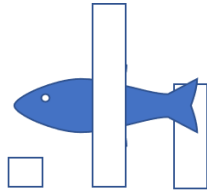


eDNA PSD Analysis Tool GUIDE



Application

<https://dhallack.shinyapps.io/appv5/>

Version:5 (October 19, 2023)

Contents

Application	1
Purpose	2
General app walk-through	2
Input data:.....	2
1. Sampling Data:	2
2. Source Data:	3
Analysis:	3
1. Prediction:.....	3
2. Sensitivity:.....	5
Working Examples	7
Two sizes example	7
Loading sampling and source data	7
Prediction - Search:	8
Prediction - Results:	9
Comments	11
References	12
Appendix A.....	12

Purpose

The purpose of this application is to assist in interpreting the particle size distribution (PSD) for environmental DNA (eDNA) in real streams. The app employs independent decay rates for various eDNA size categories, as outlined by Brandao-Dias in 2023, to link changes in PSD with either time or the distance between a source point and sample points.

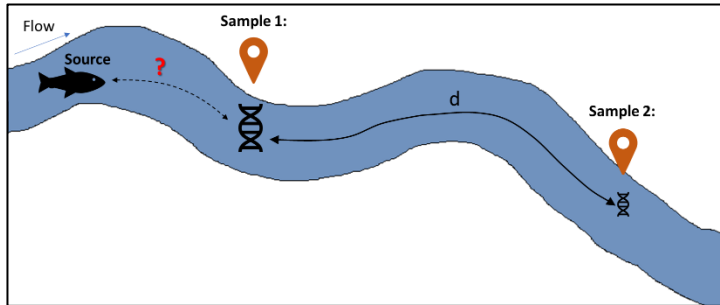


Figure 1 – Schematic of the problem

The app can estimate the distance or time from the eDNA source, using the concentration of different particle sizes found in multiple water samples and an initial expected PSD typically released from a particular species. It can also estimate the concentration or biomass released at the source.

The primary output is the probability distribution of the distance or time from the source at which the sample was collected. A second output is the probability distribution of concentration or biomass released by the source.

Overall, the app offers a tool for interpreting eDNA and its behavior in streams, which could be used for monitoring/management and to answer ecological questions.

General app walk-through

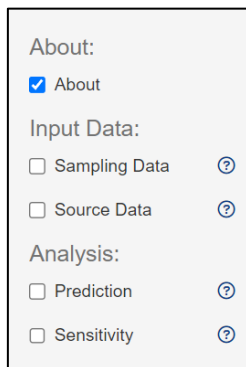


Figure 2 – Sidebar Menu

The sidebar navigation menu (Figure 2) presents all links and steps for the application. The user should follow the menu from top to bottom. Tabs can be made available or hidden under the selection using the checkboxes. Interrogation marks to the right of each step give more detailed information about its content.

The initial About page gives a general description of the problem, instructions, and physical assumptions. The application is divided into two main parts: **Input data** and **Analysis**.

Input data:

Two types of data must be entered by the user into the application:

1. Sampling Data:

The user enters eDNA decline rates or data to calculate these rates. The user can upload a CSV table with information from experiments measuring concentrations of eDNA with time or distance for multiple sizes and treatments. The CSV table should follow a specific format (Appendix A). Alternatively, the user can manually input values for removal rates and intercept and their confidence interval by size. The first option automatically generates linear regressions estimating decline rates and the 95% confidence interval, presented in a plot and a table. In this tab, the user should also choose the treatment desired for the analysis.

2. Source Data:

Here, the user should enter the expected initial concentration fractions released by a certain aquatic species. The sizes must agree with the ones specified in the “Sampling Data”. As in the “Sampling data” tab, this information can come from uploading a CSV table (format in Appendix A) or be entered manually. A PSD histogram is automatically generated comparing the source and the first sampling PSD. If the upload table has multiple treatments or replicates, the source data is averaged for each species, and a standard deviation is calculated. The error bar represents the uncertainty of this data.

Analysis:

There are two types of analysis currently available in the application:

1. Prediction:

In this panel, based on the expected initial PSD given in the “*Source Data*” tab, the app performs a multivariable search on the removal rates and Intercept confidence interval space to find a combination that can connect the source and sample points in terms of PSD. This panel is divided into two subsections: **MC Simulation** and **Prediction Distributions**.

1.1. MC Simulations

The app uses a Monte Carlo (MC) simulation approach to sample the removal rates and Intercept space, considering uniform distributions between lower and upper boundaries of the 95% confidence interval (estimated in the “*Sampling Data*” tab).

After sampling the rates and the intercept, a back extrapolation of PSDs from the sampled point until the maximum time specified is executed. The process is repeated multiple times (number specified by the user).

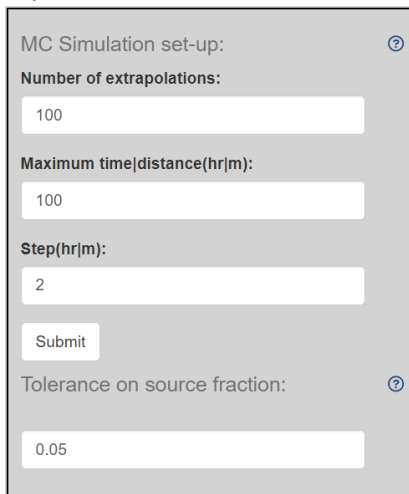


Figure 3 – Set-up MC simulation

Controls:

The user should set up parameters for a Monte Carlo simulation (Figure 3): the number of extrapolations, the maximum time of the simulation, and the step size. These parameters are case-dependent and can be adjusted dynamically according to the results. Click the “Submit” button below the parameter input sections to launch the simulation.

The tolerance on source fraction is the error admitted for the source PSD. This value should be based on the std error evaluated in the Source Data tab or by previous knowledge. The phase diagram plot displays the tolerance error as a circle around the source PSD point.

Outputs:

There are two plots in this subsection. The first plot shows the ensemble of all back-extrapolations per filter size. When a case can reach the source PSD at any point, it is highlighted in blue (Figure 4).

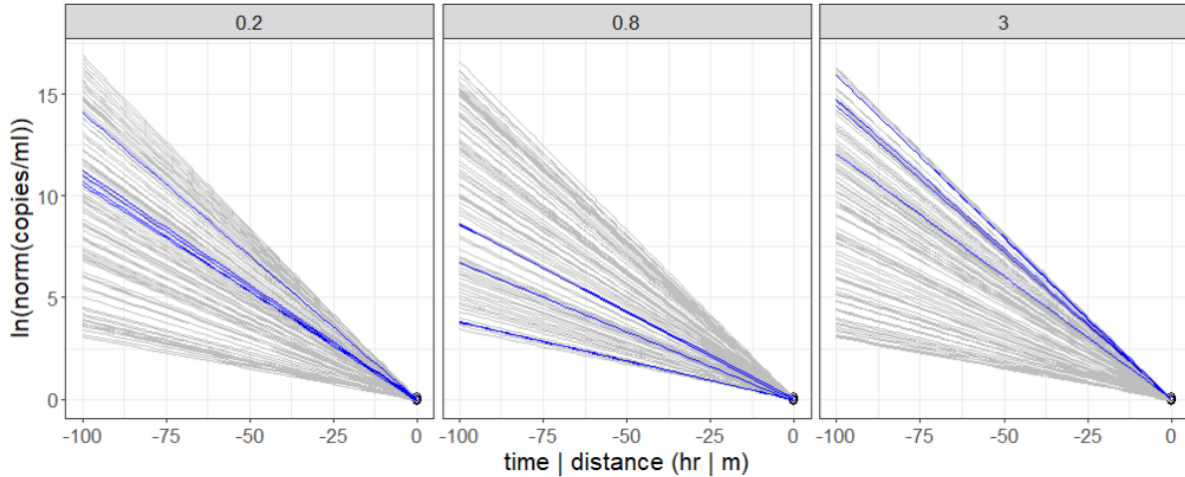


Figure 4 - Back-extrapolations searching to match source PSD

The second plot is a phase space representation of the system, where the state variables are the filter size fractions (the number of degrees of freedom of the system is one less than the total number of filters). The starting point of these trajectories is controlled by the “Intercept” sampled values, and the displacement in the phase space is controlled by the “removal rates” sampled values. For each combination, a different trajectory is generated. In the plot, all trajectories that cross the source PSD tolerance region are considered possible cases and are highlighted in blue (Figure 5).

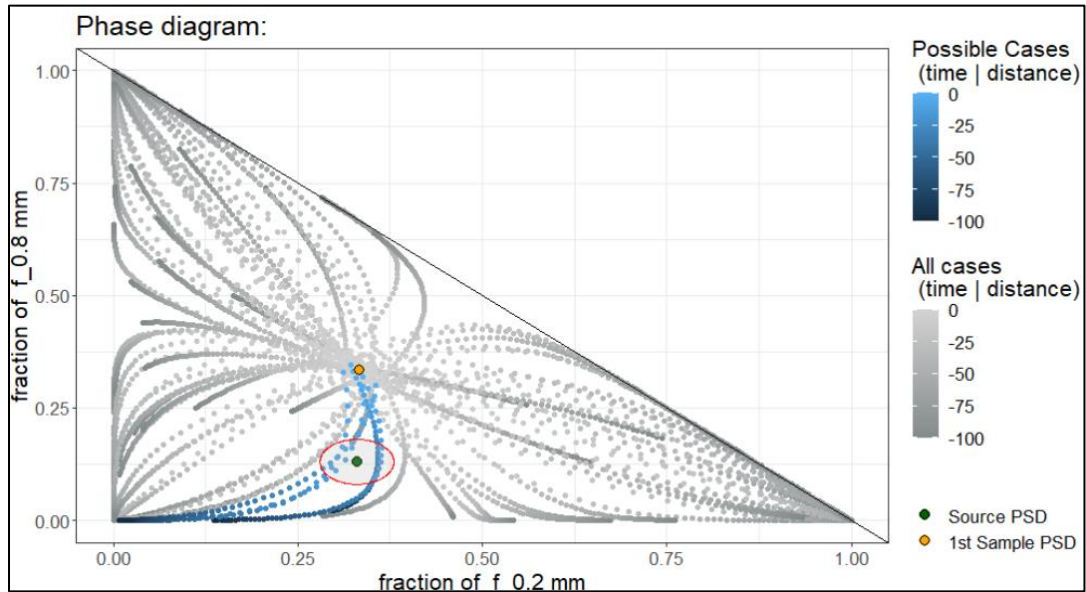


Figure 5 – Simulations in the phase space diagram. Possible cases highlighted. 3-sizes case

1.2. Prediction distributions:

This part of the panel shows histograms of source time or distance relative to the first sample and initial concentration estimations based on the selected “Possible Cases”. For each possible case trajectory, the closest point to the source PSD is chosen as the best prediction time or distance between the source and sampling.

Controls: The user can set a “*cutoff distance*” based on the maximum possible stream length (Figure 6). It eliminates solutions that give unrealistic distances between source and sample. The user can input a conversion from concentration (copies/ml) to biomass (kg) on top of the initial concentration histogram.

Outputs: The histogram distributions summarize possible solutions information for time or distance, from the source to the first sample, and for initial concentration. The boxed information gives general statistical metrics of the plotted histograms (number of data points, range, mean, standard deviation, and mode). Based on the mean and standard deviations, two theoretical probability functions are generated as references (not fitted): normal and lognormal. These can be plotted on top of the histogram by the selection of “*pdf(probability distribution function) proxy*” (Figure 6).

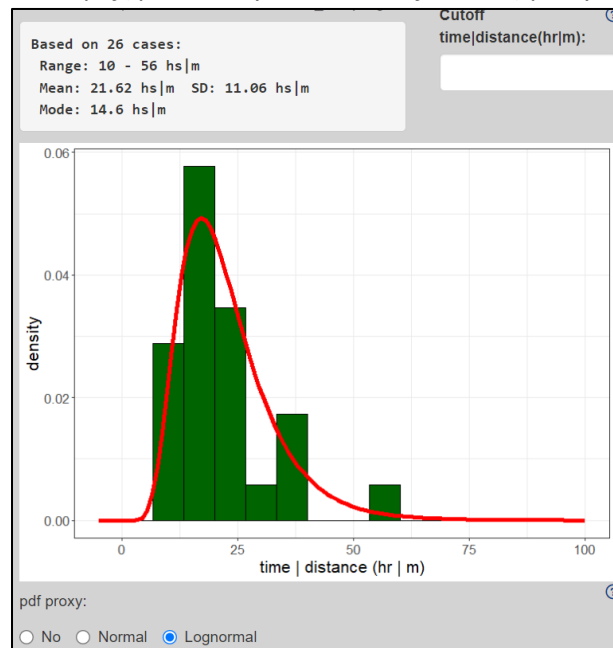


Figure 6 – Distribution histogram of time or distance of the source.

2. Sensitivity:

This panel is not part of the main workflow of the application. Still, it might be helpful to test specific



Figure 7 – Slider inputs for sensitivity

cases of combinations of removal rates or to visualize better what calculations are under the prediction tab. In this panel, the user can estimate the change in PSD a certain number of hours or meters before the first sample by specifying all the removal rates and “Intercept” values.

Controls: Using the slider buttons, the user can control the removal rates and “Intercept” of each size class and the time|distance from the first sample. The controls of the removal rates are interpolation multipliers that allow the values to go from the minimum to maximum values given in the confidence interval. The final values are shown in the “Selection table” below.

Outputs:

The two plots on the right are dynamically updated under any parameter change. The first plot shows a single back-extrapolation according to the parameters selected and the regression used to determine the confidence intervals.

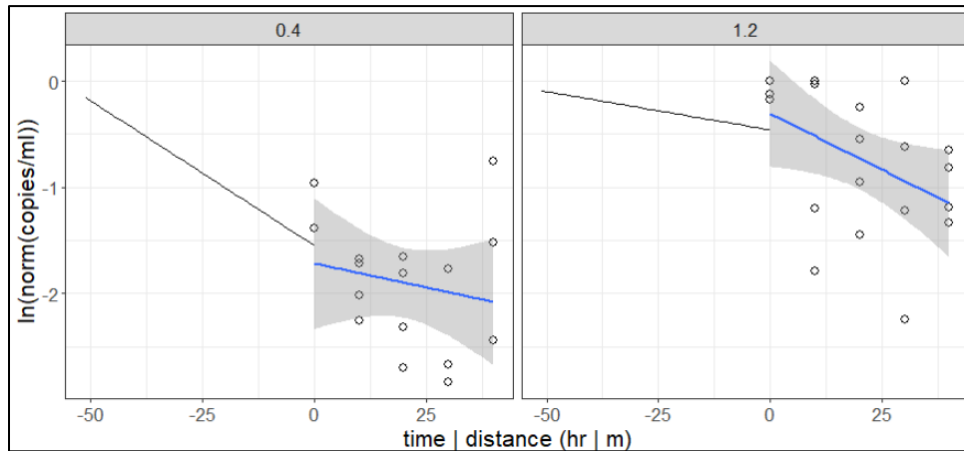


Figure 8 - Back extrapolation with parameters selected by the user

The second plot shows the change in PSD between the start and end of the extrapolation (Figure 8). The starting point ($t|d = 0$) is calculated using the “Intercept” selected values. The endpoint depends on the “removal rates” values.

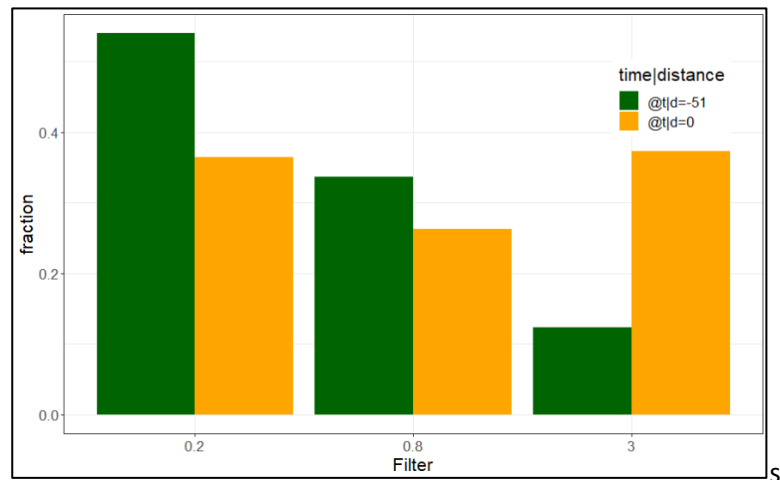


Figure 9 – Change in PSD after a time | distance specified by the user.

Additional plots at the bottom of this panel show continuous temporal behavior of concentrations and fractions of each class size during the selected period. They intend to clarify the transformation of concentrations extrapolations behaviors into the PSD changes.

Working Examples

In this section, we provide a walkthrough example using real data, and show how to interpret the main outputs of the analysis and highlight any limitations.

Two sizes example

The example is based on data collected from experimental eDNA additions conducted in the outdoor streams at the [Notre Dame Linked Experimental Ecosystem Facility \(ND-LEEF\)](#). These additions were based on previously published methods (Stream Solute Workshop 1990; Shogren et al. 2017). The filter sizes for this experiment were 0.4 and 1.2 μm . The samples were taken at 10 m intervals downstream of the eDNA addition site, which simulates a population of organisms.

For the sake of this example, we modify the data table, setting the first sample at zero and all the other samples at their distance relative to the first sample. This assumption allows us to assess the app's predictive power using real-world data.

Loading sampling and source data

A CSV table with the data following the format described by Appendix A can be loaded in the “*Sampling Data*” tab. The app will automatically perform a linear regression for each size class using the treatment the user selects. Eventual replicates for the same treatment are merged.

Here, we choose the “*SH_COB_CARP*” treatment wherein the experimental stream was covered with 90% shade cloth, lined with cobble substrate, and Common Carp (*Cyprinus carpio*) eDNA was added to the stream. The panel shows a table and a plot (Figure 10). The table summarizes the regression information per Filter Size: the average declining slope and the 95% confidence interval. The plot shows the regressions with the data per filter size. Note that there is a different regression for each Filter size. The smallest size (0.4 μm) has a larger range because its decline is not well defined based on the data. This uncertainty information is used later in the Analysis.

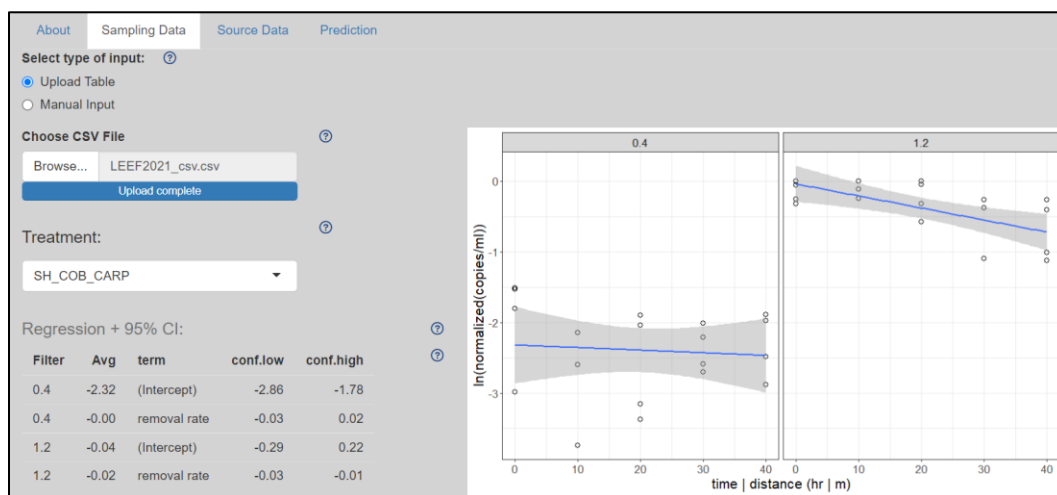


Figure 10 – Regressions based on LEEF experimental data

A CSV table containing source data can be loaded in the “*Source Data*” tab. The app will automatically calculate average and standard deviation fractions per particle size for the treatment or target the user selects. The histogram of the source fractions is plotted on the right side, named “*Source PSD*”. To compare the plot and the table also show the PSD of the first sample, which is the sample at the reference zero distance (Figure 11). The error bars are based on the standard deviation across multiple replicates.

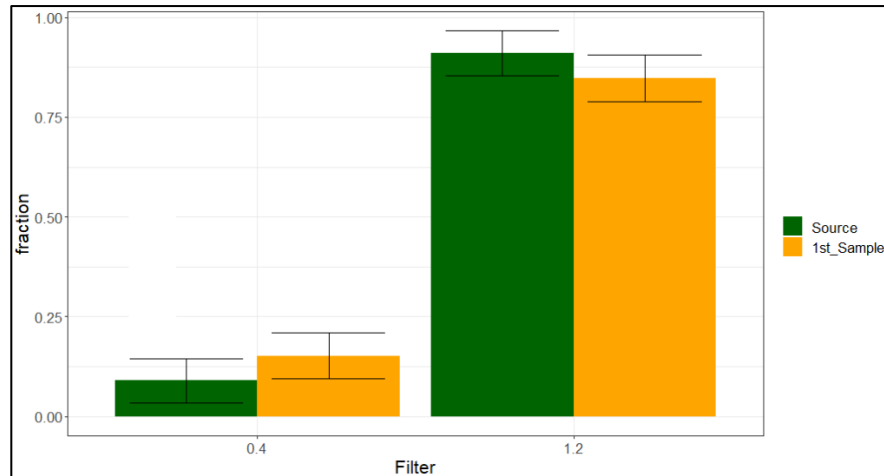


Figure 11 – PSD from source and 1st sample.

Prediction - Search:

The primary purpose of this MC simulation is to find which combinations of removal rates and intercepts would allow a possible connection between the “*Source PSD*” and the “*1st Sample PSD*”. To attain this value, random values within each parameter confidence interval are sampled, and several backward extrapolations are performed starting from the “*1st Sample*” position, which is the zero referenced point. The possible “connecting” cases are the ones in which the trajectory passes close enough to the “*Source PSD*” point, considering a tolerance criterion. This tolerance criterion should be based on the uncertainty of the “*Source PSD*” error estimated on the “*Source Data*” tab.

In this example, we have 44 possible cases that satisfy the criterion out of 100 extrapolations (Figure 12). The table below the inputs lists just the first ten possible cases that match the “*Source PSD*” criteria. The table shows the distance and the closest fraction point to the “*Source PSD*” for each possible case. The square root of the sum of the squared errors (“*SSe*”) represents the error of each case and will always be below the tolerance criterion for all possible cases.

Note that all the points on the Phase diagram stand on the 45-degree line because the sum of the fractions must always be 1. This shows that the system has only one degree of freedom.

Changes in simulation parameters are not updated dynamically. It requires clicking on the “*Submit*” button.

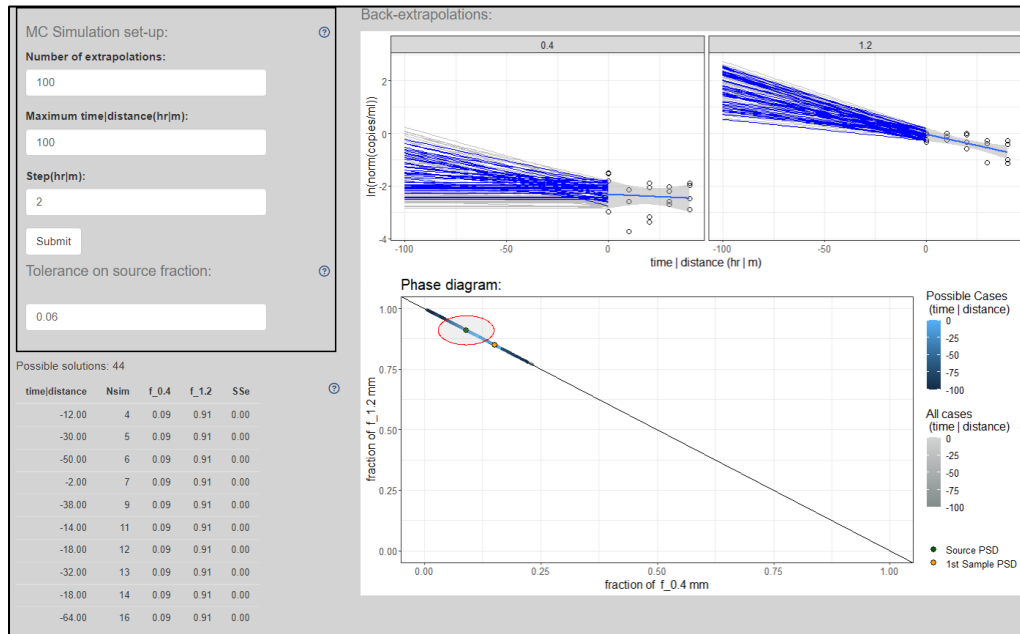


Figure 12 – MC search using 100 extrapolations trials

Prediction - Results:

In our example, with 100 extrapolations, 44 configurations reached the “Source PSD” satisfying the tolerance criterion. In the “*Prediction distributions*”, these selected scenarios are filtered and a summary of statistical metrics for distance and initial concentrations is displayed, as well as histogram distribution plots (Figure 13). Finally, we can eliminate solutions that predict unrealistic distances, for instance, a distance between the source and the sample larger than the total stream length, which would be 50m for the data in this example.

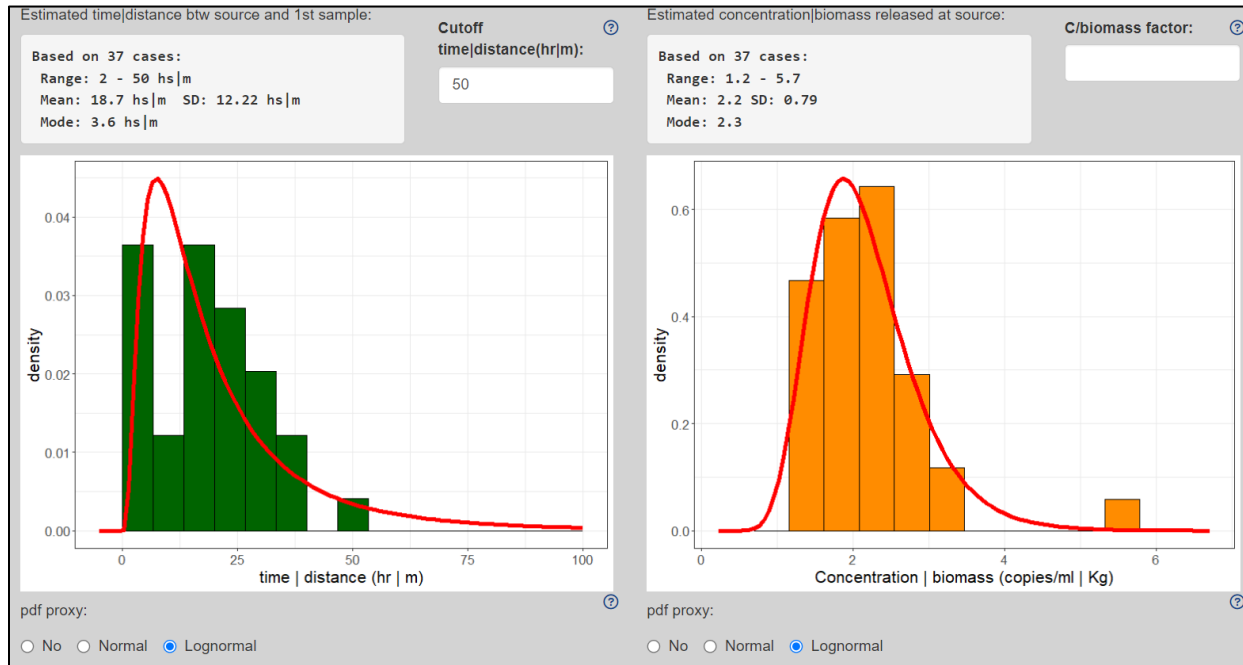


Figure 13 – Histograms of “Possible cases” after distance cutoff (100 extrapolations)

A proxy distribution is plotted on top of the histogram as a reference for the shape of a possible theoretical distribution. Our example shows that the Source could be between 2 and 50m upstream of the sample, with a mean of about 19m and a standard deviation of 12m. The range covers the actual distance, which is 10m. However, it is a very broad estimation.

The Initial concentration release could be between 1.2 and 5.7 copies/ml, with a mean of 2.2 and a standard deviation of 0.79. Since the bucket solution is diluted when it enters the stream, it is a more difficult number to check.

To refine the shape of the histograms it is necessary to increase the number of simulations(Figure 14). The mode of the distribution becomes more representative. However, the mean and standard deviation stay very similar (Figure 14). Which shows that the uncertainty on the prediction does not reduce significantly just increasing the number of simulations. Higher than 1000 extrapolations the app starts to take a significant time to run.

For more accurate answers, we need to narrow down the confidence interval for the removal rates and intercept, which might be difficult, options would be getting additional field data, a better model to represent the data, or working with three-sizes categories instead of two-sizes (Figure 5).

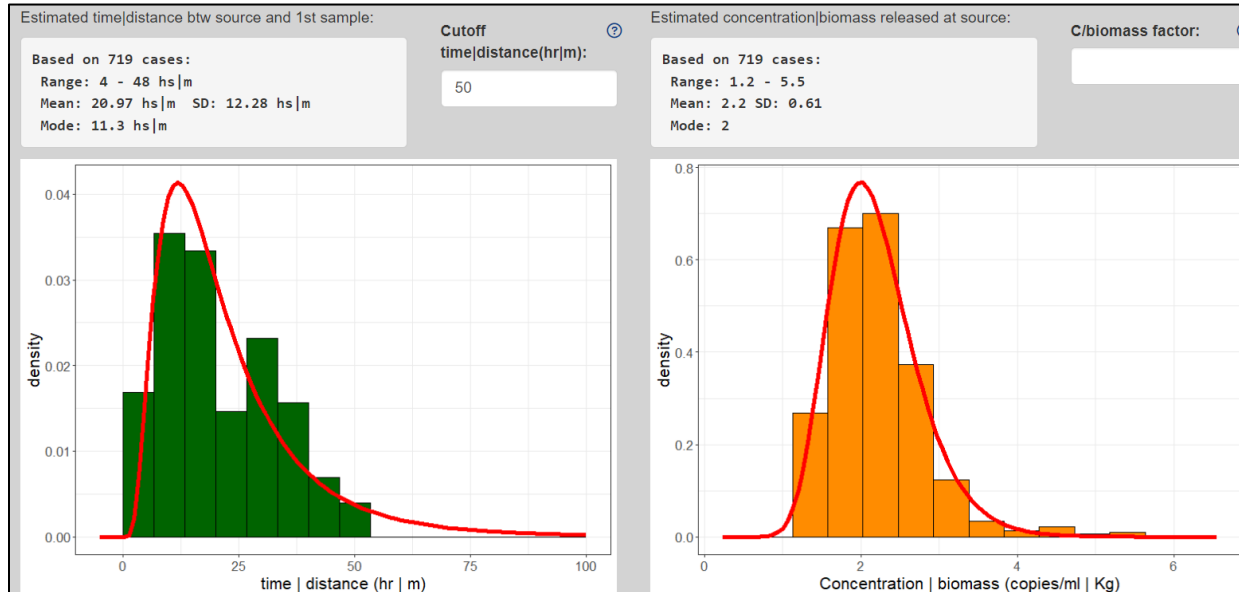


Figure 14 - Histograms using 3000 extrapolations.

Comments

The app was developed in R language using the Shiny package.

Although the app is designed to accommodate any number of class sizes, it has only been tested up from 2 to 4. Using the app with more than four would require a significantly greater number of extrapolations, which could significantly slow down the analysis process.

The current state of the app only allows one source. Multi-source analysis could be eventually added to the tool.

References

Brandão-Dias, P. F. P., Hallack, D. M. C., Snyder, E. D., Tank, J. L., Bolster, D., Volponi, S., Shogren, A. J., Lamberti, G. A., Bibby, K., & Egan, S. P. (2023). Particle size influences decay rates of environmental DNA in aquatic systems. *Molecular Ecology Resources*, 23, 756–770. <https://doi.org/10.1111/1755-0998.13751>

Shogren, A. J., Tank, J. L., Andruszkiewicz, E., Olds, B., Mahon, A. R., Jerde, C. L., & Bolster, D. (2017). Controls on eDNA movement in streams: Transport, retention, and resuspension. *Scientific Reports*, 7, 5065.

Stream Solute Workshop. Concepts and methods for assessing solute dynamics in stream ecosystems. *J. N. Am. Benth. Soc* 9, 95–119 (1990).

Appendix A

The CSV input table on “*Sampling Data*” must be in the format specified in Table 1. The names of the columns should be Treatment, Replicate, Filter, Target, Concentration, and Hours.

Table 1 – Sampling Data

	A	B	C	D	E	F
1	Treatment	Replicate	Filter	Target	Concentration	Hours
2	HB_PG_CARP	R1	1.2	CARP	1	0
3	HB_PG_CARP	R2	1.2	CARP	0.822222222	0
4	HB_PG_CARP	R3	1.2	CARP	0.646666667	0
5	HB_PG_CARP	R4	1.2	CARP	0.328689449	0
6	HB_PG_CARP	R1	0.4	CARP	0.249404176	0
7	HB_PG_CARP	R2	0.4	CARP	NA	0
8	HB_PG_CARP	R3	0.4	CARP	NA	0
9	HB_PG_CARP	R4	0.4	CARP	0.149405967	0
10	HB_PG_CARP	R1	1.2	CARP	0.166445461	10
11	HB_PG_CARP	R2	1.2	CARP	0.246666667	10

The CSV input table on “*Source Data*” must be in the format specified in Table 2. The names of the columns should be Treatment, Replicate, Filter, Target, and Concentration.

Table 2 – Source Data

	A	B	C	D	E
1	Treatment	Replicate	Filter	Target	Concentration
2	HB_BKT_COB_Post_CARP	R1	1.2	CARP	611.3207547
3	HB_BKT_COB_Post_CARP	R1	0.4	CARP	15.09433962
4	HB_BKT_COB_Post_STLHD	R1	1.2	STLHD	2077.669903
5	HB_BKT_COB_Post_STLHD	R1	0.4	STLHD	988.3495146
6	HB_BKT_MIX_Post_CARP	R1	1.2	CARP	989.5348837
7	HB_BKT_MIX_Post_CARP	R1	0.4	CARP	55.03875969
8	HB_BKT_MIX_Pre_STLHD	R1	1.2	STLHD	1164.779874
9	HB_BKT_MIX_Pre_STLHD	R1	0.4	STLHD	2817.610063