Catchall
Version 4.0
User Manual

by Linda Woodard, Sean Connolly and John Bunge

Sponsored by NSF Grant #0816638

July, 2013

To cite CatchAll:

Bunge, J., Woodard, L., Böhning, D., Foster, J., Connolly, S., and Allen, H. (2012), "Estimating population diversity with CatchAll." *Bioinformatics* 28; 1045-7. doi: 10.1093/bioinformatics/bts075.

See this paper for a brief account of the operation and statistical theory of the program.

System requirements. There are two types of programs available: the main analysis program in a variety of flavors (CatchAllName*.exe); and an interactive graphics module (CatchAll.display.xlsm) written in Excel 2007, which uses macros that need to be enabled. The graphics module runs only on a Windows platform assuming Excel 2007 or later is installed. (Apple decided not to enable macros in this version of Excel.) The GUI version of the executable (CatchAllGUI.exe) will only run under Windows. There is also a Windows command line version (CatchAllcmdW.exe). The .Net framework must be installed to run either Windows version. In addition there is a command line version (CatchAllcmdL.exe) that will run on the MAC OS and other Linux platforms, provided the appropriate version of Mono has been installed.

Input data. CatchAll is a set of two programs for analyzing data derived from experiments or observations of species abundances or multiple recapture counts. For simplicity we will use the species-abundance terminology throughout this manual, but the same methods can be applied to the total counts (row sums) of recaptures in a multiple-recapture or multiple-list study. The fundamental dataset consists of "frequency counts." This is a list of frequencies of occurrence, followed by the number of species occurring the given number of times in the sample. For example, in the following dataset,

1,295
2,63
3,30
4,6
5,4
6,6
7,1
9,6
11,1
12,2
13,1

14,1
17,1
21,1
25,1
30,1
31,1
55,1
69,1
86,1

there are 295 species with exactly one representative in the sample, called "singletons"; 63 species with two representatives; 30 with 3; ... and then there are some large or very abundant species in the "right tail"; 1 species with 55 representatives; 1 with 69, and 1 (the largest or most abundant) with 86. This structure, with a large number of rare species, and a small number of very abundant species, is typical. The dataset must be in this "comma-delimited" format, frequency,count with filename equal to datasetname.csv or datasetname.txt.

Analysis with the GUI version (CatchAllGUI.exe). To read in the data, start CatchAll (by double-clicking on CatchAll.exe); use the "Locate Input Data" button to navigate to your dataset; and double-click on the appropriate file. CatchAll then displays the first 10 lines of your dataset in a small window for verification; click "OK."

Once the data are loaded, perform the analysis by clicking one of the "Run Program" buttons. After a short time (ranging from <1 sec to < 5 minutes), the "Model Analysis Completed" button will appear; click "OK." N.B. -- the 4 Mixed Exponential Model will take longer to calculate.

A summary of the analysis appears in the "Best Models" window and the "OUTPUT FILES" window displays the pathnames for the files used by the interactive graphics program CatchAll.display.xlsm (see below for details). Other files created by the program are located in the same folder. See Output below for a detailed description of these files.

Analysis with the Windows command line version (CatchAllCmdW.exe). At least two parameters must be supplied to the Windows command line version, the input filename (complete path, if not in same directory as the executable) and the path to the directory where the output files will be written. If no such folder exists, it will be created. See **Output** below for a detailed description of these files. Optionally you can include a flag to have the program calculate the 4 Mixed Exponential Model; the default is to calculate it. N.B. -- the 4 Mixed Exponential Model will take longer to calculate.

Calculate 4 Mixed Exponential Model

```
CatchAllcmdW.exe inputfilename outputpath
```

or

```
CatchAllcmdW.exe inputfilename outputpath 1
```

Don't calculate 4 Mixed Exponential Model

```
CatchAllcmdW.exe inputfilename outputpath 0
```

Analysis with the Linux command line version (CatchAllcmdL.exe). Mono must be invoked to run this executable; at least two parameters must be supplied to the Linux/MAC command line version, the input filename (complete path, if not in same directory as the executable) and the path to the directory where the output files will be written. If no such folder exists, it will be created. See Output below for a detailed description of these files. Optionally you can include a flag to have the program calculate the 4 Mixed Exponential Model; the default is to calculate it. N.B. -- the 4 Mixed Exponential Model will take longer to calculate.

Calculate 4 Mixed Exponential Model

```
mono CatchAllcmdL.exe inputfilename outputpath
```

or

```
mono CatchAllcmdL.exe inputfilename outputpath 1
```

Don't calculate 4 Mixed Exponential Model

```
mono CatchAllcmdL.exe inputfilename outputpath 0
```

Output. Running the analysis program generates a number of files. If you use the GUI version, a folder called "Output" is created in the same directory as the input file. If you use either command line version, these files are put in the folder you designate. This folder contains the following files:

`datasetname_Analysis.csv`
This is a complete listing of all information from all analyses performed by CatchAll. See the section "Statistical Procedures" below for details.

`datasetname_BestModelsAnalysis.csv`
Column-formatted copy of summary analysis output as displayed in the main CatchAll window when using the GUI version. This is to be read into the "Summary Analysis" worksheet in CatchAll.display.xlsm (see below).

`datasetname_BestModelsFits.csv`
Fitted values for the "best models" as selected by the model selection algorithm (see "Statistical Procedures"). This is to be read into the "Best Fits Data" worksheet in CatchAll.display.xlsm (see below).

`datasetname_BubblePlot.csv`
Analysis data to generate the bubble plot display; this is to be read into the "Bubble Graph Data" worksheet in CatchAll.display.xlsm (see below).

Graphical Display. The Microsoft Excel-based module CatchAll.display.xlsm generates four displays. To view these, open CatchAll.display.xlsm by double-clicking. Near the top of the screen click Options > Enable Macros.

1. Summary analysis. This is a copy of the CatchAll output window, formatted for columns. On the worksheet "Summary Analysis," click "Import Summary Analysis" and navigate to the file `datasetname_BestModelsAnalysis.csv`. This copies the CatchAll summary output display to the worksheet, in column-formatted form.

2. Best Fits Color. This is a scatterplot showing the frequency count data as points, and the various fitted models (best, 2a, 2b, 2c) as (curved) lines. On the worksheet "Best Fits Data," click "Import Best Fit Data" and navigate to the file `datasetname_BestModelsFits.csv`. This imports the fitted values from the best selected models, and automatically generates the comparative plot on the "Best Fits Color" worksheet.

3. Bubble Graph Color. This graph shows the behavior of the estimates as τ is increased, i.e., as outliers (large frequencies) are progressively added to the data. Typically the nonparametric estimates diverge as τ increases while the parametric estimates converge. The bubble sizes are proportional to SE/2 in each case; points are not plotted (they are blanked out) if either (i) $\hat{C} > 100c$; or (ii) $SE > 10\hat{C}$. On the worksheet "Bubble Graph Data," click "Insert Bubble Graph Data" and navigate to the file `datasetname_BubblePlot.csv`. This imports the required data and automatically generates the bubble plot on the "Bubble Graph Color" worksheet.

4. Bubble Color No Non-Parametric. This is the bubble plot with the nonparametric sequence deleted, for easier visual comparison of the parametric estimates.

Note that all Excel functions are available under CatchAll.display.xlsm: in particular one can change the scale of axes; alter colors or shapes, delete plotted data sequences, etc., at will. Thus the program gives the user interactive control over the graphical displays. For more information see the appropriate Excel help screens.

Summary of main output display. Here we briefly describe the main output as found either in the GUI output screen, or equivalently in the file `datasetname_BestModelsAnalysis.csv`.

Total number of observed species: self-explanatory. *Model*: the fitted models are described in "Statistical Foundations" below (henceforth SF). *Tau*: the upper-frequency cutoff. Analysis is based on the frequency counts up to τ; the remaining counts are added *ex post facto* (see SF). *Observed Sp*: the number of species (counts) with frequencies up to τ only. *Estimated total Sp*: the final estimate of the total number of species in the population (including those with sample frequencies > τ). *SE*: standard error of the preceding estimate. *Lower CB, Upper CB*: lower and upper 95% confidence bounds. Note that the confidence interval is asymmetric, see SF. *GOF0*: a raw or naïve Pearson goodness-of-fit *p*-value. *GOF5*: Pearson goodness-of-fit *p*-value with adjacent cells concatenated to achieve minimum expected cell count of 5 (for asymptotic approximation); see SF.

Best Parm Model; Parm Model 2a, 2b, 2c. These are the parametric models (and choices of τ) selected as optimal by CatchAll, according to various goodness-of-fit criteria (see SF). If no "best" model appears it is because the stringent GOF criteria required for "best" status were not

satisfied by any model; in this case the user may consider the second-best models 2a-2c, or the other procedures including nonparametric estimates and lower bounds.

WLRM. The weighted linear regression model; see SF.

Parm Max Tau, WLRM Max Tau. The best parametric model and the WLRM computed on the entire dataset, i.e., no discarding of large outlying frequencies, so that τ = the maximum frequency in the data.

Best Discounted. The best parametric model with the low-frequency/high-diversity component removed ("discounted"), to account for uncertainty in the low-frequency sample counts such as singletons. This is usually too drastic a reduction; see SF.

Non-P 1. The statistic known as Chao1, which is a nonparametric lower bound for the total number of species (see SF). Chao1 only uses the first two frequency counts, hence $\tau = 2$.

Non-P 2. Chao's Abundance-Based Coverage Estimator ACE or its high-diversity variant ACE1, selected according to whether the coefficient of variation of the data (CV_rare) is ≤ 0.8 (ACE) or > 0.8 (ACE1), at $\tau = 10$ (or the largest $\tau \leq 10$). These are nonparametric estimates of the total number of species. See SF.

Non-P 3. Chao's Abundance-Based Coverage Estimator ACE, at $\tau = 10$ (or the largest $\tau \leq 10$), regardless of the ACE/ACE1 selection in Non-P 2 (see SF).

CatchAll
Version 3.0
Statistical Foundations

by Linda Woodard, Sean Connolly, and John Bunge
Cornell University
Sponsored by NSF Grant #0816638
June 7, 2011

1 Introduction

CatchAll is a set of programs for analyzing *frequency count* data arising from abundance- or incidence-based samples. Given the data, CatchAll estimates the total number of species (or individuals), observed + unobserved, and provides a variety of competing model fits and model assessments, along interactive graphical displays of the data, fitted models, and comparisons of estimates. The first program is CatchAll.exe, which performs the necessary statistical and numerical analysis; and the second is CatchAll.display.xlsm, which is a Microsoft Excel-based program that generates the graphical displays. We first discuss the statistical procedures underlying the main program.

2 Statistical procedures implemented by CatchAll

2.1 Parametric models

We fit a suite of five parametric models to the data. These are increasingly complex versions of the "standard model" for species estimation. For full mathematical details see, e.g., Bunge and Barger (2008); here we give a sketch intended to briefly explain the computations performed by the program. We assume that there is a fixed number of species C $(< \infty)$ in the population. The ith species contributes a random number X_i of individuals to the sample, where $X_i = 0, 1, 2, \ldots$. If $X_i = 0$ then the ith species is unobserved. X_i has a Poisson distribution with mean $E(X_i) = \lambda_i$, $i = 1, \ldots, C$, and in general we assume that $\lambda_1, \ldots, \lambda_C$ are distributed according to a *stochastic abundance model*, that is, a probability distribution with probability density function (say) $f(\lambda)$. The stochastic abundance distribution depends on some number of parameters (in our implementation there are at most 7 parameters), called θ. The observed frequency count data is then (unconditionally) distributed as *zero-truncated f-mixed Poisson*. We fit this distribution to the data via maximum likelihood, which yields an estimate $\hat{\theta}$ of the parameter (vector) θ; and from this we

obtain an estimate $p(0;\hat{\theta})$ of the *zero probability* $p(0;\theta)$, which is the probability that an arbitrary species is unobserved in the sample. Our final estimate is then

$$\hat{C} = \frac{c}{1 - p(0;\hat{\theta})},$$

where c is the number of observed species in the sample. This estimate has an associated standard error, given by

$$\text{SE}(\hat{C}) = \left(C \times \left(a_{00} - a_0^T A^{-1} a_0 \right)^{-1} \right)^{1/2},$$

where $a_{00} = (1 - p(0;\theta))/p(0;\theta)$, $a_0 = (1/p(0;\theta))\nabla_\theta(1 - p(0;\theta))$, and $A = \text{Info}(\theta, X)$, the Fisher information about θ in X, all evaluated at $\theta = \hat{\theta}$.

Actually the situation is slightly more complicated. Because frequency-count data typically exhibits a large number of rare species (graphically, a steep slope upward to the left) and a small number of very abundant species (a long right-hand tail of outliers), parametric models typically do not fit the entire dataset. Instead, some outliers must be deleted. Specifically, we fit a parametric model up to some maximum frequency τ, deleting all of the frequency-count data for frequencies $\geq \tau$, obtaining an estimate that depends on τ, $\hat{C}(\tau)$. To complete the estimate we add the number of species with counts greater than τ, $c_+(\tau)$, and the final estimate is then $\hat{C} = \hat{C}(\tau) + c_+(\tau)$. Similarly the SE is only computed on the data excluding outliers, i.e., on the frequency counts up to τ. Essentially we regard the frequencies $> \tau$ as constants or fixed points for the purposes of the analysis. This means that we compute every model at every possible value of τ and compare the results; we return to this issue below.

For confidence intervals we do not use the "Wald" or normal-approximation interval $\hat{C} \pm 1.96 * \text{SE}$, for various reasons. Instead we implement an asymmetric interval based on a lognormal transformation proposed by Chao (1987), Estimating the population size for capture-recapture data with unequal catchability, Biometrics 43(4), 783791. Let $c_-(\tau)$ denote the total number of species with frequency counts $\leq \tau$, so that $c_-(\tau) + c_+(\tau) = c$ for all τ. The lognormal-based interval is then

$$\left(c + (\hat{C}(\tau) - c_-(\tau))/d, c + (\hat{C}(\tau) - c_-(\tau)) \times d \right),$$

where

$$d = \exp\left(1.96 \left(\log\left(1 + \text{SE}^2/(\hat{C}(\tau) - c_-(\tau))^2 \right) \right)^{1/2} \right).$$

We also compute two goodness-of-fit measures. GOF0 is the p-value for the Pearson χ^2 goodness-of-fit test, comparing the observed frequencies to the expected frequencies (under the fitted model). This measure uses no adjustment for low cell counts, that is, every frequency is compared to its corresponding expected frequency. Since the χ^2 test is based on an asymptotic approximation requiring cell counts ≥ 5 (although there is not a consensus on this value), we also compute a p-value for the Pearson χ^2 test after concatenating

adjacent cells so as to achieve a minimum expected cell count of 5 (under the fitted model); this is GOF5. Since the null hypothesis in both cases is that the model fits, larger p-values support the choice of model.

We compute five progressively more complicated models, which we refer to as "order" 0, 1, 2, 3, and 4.

0. Poisson. Here the stochastic abundance distribution is a point mass at a fixed λ, i.e., all of the species sizes are assumed to be equal. This is rarely if ever realistic, and almost never fits real data, but it provides a readily computable lower bound benchmark, since heterogeneous species sizes will render this model downwardly biased.

1. Single exponential-mixed Poisson. The stochastic abundance distribution is exponential:
$$f(\lambda;\theta) = \frac{1}{\theta} e^{-\lambda/\theta},$$
$\lambda > 0$, $\theta > 0$. The mixed-Poisson distribution (of the frequency counts) is then the geometric:
$$P(X = j;\theta) = \frac{1}{1+\theta}\left(\frac{\theta}{1+\theta}\right)^j.$$
$j = 0, 1, 2, \ldots; \theta > 0$.

2. Mixture-of-two-exponentials-mixed Poisson. The stochastic abundance distribution is a mixture of two exponentials, and the mixed-Poisson distribution is then a mixture of two geometrics:
$$P(X = j;\theta) = \theta_3\left(\frac{1}{1+\theta_1}\left(\frac{\theta_1}{1+\theta_1}\right)^j\right) + (1-\theta_3)\left(\frac{1}{1+\theta_2}\left(\frac{\theta_2}{1+\theta_2}\right)^j\right),$$
$j = 0, 1, 2, \ldots; \theta_1, \theta_2 > 0; 0 < \theta_3 < 1$.

3. Mixture-of-three-exponentials-mixed Poisson. The stochastic abundance distribution is a mixture of three exponentials, and the mixed-Poisson distribution is then a mixture of three geometrics:
$$P(X = j;\theta) = \theta_4\left(\frac{1}{1+\theta_1}\left(\frac{\theta_1}{1+\theta_1}\right)^j\right) + \theta_5\left(\frac{1}{1+\theta_2}\left(\frac{\theta_2}{1+\theta_2}\right)^j\right) + (1-\theta_4-\theta_5)\left(\frac{1}{1+\theta_3}\left(\frac{\theta_3}{1+\theta_3}\right)^j\right),$$
$j = 0, 1, 2, \ldots; \theta_1, \theta_2, \theta_3 > 0; 0 < \theta_4, \theta_5 < 1$.

4. Mixture-of-four-exponentials-mixed Poisson. The stochastic abundance distribution is a mixture of four exponentials, and the mixed-Poisson distribution is then a mixture of four geometrics:

$$P(X=j;\theta)=\theta_5\left(\frac{1}{1+\theta_1}\left(\frac{\theta_1}{1+\theta_1}\right)^j\right)+\theta_6\left(\frac{1}{1+\theta_2}\left(\frac{\theta_2}{1+\theta_2}\right)^j\right)$$
$$+\theta_7\left(\frac{1}{1+\theta_3}\left(\frac{\theta_3}{1+\theta_3}\right)^j\right)+(1-\theta_5-\theta_6-\theta_7)\left(\frac{1}{1+\theta_4}\left(\frac{\theta_4}{1+\theta_4}\right)^j\right),$$

$j=0,1,2,\ldots;\theta_1,\theta_2,\theta_3,\theta_4>0;0<\theta_5,\theta_6,\theta_7<1.$

CatchAll computes all five models at every value of τ (having non-zero frequency count in the data). This generates a "combinatorial explosion" of analyses (one for each model*τ combination), which then must be sifted to find a "best" model, or at least a collection of best models. We do this according to the following algorithm, which combines statistical principles and heuristic decisions based on empirical experience.

Model selection algorithm

1. (Statistical.) Eliminate model*τ combinations for which GOF5 < 0.01.
2. (Statistical.) For each τ, select the model with minimum AICc (Akaike Information Criterion, corrected (where necessary) for small sample sizes).
3. (Heuristic.) Eliminate model*τ combinations for which estimate $>$ 100$\times$ACE1, where ACE1 is the estimate at $\tau = 10$.
4. (Heuristic.) Eliminate model*τ combinations for which SE $>$ estimate/2.
5. (Heuristic.) Then:
   - Best model: Select the largest τ for which GOF0 ≥ 0.01.
   - Model 2a: Select the τ with maximum GOF0.
   - Model 2b: Select the largest τ.
   - Model 2c: Select τ as close as possible but ≤ 10.
6. (Heuristic.)
   - If all model*τ combinations are eliminated, allow GOF5 above 0.001, but keep SE $<$ estimate/2.
   - If all combinations are still eliminated, allow GOF5 above 0.001 and allow SE up to the estimate.
   - If there are still no combinations, require calculable (computable) GOF5, but impose no restrictions on SE.

2.2 Nonparametric procedures

We compute five nonparametric estimates of C. All derive directly or indirectly from the "coverage-based" approach, under which the estimate of the total number of species is based on an estimate of the "coverage" of the sample – the proportion of the population represented by the sampled species. We compute the nonparametric estimates at every τ (as we do for the parametric estimates), but we report them only for $\tau = 10$ (or the nearest possible value), because they tend to be highly sensitive to outliers (see the section on CatchAll.display.xlsm for more discussion of this point).

1. Good-Turing, also called "homogeneous model," i.e., equal species sizes (same assumption as Model 0, Poisson, above):

$$\hat{C} = \frac{c_-(\tau)}{1 - f_1/n_-\tau} + c_+(\tau),$$

where $n_-(\tau) = \sum_{i=1}^{\tau} i f_i$.

2. Chao1:

$$\hat{C} = \begin{cases} c + f_1^2/(2f_2), & f_2 > 0 \\ c + f_1(f_1 - 1)/2, & f_2 = 0 \end{cases}.$$

This is generally regarded as a lower bound for C.

3. ACE (Abundance-based Coverage Estimator):

$$\begin{aligned} \hat{C} &= \frac{c_-(\tau)}{1 - f_1/n_-(\tau)} + c_+(\tau) + \frac{f_1}{1 - f_1/n_-(\tau)} \times \gamma^2_{rare} \\ &= \text{Good-Turing} + \frac{f_1}{1 - f_1/n_-(\tau)} \times \gamma^2_{rare}, \end{aligned}$$

where

$$\gamma^2_{rare} = \max\left(\frac{c_-(\tau)}{1 - f_1/n_-(\tau)} \frac{\sum_{i=1}^{\tau} i(i-1) f_i}{n_-(\tau)(n_-(\tau) - 1)} - 1, 0\right).$$

4. ACE1 (Abundance-based Coverage Estimator for highly heterogeneous cases):

$$\begin{aligned} \hat{C} &= \frac{c_-(\tau)}{1 - f_1/n_-(\tau)} + c_+(\tau) + \frac{f_1}{1 - f_1/n_-(\tau)} \times \gamma'^2_{rare} \\ &= \text{Good-Turing} + \frac{f_1}{1 - f_1/n_-(\tau)} \times \gamma'^2_{rare}, \end{aligned}$$

where

$$\gamma'^2_{rare} = \max\left(\gamma^2_{rare}\left(1 + \frac{f_1/n_-(\tau)}{1 - f_1/n_-(\tau)} \frac{\sum_{i=1}^{\tau} i(i-1) f_i}{n_-(\tau) - 1}\right), 0\right).$$

ACE is preferred when $\gamma_{rare} \leq 0.8$, otherwise ACE1 is preferred. CatchAll makes this selection automatically.

5. Chao-Bunge gamma-Poisson estimator:

$$\hat{C} = \frac{\sum_{i=2}^{\tau} f_i}{1 - \frac{f_1 \sum_{i=1}^{\tau} i^2 f_i}{\left(\sum_{i=1}^{\tau} i f_i\right)^2}} + c_+(\tau) = \frac{c_-(\tau) - f_1}{1 - \frac{f_1 \sum_{i=1}^{\tau} i^2 f_i}{n_-(\tau)^2}} + c_+(\tau).$$

   This is known to be consistent when the stochastic abundance model is the gamma distribution, i.e., when the sample counts follow the negative binomial distribution.

In each case we also compute a standard error based on an asymptotic approximation due to Chao (***). The variance for one of these estimators $\hat{C}$ is given by the approximate formula

$$\text{Var}(\hat{C}) \approx \sum_{i \geq 1} \sum_{j \geq 1} \frac{\partial \hat{C}}{\partial f_i} \frac{\partial \hat{C}}{\partial f_j} \hat{\text{cov}}(f_i, f_j),$$

where

$$\hat{\text{cov}}(f_i, f_j) = \begin{cases} f_i \left(1 - \frac{f_i}{\hat{C}}\right), & i = j \\ -\frac{f_i f_j}{\hat{C}}, & i \neq j. \end{cases}$$

The (empirical) standard error of $\hat{C}$ is then $\sqrt{\text{Var}(\hat{C})}$. Thus the problem is to calculate $\partial \hat{C} / \partial f_i$, which in turn depends on the formula for $\hat{C}$, in each case. We omit the specific details here.

Finally, we display two analyses of the *full* dataset, that is, with no right-truncation ($\tau = \max \tau$). These are the minimum-AICc parametric model, and the preferred ACE/ACE1 choice.

New Features in CatchAll v.3.0

John Bunge and Linda Woodard

June 7, 2011

3 The weighted linear regression model

This is an approach to analyzing frequency count data based on a novel, completely different concept from either the parametric or the coverage-based nonparametric methods discussed above. The approach is discussed in detail in Rocchetti, Bunge and Böhning (2011), "Population size estimation based upon ratios of recapture probabilities," *Annals of Applied Statistics* (in press as of this writing). Basically the frequency count data is converted to (adjusted) *ratios of successive counts*:

$$r(i) := \frac{(i+1)f_{i+1}}{f_i},$$

$i = 1, 2, \ldots$. Under the unmixed Poisson and the gamma-mixed Poisson or negative binomial models, the ratios $r(i)$ form an approximately linear function of i. It is conjectured that under mild departures from these models the linearity is preserved to some degree, i.e., the model is somewhat robust to such departures (this is a topic of current research). It is therefore reasonable to consider *linear regression* of $r(i)$ on i, that is,

$$r(i) = \frac{(i+1)f_{i+1}}{f_i} \approx \beta_0 + \beta_1 i,$$

$i = 1, 2, \ldots$. Having fit such a regression model in the usual way, one can then project the model "downwards" so as to obtain an estimate (prediction) of f_0 and hence an estimate of the total diversity. The same procedure yields standard errors and goodness-of-fit assessments.

There are three secondary considerations here.

1. This model is inherently heteroscedastic, and consequently weighted linear regression must be used (hence the name); the weights are computed automatically by CatchAll, according to the specification in Rocchetti *et al.*

2. In some cases a log-transformed regression must be used to avoid certain edge effects which can lead to negative predictions for f_0. CatchAll automatically selects between the log-transformed and the untransformed (original) models, and reports the selected mode in the summary output. However, all models are reported in the copious output.

3. As with all of the procedures implemented by CatchAll, the results depend (to varying degrees) on the right truncation point τ. CatchAll automatically selects an "optimal" τ based on goodness-of-fit criteria, and reports the corresponding results in the summary output; as usual results at all τ's are reported in the copious output. Note that, in order to avoid division by zero, the set of frequency counts used by the weighted linear regression procedure must be contiguous, that is, gaps in the frequency counts are not allowed. Hence the maximum possible τ for this model is the maximum of the contiguous frequency counts, which may not be the actual maximum frequency count.

4 Best discounted model

This procedure is intended to address the scenario in which the low-frequency counts may be inaccurately recorded. That is, for example, the number of singletons (or other very low-frequency counts) may be artificially inflated due to errors in measurement or registration, or to other experimental or observational artifacts. To address this, we compute a "best discounted model." This is based on the multiple-component parametric mixture model. Once the first CatchAll run is complete, we obtain (as detailed above) a "best fitted [parametric] model" (at a selected value of τ). This may be a mixture model of order 0, 1, 2, 3, or 4. If the selected order is 4, 3, or 2, CatchAll deletes the highest-diversity component, i.e., the component associated with fitting the lowest-frequency counts. The resulting model is one order lower (4, 3, or 2 converts to 3, 2, or 1, respectively), and this "step-down" model is reported as the "best discounted model." Specifically, the formulae are

- Step down from four-mixed to three-mixed:
  $\hat{C}^*_\tau = (1 - t5) \times \hat{C}_\tau$
  $\mathtt{SE}^* = (1 - t5) \times \mathtt{SE}$

- Step down from three-mixed to two-mixed:
  $\hat{C}^*_\tau = (1 - t4) \times \hat{C}_\tau$
  $\mathtt{SE}^* = (1 - t4) \times \mathtt{SE}$

- Step down from two-mixed to one-mixed:

$$\hat{C}^*_\tau = (1 - t3) \times \hat{C}_\tau$$
$$\mathtt{SE}^* = (1 - t3) \times \mathtt{SE},$$

where $\hat{C}^*_\tau$ is the step-down (reduced model) estimate of total diversity based on the frequency count data up to τ, and $\hat{C}_\tau$ is the original estimate of total diversity based on the frequency count data up to τ. (Counts for frequencies $> \tau$ are added in *ex post facto* as usual.) A graphical display of an example using viral diversity data is given on the next page, in Figure 1.

Figure 1: Best discounted model, stepping down from order 3 to order 2; component 1 is deleted