Analysis of Avian Species Abundance Distribution Using Christmas Bird Count Data

Shuzo Yoshihara
Mathematics/ Environmental Science Major
Senior Thesis
University Of Redlands
May 2000

Abstract

Bird population data is analyzed in this report in different ways, which focus on species abundance distribution, rather than individual species trend or spatial distribution. The population data is derived from the Christmas Bird Count survey from 1969 to 1999 at several Southern California locations. There seems to be an abundance distribution model of bird population. The model is derived from the actual data, and various techniques are applied to test the validity of this model. The main question in this report is whether it is reasonable to assume that the species abundance distribution of individual surveys can be predicted by the model.

1.0 Introduction

2.0 Christmas Bird Count Data

3.0 Tools

4.0 Data

5.0 Definition of Terms

6.0 Survey Variance and Inconsistency

7.0 Species Observations vs. Abundance

8.0 Abundance Rank

9.0 Fraction of most abundant species to total population

10.0 Observation histogram for individual survey

11.0 Conclusion

1.0 Introduction

I started this project as a part of the Salton Sea bird analysis. The Salton Sea provides one of the major remaining wetland habitats for migratory birds on Pacific Flyway. The ecology of the Sea has been changing over the years since it was formed by an accidental Colorado River flood early this century. Some of the main changes, such as rising level of salt content in the Sea water and contaminants from the agricultural draw off, could potentially have a significant negative impact on the avian ecology. It is also possible to look at the past bird survey data to see if any negative impact can be observed as a form of general decrease or sudden shift in population.

Christmas Bird Count data was first used to see if any trend in the population of birds at the Salton Sea can be observed. I found out that aside from some obvious ones, it is hard to determine how the actual population might have changed because of inconsistent survey methodology, mainly the variability of the number of observers. This report is my attempt to use this data for analysis in a way that may not be affected by this weakness of the data. One of the main advantages of this survey data is its size. Survey areas are spread out throughout the United States, and some of the surveys have been conducted for the last hundred years. This large sample size can be advantageous in statistical exercises.

2.0 Christmas Bird Count data

2.1 General background

The Christmas Bird Count is one of the largest and the longest running wildlife surveys in the world. The survey is conducted each year for one full day in late December. The survey areas are 15-mile diameter circles, which are located throughout the United States. Observers count all the birds seen within the survey circle during the day. Observers are mostly volunteers with different skill levels.

2.2 What people have used this data for

Christmas Bird Count data is commonly used for analysis of spatial distribution of specific species. It is effective to visually see broad distribution in the whole United States such as the wintering range of one species. However, a more detailed spatial distribution model is often questionable because of survey inconsistency.

2.3 Inconsistency of data

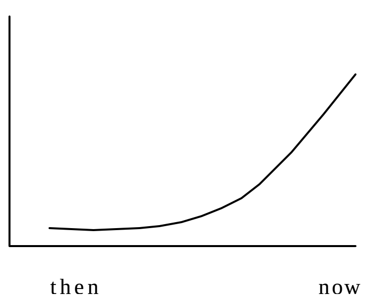
The number and the quality of observers differ in each survey. This makes direct comparison of species population in more than one survey without normalizing unreliable. If a particular species population from one survey is greater than the same species from another survey with different observers, the difference in the observed population might be the result of any one or more of the followings: different observer skill, different numbers of observers and total observation hours, the actual difference in the bird populations, or any other factors that might affect different surveys differently. In general, the quality of observers and the number of participants have increased over the years.

One common attempt to minimize this error is to divide the number of birds observed by the observation hours, which is the sum of the total hours each observer spent in the field. This normalizing approach is based on the assumption of a linear relationship between the number observed and the observation hours, which may not be true except for certain cases. For example, suppose there are 2000 geese in the field. In 1 hour, you will probably see close to 2000 geese. In 10 hours, you will still see close to 2000 geese, not 20000 geese. This normalization technique assumes that the observers would count, for example, twice as many birds in twice the length of the observation period, which is not accurate in most cases and thus it should be avoided unless there is a specific reason to believe a linear relationship.

It is possible to develop a more robust model of normalization of the observed population to estimate the actual population by including specific species and habitat parameters as well as observation data. One way to approach this model is to consider a model of the observation hours and the number of birds observed in which it starts out linear with small observation hours and then the rate diminishes and the observed number approaches the asymptote which is the actual observable population. How much and where the rate starts diminishing will depend on species, habitat, and other observation data. In this report, this normalization technique is not discussed in detail because relative abundance analysis takes account of observation hour variance.

2.4 Species population trend

In addition to determining the general ranges of wintering bird species, Christmas Bird Count data is also useful in picking up obvious trends of specific species because of the amount of the data kept since the survey began in the early 1900's. For example, it is easy to see that the numbers of pelicans and cormorants at the Salton Sea have increased dramatically over the years.



This kind of simple analysis does not require complex mathematical skills, and yet is quite effective in some cases.

3.0 Tools

Microsoft Access and Excel are used in this analysis.

4.0 Data

Christmas Bird Count data is supplied by Cornell Lab of Ornithology. The data used in this analysis is mainly from surveys in the Southern California region for the past 30 years.

Los Angeles 68-81, 82-00

Malibu 68-88, 91-00

Oceanside 69-00

Orange coastal 68-00

Palos Verdes 73-00

Salton Sea N 70-83, 86, 88-00

Salton Sea S 69-00

San Diego 68-00

Santa Barbara 68-00

Santa Catalina Island 89-00

Ventura 98-00

5.0 Definition of terms

Survey - One Christmas Bird Count survey conducted in a particular location in a particular year.

Species observation - Counting of one species in one survey. The number of observations in one survey is equal to the number of species observed in that survey.

Abundance - Number of birds of a particular species counted in a survey

6.0 Survey variance and consistency

The number of a bird species observed in one survey cannot be directly compared to the number of the same bird species observed in a different survey because of the observation inconsistency. This is true for surveys at the same location in different year, and different locations in the same year. The relative population of each species in one survey, however, can be considered relatively variance-free from survey to survey. It is not a wild assumption that the species to species ratio of the observed population to the actual population within one survey is consistent from survey to survey.

For example, survey A and survey B are conducted, in which survey A has twice as many observers as survey B. 100 song sparrows and 200 coots are observed in survey A, 50 song sparrows and 200 coots are observed in survey B. It might be difficult to determine which survey area has the higher actual population of either species, but it can be said that the number of coots relative to the number of song sparrow in survey B is likely to be higher than that of survey A.

Comparing the relative abundance distribution of a survey is similar to the above example. The underlying assumption here is that the effect of the observation variance to each species in one survey is consistent in different surveys. In other words, if there are twice as many coots as song sparrows observed in two different surveys, the ratio of the actual population of the two species in each survey is assumed to be equal, but not necessarily 2 to 1.

7.0 Species Observation vs. Abundance

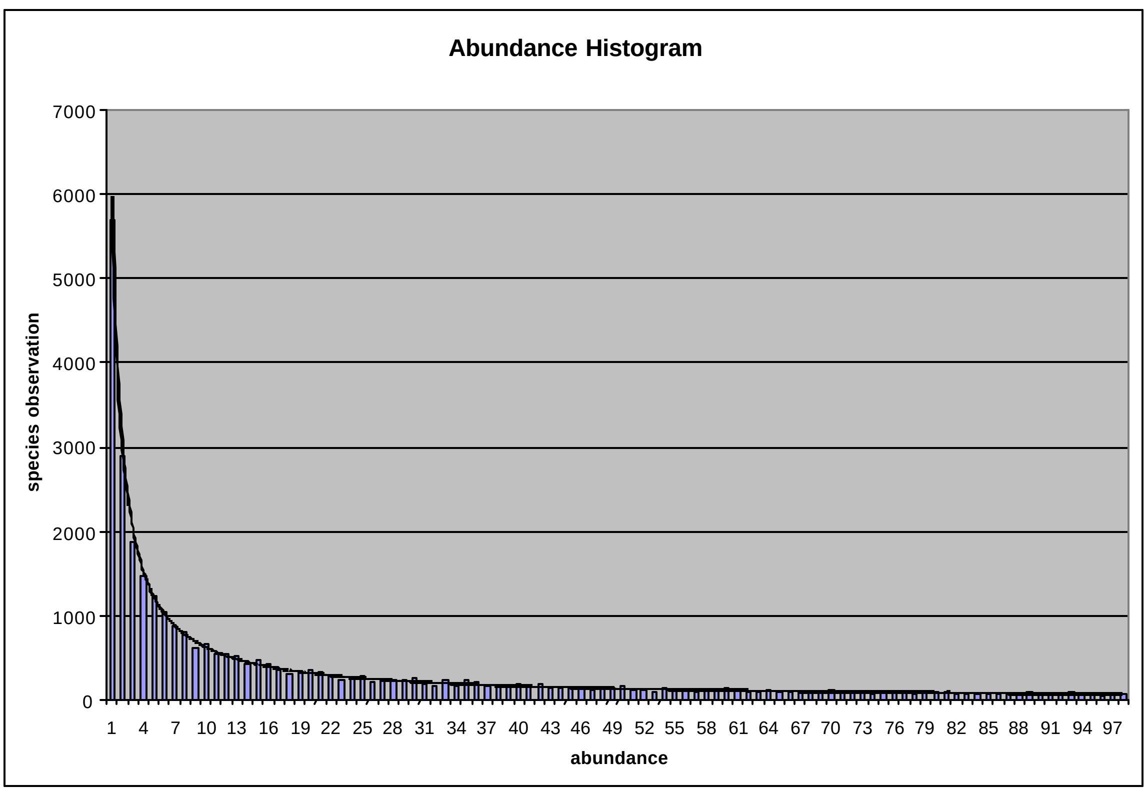
7.1 Abundance Distribution Histogram

The abundance distribution can be plotted as a histogram of the number of species observations at different abundance. The frequency at each bin, or the number observed (abundance), represents the number of species with that abundance of birds observed. The total area of the columns represents the total number of species observations. Note that the abundance goes as high as over 30,000, but only the first 100 is graphed here.

Abundance Distribution Histogram
Trendline fit for 1-100

$$Y = 5984.3x^{-0.9853}$$

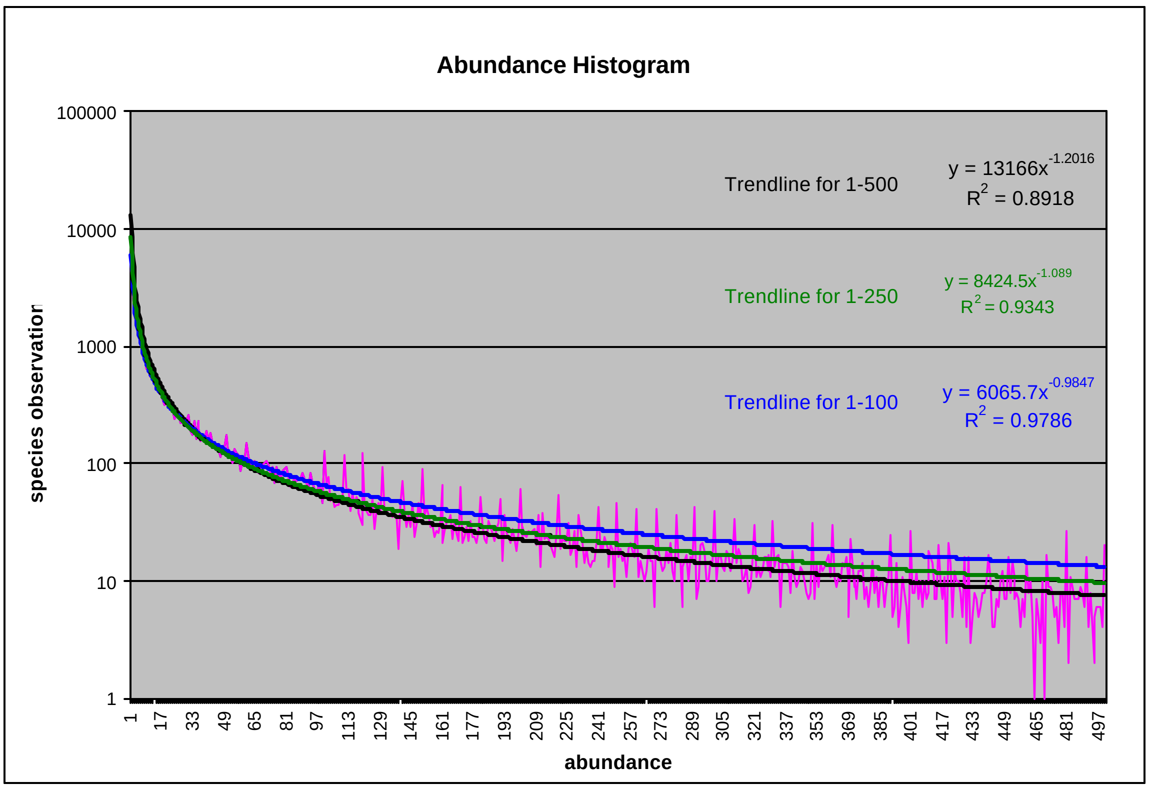
$$R^2 = 0.9856$$



This surprisingly smooth above histogram implies a possible existence of a model which explains the relationship between the number of birds observed and the number of observations in which that particular number of birds are observed. The number of species at a certain abundance is directly related to that abundance. In ecological term, it means that the ratio of the number of species at different abundance remains constant. For example, all the species observed in different surveys are divided into rare, common, and abundant species based on their relative population of each species. The ratios of the number of species in the three groups in each survey are constant. For example, they could be 50% rare, 35% common, and 15% abundant species.

7.2 Trendline analysis

Trendline analysis of this histogram results in a simple power curve of O(x) = A/x, where O is the number of observations, A is the number of instances where 1 individual of that species is observed, and X is the number of individual birds observed. This curve fits extremely well for the lower abundance region, for example, 1-100. However, it does not fit as well for the entire observation histogram, which includes the most abundant species at more than 30,000 individuals observed. The forecast by each trendline tends to overestimate the species observation value as well. The 1-100 trendline forecast is shown in blue, the 1-250 trendline forecast is shown in green in the graph below.



Although this power curve fits extremely well for the lower abundance region, the fit seems to get worse as the range increases. Changes in trendline and its R-square value, which shows the goodness of fit, of different ranges are shown in the table below.

$1-20$
$Y = 5790.6x^{-0.9625}, R^2 = 0.9935$
$1-50$
$Y = 5818.3x^{-0.9668}, R^2 = 0.9860$
$1-100$
$Y = 6065.7x^{-0.9847}, R^2 = 0.9786$
$1-200$
$Y = 7876.6x^{-1.0699}, R^2 = 0.9402$
$1-300$
$Y = 9491.8x^{-1.1214}, R^2 = 0.9209$
$1-400$
$Y = 10885x^{-1.156}, R^2 = 0.9127$
$1-500$
$Y = 13166x^{-1.2016}, R^2 = 0.8918$
$1-600$
$Y = 14080x^{-1.217}, R^2 = 0.8745$

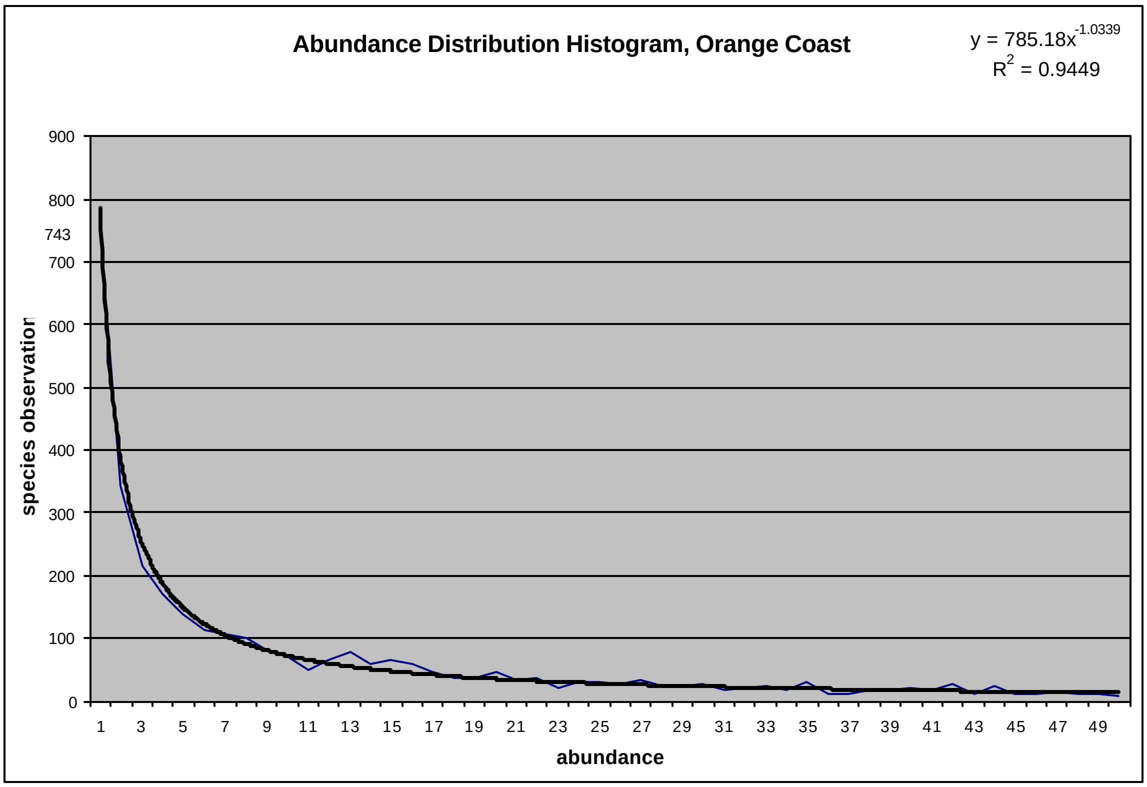
This trend of decreasing fit indicates two things. The model is much more complex than a simple power curve, and the tail, which include the higher abundance region, potentially has a completely different model such as an exponential curve. Also, note that the prediction of this power curve seems to overestimate the actual abundance distribution. The tail of the abundance histogram can be expressed as abundance rank, which will be discussed later in the report.

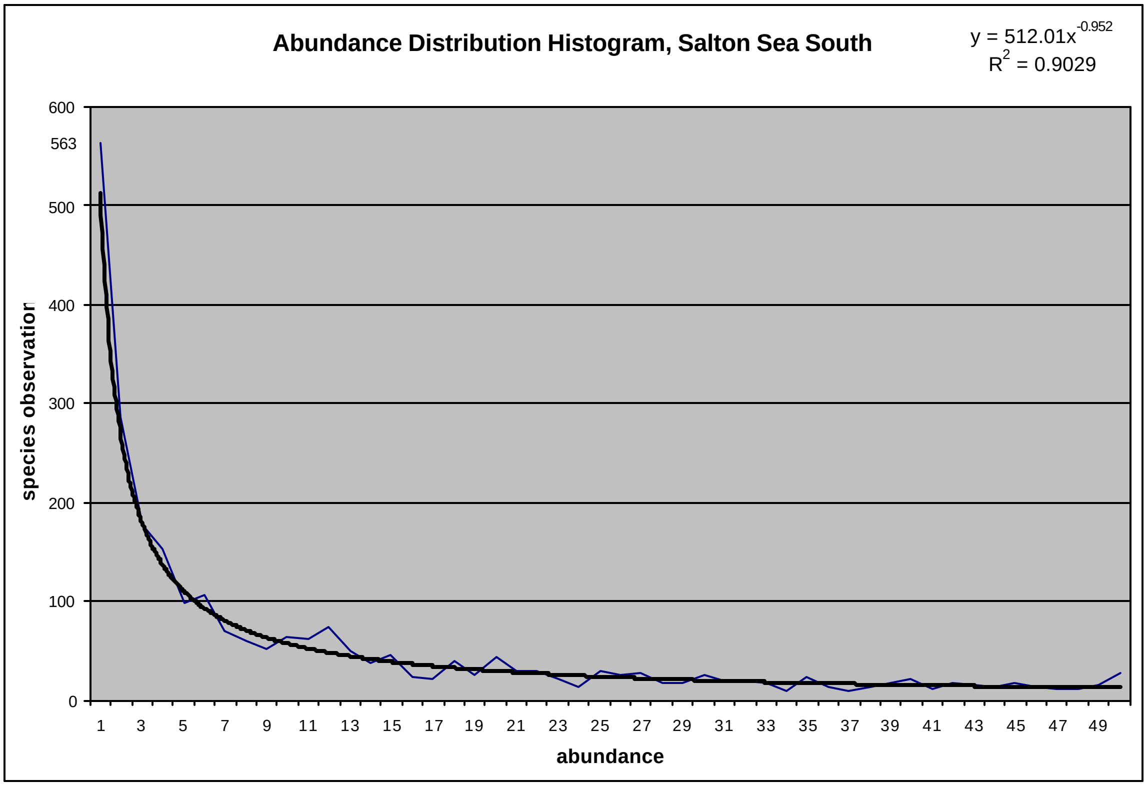
Unfortunately, more complex trendline analysis could not be performed in this report because of the software limitation in Excel. The rest of the report is mainly focused on the lower abundance region, where the power curve fits very well.

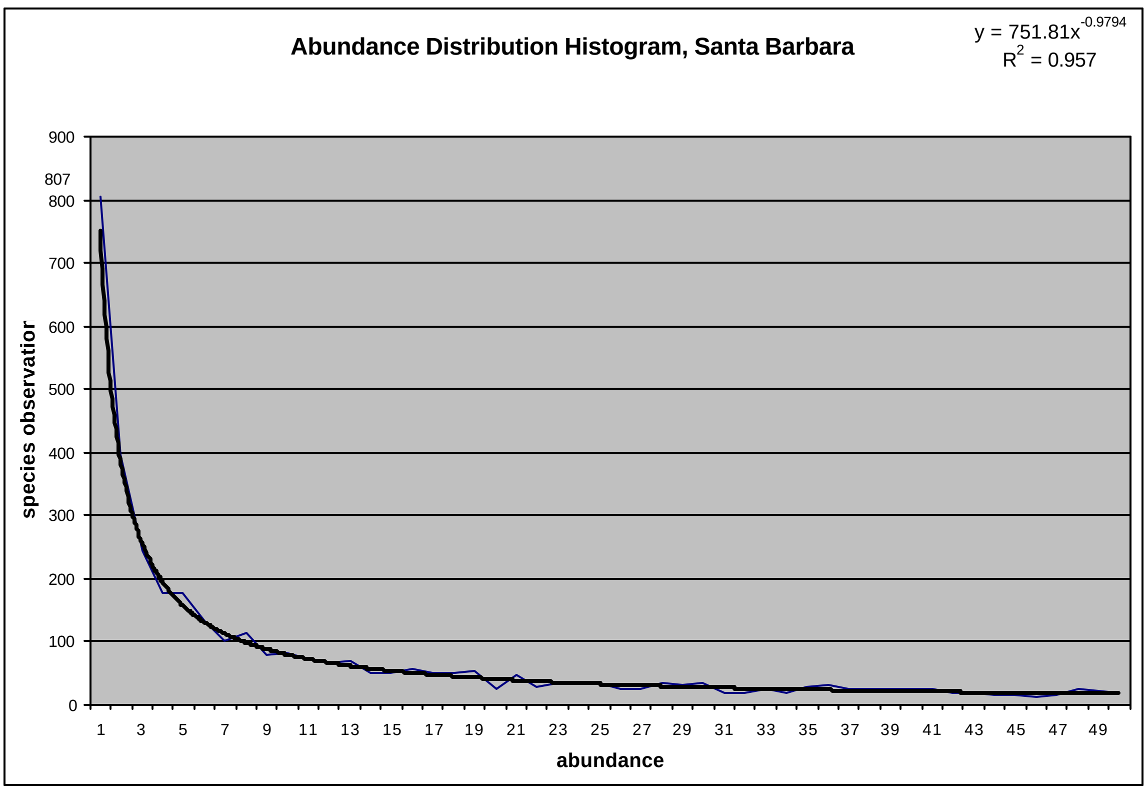
7.3 Location dependence

It is possible that different locations may result in different histograms. By using a simple query in Access, the data are divided into different locations. Then a histogram is created for each location using surveys from the same location.

Each location produces a similar power curve fit to its histogram as the overall histogram. This shows that this model is location independent. There is a small variance in the models from different locations, which is expected.



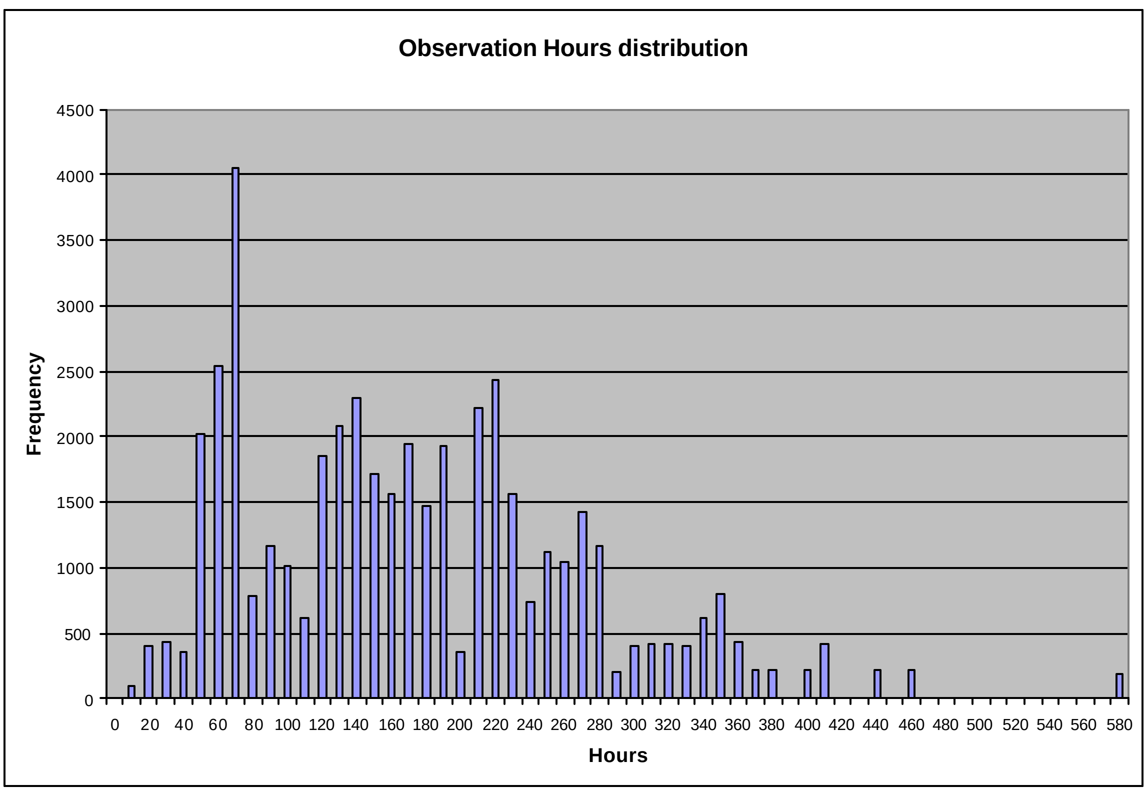

Abundance Distribution Histogram, Salton Sea South
y = 512.01x^-0.952
R^2 = 0.9029
600
563
500
400
300
200
100
0
species observation
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49
abundance


Abundance Distribution Histogram, Santa Barbara
y = 751.81x^-0.9794
R^2 = 0.957
900
807
800
700
600
500
400
300
200
100
0
species observation
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49
abundance

7.4 Observation hour dependence

The number of observation hours is one variable which is likely to affect the histogram. To test this, one histogram with observations from surveys with low observation hours is compared to another histogram from survey with high observation hours.

Although the low observation hour histogram seems to have a slightly higher variance, trendline fit is generally the same for both histograms. This is a significant result because this abundance distribution histogram model is independent of observation hours, one of the main variables in this survey data. It could also be useful in developing methodology for observation normalization.



Abundance distribution histogram
High observation hours

$1-200$

$Y = 2716.3x^{-1.060}, R^2 = 0.9151$

$1-100$

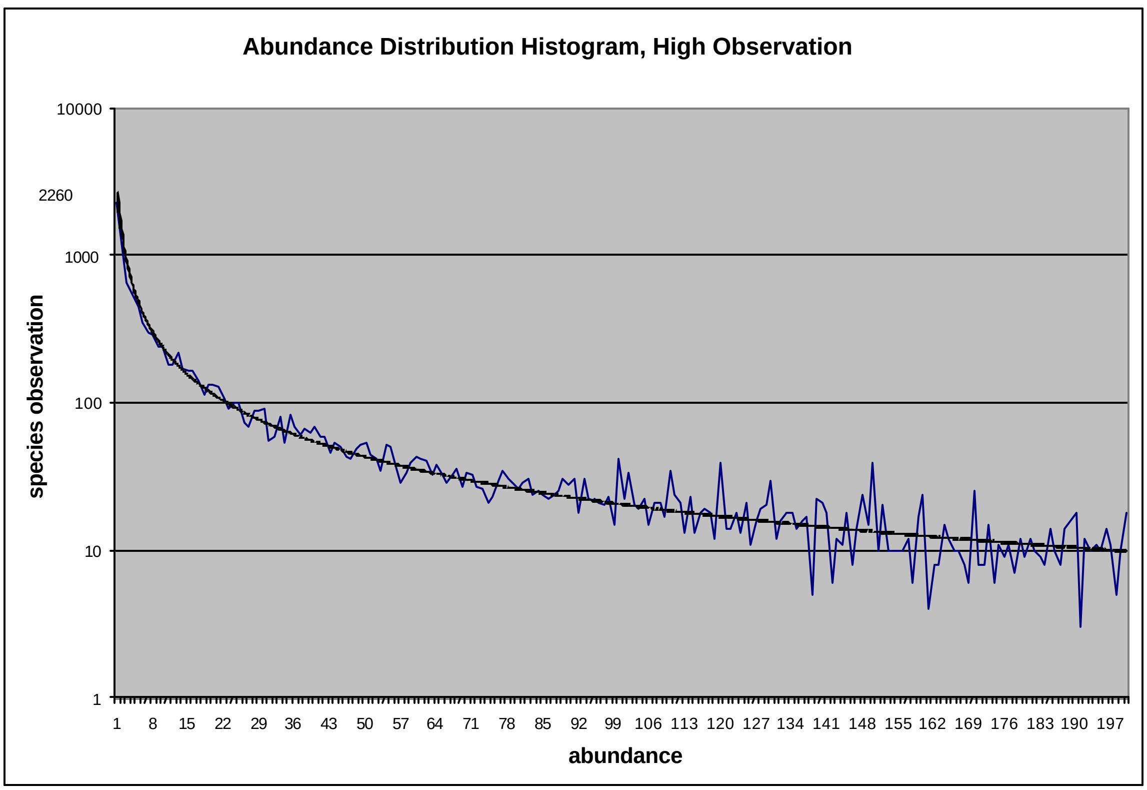
$Y = 2282.0x^{-1.0048}, R^2 = 0.9727$

$1-50$

$Y = 2123.6x^{-.09742}, R^2 = 0.9803$

$1-20$

$Y = 2019.0x^{-.09443}, R^2 = 0.9861$



Low observation hours

$1-200$

$Y = 2829.6x^{-1.1539}, R^2 = 0.8622$

$1-100$

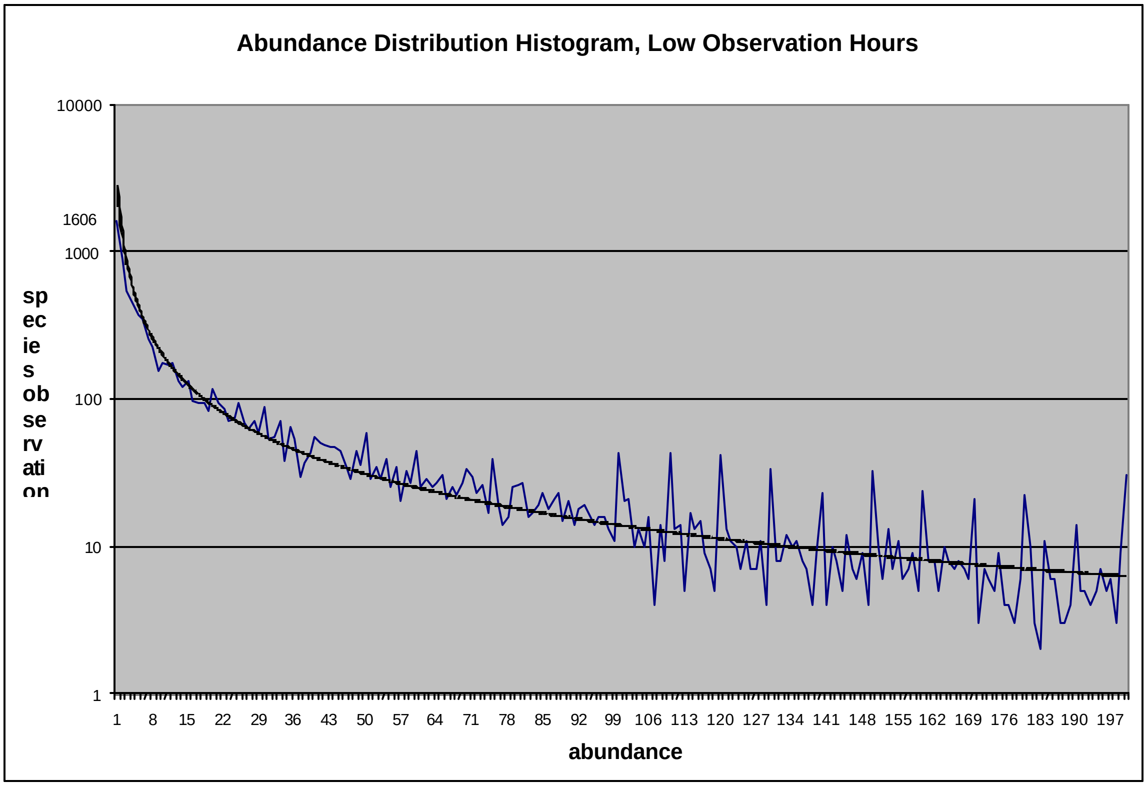
$Y = 1881.0x^{-1.0197}, R^2 = 0.9479$

$1-50$

$Y = 1667.3x^{-.09697}, R^2 = 0.9605$

$1-20$

$Y = 1702.6x^{-.09820}, R^2 = 0.9784$



7.5 Variance analysis

The difference between the actual value and the expected value from the trendline power curve is measured. For this analysis, the curve is assumed to be y = A/x, where y value is the number of species observations at a certain abundance and x value is that abundance. Thus the expected value of the product of x and y is a constant. If the curve accurately describes the abundance distribution model, the error is expected to be normally distributed, and also consistent throughout the abundance region where the curve is fitted. The abundance region of 1-100 is used in this analysis. The error is calculated by the following equation:

1) $O_i = \frac{A}{x_i} + e_o$, where A = constant, O_i = the number of species observations at abundance i, x_i = abundance i, *and* e_o = random error.

This error represented by e_o is not linear. The equation can be modified to calculate the error linearly so the expected distribution is normal.

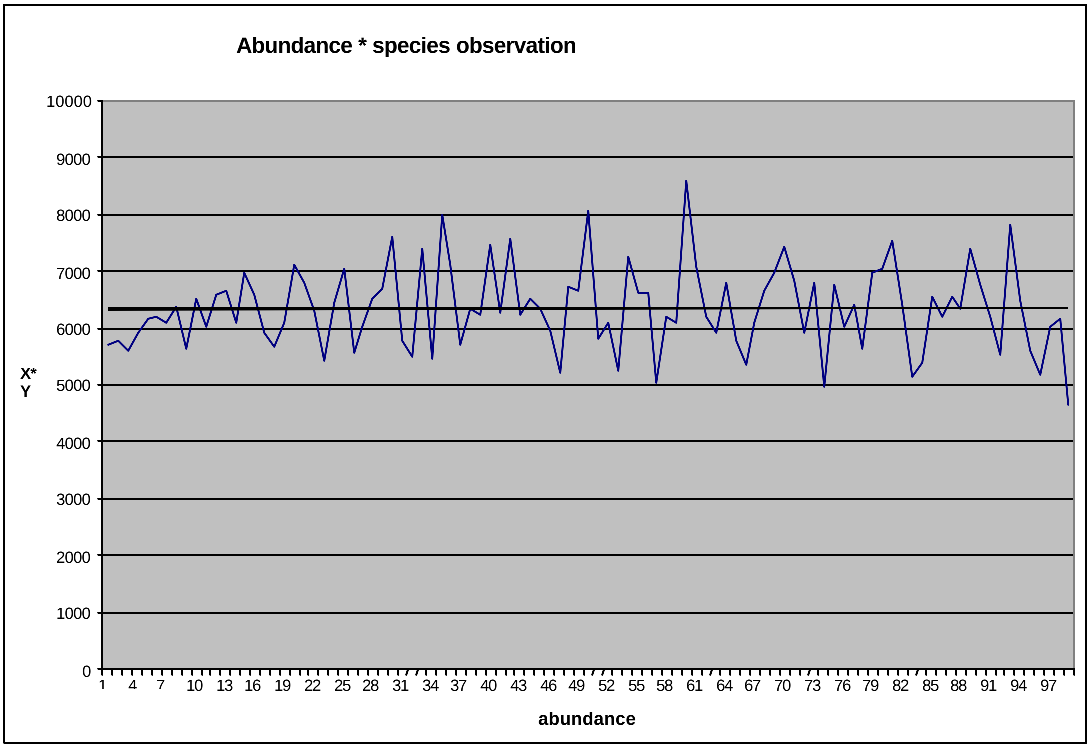
2) $A = O_i x_i + e_i$, where $A = average(O_i x_i)$

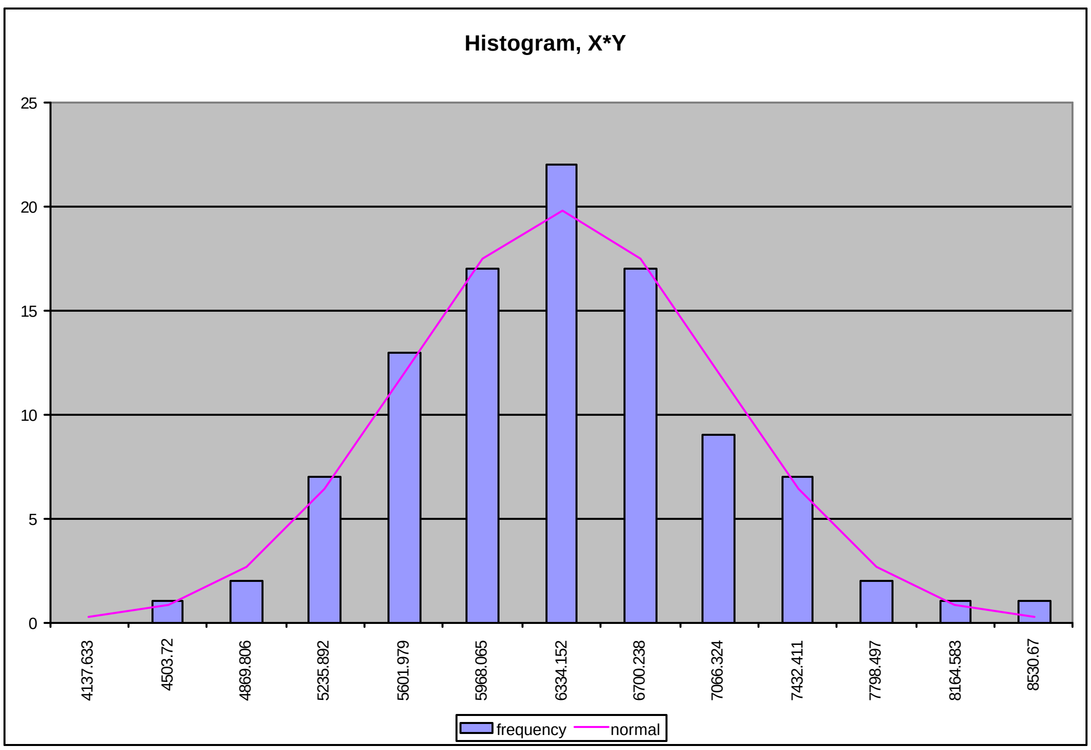
The graph in the following page is $O_i x_i$ with the horizontal line representing A. The normal curve in the histogram is generated with the mean and the variance of $O_i x_i$.

Mean = 6334

Standard Deviation = 732

The error is normally distributed as expected, which supports that the validity of the model. Also, the error does not seem to be skewed to one side of abundance. It is possible to measure the skewness by homoschedasity test.

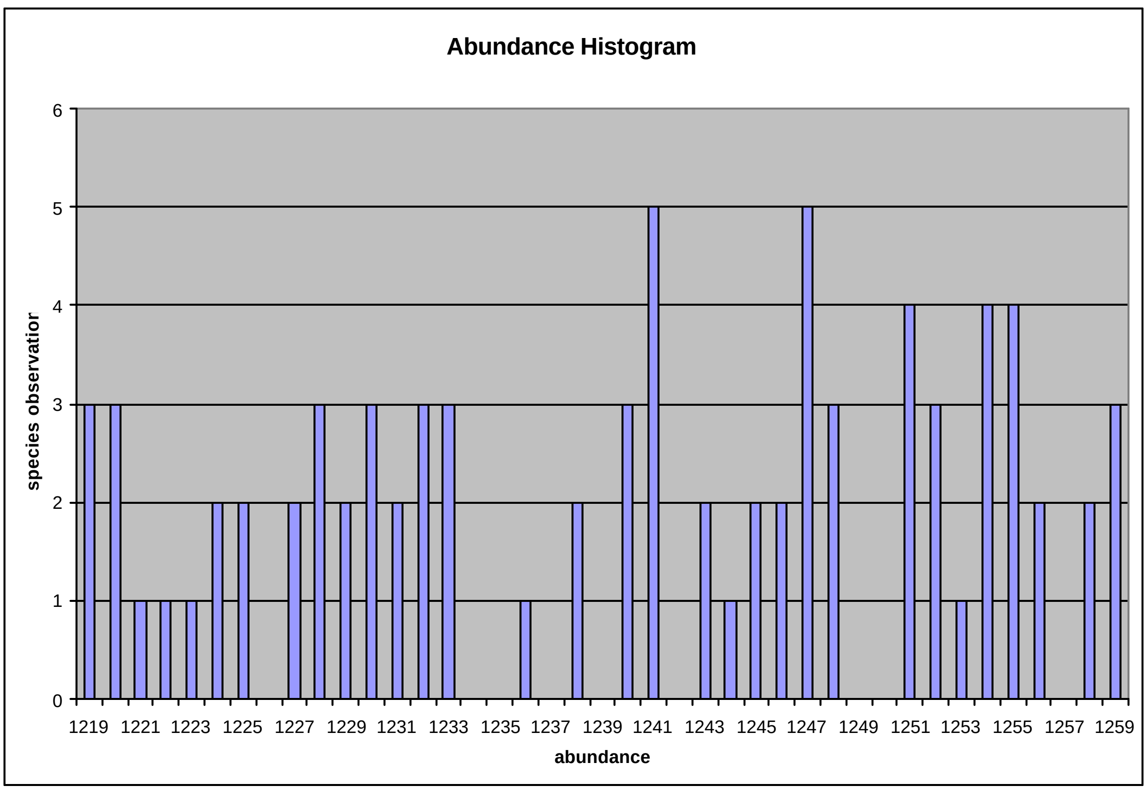

Abundance * species observation
X* Y
10000
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
1 4 7 10 13 16 19 22 25 28 31 34 37 40 43 46 49 52 55 58 61 64 67 70 73 76 79 82 85 88 91 94 97
abundance


Histogram, X*Y
25
20
15
10
5
0
4137.633
4503.72
4869.806
5235.892
5601.979
5968.065
6334.152
6700.238
7066.324
7432.411
7798.497
8164.583
8530.67
frequency
normal

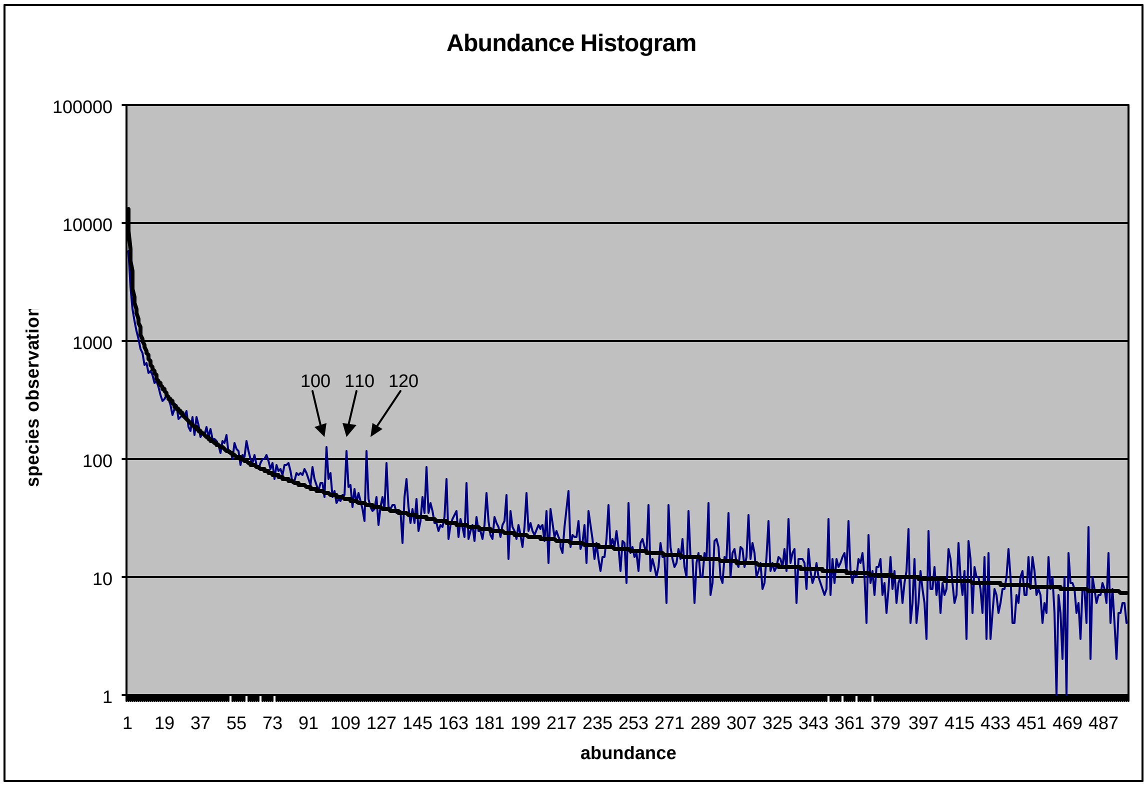
7.6 Other minor problems

This section turned out to be of a very minor impact on the whole report. I have included this because of some interesting statistical techniques.

Although there are more than 40,000 data points in this histogram, there is still a problem caused by lack of data. At a very high abundance, there are certain bins in the histogram which have frequency of zero. For example, there are no instances when 1235 individuals of any species are observed. This gap in the histogram prevents trendline analysis because Excel does not plot trendline of certain curves if there are any data points which are zero.



Also, there is one unnatural variance in this histogram which can be problematic when testing for a goodness of fit. The numbers over one hundred tend to be rounded off to a multiple of 10 by the observers. There are far more species observations of 200 individuals than 199 individuals, for example. This affects the goodness of fit and the variance of the curve. This trend can be seen from a graph easily.



Data smoothing

There are several ways to fix these problems and both involve smoothing of this data.

Wider bin

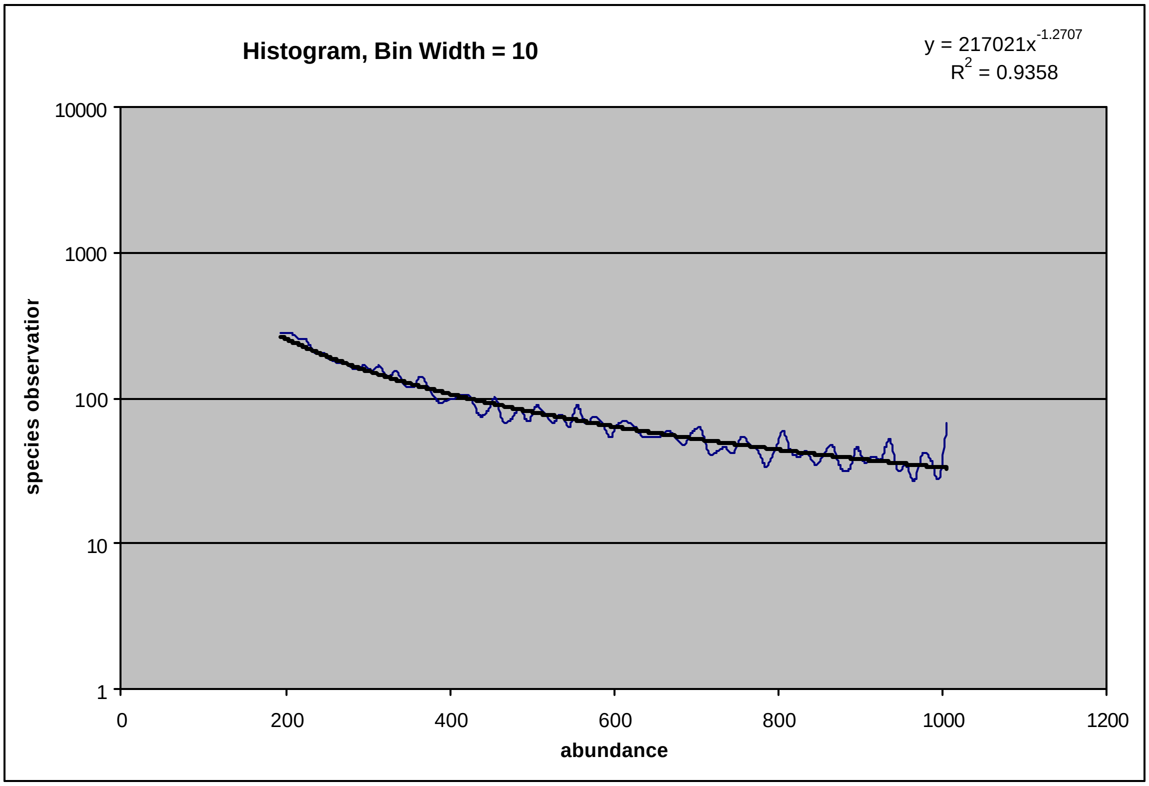
The first one is to increase the bin width of the histogram. This will prevent the first problem as long as the bin is wide enough so that any range of consecutive zeros is less than the bin width. It is usually the case that the gap becomes wider towards the more abundant region so variable bin width can be effective as well. This approach works well for the second problem if the bin is centered at the "nice" numbers.

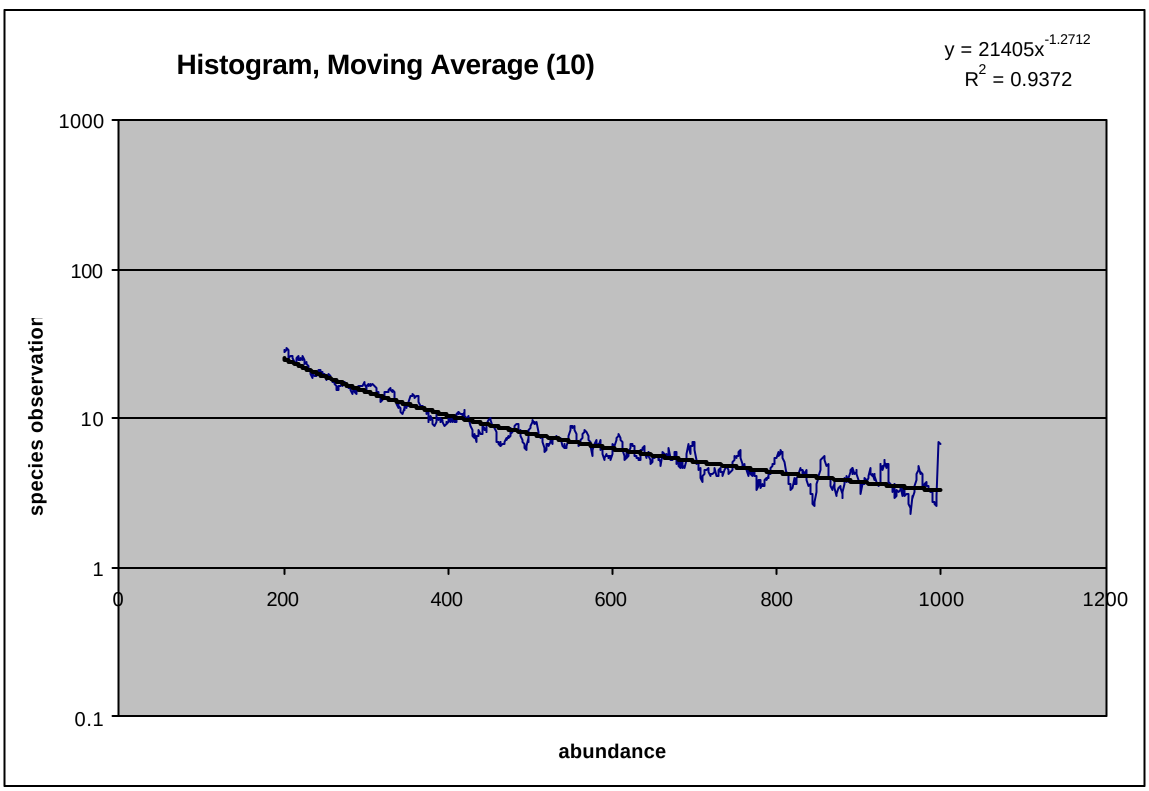
Moving average

Another method is called moving average. This approach is used for smoothing natural variance by actually changing the data. Each column is replaced by the average of a specified amount of neighboring columns. This method works well for the problem because any zero in the histogram is a result of natural variance. This however does not handle the second problem as well especially if the range is small because it is somewhat predictable error. It also assumes a linear curve within the range, which may not be true at the lower abundance region and if the range is too wide. The moving average formula is shown below.

$$MA(O_i) = \frac{\sum_{i-n}^{i+n} O_k}{2n+1}$$, range = $2n+1$

The results of both methods are similar.

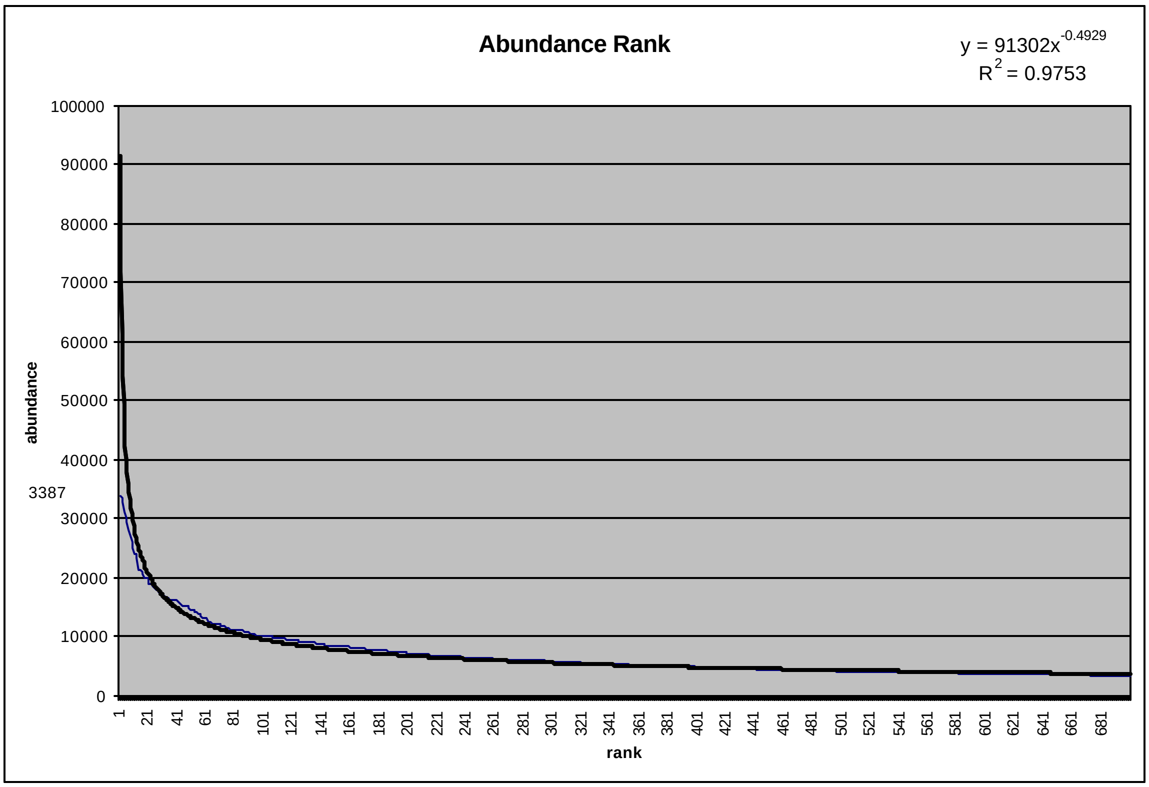

Histogram, Bin Width = 10
y = 217021x-1.2707
R2 = 0.9358
10000
1000
100
10
1
0
200
400
600
800
1000
1200
species observatior
abundance


Histogram, Moving Average (10)
y = 21405x-1.2712
R2 = 0.9372
1000
100
10
1
0.1
0
200
400
600
800
1000
1200
species observatior
abundance

8.0 Abundance Rank

8.1 Basic model

The abundance rank graph can be considered as the inverse of the abundance histogram. The number of birds observed is the independent variable in this graph, and it is plotted in order from the highest. This graph is useful to analyze the higher abundance region which is essentially the tail of the observation histogram.



The power curve does not fit quite well although this graph looks very similar to the observation histogram. The fit of the curve to the data is very bad especially at the beginning. This could be the result of high variance in the counting of very abundant birds or the physical limit of the number of birds that can be counted in one day.

*While I was writing this, I realized that this abundance rank might work better in individual surveys (is this approach valid in the sum of observations from different surveys?) but I have not tried this. The abundance rank of an individual survey can potentially result in a quite different curve than the overall abundance rank.

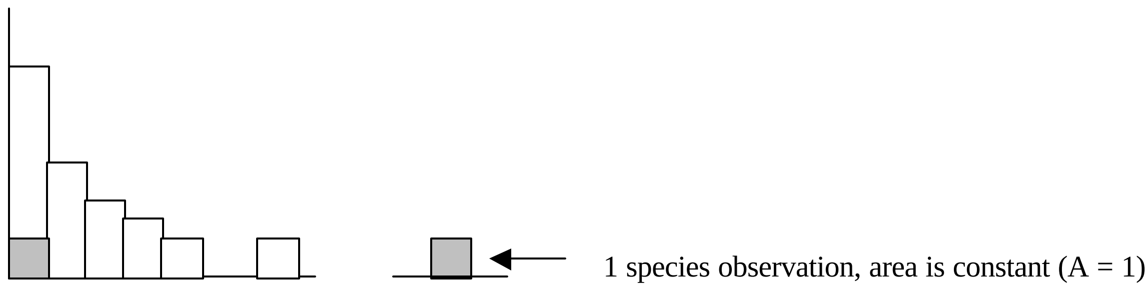
Trendline analysis similar to the one on the abundance histogram is performed. The actual fit seems to be worse than what the R^2 value indicates, especially at the beginning end of the graph. The negative power of the curve increases as the range increases, which is the opposite result to the abundance histogram. This is expected because of the inverse relationship between the abundance histogram and the abundance rank.

$1-50$
$Y = 43774x^{-0.272}, R^2 = 0.9546$
$1-150$
$Y = 56155x^{-0.3674}, R^2 = 0.9628$
$1-550$
$Y = 83567x^{-0.4731}, R^2 = 0.9735$
$1-1050$
$Y = 106566x^{-0.525}, R^2 = 0.9783$

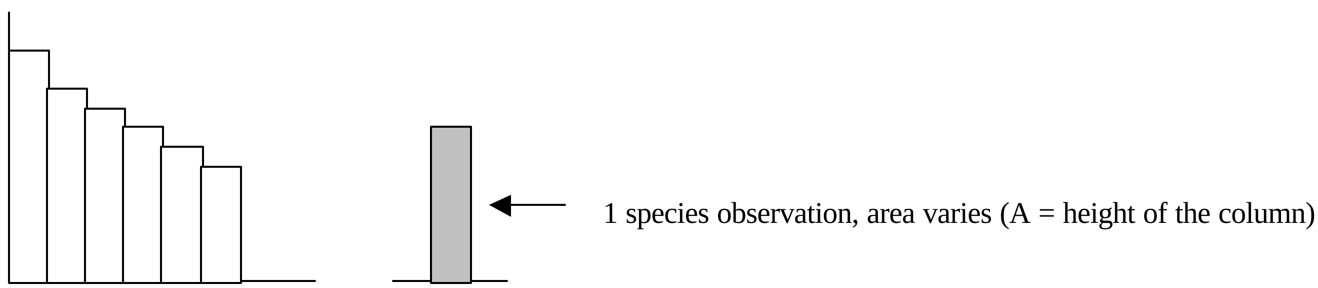
$1-5050$
$Y = 339158x^{-0.722}, R^2 = 0.9673$
$1-10050$
$Y = 1\times10^6 x^{-0.9203}, R^2 = 0.9402$
$1-20050$
$Y = 2\times10^7 x^{-1.2426}, R^2 = 0.9152$
$1-30050$
$Y = 2\times10^8 x^{-1.5531}, R^2 = 0.8889$

This abundance rank graph is not exactly the inverse of the observation histogram because it does not measure the number of occurrences of each abundance. The tail of this graph shows multiple columns with the same height, which corresponds to the frequency of the abundance in the histogram. Furthermore, while the area of one column in the observation histogram directly corresponds to the number of species observations, one observation in the abundance rank graph is represented by the area of each column, which represents the abundance and is not constant. If this is the cause of the bad fit, it can be fixed by dividing each column width by its height so the area of each column is always one.

Abundance histogram

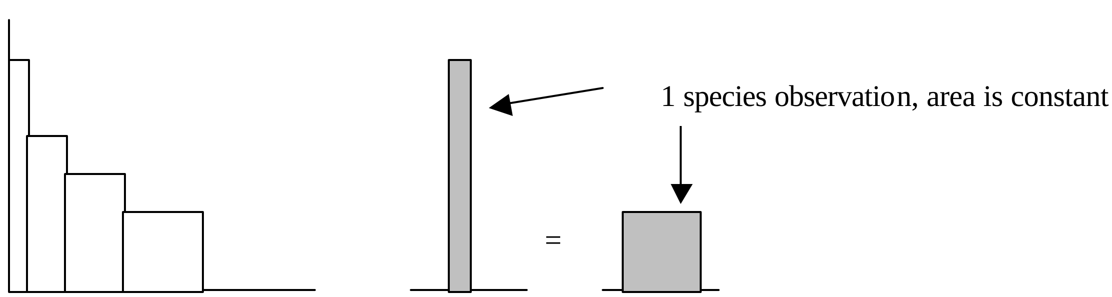


Abundance rank



Adjusted abundance rank

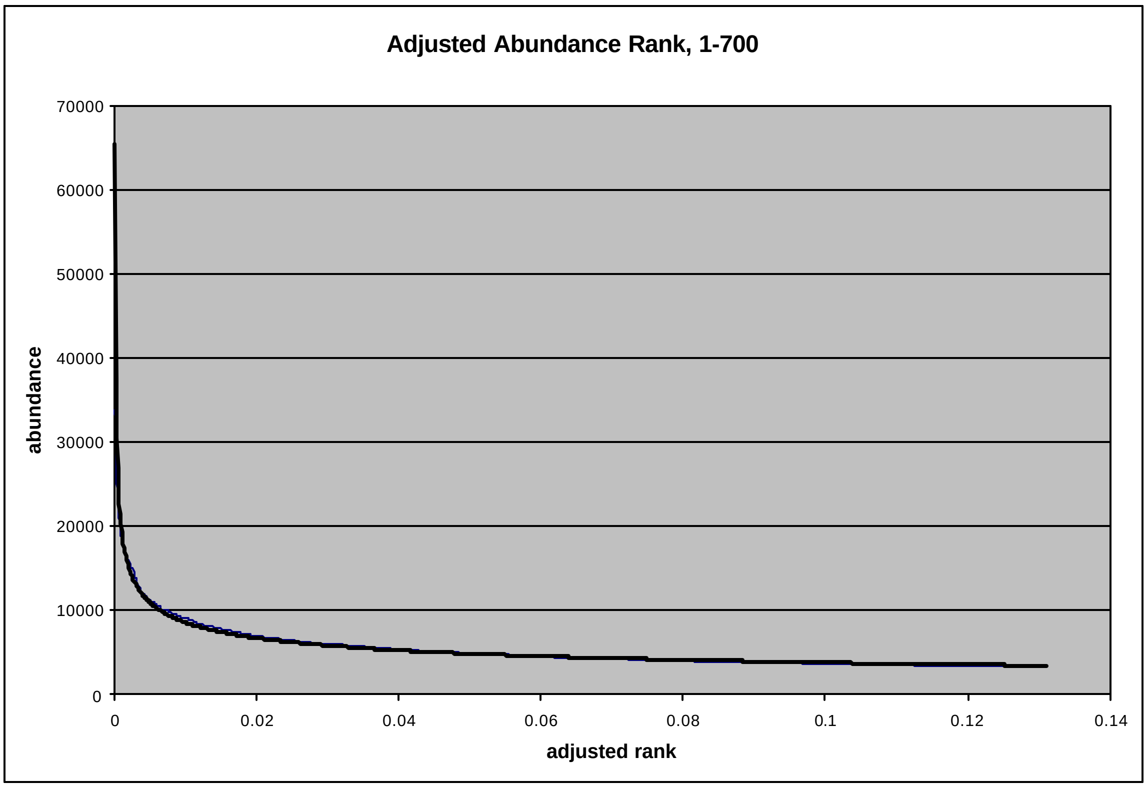
$dx'_i = \dfrac{dx}{H_i}$, where dx = width of the column before = 1, H_i = ith abundance, dx'_i = adjusted width of the column at ith abundance



The width of the column is the smallest at the highest abundance. The area of each column which represents 1 species observation is now 1.

8.2 Adjusted width abundance rank

This graph shows considerably better fit to a power curve. As in the observation histogram, the power curve fits the best for the left side of the graph, and the trendline fit becomes worse with more data points included although the fit is still very good. This result in trendline behavior is similar to the histogram indicating a similar model.



$1-50$

$Y = 3803.8x^{-0.2297}, R^2 = 0.9712$

$1-150$

$Y = 2455.4x^{-0.288}, R^2 = 0.9802$

$1-550$

$Y = 1796.6x^{-0.3402}, R^2 = 0.9887$

$1-1050$

$Y = 1620x^{-0.3621}, R^2 = 0.9918$

$1-5050$

$Y = 1268.6x^{-0.4468}, R^2 = 0.9871$

$1-10050$

$Y = 1155.5x^{-0.5268}, R^2 = 0.9778$

$1-20050$

$Y = 1171.4x^{-0.626}, R^2 = 0.9779$

$1-30050$

$Y = 1304.1x^{-0.7041}, R^2 = 0.9767$

There is also a slightly different way of approaching trendline analysis (something that I should have done for all of them but I ran out of time). Instead of fitting a curve from the left most point to a different abundant point, different curves are fitted to different abundant regions. In this way, the curve which fits the whole data can be considered as a composite of several curves each representing the best fit for that interval.

This approach results in a much better curve fit. R-square is very high indicating that each curve is optimized for each interval with different power. It seems that the larger interval suggests better R-square value as well, possibly because of less magnification of data variance. One interesting result of this analysis is the decreasing trend of the power as the interval shifts to the right. The power seems to be approaching -1, which is in agreement with the abundance histogram where the power is -1 at the lowest abundant region.

$$1-20$$
$$Y = 5079.4x^{-0.1961}, R^2 = 0.9115$$
$$21-50$$
$$Y = 3682.2x^{-0.2332}, R^2 = 0.9848$$
$$51-100$$
$$Y = 1587x^{-0.3689}, R^2 = 0.9806$$
$$101-200$$
$$Y = 1785.3x^{-0.3515}, R^2 = 0.9968$$
$$201-500$$
$$Y = 1501.7x^{-0.3951}, R^2 = 0.9982$$

$$501-1000$$
$$Y = 1472.9x^{-0.4027}, R^2 = 0.9981$$
$$1001-5000$$
$$Y = 1302.6x^{-0.5249}, R^2 = 0.9980$$
$$5001-10000$$
$$Y = 1650.9x^{-0.7119}, R^2 = 0.9998$$
$$10001-30000$$
$$Y = 2974.3x^{-0.8748}, R^2 = 0.9973$$

This decreasing power curve solves the bad fit of both abundance histogram and abundance rank, which are approximately inverses of each other. In both cases, constant power curve overestimates the actual species observation frequency at high abundance. Decreasing power lowers the curve to adjust to the lower frequency at high abundance.

8.3 Curve of the best fit

For both the abundance histogram and the abundance rank, the curve can be expressed as the following:

Abundance histogram

1) $$Y = \frac{A}{x^{w(x)}}$$

Abundance rank

2) $$Y = \frac{B}{x^{1/w(x)}}$$

where $w(x)$ is continuously increasing toward the high abundance region

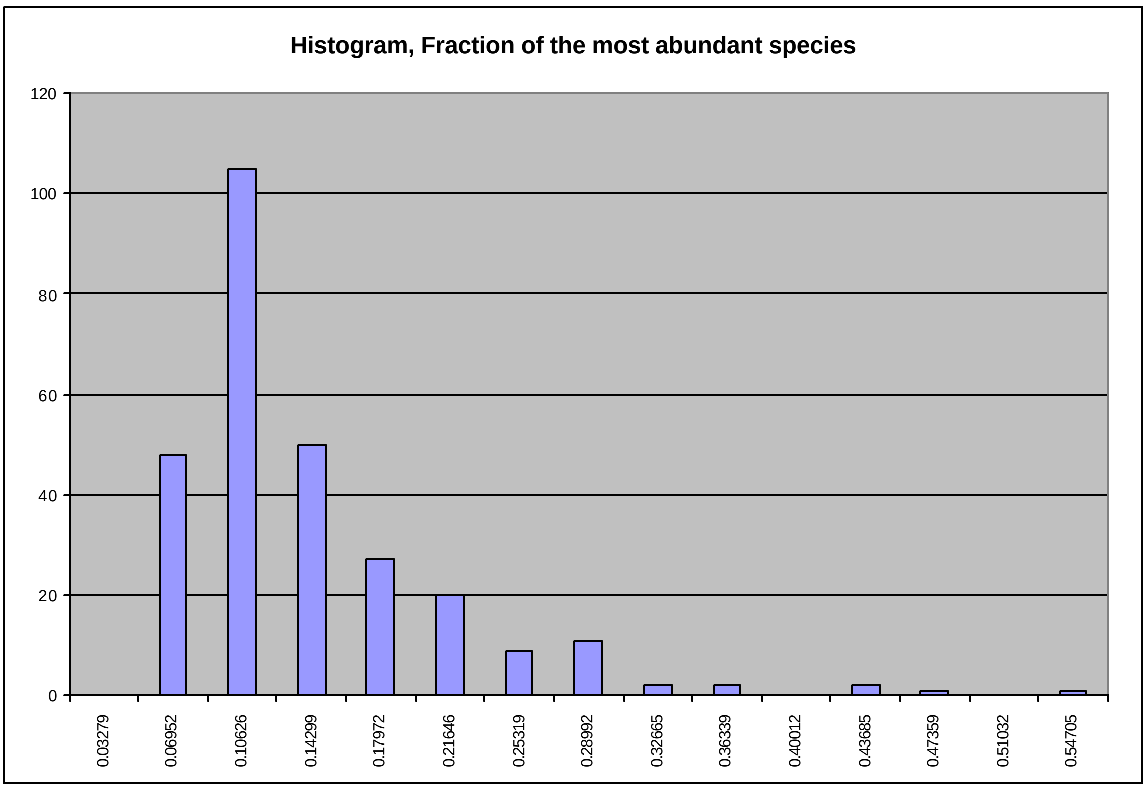
$Y = A/x$ is a special case of this where $w(x) = 1$.

9.0 Fraction of most abundant species to total population

9.1 Basic model

So far, most of the statistics analysis used here treat data from different surveys as the same observation data. They do not distinguish between observations from different surveys. It is possible to test if there are any differences between surveys and locations, and one such statistic is the fraction of the population of the most abundant bird to the total population observed in the same survey. It can also be considered as the measure of dominance of one species in a particular survey. The fraction is calculated for each survey by the following equation:

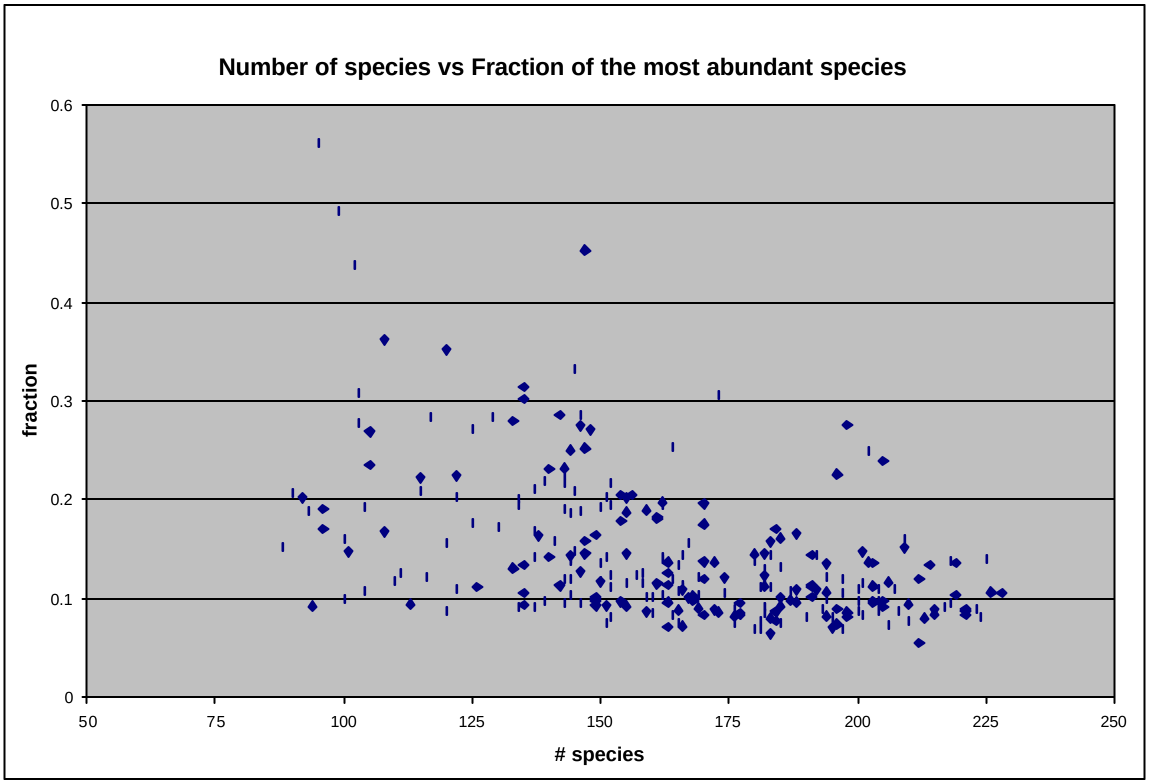
$p_k = \dfrac{\max(x_i)}{\sum x_i}$, where x_i = abundance for species i in survey k.



The data ranges from 0 to 1, and the shape of the curve resembles the beta distribution. This graph seems to have a high variance, which can be considered to be caused by either the difference in the number of species, or by different locations.

9.2 By the number of species

The graph of the number of species against the fraction of the most abundant bird shows no observable trend. The points seem to be spread out evenly, suggesting there is no correlation between these two.

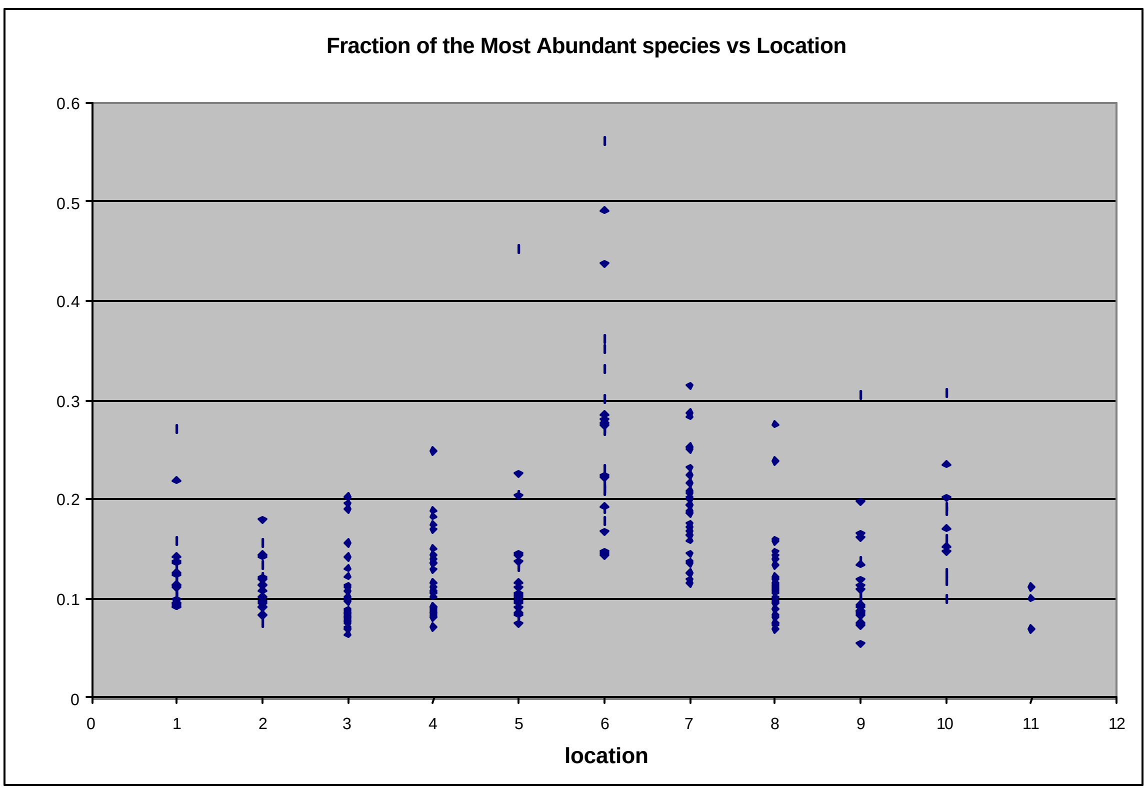


9.3 By location

Both Salton Sea survey locations appear to be significantly different from other locations with higher means and higher variances than others. This may be caused by the large ecological scale of the Salton Sea. The Salton Sea provides larger wetlands habitats than other locations, resulting in very large number of particular species which find the Salton Sea attractive. All locations seem to have a similar distribution pattern with a peak around .12 and gradually decreasing toward higher fraction.

One interesting feature of the graph is that most of the locations have a few surveys with a much higher fraction than normal. This may be linked to a phenomenon similar to population explosion which seems to occur periodically in particular species. A quick look at the graph of fraction against time of some of the locations shows that peaks are spread out, and variance seems to be consistent over the years. Time series analysis could be effective for this as well.

Location	Name	Mean	Standard Dev
1	Los Angeles	0.122	0.038
2	Malibu	0.112	0.024
3	Oceanside	0.107	0.041
4	Orange Coast	0.121	0.040
5	Palos Verdes	0.134	0.078
6	Salton Sea (north)	0.269	0.103
7	Salton Sea (south)	0.193	0.051
8	San Diego	0.119	0.047
9	Santa Barbara	0.109	0.048
10	Santa Catalina	0.175	0.057
11	Ventura	0.094	0.022
	Overall	0.143	0.073



10.0 Observation histogram for individual survey

A bigger question for the observation histogram is the validity of the model for the individual surveys. It seems that the model exists for the sum of all the surveys, but whether the same model holds true for individual surveys is another question. Individual surveys do not have as obvious a trendline fit as the overall histogram because of the smaller sample size. For this reason, it is very difficult to determine the goodness of fit to this model by looking at each survey one by one. Some data might fit well, while some might not, regardless of the validity of the model. One way to test this is to measure the distribution of the goodness of fit of each survey. The goodness of fit is expected to have random error if the model is accurate.

(Now I am not sure if this makes sense. Distribution of goodness of fit? What I mean here is if it is possible to categorize each survey into different goodness of fit, such as negative really bad, negative bad, negative ok, good, positive ok, positive bad, positive really bad, then this is normally distributed. Well, goodness of fit is absolute value, but consider this: there are two surveys. Both are tested against the average of the two. Both of them should have the same goodness of fit (in effect, they cancel each other's goodness of fit), but one is positive and the other is negative. Perhaps I do not mean goodness of fit, but some other statistic like mean of differences for each column...)

10.1 Abundance density histogram

The overall histogram is essentially the sum of all the surveys. Each column of the histogram is divided by the total number of surveys, and the resulting histogram represents the distribution of the average survey. Furthermore, its total area is normalized to 1, and it can be called normal abundance density histogram. The tail of the histogram is cut off at the highest number of birds observed. Each column represents a fraction of the total number of species observed to the expected number of species at that abundance. Expected number of species for each abundance in a survey can be obtained by multiplying the column by the total number of species observed in the survey. This average abundance density histogram is the expected distribution for a certain number of species observed.

<u>Average abundance histogram</u>

1) $$\alpha_i = \left(\frac{Y_i}{N}\right)$$

where α_i = average species observation at abundance *i per survey*, Y_i = species observation at abundance *i*, N = the total number of surveys.

<u>Normal abundance density histogram</u>

2) $$d_i = \frac{\alpha_i}{\sum \alpha_j}$$

where d_i = species observation density at abundance *i*, $\sum \alpha_j$ = the average total number of species per survey.

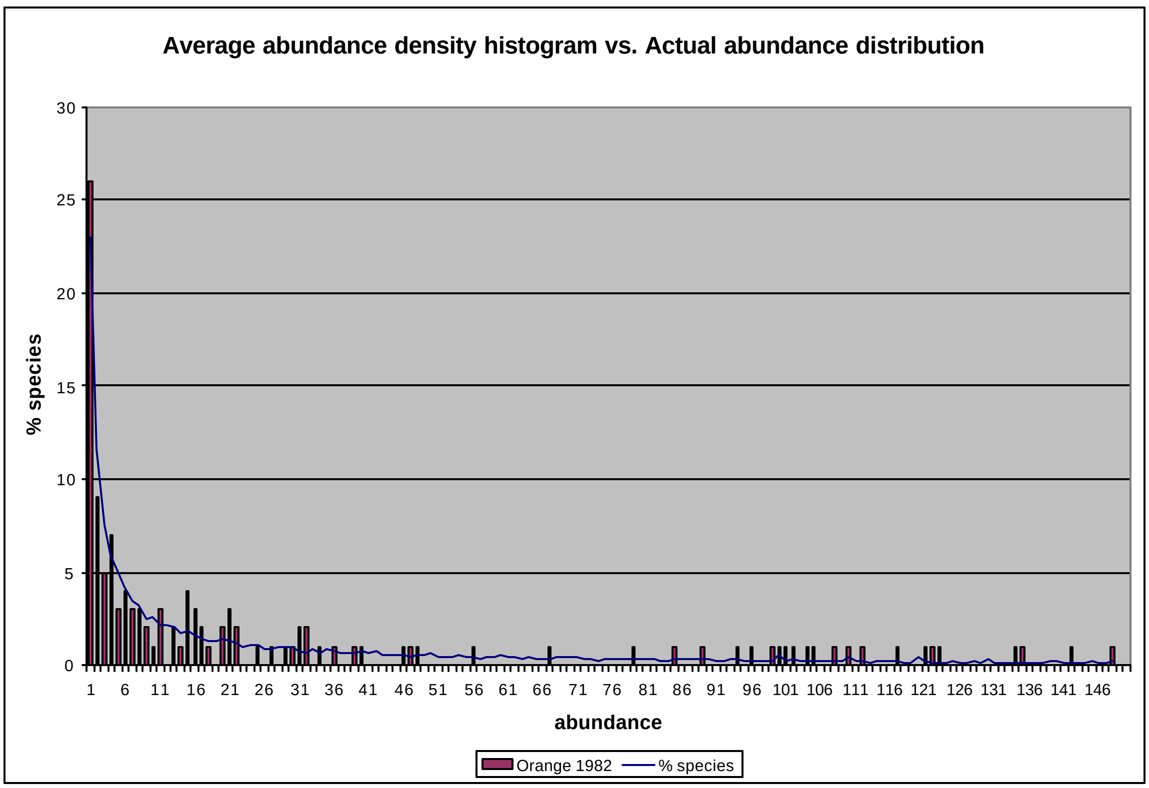
<u>Abundance density histogram</u>

3) $$a_i = d_i M$$

where a_i = the expected species observation at abundance *i* when there are *M* many total number of species.

The sum of d_i is equal to 1 and the sum of a_i is equal to the total number of species, M. The average abundance density histogram is the expected abundance distribution of a survey with a particular number of species. The abundance density histogram represents the expected abundance distribution of a survey with the total number of species M. However, this histogram looks significantly different from distribution of any one survey because the expected species observation at each abundance is the probability that any one observation will fall onto that abundance.

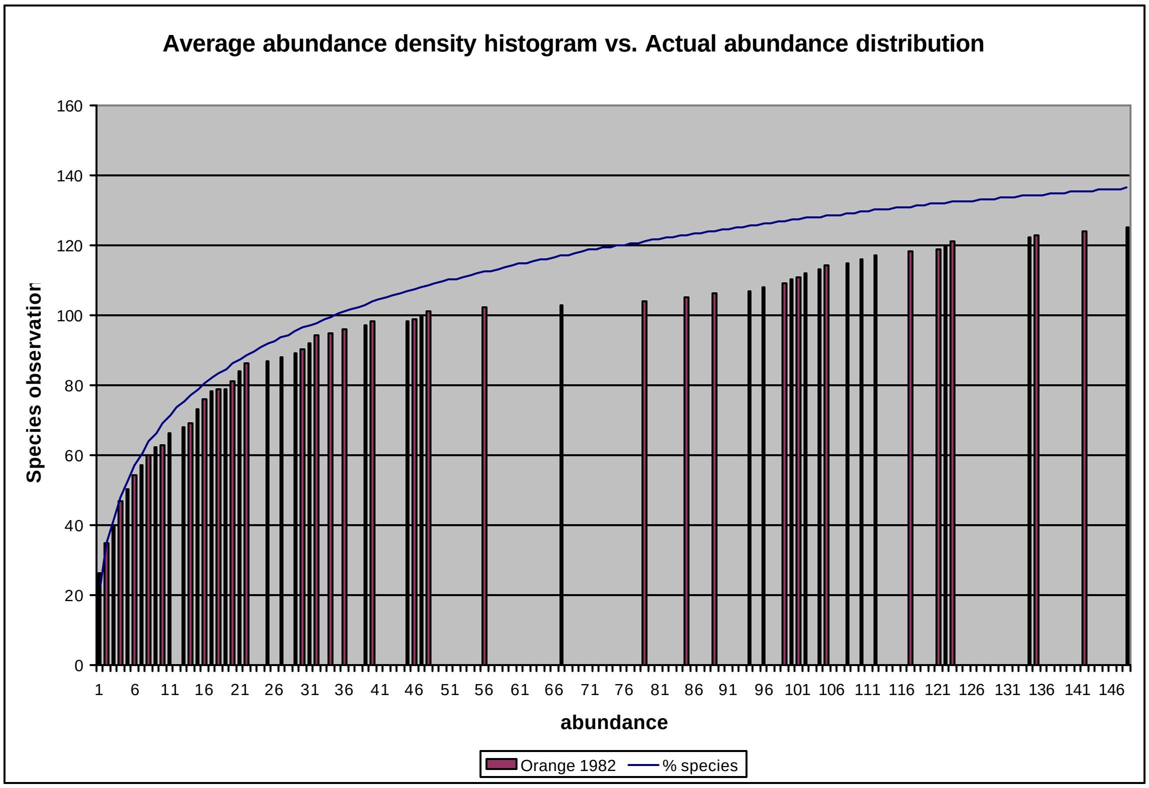
Example: Orange coast 1982



The goodness of fit of individual surveys cannot be directly tested to the average abundance density histogram because the actual data are discrete while the histogram is continuous. Each column in the histogram represents the expected number of species for that abundance, which might not be an integer. The total area of the average histogram and the individual survey histogram both represents the number of species so the areas are equal. One way to measure the goodness of fit of a specific abundance is to compare the area of all the columns less than that abundance.

10.2 Cumulative histogram

The width of the columns in the histogram is always 1, so the area can be represented as cumulative histogram. In this histogram, the frequency at each abundance represents the number of species with less than that abundance observed. The curve is always increasing and approaching the total number of species. The expected species abundance curve and the actual survey data curve will meet at the highest abundance, where the number of cumulative species observations equals to the total number of species.



In the cumulative histogram, it is possible to directly calculate the difference in the frequency at each abundance of a survey from the expected value at that abundance. There are several ways to measure the goodness of fit.

10.3 Chi-square analysis technique

The Chi-square test is used to measure how well the model fits to the data. The statistic is given by:

$$X^2 = \sum_{i=1}^{k} \frac{(n_i - E(n_i))^2}{E(n_i)} = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}$$

In this test,

k = the number of species of different abundance

n_i = number of species at *i*th abundance

n = total number of species ($n = \sum^{k} n_i$)

p_i = the abundance density at *i*th abundance

$E(n_i)$ = expected number of species at *i*th abundance

*i*th abundance = abundance of the survey at *i*th rank

I have not finished the actual test of either method because of my limited programming skill in Access and Excel. There are more than 250 surveys and each test is sufficiently complicated that generating histogram and testing each survey without help of some automation by programming is extremely long and tedious.

The goodness of fit statistic is expected to have a Chi-square distribution if each individual survey actually comes from the model because random errors in surveys are normally distributed. The distribution of the goodness of fit statistic of all the sample surveys can then be tested to see whether they have a Chi-square distribution. If this is true, species abundance distribution can be predicted and the result of this abundance distribution can be compared to the expected distribution to see if the particular survey represents something unusual.

Sample analysis is shown in the next page.

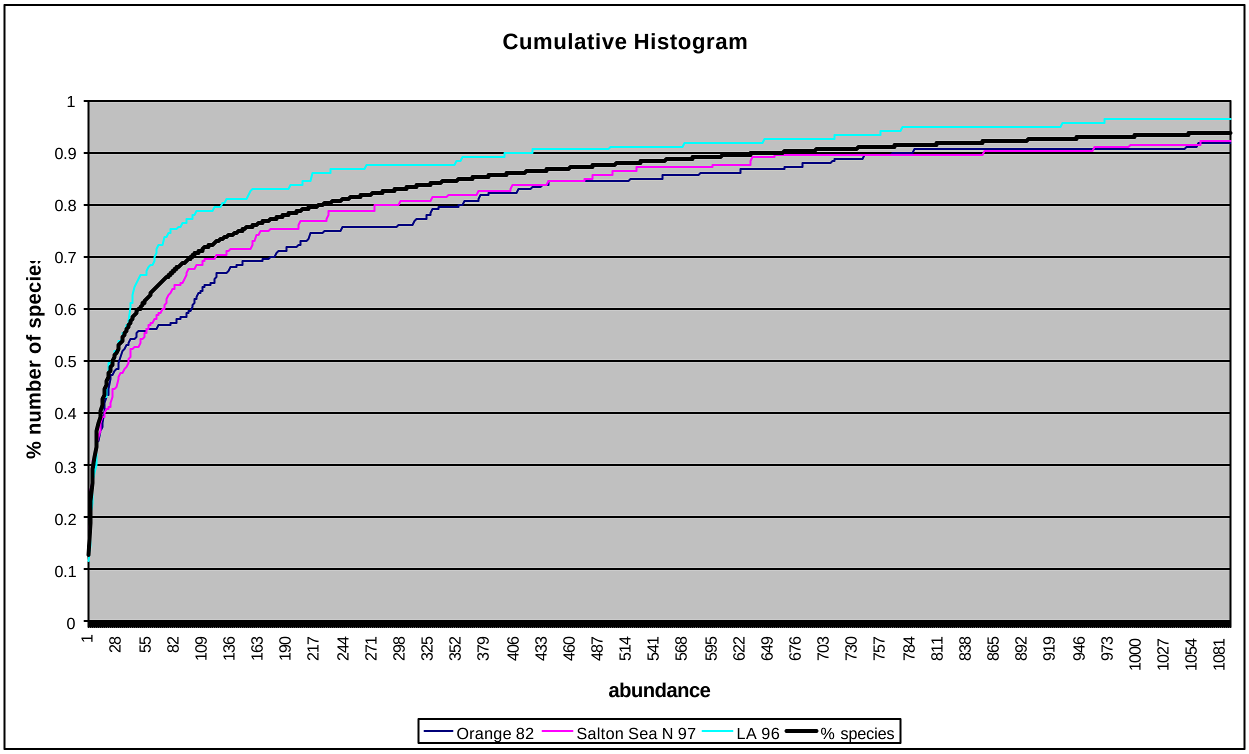
Sample Analysis

The following three surveys are randomly chosen and analyzed.

Orange 1982

Salton Sea North 1997

Los Angeles 1996



The result of Chi-square analysis is shown in the follwing table.

Survey	Chi-square
Orange 82	51.078
Salton Sea N 97	26.461
Los Angeles 96	16.705

Low Chi-square value indicates good fit. It is hard to interpret anything from this result because sample size is so small, but the low Chi-square value of Los Angeles 96 survey seems to disagree with the apparent mediocre fit of the same survey seen in the graph.

10.4 Location analysis

The expected number of species and highest population vary among different habitats. The result of the goodness of fit test can be divided into different locations and tested if the results are different in different location. It is unlikely that different locations are hugely different because all of the locations are in the similar Southern California region, but there might be measurable differences.

Depending on the still unconducted tests of goodness of fit distribution, it is probably reasonable to assume that the abundance distribution of a survey can be described by the same model as the overall abundance histogram, with the addition of location variance and a natural survey variance, which is normally distributed with mean effect of zero. The following is the equation for species abundance distribution for individual surveys:

Abundance distribution(location) = E(abundance distribution) + location variance + natural survey variance
Where E(abundance distribution) is the expected abundance distribution of that particular survey, location variance is constant in the same location, and natural survey variance is normally distributed with mean zero.

E(abundance distribution) in this report seems to be $Y = \frac{A}{x^{w(x)}}$. This could look significantly different in other regions such as in the mountains, where the number of birds and species variety are much lower in general during winter.

12.0 Conclusion

The existence of the species abundance distribution model is implied by a very good power curve fit to the overall abundance histogram, which includes observations from all the surveys. Different power curve fits in the higher abundance region and the lower abundance region indicate that they seem to be explained by different models. Using both the abundance histogram and the adjusted abundance rank allows examining both ends of abundance easier. The analysis of the model is not complete due to its complexity.

One of the major advantages of this model is that it is relatively unaffected by survey to survey observation hour variance. This is significant because it allows direct comparison of species abundance distribution in different surveys without worrying about observation difference. This is not possible in spatial distribution or trend analysis without proper normalization of observation variance from survey to survey.

It also seems that the model can be applied to individual surveys as well as a collection of different surveys. Goodness of fit distribution analysis, which has not been done because of lack of more powerful programming at this point, tests the variance of the abundance distribution model fit of different surveys. This is expected to have a normal distribution if the model is valid for individual surveys.

The model is somewhat location dependent, as shown in the fraction of the most abundance species analysis. This makes sense considering possibly different species distributions in different habitats.

The most significant result of this analysis is that there is a model that seems to fit all the surveys. This implies that the relative abundance species distribution is consistent from survey to survey, at least in similar habitat in Southern California. In other words, the ratio of relatively rare, common, and abundant species in any given survey in the region is consistent.