estimators (Chao 2 and jackknife) using three available data sets and chose the Chao 2 estimator based on performance. The parametric models had serious goodness of fit problems and were rejected. The second order jackknife consistently overestimated species richness relative to Chao 2. Colwell and Coddington (1994) also found that the jackknife had a slight positive bias in their analysis of an extensive data set. For this reason, it was not considered further.

For any set of samples, the Chao 2 estimator can be calculated as

$$S_2^* = S_{obs} + (L^2 / 2M)$$

where  $S_2^*$  is the estimated species richness,  $S_{obs}$  is the observed number of species in the samples,  $L$  is the number of species that occur in only one sample, and  $M$  is the number of species that occur in exactly two samples. The variance of  $S_2^*$  can be estimated as

$$\text{var}(S_2^*) = M \left[ \left( \frac{L/M}{4} \right)^4 + (L/M)^3 + \left( \frac{L/M}{2} \right)^2 \right]$$

$S_2^*$  can be used in a sequential manner as each sample is added to a pooled set. As in the case of generating species accumulation curves, the order that samples are added affects the shape of the curve of  $S_2^*$  vs. the number of pooled samples. The analysis thus requires generating an ensemble by randomly permuting sample order 100-200 times and calculating the mean  $S_2^*$  for the ensemble. The curve of  $S_2^*$  vs. the number of pooled samples increases initially with sample size until about the square root of twice the total fauna is observed (Colwell and Coddington 1994). At that point the estimator should level off and become independent of sample size (Colwell and Coddington 1994). Evidence that the estimator has leveled off and become stable is necessary before it can be used with confidence.

### Data Sets and Biotope Identification

We carried out a statistical analysis of species accumulation curves and species richness estimates for samples within each biotope in the Kingston (Area 3) and Tappan Zee (Area 1) pilot areas and an area east of Shelter Island, NY. These three sampling areas spanned a wide range of habitats from freshwater to almost marine conditions and each varied considerably by sediment type.

The basic sampling design was similar in all three areas. Multibeam bathymetry, side-scan sonar and other geophysical data were used to identify the major geophysical provinces in each area (Strayer et al. 2005, Maher and Cerrato 2002, Cerrato unpublished data). Station locations for animal and sediment grabs were then determined by stratified random sampling, with an equal number of stations allocated to each province. Strayer et al. (2005) collected 3 pooled petite PONAR grab samples at each station. Maher and Cerrato (2002) and Cerrato (unpublished data) collected two replicate 0.04 m<sup>2</sup> modified van Veen grabs at each station. Samples for all studies were washed through a 0.5 mm sieve. Samples for grain-size analysis were also collected.

Each study area was subdivided into biotopes using an iterative application of Redundancy Analysis (RDA) and K-means cluster analysis (Jongman et al. 1995; Legendre and Legendre 1998). Environmental data for RDA consisted of nominal variables designating the geophysical provinces and grain-size data.

Faunal data for RDA were Hellinger transformed abundances (Legendre and Gallagher 2001). Constrained ordination sample scores from the first four RDA axes were used as input variables for K-means cluster analysis. The optimal number of clusters of stations (i.e., biotopes) was determined by running the analysis assuming different numbers of groups and identifying the best fit using the C-H statistic. Nominal biotopes were then substituted for the nominal geophysical province variables and the RDA and K-means analysis was repeated until biotope membership stopped changing (usually only 1 iteration was necessary).

### Species Richness Data Analysis

Statistical analysis of species accumulation curves and species richness estimates was carried out for samples within each biotope. When replicate samples were available at a station (Area 1 and Shelter Island), the analysis was run both using the replicates as individual samples and by averaging the replicates for each station. The fraction of species collected was estimated by calculating  $S_{obs}/S_2^*$ . Analysis for each biotope in the studies provided examples of the sampling effort needed to collect a reasonable fraction of the species present in an area. The results were then examined to set guidelines for sampling effort.

### Results

Species richness could be estimated for about two-thirds of the biotopes (Tables 1-3). In the remainder, too few samples were available, and the Chao 2 estimator did not level off when plotted against sample number. Figure 1 shows examples of biotopes where the Chao 2 index successfully provided (A) and failed to provide (B) an estimate of species richness. Also shown are species accumulation curves for the two biotopes. Calculations based on replicate samples tended to yield slightly lower species richness estimates (~3.7% on average) than when the samples were averaged for each station (compare individual tables in Tables 1 & 3).

Biotopes within Area 1 were the most problematic, with estimates of species richness possible for only 2 of the 5 biotopes. In contrast, 4 of 6 attempts were successful at Area 3 and 4 of 5 attempts at Shelter Island. Four of the 6 failed attempts corresponded to biotopes with the smallest number of samples within an area (Biotopes 2 & 5 in Area 1, Biotope 4 in Area 3, and Biotope A in Shelter Island). However, small sample size did not always indicate failure; 3 of the 4 biotopes with only 5 samples in Area 1 yielded species richness estimates. Note also that Shelter Island biotopes had 2-3 times more observed species than the Hudson River areas, and that a high number of observed species does not guarantee convergence of the Chao 2 estimator (e.g., Biotope A).

In eight of the ten data sets where species richness estimates were possible, a fixed collection effort of 10 samples yielded at least 70% to more than 85% of the estimated species present (Figures 2-4). The two exceptions were biotope CEG at Shelter Island where 10 samples resulted in the collection of 54% of the species, and Biotope 4 in Area 1 with 58% collected. Both these biotopes were the most diverse assemblage examined in their respective area. Although only three large-sample examples were available (Biotopes 1 and 4 in Area 1, Biotope CEG at Shelter Island), a 15-sample collection effort increased the fraction of species collected by an additional 6 to 8%. Overall, the variable sampling effort conducted in the three areas yielded 69 to 91% of the estimated species within each biotope (Tables 1-3).

## Discussion

We examined a diverse assemblage of biotopes from freshwater to almost marine conditions and of varying bottom types. We expected species accumulation curves and species richness to vary by salinity regime, sediment type, and probably several other factors. Biotopes at Shelter Island, the most marine area (29-30 ppt), had more than twice the number of species than either the mesohaline Area 1 (8-11 ppt) or the freshwater Area 3. Because of the high species richness, species accumulation curves, when plotted as a fraction of species sampled, tended to increase more slowly at Shelter Island compared to the other two areas (Figures 2-4). For example with 5 samples, an estimated 41-59% of the species were collected at Shelter Island, but 47-65% at Area 3 and 48-66% at Area 1. This difference is small and may not be significant, especially when considering the error in estimating species richness (see standard deviations in Tables 1-3). This pattern should, of course, be tested as additional data sets are analyzed, but it suggests that some site-to-site adjustment in sampling effort may be necessary.

For any fixed level of sampling effort within an area, coarse sediment, low species richness biotopes tended to have a greater estimated fraction of species sampled than finer-grained, high species richness biotopes (Figures 2-4). With a sample size of 20, for example, an estimated  $91 \pm 10\%$  of the species were collected in the coarse central channel deposits at Area 1 compared to  $69 \pm 30\%$  in the mudflats (Figure 2). Caution is again advised in interpreting this pattern, both because of the size of the error estimates and because a fraction of the biotopes failed to produce species richness estimates. This pattern is, however, consistent across the three data sets and suggests that within a study area, coarser bottom types with low biodiversity may require lower sampling effort than finer-grained, more diverse biotopes. Further examination of additional data sets is again warranted.

The failure of the Chao 2 estimator to produce a species richness estimate was due primarily to the small number of samples available for some biotopes. Although an absolute minimum sample size was not apparent, biotopes with failed estimates tended to be the smallest sample sizes within an area (4 of 6 failures). Other possible causes for failed estimates also need to be considered. Since the Chao 2 estimator is based on the number of rare species in a set of samples, it would be sensitive to the care taken in processing samples and a taxonomist's ability to recognize a rare species occurrence. In addition, misclassification of samples into a biotope is also a concern, since non-homogeneity would increase the number of "rare" species. One potential example of this may be Biotope 3 in the Area 1 data set. This biotope was represented by 26 samples from 4 original geophysical provinces and was one of the largest sample sets in the study. In the future, a test for homogeneity should be developed and used with the species richness estimator. Although the details have not been worked out, such a test can probably be based on comparing species accumulation to rarefaction curves (Colwell and Coddington 1994). Rarefaction can also produce a plot of expected number of species vs. number of samples, but does so assuming random sampling of the pooled data without replacement.

The slightly lower species richness estimates (~3.7% on average) when calculations were based on replicate samples compared to when the samples were averaged for each station (compare individual tables in Tables 2 & 3) is easily explained. Replicate samples at a station are not completely independent of one another but were treated as if they were. The Chao 2 estimator depends on the number of rare species present in the set of samples. A species that occurred at only one sampling station within a biotope would increase  $M$  in the Chao 2 formula when 2 replicates were used, but it would increase  $L$  when station averages were calculated. Since  $M$  is in the denominator and  $L$  is in the numerator of the Chao 2 formula, only one or two such occurrences would account for the observed differences. Since the differences were small, either replicates or station averages can be used to examine sampling effort.

So, how many faunal samples should be collected in each biotope to characterize community structure? Based on the data sets analyzed, 10 samples would result in the collection of 70 to >85% of the species in most biotopes but only about 50-60% of the species in the most diverse assemblages within an area. Collecting >70% of the species is likely to be sufficient to characterize community structure, but 50-60% seems inadequate. From the limited examples of large-sample, species rich biotopes (Biotope 4 in Area 1 and Biotope CEG in Shelter Island), 20 samples might be needed for  $S_{obs} / S_2^*$  to exceed 70%. An effort beyond 20 samples for one biotope is probably not practical in a survey. The analysis suggests, therefore, that sampling levels may need to be two-tiered with at least 10 samples in a biotope but with twice that effort allocated to one or two of the most diverse biotopes. *A priori* identification of these diverse biotopes from the geophysical and grain-size data would be useful but is beyond the scope of the present study. It does appear to be possible, however, since at least in the three areas examined, biotopes with finer-grained sediments tended to be the most diverse.

## References

- Bell, R.E., R.D. Flood, S.M. Carbotte, W.B.F. Ryan, C. McHugh, M. Cormier, R. Versteeg, D. Chayes, H. Bokuniewicz, and V. Ferrini (2000) Hudson River Estuary Program Benthic Mapping Project Revised Final Report. Submitted to the New York State Department of Environmental Conservation. Lamont-Doherty Earth Observatory.
- Colwell, R.K. and J.A. Coddington (1994). Estimating terrestrial biodiversity through extrapolation. *Phil. Trans. Roy. Soc. Lon. B.* 345: 101-118.
- Jongman, R.H.G., C.J.F. ter Braak, and O.F.R. van Tongeren (1995) *Data Analysis in Community and Landscape Ecology*. Cambridge University Press, Cambridge. 299 pp.
- Legendre, P. and E.D. Gallagher. 2001. Ecologically meaningful transformations for ordination of species data. *Oecologia*: 129: 271-280.
- Legendre, P. and L. Legendre (1998) *Numerical Ecology*. Elsevier Science, Amsterdam. 853 pp.
- Maher, N.P. and R.M. Cerrato (2004) *Feasibility of Mapping Benthic Biotopes in the Hudson River*. Final Report to Lamont-Doherty Earth Observatory, Columbia University. Marine Sciences Research Center, Stony Brook University, Stony Brook, NY.
- Strayer, D.L., H.M. Malcolm, R.E. Bell, S.M. Carbotte, and F.O. Nitsche (2006) Using geophysical information to define benthic habitats in a large river. *Freshwater Biol.* 51: 25-38.
- Maher, Nicole P., 2006. *A New Approach to Benthic Biotope Identification and Mapping*. Ph.D. Thesis. Marine Sciences Research Center, Stony Brook University, Stony Brook, NY. 181 pp. (copies of thesis can be obtained from Dissertation Express <http://disexpress.umi.com> Order # 3258909)

Table 1. Species richness estimates ( $S_2^*$ ) using the Chao 2 index on data from Area 3. The index was applied to replicate samples for each biotope.  $S_{obs}$  is the number of observed species.

Species richness based on replicate samples

Biotope	Number of Samples	$S_{obs}$	Species Richness $S_2^*$ (±s)	$S_{obs}/S_2^* * 100$	Bottom Type
1	10	23	-	-	Debris field and channel deposits - sand
2	5	28	38.1 (8.8)	74	Margin tributary - coarse
3	5	27	39.0 (10.4)	69	Margin - fine
4	5	10	-	-	Debris field and channel sediment waves
5	5	12	14.3 (3.4)	85	Channel sediment waves
6	10	26	30.5 (4.9)	85	Margin - sand and dredged channel

Note: "-" indicates that too few samples were available to estimate species richness

Table 2. Species richness estimates ( $S_2^*$ ) using the Chao 2 index on data from Area 1. The index was applied separately to replicate samples and to station averages.  $S_{obs}$  is the number of observed species. A) Species richness based on replicate samples ignoring stations. B) Species richness based on station average data.

Species richness based on replicate samples

Biotope	Number of Samples	$S_{obs}$	Species Richness $S_2^*$ (+s)	$S_{obs}/S_2^* * 100$	Bottom Type
1	20	30	33.1 (3.4)	91	Coarse Central Channel
2	8	19	-	-	Oyster Reef
3	26	25	-	-	West Channel Margin
4	34	30	38.2 (10.0)	79	Mud Flats
5	12	23	-	-	East Channel Margin

Species richness based on station averages

Biotope	Number of Stations	$S_{obs}$	Species Richness $S_2^*$ (+s)	$S_{obs}/S_2^* * 100$
1	10	30	32.5 (2.5)	92
2	4	19	-	-
3	13	25	-	-
4	17	30	40.1 (10.7)	75
5	6	23	29.0 (7.4)	79

Note: "-" indicates that too few samples were available to estimate species richness

Table 3. Species richness estimates ( $S_2^*$ ) using the Chao 2 index on data from Shelter Island. The index was applied separately to replicate samples and to station averages.  $S_{obs}$  is the number of observed species. A) Species richness based on replicate samples ignoring stations. B) Species richness based on station average data.

Species richness based on replicate samples

Biotope	Number of Samples	$S_{obs}$	Species Richness $S_2$ ( $\pm s$ )	$S_{obs}/S_2$ *100	Bottom Type
CEG	26	115	153 (19)	75	Muddy sand with shell
D	12	75	94.5 (10.5)	79	Shell hash
F	10	49	63.7 (9.6)	77	Sand
B	12	77	93.5 (9.0)	82	Macroalgae and Sand
A	10	62	-	-	Coarse sand and shell

Species richness based on station averages

Biotope	Number of Stations	$S_{obs}$	Species Richness $S_2^*$ ( $\pm s$ )	$S_{obs}/S_2$ *100
CEG	13	112	158 (22)	71
D	6	75	103 (15)	73
F	5	49	63.7 (8.3)	77
B	6	77	102 (13)	75
A	5	62	-	-

Note: Too few samples were available for biotope A to estimate species richness

Figure 1. Examples of species accumulation curves ( $S_{\text{obs}}$ ) and estimated species richness based on the Chao 2 index ( $S_2^*$ ). The  $S_2^*$  curve must level off for the index to yield a valid estimate of species richness. A) For the Area 1 Biotope 1 data set, the Cho 2 index has leveled off. B) For the Area 1 Biotope 3 data set, the Cho 2 index is still increasing. Each curve represents the average of 100 permutations of sample order.

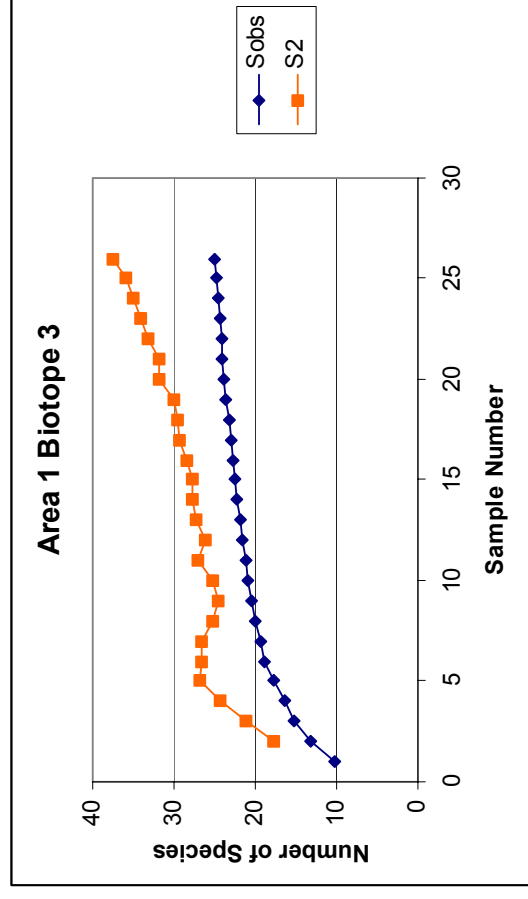
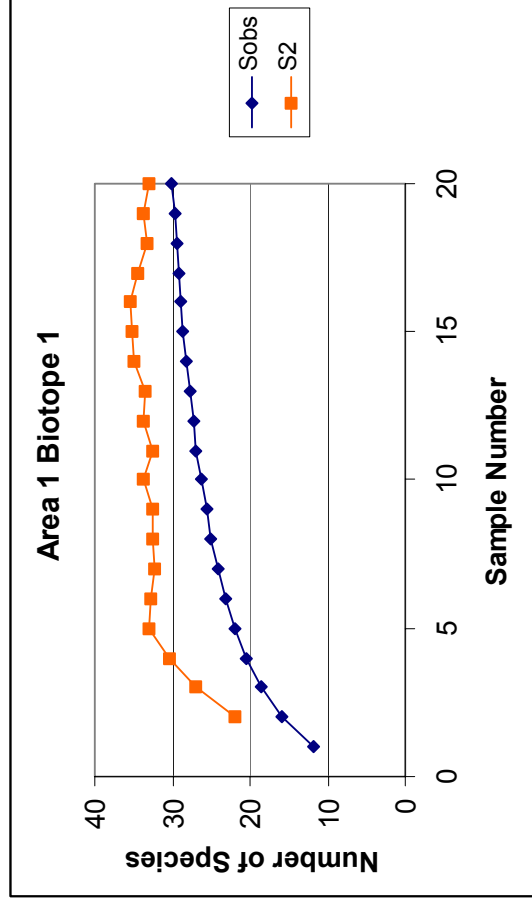


Figure 2. Estimated fraction of species collected versus sample size for Area 3. Each curve represents the average of 100 permutations of sample order. Only biotopes with valid species richness estimates are shown.

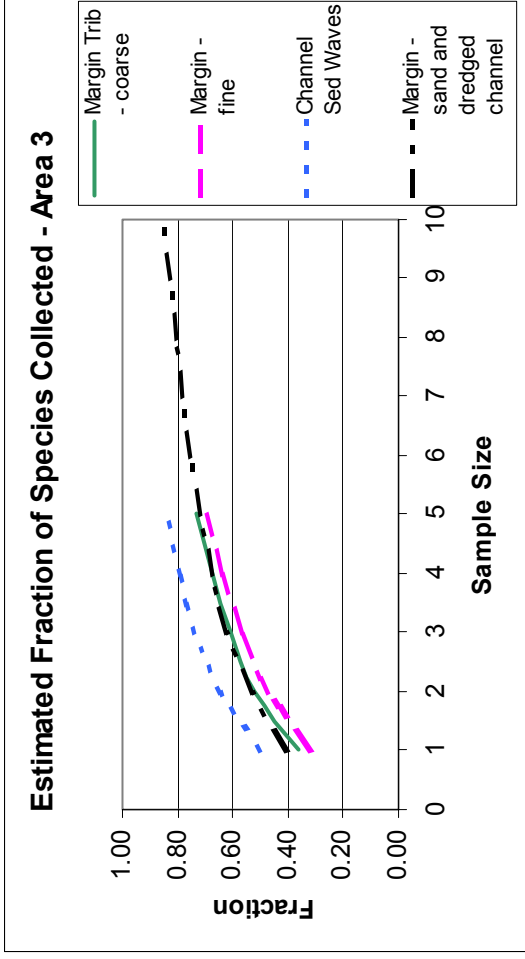


Figure 3. Estimated fraction of species collected versus sample size for Area 1. Each curve represents the average of 100 permutations of sample order. Only biotopes with valid species richness estimates are shown. A) Species richness based on replicate samples ignoring stations. B) Species richness based on station average data.

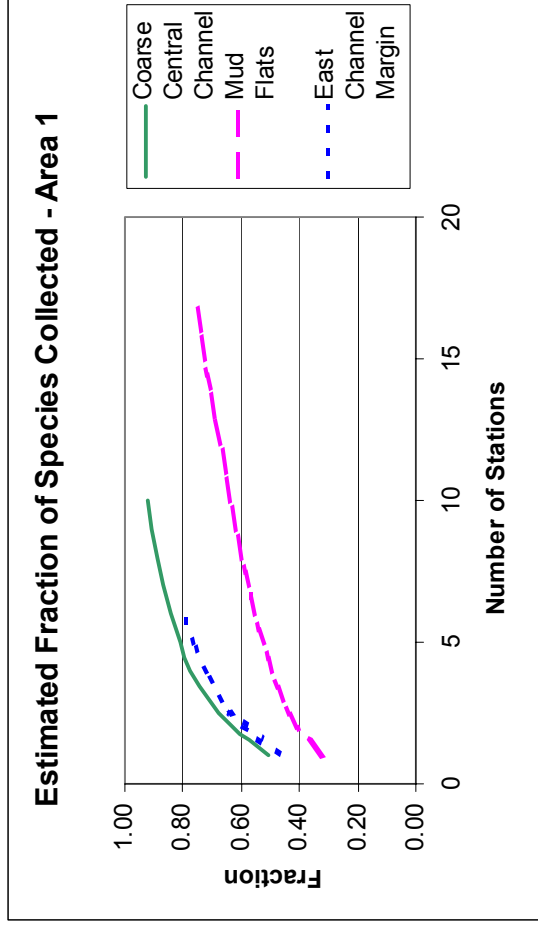
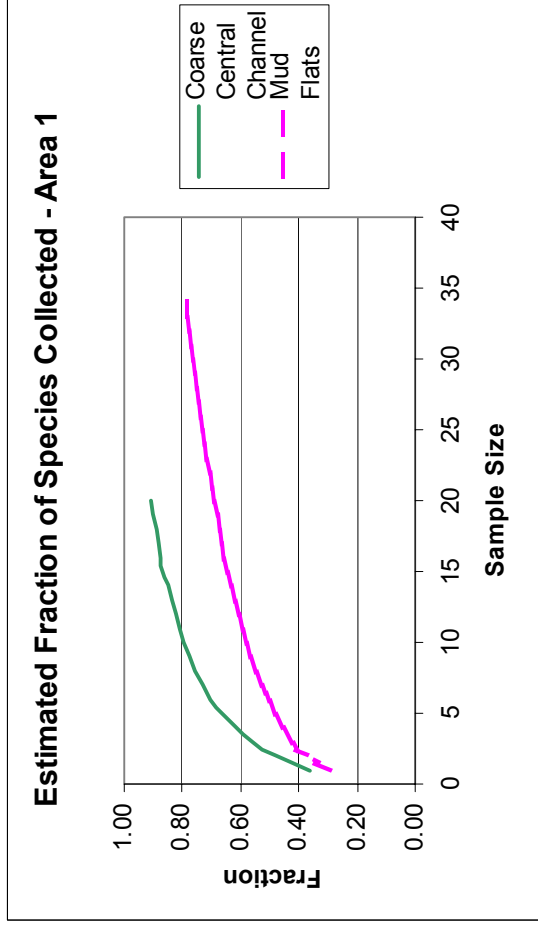


Figure 4. Estimated fraction of species collected versus sample size for Shelter Island. Each curve represents the average of 100 permutations of sample order. Only biotopes with valid species richness estimates are shown. A) Species richness based on replicate samples ignoring stations. B) Species richness based on station average data.

